Essays on Estimation, Calibration and Inference for Simulation Models

Ph.D. Candidate: Mario Martinoli
Doctoral Advisor: Prof. Raffaello Seri

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in
Methods and Models for Economic Decisions
Università degli Studi dell'Insubria - Dipartimento di Economia (DiECO)
2021

Abstract

The last century has seen a growing interest in complexity in economics and social sciences. The need to model the complex and emergent dynamics of a system has spurred many researchers towards the exploration of estimation techniques that could be used for increasingly complex models, fostering the rise of simulation-based econometric methods. This dissertation aims at contributing to the blossoming literature concerning estimation, calibration and inference for simulated models.

Chapter 1, jointly written with Raffaello Seri and Davide Secchi, gives a historical review of the different simulation methods, from the first experiments with early computers to sophisticated agent-based models (ABM). In this chapter we focus on three fundamental aspect governing the dynamics of the system: randomness, causation and emergence.

Chapter 2 critically review the different approaches developed for the estimation, calibration and validation of simulation models. We begin clarifying the concept of identification and estimation and we expose the problems related to the dependence of the simulated data on the parameters to be estimated. Then, we formalize the meaning of calibration, estimation and validation of simulated models. Subsequently, considering the classical simulation-based econometric frameworks, we detail the characteristics of the most popular techniques (e.g., indirect inference, method of simulated moments, simulated minimum-distance, simulated maximum likelihood, and approximate Bayesian computation). We then make a comparison of the different methods, explaining their main advantages and weaknesses. In the last section of the chapter, we shift our attention on the estimation, calibration and validation of ABM.

Chapter 3, co-authored with Raffaello Seri, Davide Secchi and Samuele Centorrino, considers the issues of calibrating and validating a theoretical model. We aim at selecting the parameters that better approximate the data when the researcher has to choose among a finite number of alternatives. Given a pre-specified loss function, we propose to build a Model Confidence Sets (MCS, see [223]) to restrict the number of plausible alternatives, and measure the uncertainty associated to the preferred model. We further suggest an asymptotically exact logarithmic approximation of the probability of choosing a certain configuration of parameters. A numerical procedure for the computation of the latter is provided and its results are shown to be consistent with Model Confidence Sets. The implementation of our framework is showcased using a model of inquisitiveness in ad hoc teams (see [30]).

The similarity between simulated and real-world observations is generally computed minimizing their statistical distance (see Chapter 2 of this thesis). Therefore, a natural implication of estimation and calibration concerns the study of the asymptotic properties of divergence measures. Chapter 4, that is jointly written with Raffaello Seri, is devoted to the estimation of the entropy of a discretely supported time series through a plug-in estimator. We demonstrate the almost-sure convergence of the observed entropy H_N to a limit H_∞. We show that the widely used bias correction proposed by [418] is incorrect and we fix it in order to remove the $O\left(N^{-1}\right)$ part of the bias. We provide the asymptotic distributional results under α-mixing for the general case and under *degeneracy* (i.e. when the values taken by the marginal distribution of the process are equiprobable). We introduce

the estimators of bias, variance and distribution under degeneracy, and we provide results on the errors in the estimation. At last, we propose an application of the entropy to a goodness-of-fit test for the marginal distribution of the process. Some simulation experiments and numerical examples are provided to support the theoretical results.

We conclude the thesis presenting two new estimation methods that can be used to overcome some drawbacks peculiar to classic simulation-based estimators.

In Chapter 5, co-authored with Raffaello Seri, we exploit a nonparametric sieve regression (see [208, 354, 88]), estimated through ordinary least squares (OLS), to find the parameters of a simulation model producing moments that are similar to real-world statistics. We run a simulation model for several parameter values, we compute the statistics on each run, and we estimate nonparametrically the function linking the generated statistics and the associated parameters. Using the real-world statistics as explanatory variables in the previous nonparametric regression, we estimate the parameters of the model. Differently from simulated minimum-distance techniques (e.g., indirect inference and simulated method of moments), our setup does not involve any objective function and no optimization algorithm is required. This leads to several advantages when compared to classic simulation-based econometric methods. In the end, we explicitly and rigorously characterize the asymptotic theory of the estimator, including the order of the bias, confidence intervals and hypotheses tests. Ultimately, we evaluate the approach through a small simulation study.

Finally, Chapter 6, also jointly written with Raffaello Seri, is similar in spirit to the work by [86]. The statistical framework is close to the one exposed in Chapter 5 but, instead of OLS, we estimate the parameters of a simulated models via a nonparametric least absolute shrinkage and selection operator (Lasso) regression. The nonparametric element is introduced to capture the nonlinear relations between the statistics and the parameters. This implies some advantages, when comparing the method to the previous chapter, that will be clarified in the following. First of all, the Lasso allows the joint estimation and automatic selection of a subset of the basis functions used to model the nonparametric function linking the statistics and the parameters to be estimated. Second, in Lasso regression the number of basis function in the dictionary is not upper bounded by the number of points chosen out of the parameter space, while in OLS it is. Third, the oracle property of the Lasso suggests that the researcher may run this algorithm, identify the nonnull coefficients of the regression, and run a nonparametric sieve regression estimated by OLS containing only the retained coefficients, thus making the inferential tools of Chapter 5 available in the present situation. Furthermore, we explicitly and rigorously characterize the asymptotic behavior of the estimator. We end the chapter with a small simulation study showing the correct behavior of the method.

Contents

Acknowledgements

> “Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise.”
> (John W. Tukey, *The Future of Data Analysis*)

When at the end of my path someone will ask me: “how was the PhD?”, the first answer will be: “blood, toil, tears and sweat”, but the right answer should be: “an amazing journey to the heart of knowledge”.

I am deeply grateful to my supervisor and mentor, Raffaello Seri, for his guidance, his support, his infinite knowledge and all the time he dedicated to me. He taught me to think outside the box. Much more than a professor, he is a friend.

I am thankful to Samuele Centorrino for his supervision, his stimulating advice and his recommendations during my visiting period at Stony Brook University.

I am also in debt with Davide Secchi for introducing me to the world of simulations and agent-based models.

I would also like to thank Bernhard Beckermann, Uwe Cantner, Ernesto Carrella, Eugenio Caverzasi, Francesca Chiaromonte, Fulvio Corsi, Davide Fiaschi, Alan Kirman, Francesco Lamperti, Pascal Lavergne, Federico Martellosio, Giovanni Mellace, Giovanni Migliorati, Alessio Moneta, Matthias Müller, Giuseppe Ragusa, Emmanuel Rio, Andrea Vezzulli, and all the academic staff of DiECO and the Department of Economics of Stony Brook University for their advice and feedback.

I am also grateful to my PhD mates and colleagues Anna, Cecilia, Daniele, Daniele, Evangelia, Luca, Matteo, Zeid. It was an honor to share this experience with you.

A special thank goes to Camilo, Laura and Marcos for letting me live an unforgettable experience in the United States. You made me feel at home.

Last but not least, an infinite thank to my family for supporting and bearing me in every difficult moment, and thanks to my friends, I know you will always be there.

To my beloved niblings.

Chapter 1

Randomness, Emergence and Causation: A Historical Perspective of Simulation in the Social Sciences[1]

This chapter is a review of some simulation models, with special reference to social sciences. Three critical aspects are identified, i.e. randomness, emergence and causation, that may help understand the evolution and the main characteristics of these simulation models. Several examples illustrate the concepts of the paper.

1.1 Introduction

The following pages present a selection of computational models and techniques that have been used in the last 70 years and provides an overview of how the field has evolved. In an era of cheap and fast computation, it is particularly important to look back at the history to understand the specific reasons that make current advanced techniques so remarkably relevant to social scientists.

We do not pretend to present a comprehensive overview of simulation modeling techniques and systems, but have selected those that we believe have contributed the most to build current simulation approaches. In so doing, we identify a thread, concerning *randomness*, *emergence* and *causation*, that specifies some of the most relevant characteristics of current techniques. In an attempt to make these comments as much visible as possible, they are found under the header "Intermezzo", as it interrupts the flow of the presentation and weaves together the different methods.

Before moving forward, a few words needs to be spent on these three aspects and why they are so important as to appear as the 'fil rouge' of the chapter. Let us start from *randomness*. Any simulation of a social system has to be able to reproduce elements that appear unpredictably and

[1] This chapter is included in ”Seri, R., Secchi, D., and Martinoli, M. (2020). Randomness, emergence and causation: A historical perspective of simulation in the social sciences. In Complexity and Emergence. Springer Science+Business Media.“

without any apparent connection to the phenomenon under analysis (or to an outcome variable). The reason is connected to Laplace's demon (see [286, p. 2]): an omniscient intellect having complete knowledge of all the forces and positions of the items composing the natural world as well as unlimited computing possibilities would be able to predict without error all future events. Randomness is a way of accounting for our ignorance of these initial conditions and for our computational limits. That is why a model not taking into account the possibility of unpredictable and/or external events would fall short of capturing the inherent complexity of most phenomena. The second aspect, *emergence*, is tied to the assumption that social systems are complex ([116, 161]). When this assumption holds for the computational simulation that models a social system, then uncertainty, ambiguity, and unpredictability are key features of that research effort. Some have argued that a successful simulation is one that presents the modeler with counter-intuitive and surprising results (see, e.g., [170]). We do not subscribe to this view, because it is too radical and only fits certain types of simulations. However, we can certainly support the idea that simulating a social system means to allow for a true intellectual enquiry, where results are not entirely discernible by simply looking at the code. On this respect, "a model is not a model" (opposite to what some argue in a recent editorial, see [267]). The third aspect is that of *causation*. One of the defining aspects of a computational simulation are the mechanisms that specify how its component parts behave. The interactions of these parts are reflected in the values taken by the aggregate variables describing the system that, on their turn, impact the single components. The components and the aggregates are thus linked by up- and down-ward relations. In a social system, both causal directions need to be present to explain most phenomena. The history of computational simulation in the social sciences has always bounced back and forth between these two levels, and settled on recent techniques that could account for both.

The methods we are going to present are very heterogeneous. Some of them are specified at the level of the individual, others are aggregate. Some of them require interactions between agents, others don't. Some of them are deterministic, others contain random elements. However, all of them share the same two characteristics: (a) objects (individuals or quantities) are reduced to a finite number of idealized types (that may vary in quality); (b) objects are specified by relations partially or fully connecting them together. The historical review that we are going to present will show how the two mechanisms that have governed the evolution of these methods are indeed the identification of basic units of analysis, at whatever level they are defined, and the determination of the mechanisms that connect them. The solutions that have been proposed to these two questions have led to the development of several simulation methods.

Now we come to the structure of the chapter. In Section 1.2, we review some computational experiments involving early computers. In Sections 1.3 and 1.4 we respectively review System Dynamics and Discrete-Event Simulation, two methods of inquiry considering the aggregate behavior of a system. Then, we review some Microsimulation techniques, in Economics and Political Science, in Section 1.5. Section 1.6 covers Cellular Automata, while Section 1.7 introduces Agent-Based Models. Section 8 wraps up the main conclusions.

1.2 Experiments with Early Computers

1.2.1 ENIAC

One of the first electronic computers—the ENIAC, Electronic Numerical Integrator and Computer—was built at the beginning of 1945 at the University of Pennsylvania in Philadelphia ([337, p. 125]). In the spring of 1946 at Los Alamos, Stan Ulam suggested that ENIAC could be used to resuscitate some statistical sampling techniques that "had fallen into desuetude because of the length and tediousness of the calculations" ([337, p. 126]). He discussed the idea with John von Neumann, who sent, on March 11, 1947, a letter to the leader of the Theoretical Division of the Los Alamos National Laboratory, Robert Richtmyer, with "a detailed outline of a possible statistical approach to solving the problem of neutron diffusion in fissionable material" ([337, p. 127]):

> The idea then was to trace out the history of a given neutron, using random digits to select the outcomes of the various interactions along the way. [...] von Neumann suggested that [...] "each neutron is represented by [an 80-entry punched computer] card ... which carries its characteristics," that is, such things as the zone of material the neutron was in, its radial position, whether it was moving inward or outward, its velocity, and the time. The card also carried "the necessary random values" that were used to determine at the next step in the history such things as path length and direction, type of collision, velocity after scattering—up to seven variables in all. A "new" neutron was started (by assigning values to a new card) whenever the neutron under consideration was scattered or whenever it passed into another shell; cards were started for several neutrons if the original neutron initiated a fission. ([141, p. 133])

This led Nicholas Constantine Metropolis and Stanislaw Ulam to introduce, in 1949, the name of *Monte Carlo method* ([338]) for a statistical sampling method:

> I [Metropolis] suggested an obvious name for the statistical method—a suggestion not unrelated to the fact that Stan had an uncle who would borrow money from relatives because he "just had to go to Monte Carlo." ([337, p. 127])

The name of Monte Carlo method is nowadays generally used to denote a rather heterogeneous array of techniques for solving mathematical problems by:

- reducing their solution to the computation of an expectation with respect to a random variable and

- approximating this expectation with the empirical average based on a sample of realizations of the random variable.

The simplest example is the integration of a function on a bounded domain.

Example 1.1. [Monte Carlo integration] The aim is to compute the integral of a function f defined on a bounded domain that we identify, without loss of generality, with the unit interval $[0, 1]$. The integral is $\int_0^1 f(x)\,\mathrm{d}x$. A solution is to remark that a random variable X uniformly distributed over the interval $[0,1]$ has probability density function 1, so that $\int_0^1 f(x)\,\mathrm{d}x = \mathbb{E}f(X)$. This means that, if we have a sample $\{x_1, \ldots, x_n\}$ of realizations from X, we can approximate $\int_0^1 f(x)\,\mathrm{d}x = \mathbb{E}f(X)$ through $\frac{1}{n}\sum_{i=1}^n f(x_i)$.

Modern statistical sampling methods largely predate the Monte Carlo method.

Example 1.2. [Buffon's needle] An often-misquoted antecedent is Buffon's needle, a mathematical problem requiring to compute the probability that a needle of length l, randomly cast on a floor with equally-spaced parallel lines at a distance d, lands on a line. The problem was presented by Buffon in 1733 at the Académie Royale des Sciences of Paris (see [77]) and solved in [78, pp. 100-105] (see also[285, pp. 359-360]). The original aim of Buffon was to look for an explicit solution. The probability itself is ${}^{2l}/_{\pi d}$ (when $l \leq d$) and this explains why several authors have performed the experiment repeatedly to provide, through an empirical approximation to the probability, an approximation to π. Despite the problem lends itself to a sampling solution, there is no evidence that Buffon ever tried to do that. Notwithstanding this, many authors (among which, e.g., [322, p. 120]) have written that Buffon proposed a sampling solution: what probably has misled them is the fact that [121] and [122, pp. 170-171] presented the needle problem together with another problem that Buffon studied with 2048 trials (this is confirmed by the fact that [322, p. 120] states that Buffon tried 2048 tosses in the needle problem). Other authors (see [461, p. 117]) have attributed the sampling solution to [285, p. 360]: despite Laplace is indeed talking about the limit for a large number of draws, he is probably making a reference to a frequentist argument rather than to a real sampling solution to the problem. However, the references in [121] (reprinted in [122, pp. 170-171] and [288]) suggest that the sampling approximation of π through Buffon's needle was in use as early as 1855.

1.2.2 Intermezzo

As no computer-based method for building random numbers was (and still is) known, John von Neumann went on to study algorithms to generate *pseudo-random numbers*, i.e. numbers with characteristics similar to those of random numbers. A famous but somewhat trite quotation is:

> Any one who considers arithmetical methods of producing random digits is, of course, in a state of sin. For, as has been pointed out several times, there is no such thing as a random number—there are only methods to produce random numbers, and a strict arithmetic procedure of course is not such a method. ([488, p. 36])

The interpretation that is generally given to this sentence is inaccurate. Von Neumann was not being skeptical, as often interpreted, of the usefulness of *pseudo-random numbers*. He was just suggesting that "'cooking recipes' for making digits [...] probably [...] can not be justified, but should merely be judged by their results." ([488, p. 36]). This is the main way in which *random-number generators* (RNG, though a better name would be *pseudo-random-number generators*, PRNG) are evaluated today, through batteries of statistical tests in which their behavior is compared with the theoretical behavior of true random numbers (e.g., [330, 74]). As computer simulation was, at that time, very difficult, random digits were collected in publications among which the famous *A Million Random Digits with 100,000 Normal Deviates* ([106]), published in 1955.

However, even before and around the construction of ENIAC several computational experiments were being performed using analog computers, and they often used random numbers too.

1.2.3 FERMIAC

In the 1930s, when he was still in Rome, Enrico Fermi was studying neutron transport: when a neutron (a sub-atomic particle) passes through matter, it can interact with other particles, or it can not. However, the aggregate behavior of neutrons seemed out of reach. Fermi assumed that each neutron was like an agent whose behavior was dictated by the sampling of a random number. He then computed the aggregate results for a large numbers of neutrons on a mechanical calculator:

> Fermi had invented, but of course not named, the present Monte Carlo method when he was studying the moderation of neutrons in Rome. ([438, p. 221])

According to his student Emilio Segrè (see [337, p. 128]), Fermi kept the technique secret and used it to solve several problems. His colleagues were often astonished by the precision of his computations. These back-of-the-envelope calculations contributed to create the myth of so-called Fermi estimates.

In the late 1947 the ENIAC (see Section 1.2.1) was moved at the Ballistics Research Laboratory in Maryland. Fermi built an analog simulator of neutron transport, called the FERMIAC (a pun on ENIAC; an image can be found in [337, p. 129]). In the FERMIAC, neutrons were modelled as agents in a planar region, whose behavior was affected when a material boundary was crossed.

1.2.4 MONIAC

In 1949, at the London School of Economics (LSE), Bill Phillips (Alban William Housego “A. W.” “Bill” Phillips, who later introduced the Phillips curve) built an hydraulic machine called *Monetary National Income Analog Computer* or MONIAC (see [361]). The name was a pun on “money” and “ENIAC”. It was a system of tanks and valves through which water flowed, in a simulation of money circulation in the UK economy. The example of the MONIAC looks very distant from the ones we will see below, but it contains the features outlined above: here the tanks are the objects; the valves are the relations connecting the objects.

Both the FERMIAC and the MONIAC were example of *analog* or *analogue computers*, i.e. machines using physical (electrical, mechanical, or hydraulic) phenomena expressed in terms of variables measured on a continuous scale to model a phenomenon. Analog computers were very common when no other computing method was available. The difference with respect to digital computers is that the latter use information stored in discrete form: the earliest digital computers were *program-controlled*, i.e. they were programmed by modifying the physical structure (plugs, wires, etc.) of the machine; modern computers are *stored-program*, as the program is stored in memory, without hardware modifications.

1.3 System Dynamics

System dynamics (SD) is an approach to the dynamical study of systems composed of objects in interaction. The idea of SD is to model the change over time of some quantities through feedback loops, accumulation of flows into stocks, and identification of inflows and outflows. The system is generally represented first graphically as a diagram and then mathematically as a system of differential (or difference) equations, that are then solved numerically by a computer program. Its central insight is the fact that the structure connecting the components is sometimes more important than the components themselves in determining the behavior of the system. It was founded, as a branch of systems theory ([50]), by Jay Wright Forrester in the 1950s (see, e.g., [169] or the historical accounts in [171] and [284]). At the beginning, SD was developed to analyze complex business problems, in connection to the author's position at the MIT Sloan School of Management. It has been applied to several problems ever since. SD goes through a series of steps to transform a verbal description of the phenomenon under scrutiny into a mathematical model (see, for example, [464] for a worked-out example on new product adoption).

Example 1.3. [Lotka-Volterra Model] An early example of a system of differential equation in which the elements appearing in the system can be interpreted as feedback loops is the Lotka-Volterra Model (LVM), developed by Lotka ([308, 309, pp. 92-94]) and Volterra ([485, 486]) in some seminal works. The LVM is a predator-prey model describing the dynamics of two species, i.e. predators and preys, interacting in an ecological environment. The presence of feedback loops is made clear both by Lotka (see the graphical representation in [308, pp. 411-412]) and by Volterra:[2] "conviene [...] di schematizzare il fenomeno isolando le azioni che si vogliono esaminare e supponendole funzionare da sole, trascurando le altre." ([486, p. 31]). A SD approach is in [127], while the final behavior of the system is illustrated in Figure 1.3.1.

Example 1.4. [The Limits to Growth] The Club of Rome is a think tank founded in 1968 in Rome as "an informal association of independent leading personalities from politics, business and science, men and women who are long-term thinkers interested in contributing in a systemic interdisciplinary and holistic manner to a better world." Their 1972 book The Limits to Growth ([336]) used system dynamics to study the world economy and population and raised considerable interest and concern about its sustainability.

Example 1.5. [The Lorenz system] In 1963, Edward Norton Lorenz studied atmospheric convection through differential equations (see [306]). He realized that a small change in the initial conditions could have long-term effects on the behavior of the system. The idea is to take two starting points on two nearby trajectories: moving along them, they will eventually diverge. Similar insights had already been advanced by Henri Poincaré in 1890 while studying the three-body problem and by Jacques Hadamard while studying motion on surfaces of negative curvature, but had little impact on the literature. Lorenz's discovery, instead, sparked a small revolution. It led to the identification of so-called *deterministic chaos*, *chaos theory* or, simply, *chaos*, i.e. sensitivity to initial conditions

[2] In English, "it is more effective [...] to schematize the phenomenon by isolating the actions that one wants to examine and, assuming they behave independently, irrespectively of the others" (our translation).

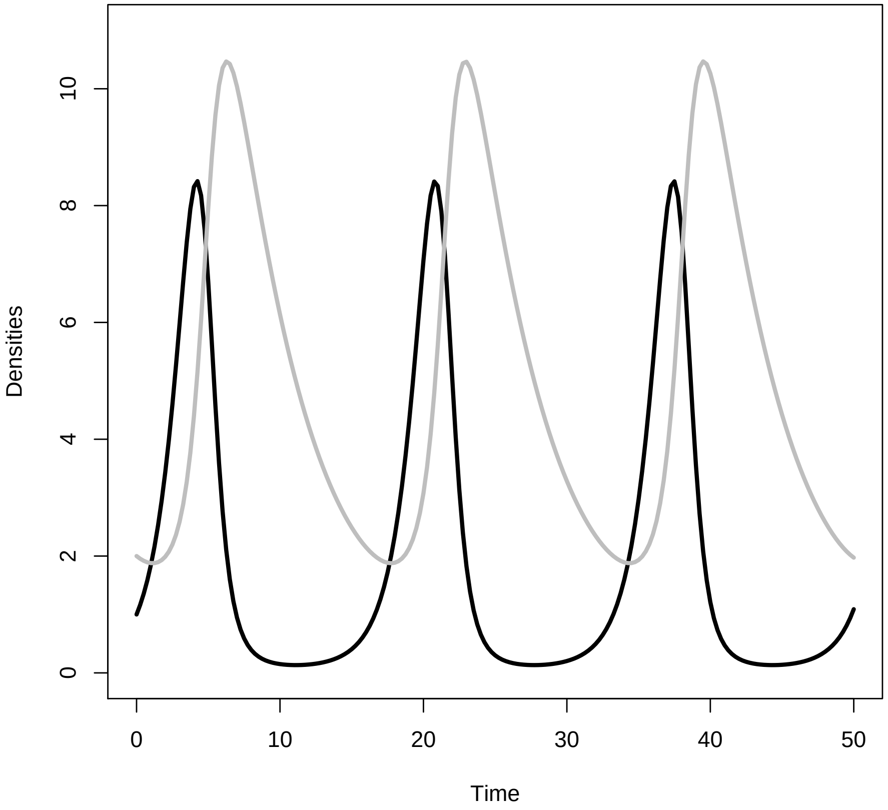

Figure 1.3.1: Population densities of the two species in the Lotka-Volterra model (predator in grey, prey in black).

in deterministic systems (often called *dynamical systems*). Lorenz coined the term *butterfly effect* for this phenomenon. An oft-quoted sentence is taken from the title of Lorenz's talk at the 139th meeting of the American Association for the Advancement of Science in 1972: "Does the flap of a butterfly's wings in Brazil set off a tornado in Texas?"

Instances of chaos in models from several domains, and SD among them, were described. As an example, in 1986, Erik Mosekilde and Javier Aracil received the Jay W. Forrester Award for their work on chaos in SD.

Example 1.6. [A Sound of Thunder] In the June 28, 1952, issue of Collier's magazine, a science fiction short story by Ray Bradbury was published under the title *A Sound of Thunder* ([68]). It described a time travel into the past whose impact on the future goes awry because of a butterfly (no spoilers). This story is sometimes miscredited with the origin of the name butterfly effect but, despite being a wonderful example of the very concept, it had no bearing on its development.

1.3.1 Intermezzo

While chaos is extremely important from a theoretical point of view, its relevance in real examples is difficult to work out:

> An essential point made by Poincaré is that chance and determinism are reconciled by long-term unpredictability. Here it is, in one crisp sentence: A very small cause, which escapes us, determines a considerable effect which we cannot ignore, and we then say that this effect is due to chance. ([419, p. 48])

The problem with chaos is that the dependence on the initial conditions makes difficult to forecast the future of the system, as initial conditions are always observed with a small error. This is why chaotic dynamical systems may be modeled as stochastic processes:[3]

> En dernière analyse, le hasard réside donc [...] dans l'œil de l'observateur. ([144, p. 14])

For us, what matters most is that chaos is a property of the system that is not shared by its components when considered in isolation. Properties like this are called *emergent*:

> The ability to reduce everything to simple fundamental laws does not imply the ability to start from those laws and reconstruct the universe. [...] The constructionist hypothesis breaks down when confronted with the twin difficulties of scale and complexity. [... A]t each level of complexity entirely new properties appear. [...] Psychology is not applied biology, nor is biology applied chemistry. [... T]he whole becomes not only more than but very different from the sum of its parts. ([11, pp. 393-395])

Emergent properties arise when the system as a whole displays a behavior that is not explicit in its single components. As put forth in [11], a system of interacting quantities/agents is not only more

[3] In English: "So, in the end chance lies [...] in the eye of the observer" ([145, p. 4]).

than the sum of its components, it is different from their sum. We will see below some examples of *emergence* (some authors use *supervenience* for a related concept).

For the moment we review some of the history of the concept. As remarked in [279, p. 49], one of the first disciplines to embrace emergence as its central phenomenon was Economics, through the work of Adam Smith:[4]

> [E]very individual [...] neither intends to promote the publick interest, nor knows how much he is promoting it. [... H]e intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention. [...] By pursuing his own interest he frequently promotes that of the society more effectually than when he really intends to promote it. ([459, p. 35])

The philosopher John Stuart Mill wrote, when dealing with failures of the principle of the Composition of Forces:

> The chemical combination of two substances produces, as is well known, a third substance with properties entirely different from those of either of the two substances separately, or of both of them taken together. Not a trace of the properties of hydrogen or of oxygen is observable in those of their compound, water. [... W]e are not, at least in the present state of our knowledge, able to foresee what result will follow from any new combination, until we have tried it by specific experiment. ([341, pp. 426-427])

The metaphor of water was a recurrent one in the work of early emergentists ([60, p. 37]). In Biology, Thomas Henry Huxley, in the book [243, pp. 16-17], introduced the idea that "there is no sort of parity between the properties of the components and the properties of the resultant": he used the "aquosity" of the oxide of hydrogen (i.e. water) as a comparison for the "vitality" of living systems, and he admonished that those who say that "the properties of water may be properly said to result from the nature and disposition of its component molecules" are "placing [their] feet on the first rung of a ladder which, in most people's estimation, is the reverse of Jacob's, and leads to the antipodes of heaven." Another often-quoted antecedent involving a different discipline is the recognition by the French sociologist Émile Durkheim that social facts cannot be reduced to the agents that are involved in them:[5]

> La dureté du bronze n'est ni dans le cuivre ni dans l'étain ni dans le plomb qui ont servi à le former et qui sont des corps mous ou flexibles ; elle est dans leur mélange. La

[4] The sentence is often misquoted replacing the obsolete "publick" with the more modern "public."

[5] The article containing this quotation became the Préface of the second edition of *Les Règles de la méthode sociologique* ([139]), and is generally quoted as such (despite the article is antecedent); the sentence is not in the first, 1895, edition. In English:

> The solidity of bronze lies neither in the copper, nor in the tin, nor in the lead which have been used to form it, which are all soft or malleable bodies. The solidity arises from the mixing of the two. The liquidity of water, its nutritive and other properties, are not in the two gases of which it is composed, but in the complex substance they form by coming together. [... Social facts] reside in the society itself that produces them and not in its parts, namely, its members. ([140, pp. 39-40])

It is difficult to say whether Durkheim was aware of Huxley's example, but he was surely well acquainted with the work of Huxley's friend, Herbert Spencer (see [146]), on social organisms. By the way, the metaphor of water is used in [146, p. 96] to describe the shorthand system developed by William George Spencer, Herbert Spencer's father.

> fluidité de l'eau, ses propriétés alimentaires et autres ne sont pas dans les deux gaz dont elle est composée, mais dans la substance complexe qu'ils forment par leur association. [... Les faits sociaux] résident dans la société même qui les produit, et non dans ses parties, c'est-à-dire dans ses membres. ([138, p. 9])

Several other historical examples are in [333, pp. 63-64, 863]; outside Biology, the economist Elinor Ostrom quoted this book ([369, p. 44]) as one of her major sources of inspiration.[6]

1.4 Discrete-Event Simulation

Discrete-Event Simulation (DES) is a kind of simulation developed in the 1950s that:

> utilizes a mathematical/logical model of a physical system that portrays state changes at precise points in simulated time. Both the nature of the state change and the time at which the change occurs mandate precise description. Customers waiting for service, the management of parts inventories, or military combat are typical application domains for discrete event simulation. ([351, p. 370] or [350, p. 149])

Strictly speaking, the name DES denotes (almost) any simulation taking place in discrete time, but this has several consequences on how it is performed. The system has finitely many components, with finitely many states. These components interact through events having no duration. In general, the state of the system is described by a *state variable.*

Example 1.7. [Queueing Systems] The most classical example of a DES is a queue. As an example, individuals from a *calling population* arrive at random times in front of one or more *servers*, servicing them in *FIFO (first in first out)* order with random serving times; if all servers are busy, a *waiting line* creates and its length is the state variable. Several queueing systems are easily solvable, others are not, and require simulation to be solved. An example may be found in [24, Section 2.1]. The two tables provide some random values for the arrival times and for the service times. Through the rules, one can get the number of customers in the system.

1.5 Microsimulation

1.5.1 Microsimulation in Economics

Microsimulation was introduced in 1957 by Guy Henderson Orcutt in [367]. Here is a definition adapted from the one provided by the International Microsimulation Association:

> Microsimulation refers to a wide variety of modeling techniques that operate at the level of individual units (such as persons, firms, or vehicles), with rules applied to simulate

[6] These pages by Mayr contain some mistakes. First, the book of Lloyd Morgan cited by Mayr is probably the one from 1923 ([347]), not from 1894, as emergence starts appearing in his work from 1912 (see [60, p. 59]). Second, the quotation just after that is not by Morgan but is taken from the book [380, p. 72] where it is used to illustrate the reasoning in [348, p. 59].

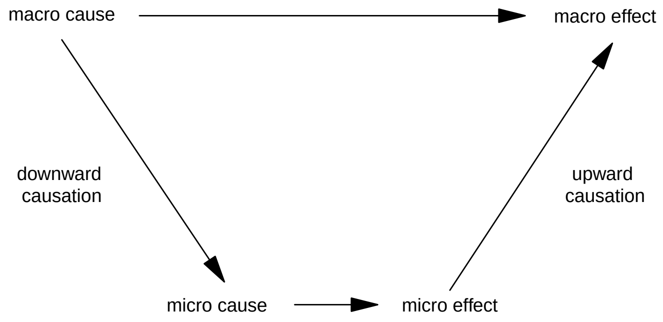

Figure 1.5.1: Coleman's boat as represented in [99, pp. 8 or 10]

> changes in state or behavior. These rules may be deterministic or stochastic, with the result being an estimate of the outcomes of applying these rules, possibly over many steps involving many interactions. These estimates are also at the micro level, allowing analysis of the distribution of the outcomes and changes to them, as well as the calculation of any relevant aggregate. ([159, p. 2142])

Microsimulation bears resemblances with agent-based modeling (see Section 1.7) but they "have remained very distinct fields in the literature with microsimulation methods drawing heavily on micro-data" ([159, p. 2142]). The reliance on micro-data for the construction of rules of behavior—that is generally considered a positive, when not the defining feature, of this method—has somewhat limited the scope of application of microsimulation to situations in which these data are available.

1.5.2 Intermezzo

What is missing from Microsimulation can be illustrated using the so-called Coleman's boat reproduced in Figure 1.5.1 and 1.5.2.

The simplest form of this diagram (see [99, pp. 8] or [52]) is shown in Figure 1.5.1. It illustrates the causal paths between micro- and macro-level phenomena: a macro- level cause influences agents at a micro-level and this in turn influences the macro-level. Here, "the macro level is an abstraction, nevertheless an important one" ([99, p. 12]). A slightly different representation, from [227, p. 59], is in Figure 1.5.2. The grey area represents the macro-level, while the white one represents the micro-level. Applications to simulation methods are in [476, p. 35], [227, p. 59] and [228].

Before turning to the explanation of these graphs, we remark that some authors (see [79, p. 454] and [327, p. 42]) prefer to refer to this representation as a *Boudon-Coleman diagram* (see [399] for a study of the antecedents of Coleman's boat) while others present modifications of the boat without an upper macro-macro path ([98, p. 1322]).

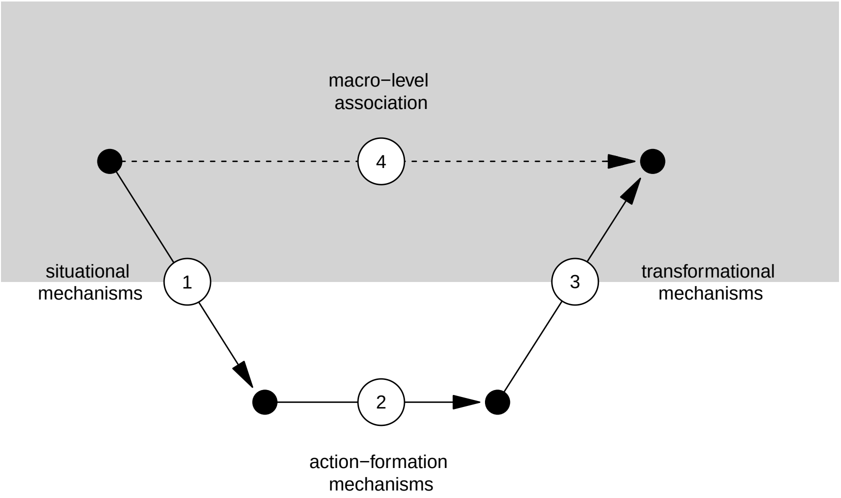

Figure 1.5.2: Coleman's boat as represented in [227, p. 59]

The mechanisms "by which social structures constrain individuals' action and cultural environments shape their desires and beliefs" ([227, p. 59]) represented by path 1 are called *situational.* They represent *downward causation. Action-formation* mechanisms (path 2) "[link] individuals' desires, beliefs, etc., to their actions" ([227, p. 59]). The mechanisms "by which individuals, through their actions and interactions, generate various intended and unintended social outcomes" ([227, p. 59]) are called *transformational* (path 3). They represent *upward causation.* Path 4 does not represent causality as "explanations that simply relate macro properties to each other [...] are unsatisfactory" ([227, p. 59]). One can iterate the "boat" over time: at each step of the simulation, there is an upward causation path from the agents towards the macro-level, and a downward causation path from the macro-level to the behavior of the agents.

The link of emergence with Coleman's boat is that emergent phenomena can generally be identified with macro behaviors induced by the mechanisms taking place in the bilges of the boat. Transformational mechanisms are especially relevant as they are the final step through which upward causation generates emergent phenomena. It is no surprise, therefore, that simulation methods built from the characteristics of the single agents may be better at modeling emergence (see [445, p. 230]).

Now, returning to Microsimulation, Economics has at least two mechanisms of individual market coordination that are coherent with both upward and downward causation: *general equilibrium* and *partial equilibrium.* The former takes place when equilibrium between demand and supply is achieved on all markets inside an economy at the same time, and changes in one market affect all other markets. The latter happens when one market is considered in isolation and is supposed not to affect the other markets. Both mechanisms predict that individuals, without coordination but only through *tâtonnement*, select prices achieving a macro-level equilibrium characterized by *market clearing*, i.e. full allocation of goods in any market of the economy. As individuals face these prices, this constitutes a source of downward causation, from the macro level to the micro one. But the application of partial and general equilibria in simulated models has two problems. First, it is not credible that these concepts of equilibrium hold exactly true, as perfect market clearing seems to be the exception rather than the norm. Second, in models representing a proper subset of the economy, it is difficult to imagine quantitative mechanisms of downward causation.

Coleman's boat can also be useful to classify simulation models. Indeed, some authors ([188, 317]) identify three categories of models:

- Macrosimulations (e.g., System Dynamics, see Section 1.3, Discrete-Event Simulation, see Section 1.4) focus on an aggregate level and operate at the level of the deck of Coleman's boat;
- Microsimulations (e.g., Microsimulation, see Section 1.5.1, Simulation of Voting Behavior, see Section 1.5.3) focus at the individual level and take place in the bilges of Coleman's boat;
- the third category is composed of models in which there is an iteration between the two levels. These are identified with so-called Agent-Based Models (see Section 1.7).

1.5.3 Early Simulation of Voting Behavior

Some models similar to economic Microsimulations can be found in the early literature on simulation of voting behavior. We include these models in this review because of their accent on agents' heterogeneity.

[392] presented a model they had developed for the Democratic Party in the 1960 US presidential campaign, the so-called Simulmatics project. The original model used the positions of 480 types of voters (that were consolidated to 15 in the published paper) on 52 issues. The data for the model were based on over 100,000 interviews in polls collected over 10 years by polling firms. The researchers advised Kennedy that he would benefit from taking a strong stance in favor of civil rights and from openly dealing with his Catholic religion. The paper was so influential that in 1964 Eugene Leonard Burdick, a political scientist and novelist, wrote a novel, called The 480 (see [80]), criticizing the fact that the use of computer models made easy to choose strategies to maximize votes and manipulate electors.

In 1965, Ithiel de Sola Pool, Robert P. Abelson and Samuel L. Popkin published *Candidates, Issues, and Strategies: A computer simulation of the 1960 and 1964 presidential elections*, a book describing in detail their model (see [393]). [1] considered a simulation model of voting in the fluoridation referendum (i.e. whether tap water should be compulsorily fluoridated or not). The model had 500 agents behaving according to 51 rules (22 about information processing, 27 on information exchange, 2 for voting behavior).

1.6 Cellular Automata

Cellular automata (CA, sing. *automaton*) are systems composed of individuals taking on one of a discrete number of states, arranged in fixed cells (hence the name *cellular*) on a grid, interacting according to deterministic rules depending on the neighboring agents' state (hence the name *automata*). A more formal definition is the one in the Stanford Encyclopedia of Philosophy:

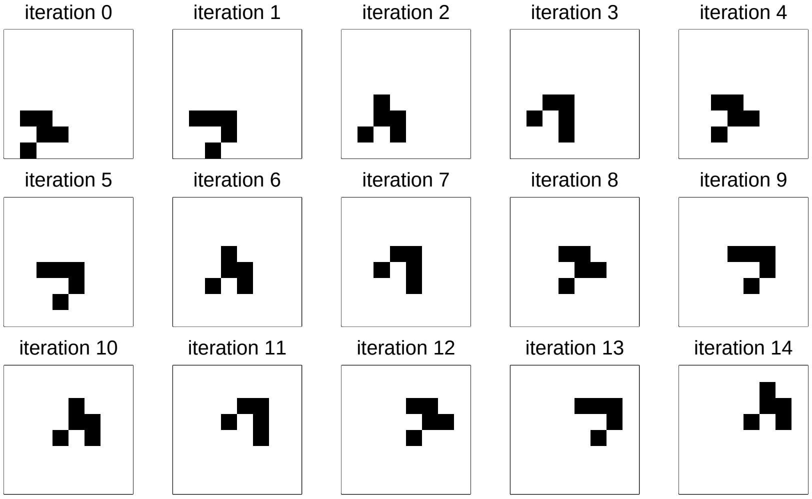

Figure 1.6.1: Game of Life: evolution of a glider on a 8 × 8 checkerboard.

> CA are (typically) spatially and temporally discrete: they are composed of a finite or denumerable set of homogeneous, simple units, the atoms or cells. At each time unit, the cells instantiate one of a finite set of states. They evolve in parallel at discrete time steps, following state update functions or dynamical transition rules: the update of a cell state obtains by taking into account the states of cells in its local neighborhood (there are, therefore, no actions at a distance). ([52])

They have been used both as specific examples of real-world phenomena and as abstract examples of how complex behavior can arise from simple rules. Note that the rules of behavior of cellular automata are deterministic and fixed. They were first formalized by Stanislaw Ulam and John von Neumann in the 40s, while the former was working on the growth of crystals and the latter on self-replicating systems. The work of von Neumann culminated in the classic [487] (note that the symposium for which the paper was written was held in 1948), while the work of Ulam dates [478]. However, it was only in the 1970s that CA rose to prominence with the following example.

Example 1.8. [Game of Life]

In 1970, in [181], Martin Gardner popularized "a fantastic solitaire pastime" invented by John Horton Conway. This is indeed a cellular automaton with very simple rules:

- each cell can be either occupied by a living creature or empty;
- the creature in the cell dies if it has 1 or 4+ neighbors (resp. of loneliness and overcrowding);
- an empty cell comes to life if it has 3 living neighbors.

The neighbors of a cell are the ones in the *Moore neighborhood*, i.e. the set of 8 cells in contact through a side or a corner with the cell (a *von Neumann neighborhood*, instead, is the set of 4 cells

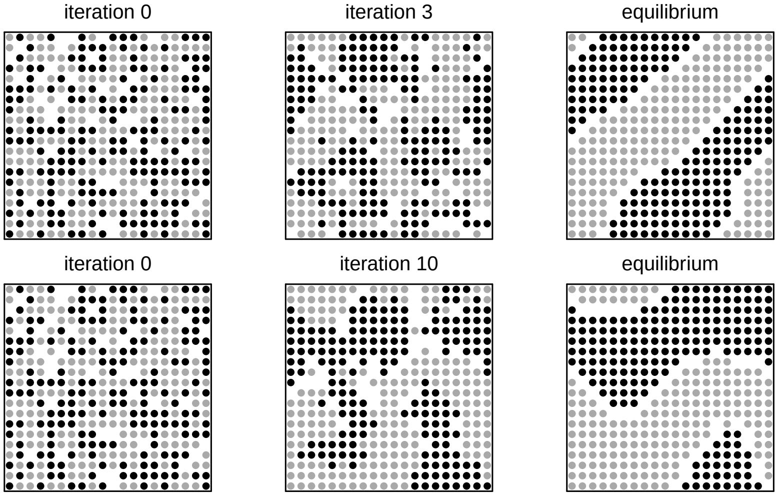

Figure 1.6.2: Schelling Segregation Model for different values of x: the row above shows the case $x = 0.5$, the row below the case $x = 0.9$; the left column displays a random initial configuration that is equal for both values of x; the second column shows the model after 3 (above) and 10 (below) iterations; the last column shows what happens after an arbitrarily large number of iterations.

in contact through a side with the cell). The game is generally applied starting from a configuration of activated cells. At the beginning, Conway thought that such a system could not create a universe in constant expansion, but he was soon proved to be wrong (see the *glider* in [45, p. 131] or Figure 1.6.1 and the *glider gun* in [45, p. 935]). The number of configurations that have been explored is incredibly large (see [45, Chapter 25]).

Around the same years, Thomas Crombie Schelling introduced a model dealing with segregation, i.e. the enforced separation of different ethnic groups in a community. By taking inspiration from James Sakoda, who created a set of so-called checkboard models (see [229] for the detailed story), Schelling's *Segregation Model* ([427, 428]) showed that a personal slight preference towards a less diverse neighborhood could create in the long run a segregated community.

Example 1.9. [Schelling Segregation Model]

Schelling ([427, 428]) proposed a model in which two types of individuals, say, A and B, are located on a one-dimensional or two-dimensional grid. Some of the cells may be empty. At each step, each individual counts how many in their Moore neighborhood are like them: if the proportion is smaller than a threshold value x, they move to a new position. This position is chosen deterministically: it is the nearest empty position satisfying their threshold. Schelling did not quote explicitly cellular automata in his paper, but to keep the paper inside the framework was compelled to introduce awkward deterministic rules (as an example, individuals choose to move according to a certain order in the grid). However, this is not the version of Schelling's model that is generally used: in the latter, some randomness is generally introduced in the relocation of moving individuals. This small step gets this cellular automaton near to agent-based models (see Section 1.7). This instance of the model is displayed in Figure 1.6.2. The evolution of the system for different values of x has been characterized:

> Initially the system quickly develops small clusters, but then a slow evolution toward larger clusters follows. [... T]he system evolves toward one big cluster or very few clusters. In the case of $x = 1/2$ the cluster surface tends to form flat surfaces [...] In $x \neq 1/2$ cases the surface is bumpy and irregular ([484, p. 19263])

1.6.1 Intermezzo

In this model, segregation is an example of *emergence*. We discuss in the following the implications of emergence and its importance for the model and for the development of simulation in the social sciences.

Example 1.10. [Schelling Segregation Model] In Schelling's model, segregation is an emergent property, as nobody necessarily wants it to take place. Schelling stated this opposition in the famous title of one of his books, *Micromotives* and *Macrobehavior* ([429]), whose blurb summarizes the idea as follows: "small and seemingly meaningless decisions and actions by individuals often lead to significant unintended consequences for a large group." The description in the blurb replicates the definition of the upward causation path in Coleman's boat.

Note the striking difference with respect to two economic frameworks that gained traction in the last decades.

On the one hand, in modern Macroeconomics, several models are based on a *representative agent*, i.e. an agent that represents the whole economy (see, e.g., [219]). This became a central element of *Real Business Cycle* (*RBC*) theory, first, and, after that, of *Dynamic Stochastic General Equilibrium Models* (*DSGE*) introduced in the two seminal papers [280] and [304]. DSGE are, in general, macroeconomic models featuring an economy in general equilibrium; moreover, the models are microfounded, i.e. they are not formulated in terms of aggregate quantities but they derive their behavior by aggregating microeconomic individual models. In the case of DSGE, their behavior can be reduced to that of a representative agent maximizing expected utility. In [215], the authors provide an interesting point of view focused on the causal structure of these dynamic models:

> These types of models are nowadays the most widely used to draw and to evaluate policy claims because they bear the advantage of simultaneously addressing two critical issues about causal structures. On the one hand, under the acceptance of the rational expectation hypothesis, the structure modeled by the RBC/DSGE approach remains invariant under policy intervention because it takes into account the forward-looking behavior of the economic agents. On the other hand, the theoretical structure has an empirical counterpart in which the distinction between endogenous and exogenous variables is eschewed. ([215, p. 126])

The emphasis on the representative agent implies that any characteristic of the economy is a characteristic of the agent, and no emergence seems possible (see [279, p. 51]).

On the other hand, in Microeconomics, as well as in other Natural and Social Sciences, some market and non-market interactions, in which tactical and strategic factors are preeminent, are studied through the lens of *game theory*, a branch of economics/mathematics introduced by Oskar Morgenstern and John von Neumann ([489]) in which agents interact taking into account other agents' reactions. Here emergence is possible as a consequence of strategic interaction between the agents.

These two situations describe a whole spectrum of models, from one in which no emergence is possible to one in which emergence is a consequence of strategic interaction. But Schelling's model is different as emergence is a consequence of the myopic behavior of individuals in a dynamic context. There is no planning at all. (One could even show that, if x is very high, no stable equilibrium is possible: if individuals have strong preferences against diversity, they do not get what they want!)

1.7 Agent-Based Models

An *Agent-Based Model* (henceforth *ABM*) is a computational model whose unit is the *agent*, an autonomous individual behaving in a given *environment* according to established *rules* ([432]). The agent is the unit of analysis and it can be anything the modeler is interested in, from a neutron to a country. Its general features can be characterized by:

- *autonomy*: each agent is modeled independently from the others and it can develop in ways that are not predictable solely by looking at the initial conditions set;

- *interaction*: interactions with other agents may modify the characteristics of the agent and the way in which it perceives the environment and the self;

- *complexity*: some characteristics "emerge" during the interactions.[7]

Agents interact in a limited space that is usually referred to as *environment*, such that the position of agents can represent either their physical location (see Schelling's Segregation Model in Section 1.6) or their psychological state of mind (see the Garbage Can Model below).

The *rules* are the norms that regulate what happens in the model and are sometimes identified as mechanisms. They can be:

- *behavioral*: they define what each agent should be doing in general and/or as a function of their characteristics;

- *interactional*: they define what happens to an agent and/or to the environment when they interact;

- *time-dependent*: rules may modify agents' characteristics, other rules, or the shape of the environment as time—however defined in the simulation—goes by;

[7] It is a bit odd to attach this aspect to agents, but we want to highlight that agents can be characterized as complex; see below and [142].

- *developmental*: rules set the conditions for agents (and/or the environment) to change, evolve or die.

The next step we deem appropriate at this point in the chapter is to try and explain the difference between ABM and other simulation frameworks, especially because ABM is the latest and most advanced of all known techniques so far. A classification of simulation models can be based on the following dichotomies:

- The backings of the model can be based on equations or on properties of the objects.
- The approach can be either at the macro- or at the micro-level.
- Agents can be homogeneous or heterogeneous.
- Rules can be homogeneous or heterogeneous.
- The environment can be either static or dynamic.

While the other concepts have already been explained, it can be interesting to spend a word on backings. Most simulation models involve behaviors that are dictated by equations that connect the different elements of the model. This is clearly true for simulation models working at the aggregate level, but also some microsimulation models are based on actions described through the application of equations on individual-level variables. Some models, however, start from the specification of the characteristics of the objects, be they agents and/or rules, and let them interact more or less freely. A first consideration is that this further increases the distance between the observed behavior of the components of the model and the elements governing it. Indeed, while the result of an equation is often rather predictable, it is not the case for the interactions of objects possessing their own characteristics and behaving on their basis. For this reason, models whose backings are based on objects often offer more opportunities for the development of those features that are not possessed by their own components but that are born out of the interactions, i.e. exactly emergent properties. This is not to say that equation-backed models cannot exhibit emergence—as the examples above show, they can and they often do—but only that emergence is often more unpredictable in object-backed models. A second, related, point is that object-backed models are often built using a *bottom-up* approach, i.e. starting from the properties of the objects, and they recover aggregate features only as a result of the interactions. From this point of view, they start from the bilges of Coleman's boat (see Figures 1.5.1 and 1.5.2) but they also involve its deck.[8]

From Table 1.7.1 it is probably more apparent to understand why ABM is considered the most advanced computational simulation approach as of yet. In fact, by comparing core components of the simulation approaches reviewed so far, it becomes clear how ABM stands at odds with most of them. Of course, the agent-based approach has taken from past simulation techniques, but its

[8] It is worth noting that ABM can also be based by equations, better, by a mix of equations and object-based modeling. Actually, we are not aware of ABM that do not have any equation embedded in their coding. The difference of this approach is in the ability to mix and mash both object and equation-based techniques.

Table 1.7.1: A comparison of computational simulation models.

	Backings		Approach		Agents		Rules		Environment	
	Equation	Object	Macro	Micro	Homogeneous	Heterogeneous	Homogeneous	Heterogeneous	Static	Dynamic
Early Comp.	X		X	X	X		X		X	
SD	X		X			X	X		X	
DES	X		X		X		X		X	
Microsim.	X			X	X		X			X
Voting Models	X			X	X		X		X	
CA	X	X		X	X		X		X	
ABM		X	X	X		X		X		X

comprehensive reach makes it stand as a jump ahead. Omitted from the table and not explicitly mentioned (only cursorily in the introduction) in this chapter so far are the surrounding conditions that make ABM a viable option. We are referring to the surge of computational power and to the possibility that even home computers are capable of performing complicated operations, unthinkable twenty years ago. This means that, for example, having heterogeneous agents in a simulation came at very high costs before the middle of the 1990s, while it is relatively (computationally) cheap to allow them today. This technical hardware and software innovation opened up for the possibility of a different approach to simulation modeling.

In the following pages, we introduce an example concerning the ABM version of the celebrated Garbage Can Model.

Example 1.11. [Garbage Can Model] The *Garbage Can Model* (GCM) of [96] is a model of decision making in an *organized anarchy*, i.e. an organization characterized by the three properties of problematic preferences, unclear technology, and fluid participation.

> [A]n organization is a collection of choices looking for problems, issues and feelings looking for decision situations in which they might be aired, solutions looking for issues to which they might be the answer, and decision makers looking for work. ([96, p. 2])

The GCM has been extremely influential in organizational behavior. The model was implemented using one of the earliest computer languages developed by IBM, FORTRAN, and it was not an ABM. Yet, by using an ABM jargon, four types of agents can be identified: (a) problems, (b) opportunities, (c) solutions, and (d) decision makers. The overall goal of the model is to determine whether a formal (hierarchic) organizational structure provides the institutional backbone for problem solving that is better than an informal (anarchic) organizational structure, or not. In the first case, the four types of agent interact following a specified sequence while in the other they interact at random. There are two ways in which participants make decisions in the organization. One is *by resolution*: it happens when problems are solved once participants match opportunities to the right solutions, i.e. when the right combination of the four agents are on the same patch at the same time. The other is *by* *oversight*: it is when solutions and opportunities are available to participants but no problems are actually solved.

Not all problems are solved automatically, just by having opportunities, decision makers, and solutions available. In fact, all problems have difficulty levels, participants have abilities, and solutions have a certain degree of efficiency. The problem is solved if the match of the participant with an opportunity and a solution is greater than its difficulty.

The findings of the original model are the following: resolution is not the most common style of decision making; hierarchies reduce the number of unresolved problems but increase problem latency; important and early problems are more likely to be solved.

The model has later been implemented by [162, 163] in ABM form. In the agent-based version of the model, there are three types of structures:

1. *Anarchy.* There is no hierarchy so that abilities, efficiencies, and difficulties are randomly distributed among agents.

2. *Hierarchy-competence.* Abilities, efficiencies, and difficulties increase as one moves up the hierarchical ladder.

3. *Hierarchy-incompetence.* Abilities, efficiencies, and difficulties decrease as one moves up the hierarchical ladder.

Finally, the model implements two modes of (not) dealing with problems (i.e. *flights*):

1. *buck passing*: when one participant has the alternative of passing the decision on a problem to another participant;

2. *postpone*: when problems are kept on hold by participants and eventually solved at an unspecified future time.

The results of the model are summarized as follows by the authors:

> The [...] properties point to very interesting features of organizational decision-making. [...]
>
> 1. Decisions by oversight are very common, much more common than decisions made in order to solve problems. This result suggests that the rational mode of decision-making is a very rare case. Most decisions are socially induced acts, made with the purpose of obtaining legitimacy by conforming to required rituals.
>
> 2. If there is a hierarchy, then top executives are busy with gaining legitimacy for their organization by means of decisions by oversight, whereas the bottom line cares about solving problems.
>
> 3. Organizations make themselves busy with a few problems that present themselves again and again. So participants have the impression of always facing the same problems. ([163, p. 123])

1.7.1 A Menagerie of Names

ABM have been used in several disciplines: Biology/Ecology ([210, 209, 397]), Computer Science ([128, 362]), Sociology ([20, 317, 477, 462]), Management ([3, 347]) and Organizational Behavior/Organization Science ([30, 160, 436]), Political Science ([119]), Psychology ([102, 103]) and Cognitive Sciences ([318]), Population Studies ([212]), Economics ([470, 21]) and Finance ([493]), Transportation Research ([319]).

As expected, their very general nature implies that ABM can be adjusted in several ways according to the discipline. The different definitions of ABM are exposed in the following parts of the chapter. Since Social Sciences include many different disciplines, we decided to restrict our attention only to a part of these and to compare them with a selection of other disciplines, from the area of Natural Sciences and Techniques (e.g., Physics, Engineering, and Biology). In particular, we dedicate special attention to Economics and Management, highlighting the peculiarities of agents as heterogeneous individuals, with bounded rationality, interacting among themselves.

Biology/Ecology

In Biology, an ABM is sometimes called an *Individual-Based Model* (*IBM*), but there is considerable confusion as to the definition. Some authors consider individual-based and agent-based as synonyms (see, e.g., [397, p. 3]). Others reserve the term individual-based when individuals are simpler and rules are formulated probabilistically at the individual level (see, e.g., [59, p. 338]).

Example 1.12. [Conservation Biology] In this branch of biology, and especially in *population viability analysis* (*PVA*), i.e. the quantitative assessment of risk of extinction concerning a population or a species, most classical models fail to take into account the spatial nature of habitats: spatially dispersed animals, like the giraffe (see [107, 246]), may be at higher risk than commonly thought. *Metapopulation models* were the first to con- sider several populations interacting in separated locations, but they generally constrain the locations to be discrete. ABM offer a continuous improvement.

Computer Science

In Computer Science, one finds the concept of *Multi-Agent System* (*MAS*, see [128] for a survey and [362] for the relation between MAS and ABM), used to denote a computer system of intelligent agents in interaction inside an environment.

The aim is not to study the behavior of agents, but to solve a problem or compute a quantity. An agent is seen here only as a computing entity (and this is why, at odds with most ABM, it must be intelligent). As agents share knowledge or computing power, they approach the solution of the problem. From this point of view, MAS can be seen as a subfield of *Distributed Artificial Intelligence* (*DAI*):

> DAI is the study, construction and application of multiagent systems, that is, systems in which several interacting intelligent agents pursue some set of goals or perform some set of tasks. ([496, p. 1])

Another related, but slightly more general concept, is that of *Artificial Adaptive Agents* (*AAA*). This name is often connected with so-called *Complex Adaptive Systems* (*CAS*, e.g., [343]):

> Such a system is complex in a special sense: (i) It consists of a network of interacting agents (processes, elements); (ii) it exhibits a dynamic, aggregate behavior that emerges from the individual activities of the agents; and (iii) its aggregate behavior can be described without a detailed knowledge of the behavior of the individual agents. An agent in such a system is adaptive if it satisfies an additional pair of criteria: the actions of the agent in its environment can be assigned a value (performance, utility, payoff, fitness, or the like); and the agent behaves so as to increase this value over time. A complex adaptive system, then, is a complex system containing adaptive agents, networked so that the environment of each adaptive agent includes other agents in the system. ([236, p. 365])

What differentiates CAS from ABM is the emphasis, that is generally lacking in the latter, on adaptivity, but the two approaches are not mutually exclusive (see [364] for an example).

Physics

Physics does not generally require the introduction of agents as sentient and autonomous entities. Optimization provides an example of the difference between social sciences and physics:

> In physics and the natural sciences, maximization typically occurs without a deliberate "maximizer." [... M]aximizing behavior differs from nonvolitional maximization because of the fundamental relevance of the choice act. Fermat's "principle of least time" in optics was a fine minimization exercise (and correspondingly, one of maximization). It was not, however, a case of maximizing behavior, since no volitional choice is involved (we presume) in the use of the minimal-time path by light. ([439, p. 745])

A notable exception is the area of complex systems, that is at the border of Physics. Outside this area, the generic name of Monte Carlo is used for models that would elsewhere give rise to ABM.

Example 1.13. [Why the Brazil Nuts are on Top] According to physical intuition, when we shake a box of nuts, the largest nuts should go on the bottom while the smallest should float; the evidence suggests that the contrary is true. In [415, 416] the authors build a Monte Carlo model to show how and when this happens.

Economics

In Economics, the most appreciated features of ABM are that they allow for heterogeneity of the agents, both in their types and their characteristics, for interactions taking place in non-trivial, often dynamic, networks ([186]), and for a wide range of individual behaviors, from perfect foresight to bounded rationality ([394]).

The name *Agent-based Computational Economics* (*ACE*) is sometimes used to denote the computational study of economic processes modeled as dynamic systems of interacting agents. According to [290, p. 246], an agent refers, in general, to "an encapsulated collection of data and methods representing an entity residing in a computationally constructed world."

ACE may overcome some critiques moved to DSGE models in macroeconomics (see [262, 263, 82]):[9]

> The advantage of the ACE approach for macroeconomics in particular is that it removes the tractability limitations that so limit analytic macroeconomics. ACE modeling allows researchers to choose a form of microeconomics appropriate for the issues at hand, including breadth of agent types, number of agents of each type, and nested hierarchical arrangements of agents. It also allows researchers to consider the interactions among agents simultaneously with agent decisions, and to study the dynamic macro interplay among agents. Researchers can relatively easily develop ACE models with large numbers of heterogeneous agents, and no equilibrium conditions have to be imposed. Multiple equilibria can be considered, since equilibrium is a potential outcome rather than an imposed requirement. Stability and robustness analysis can be done simultaneously with analysis of solutions. ([97])

A second difference between ACE and DSGE concerns the agent's expectations, that in DSGE are generally *rational*. This is a mechanism of expectation formation introduced in [349], according to which ex-ante expectations concerning the future value of a variable differ from its real value by a zero-mean random term. This can be justified supposing that agents know the model representing the economy, from which the alternative name of *model-consistent expectations*, i.e. individuals and researchers share the same model of the economy: "Muth's notion was that the professors, even if correct in their model of man, could do no better in predicting than could the hog farmer or steelmaker or insurance company. [...] The professors declare themselves willing to attribute to economic actors at least as much common sense as is embodied in professional theories." ([334, p. 53]) Now, this kind of *model-consistent rationality*, according to which agents are able to analyze the economy as economists, cannot be generally assumed in ACE (and more generally in ABM) because of the very way in which models are built.

ACE are widely used in Finance, although, as pointed out by [238, 289, 239, 126], in Financial Economics several features of ABM are not used (interactions taking place over networks, coexistence of several kinds of agents, etc.) and the attention is more focused on the heterogeneity and bounded rationality of consumers.

Another class of models is formed by *history-friendly models* (HFM, see [323, 324]):

> "[H]istory-friendly" models (HFMs) of industrial evolution [...] are variants of ABMs which aim at capturing in stylized form qualitative theories about mechanisms and factors affecting industrial evolution, technological advances and institutional changes. HFMs consist of three steps: appreciative theories of the history of a specific industry, history-replicating simulations, and history- divergent simulations. In HFMs, model building and calibration are conducted with the guidance of the history. ([504, p. 45])

[9] We refer to [154, 295, 264, 153, 216] for a detailed discussion on the differences between ACE and DSGE.

As explained above, an important step in HFM is *calibration*, namely the search for values of the parameters of the model producing an output that is approximately similar to a set of real data.

Management and organization research

The status of ABM among the management disciplines is still controversial. In fact, and in spite of the few attempts made so far ([160, 344, 432, 437]), there is still no clear definition of what ABM are for scholars in the area of Management and Organizational Research (MOR). If one excludes the more engineering-related area of MOR, that is Operations and Supply Chain Management, there are very few examples of ABM.

In a review of the literature, Wall ([492]) divides the contributions in two groups, those related to exploration/exploitation and those dealing with differentiation and integration. Models pertaining to the former originate from James G. March's famous categorization of possible opposite decisions an organization usually faces ([328]) between—to make a very long story short—putting existing resources to work (exploitation) or seek additional resources (exploration). The latter dichotomy is engrained into very old discourses within the MOR literature, and relates to the basic decision to "make or buy" ([287]). One of the most interesting findings of the review—although not discussed openly—is that almost all models are of a special kind: they are NK models (see below). According to another recent study ([35]), it seems there is a trend in MOR where scholars engage in one of the simplest kinds of ABM while (a) not calling them as such, and (b) de facto establishing a parallel literature. Using NK models to address MOR topics is probably a sign that the field is struggling with concepts such as complexity, emergence, and randomness.

But, outside of this quite remarkable trend, there still are areas of MOR where ABM have appeared. This is the case, for example, of those ABM dealing with routines (e.g., [346, 71]), work team dynamics (e.g., [331, 30]), and, more broadly, with organizational behavior ([436]).

Example 1.14. [Adaptation on a Rugged Landscape] Given their prominence in MOR, it makes sense to introduce NK models in an example. Before shortly describing the example, it is worth dedicating a few words on this typology of models. They were introduced by evolutionary biologist Stuart A. Kauffman in the late 1980s ([256, 257]) to study fitness and adaptation. Using Kauffman and Weinberger's words:

> The distribution of the fitness values over the space of genotypes constitutes the fitness landscape. [...] The space consists of all 20 N proteins, length N, arranged such that each protein is a vertex next to all 19 N single mutant variants obtained by replacing one amino acid at one position by one of the 19 remaining possible coded amino acids. Each protein in the space is assigned some "fitness" with respect to a specific property, such as binding a specific ligand, where "fitness" can be defined as the affinity of binding. An adaptive walk can be conceived as a process which begins at a single protein in the space and passes via ever fitter 1-mutant variants. [...] N is the number of "sites" in the model genotype or protein, while K is the number of sites whose alternative states, "alleles" or amino acids, bear on the fitness contribution of each site. Thus K measures the richness of epistatic interactions among sites. ([257, pp. 211-212])

Outside of Biology, a modeler could attribute a diversity of characteristics to the two main parameters N and K, preserving their relations, and adapting to the study of different types of “fitness”. This was the intuition of Daniel A. Levinthal ([297]) who was probably the first to introduce the MOR community to NK models. He studied how organizations adapt to different forms (organizational design). In the model, there are N organizational attributes and K other attributes that affect an organization’s fitness, i.e. the extent to which interaction affects adaptation. Dependence on initial conditions was one of the main findings of this simulation model.

1.8 Conclusions

The aim of the chapter was to show the central role played by randomness, emergence and causation for the development of different groups of simulation models. According to the state of the art of the literature, we have outlined a short and necessarily partial history of simulation models, with special attention to Social Sciences. The models that we have covered are some early works with analog and digital computers, System Dynamics, Discrete-Event Simulation, Microsimulation in Economics and Political Science, Cellular Automata and Agent-Based Models.

To conclude, ABM can be considered the most advanced computational simulation approach so far. Indeed, although this approach has taken from past simulation techniques, its comprehensive reach makes ABM stand as a jump ahead.

Chapter 2

Estimation, calibration and validation of simulated models: a literature review

The aim of this chapter is to provide a comprehensive literature review on the estimation, calibration and validation methods for simulation models. We start exposing the definition of estimation and we examine the dependence of the simulated data on the parameters to be estimated. Then, we give an interpretation of the meanings of calibration, estimation and validation of simulated models. Subsequently, we consider the classical approach for simulation-based econometric models, and we detail the characteristics of indirect inference, method of simulated moments, simulated minimum-distance, simulated maximum likelihood and approximate Bayesian computation. We compare the different methods and we explain the main advantages and weaknesses of these techniques. Finally, in the last section, we focus on the estimation, calibration and validation of agent-based models.

2.1 Introduction

The choice of the parameters of an economic model is generally carried out via a set of techniques that go under the name *estimation.* The preliminary step of the estimation process is *identification*, that is the search of conditions under which the distribution of the data at the true parameter is different from that at any other possible parameter (see [355, p. 2124]). This definition is the classic one and is called *point* or *global identification* (see [270, 417]).[1] This assumption represents the cornerstone of the whole estimation process, and must hold true in order to prove the consistency of the estimator, e.g., the convergence in probability of the estimator to the true value.

According to [355, pp. 2114-2115], many estimators belong to the general class of *extremum* estimators (see [164, 165, 491, 242, 247, 325]), and rely on the optimization of an objective function

[1] Many other definitions are given in the literature (i.e., set identification, causal identification, local identification, generic identification, weak identification, etc.); a detailed discussion can be found in [298].

depending on the data. In this class, we include *maximum likelihood* (see [164, 165, 491]), the *generalized method of moments* (see [221]), the *minimum-distance method* (see [389, 326]) and *nonlinear least squares* (see [247, 325]).

When the objective function cannot be evaluated analytically or numerically (see [465]), as an example because the model is characterized by an intractable likelihood function (i.e., lack of closed-form solution), the above-listed estimators are useless, and the researcher has to invoke their simulated counterparts. In these methods, additional data are simulated by the researcher and used to compute a surrogate objective function for each one of a series of values of $\boldsymbol{\theta}$.

A problem concerning the estimation of simulated models is whether or not the dependence of the surrogate objective function on $\boldsymbol{\theta}$ is smooth enough. The condition that is generally used to achieve this is called *stochastic equicontinuity* (see, e.g., [335], [372] and [355, pp. 2136-2137]) and it boils down to the requirement that the (pseudo-)random numbers involved in the approximation of the surrogate objective function are the same for different values of $\boldsymbol{\theta}$. This implies that the surrogate objective function depends continuously on $\boldsymbol{\theta}$, while its violation implies that the function is "rough" or "rugged". This phenomenon, called *chattering*, was outlined, e.g., by [335, p. 999]. In his contribution, the author specifies that "a simulator must avoid 'chatter' as $\boldsymbol{\theta}$ varies; this will generally require that the Monte Carlo random numbers used to construct $f(\boldsymbol{\theta})$ *not* be redrawn when $\boldsymbol{\theta}$ is changed". This concept was then addressed by [195, p. 16]. The authors pointed out that "[w]hen $\boldsymbol{\theta}$ changes it is possible to keep the same drawing [of the random numbers ... I]t is *necessary* to keep these basic drawings fixed when $\boldsymbol{\theta}$ changes, in order to have good numerical and statistical properties of the estimators based on these simulations" (emphasis of the authors). This was also stressed by [272, p. 78] and [143, p. 346] in two more recent contributions. However, the recommendations given by [335] and [195] do not hold for some specific simulation-based models such as agent-based models (ABM).

There are two different aspects related to chattering. The first is a theoretical aspect, linked to the limit of the objective function in an optimization problem; the second is a computational problem linked to the type of optimization routine that is used, whether it is derivative-free or derivative-based. Let us start from the second problem. Consider a derivative-based algorithm. As, based on the theoretical problem, the function is rugged and nowhere differentiable, the derivatives do not exist. However, as most optimization algorithms generally calculate numerical derivatives using the representation of the derivative as a limit of the incremental ratio, the algorithms end up using finite derivatives whose value is random and determined by the roughness of the function at the point. It is not clear whether and how an optimization algorithm using numerical derivatives in this situation converges. Now, suppose we use derivative-free algorithms. In this case, we can imagine that the algorithm converges to the global minimum of the objective function seen in the previous theoretical point. But there is no proof that the asymptotic distribution of this theoretical limit does not depend on chattering. This theoretical problem persists even in the absence of computational issues.[2]

[2] In Chapter 5 and Chapter 6 of the thesis we provide two estimation frameworks that do not involve any optimization function, in order to deal with the violation of the stochastic equicontinuity hypothesis, and we discuss their asymptotic properties.

Due to the problems related with the estimation of nonlinear, highly parametrized and complex models, some researchers shifted their attention from estimation to a rather heterogeneous groups of techniques collected under the header of *calibration*. Therefore, the latter has become a very useful tool to determine the parameter of simulated models.

During the last four decades, many efforts have been devoted to clarify the meanings of estimation, calibration and validation in simulation models and the relations linking these approaches. One of the first definitions of calibration is given by [280]. In this seminal paper, the authors argue that calibration proceeds by setting model parameters to certain values chosen according to prevailing theoretical and/or empirical evidence. A comparison between calibration and estimation can be attributed to [222] who point out that the difference between calibration and estimation is “artificial at best”, stating that calibration is a way of doing estimation without consider the sampling error in the sample mean (see also [406, 444]). From our viewpoint, the researcher doing calibration aims at attributing values to the parameters of the model producing an output that is approximately similar to a set of real data. An interesting perspective is given by [63] who define *verification* as the process of making sure that the implementation is coherent with the structure of a model (see [365, 112]), calibration as the moment in which the parameters are adjusted to reproduce the data (see [368]) and *validation* for the comparison of the output of a calibrated model with an external source of data (see [315, 151]).

In the following we will first outline the techniques used to estimate “classical” simulated models (e.g., models characterized by recursive equations), and then we will shift our focus to the estimation, calibration and validation of agent-based models.

2.2 Estimation methods for simulated models

The common practice in the estimation of simulated models is to apply econometric techniques to select appropriate parameter values (see [444, p. 3]). The five main approaches that we will see are: *indirect inference* ([196, 460, 195]), *method of simulated moments* ([335, 372, 133]), *simulated minimum-distance* ([217, 463, 191, 48]), *approximated likelihood* ([292, 293, 157, 2]) and *nonparametric simulated maximum likelihood* ([156, 272]), *approximate Bayesian computation* ([148, 25, 131, 175, 168, 458]).

In order to better explain the different methods, let us introduce the vector of real data $\mathbf{y}$ and the vector of simulated data $\mathbf{z}(\boldsymbol{\theta})$, where the parameters $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^P$ and Θ represents the parameter space.

Indirect inference (II) was first introduced by [196], [460] and [195]. This method identifies estimates of model parameters using an auxiliary model as an intermediate step. The parameters $\boldsymbol{\theta}$ of the auxiliary model are estimated both on the real dataset $\mathbf{y}$ and on data simulated from the original model $\mathbf{z}(\boldsymbol{\theta})$, and the distance between the estimated parameters in the two cases, i.e. the (empirical) real dataset and the simulated data, is minimized. The method is particularly suitable for all those cases in which the likelihood function of the original model is difficult to obtain, but getting simulated data is simple. The technique relies on the fact that, when the number of observations increases, the estimated values under the auxiliary model approach a pseudo-true value. The method allows to obtain consistent and asymptotically normal but generally inefficient estimators. Among the recent literature, we quote a paper by [130] who calibrate DSGE via II, and a contribution by [75] who develop Generalized Indirect Inference (GII). The idea of GII is to apply two different descriptive statistical models to the simulated and the observed datasets. As long as the two descriptive models share the same vector of pseudo-true values, the GII estimator is consistent and asymptotically equivalent to the II estimator (see [75] for further details). Since GII smooths the objective function exploiting a smoothed function of the latent utilities as the dependent variable in the auxiliary model, this method can be used to overcome some drawbacks met in the estimation of discrete choice models (DCM). Indeed, as pointed out by [75, p. 178], in DCM the researcher tackles the computational issue of optimizing a multidimensional step function using slower derivative-free methods. This is very time-consuming and puts severe constraints on the size of the structural models that can be feasibly estimated.

The method of simulated moments (MSM), developed by [335] and [372] in two seminal works, can be considered as an extension of the Generalized Method of Moments (GMM).[3] In GMM, the objective function to be minimized is a quadratic distance between the empirical moments based on the real data and the theoretical moments depending on the parameters $\boldsymbol{\theta}$. In MSM, the theoretical moments are replaced by the empirical moments based on the simulated data $\mathbf{z}(\boldsymbol{\theta})$. Recently, two interesting contributions were provided by [420], who exploits MSM for the estimation of nonlinear DSGE models, and by [143], who compare the performance of maximum likelihood and MSM for dynamic DCM. Although MSM is less computational intensive than II, it can be difficult to identify which moment to match (see [179] for a discussion of these issues in GMM, and [499] for an illustration of the effort that is required to identify moments that characterize the process under scrutiny). As argued by [195, Sec. 2.1], the researcher must choose an adequate calibration criterion to obtain consistency of the MSM estimator. Two kinds of criteria can be considered: (i) path calibration, which measures the difference between the trajectories of $\mathbf{y}$ and $\mathbf{z}(\boldsymbol{\theta})$, and (ii) moment calibration which is based on the difference between the empirical moments computed on $\mathbf{y}$ and $\mathbf{z}(\boldsymbol{\theta})$. While the second technique generally leads to a consistent estimator, path calibration does not yield consistency. This is made clear by the following example, borrowed from [195, p. 20]. Let:

$$\widetilde{\boldsymbol{\mu}}_n = \arg\min_{\boldsymbol{\mu}} \sum_{i=1}^{n} \left(z_i\left(\boldsymbol{\mu}\right) - y_i\right)^2$$

be the path calibrated estimator of $\boldsymbol{\mu}$ and let

$$\begin{aligned}\widetilde{\boldsymbol{\mu}}_\infty &= \arg\min_{\boldsymbol{\mu}} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(z_i\left(\boldsymbol{\mu}\right) - y_i\right)^2 \\ &= \arg\min_{\boldsymbol{\mu}} \left\{ \mathbb{V}\left(\mathbf{z}\left(\boldsymbol{\mu}\right)\right) + \mathbb{V}\left(\mathbf{y}\right) + \left[\mathbb{E}\left(\mathbf{z}\left(\boldsymbol{\mu}\right)\right) - \mathbb{E}\left(\mathbf{y}\right)\right]^2 \right\}\end{aligned}$$

be the asymptotic solution of $\widetilde{\boldsymbol{\mu}}_n$ satisfying the first order conditions, where the variance $\mathbb{V}(\cdot)$ is a generic function of $\boldsymbol{\mu}$. It is well-known that $\widetilde{\boldsymbol{\mu}}_n$ is consistent if and only if $\widetilde{\boldsymbol{\mu}}_\infty = \boldsymbol{\mu}_0$, however this

[3] We refer to [221] and [355] for a detailed treatment of GMM.

condition is not always satisfied. Indeed, if we consider $\mathbb{E}\left(\mathbf{y}\right) = \boldsymbol{\mu}_0$ and $\mathbb{V}\left(\mathbf{y}\right) = \boldsymbol{\mu}_0^2$, we get:

$$\widetilde{\boldsymbol{\mu}}_\infty = \arg\min_{\boldsymbol{\mu}} \left[\boldsymbol{\mu}^2 + \boldsymbol{\mu}_0^2 + \left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)^2\right]$$

and $\widetilde{\boldsymbol{\mu}}_\infty = \frac{\boldsymbol{\mu}_0}{2} \neq \boldsymbol{\mu}_0$.

The simulated minimum-distance method is a simulated version of the minimum-distance method (MDM). While MDM aims at minimizing the distance between the empirical cumulative distribution function (cdf) of the real data $\widehat{\mathbb{P}}_{\mathbf{y}}$ and a theoretical cdf depending on parameters $\boldsymbol{\theta}$ (see, e.g., [389, 326, 42, 73]), its simulated counterpart replaces the theoretical cdf with the empirical cdf based on the simulated data $\widehat{\mathbb{P}}_{\mathbf{z}(\boldsymbol{\theta})}$. The simulated minimum-distance method is well suited when the researcher can simulate data from the model distribution but not evaluate its density (see, e.g., [48, p. 3]).

The agreement between $\widehat{\mathbb{P}}_{\mathbf{y}}$ and $\widehat{\mathbb{P}}_{\mathbf{z}(\boldsymbol{\theta})}$ must be assessed with respect to some norm or metric $d\left(\widehat{\mathbb{P}}_{\mathbf{y}}, \widehat{\mathbb{P}}_{\mathbf{z}(\boldsymbol{\theta})}\right)$. We recall that a distance or metric between two arguments is a symmetric function that respects the triangle inequality and is zero if and only if its two arguments are equal. In this context, a norm is a function of the difference of the two probabilities, $\widehat{\mathbb{P}}_{\mathbf{y}} - \widehat{\mathbb{P}}_{\mathbf{z}(\boldsymbol{\theta})}$, that respects the above properties and satisfies absolute homogeneity, i.e. $d\left(a \cdot \left(\widehat{\mathbb{P}}_{\mathbf{y}} - \widehat{\mathbb{P}}_{\mathbf{z}(\boldsymbol{\theta})}\right)\right) = |a| \cdot d\left(\widehat{\mathbb{P}}_{\mathbf{y}} - \widehat{\mathbb{P}}_{\mathbf{z}(\boldsymbol{\theta})}\right)$. Most of the norms used for estimation are L_p norms. According to [178, p. 2], "[t]he sup [i.e. L_∞] norm is not a good choice because simulation methods often introduce granularities that destroy its effectiveness. [... T]he two natural candidates that provide some smoothing are the L_1 and L_2 norms. Of them, the L_2 is analytically more tractable and is also more traditional."

Distances are more commonly used than norms. Several distances can be considered in order to calibrate the marginal distribution of the process, its profile over time or its complete distribution. The most promising metrics concern the distributions of the processes involved (see [184, 402, 32, 34, 66, 426]). A key role is played by f-divergences between probability measures introduced by [114, 115] and [9] independently. A formal definition of f-divergences can be given following [301, p. 4398]. Let $\mathcal{F}$ be the class of convex functions $f : (0, \infty) \to \mathbb{R}$. For every $f \in \mathcal{F}$, we define the f-divergence as $d_f\left(\mathbb{P}_{\mathbf{y}}, \mathbb{P}_{\mathbf{z}(\boldsymbol{\theta})}\right) = \int f\left(\frac{\mathrm{d}\mathbb{P}_{\mathbf{y}}}{\mathrm{d}\mathbb{P}_{\mathbf{z}(\boldsymbol{\theta})}}\right) \mathrm{d}\mathbb{P}_{\mathbf{z}(\boldsymbol{\theta})}$, if $\mathbb{P}_{\mathbf{y}} \ll \mathbb{P}_{\mathbf{z}(\boldsymbol{\theta})}$ and $0f\left({}^0\!/\!_0\right) = 0$. Examples of f-divergences in statistics and information theory are the information divergence, the Pearson divergence, the Hellinger distance and the total variation. An interesting contribution in this field is due to [426] (see also [402] for a survey of these divergences). The work of [426] is devoted to proving bounds among f-divergences, which are particularly useful for the derivation of the rate of convergence of probability measures. These discrepancies can be also used to make estimation and inference (see, e.g., [301, Sec. VIII] and [421] who consider the distributional distance between a pair of processes developed by [203]).

Another important branch of application of the minimum-distance approach, closely related to estimation and calibration, is *sensitivity analysis* (SA). This stream of literature aims at studying the structure and the uncertainty associated to the model output when a change occurs in the model input (see [395] for a broad application of SA to different fields of research). Many approaches are possible, the most recent advances being summarized in [424, 34, 66, 472, 65, 469]. In the next section we will clarify the main aspects of sensitivity analysis for simulated models, with special attention to ABM.

In methods based on simulated/approximated likelihood functions, in order to overcome the lack of a closed-form solution, the objective function of the model is replaced with an approximation. The approximated likelihood is then evaluated instead of the exact likelihood (see [292, 293, 157, 2, 156, 272] among others). The development of the theory of simulated maximum likelihood is due to [292, 293], who provide a generalization of [371]. In [292, 293], the author characterizes the asymptotic theory for simulated maximum likelihood estimators of discrete response models, exploiting the theory of generalized U-statistics. The estimators are shown to be consistent and asymptotically normal (see [292]) and the bias introduced by the nonlinearity of the derivatives of the log likelihood function and the simulation errors is identified and a correction is proposed (see [293]). More recently, the literature has focused on the estimation of computed dynamic economic models via simulated maximum likelihood. [157, 2] study the effects on statistical inference of using an approximated likelihood instead of the exact likelihood function. In particular, the authors show: (i) the convergence of the approximated likelihood to the exact likelihood, (ii) that the errors in the approximated likelihood function accumulate as the sample size grows and (iii) the need to reduce the size of the error in the approximated policy function, as the sample size increase, in order to guarantee the convergence of the estimates.[4] A nonparametric approach to the estimation of dynamic models is finally provided by [156, 272]. The two works are similar in spirit but, while [156] start from a fully parametric model, whose reduced form can be simulated, and then approximate the unknown likelihood function with a kernel-based nonparametric estimator based on simulations of the endogenous variables of the model, [272] use simulated observations to nonparametrically estimate the unknown density by kernel methods and then construct a likelihood function to be maximized. Both [156] and [272] provide the asymptotic properties of the estimators. The strength of these nonparametric approaches relies on the fact that they can be applied to a very general class of models.

Beside the frequentist approach, some techniques based on Bayesian statistics have been developed. The philosophical principle fostering the development of Bayesian methods for simulation models relies on its axiomatic approach. The axiomatic view of Bayesian statistics is well highlighted by [457, p. 108], who assumes that any decision-maker is characterized by a specific probability distribution according to their choices made under uncertainty, and the probability distribution is updated following a Bayes' rule as new evidence emerges. In simulated models, approximate Bayesian computation (ABC) aims at providing draws from an approximation to the exact posterior density $\pi\left(\boldsymbol{\theta}\left|\mathbf{y}\right.\right) \propto p\left(\mathbf{y}\left|\boldsymbol{\theta}\right.\right)\pi\left(\boldsymbol{\theta}\right)$, when parameters and pseudo-data can be simulated from the simulated posterior density $p\left(\mathbf{z}\left(\boldsymbol{\theta}\right)\left|\boldsymbol{\theta}\right.\right)\pi\left(\boldsymbol{\theta}\right)$, but the simulated conditional density $p\left(\mathbf{z}\left(\boldsymbol{\theta}\right)\left|\boldsymbol{\theta}\right.\right)$ is intractable. The ABC algorithm is composed by three steps: (i) simulate $\boldsymbol{\theta}$ from $\pi\left(\boldsymbol{\theta}\right)$, (ii) simulate $\mathbf{z}\left(\boldsymbol{\theta}\right)$ from the likelihood $p\left(\cdot\left|\boldsymbol{\theta}\right.\right)$ and (iii) select $\boldsymbol{\theta}$ such that the distance function $d\left(\eta\left(\mathbf{y}\right),\eta\left(\mathbf{z}\left(\boldsymbol{\theta}\right)\right)\right) \leq \varepsilon$, where $\eta\left(\cdot\right)$ is a statistic and $\varepsilon > 0$ is the tolerance level (see [113, 490, 175, Sec. 5.1] for a complete description of the ABC algorithm). One of the first contribution in this field is due to [148], who develop a Bayesian Markov-chain Monte Carlo (MCMC) framework for the estimation of parame-

[4] The framework proposed by [157, 2] can be seen as a generalization of [425]. The policy function of [157, 2] concerns an objective function used to conduct policy analysis.

ters of stochastic volatility models characterized by discrete observations. However, [148] does not provide a full asymptotic theory. The large-sample properties for ABC were first developed by [25], who proved the rate of convergence of ABC under some mild conditions, and subsequently refined by [175] and [168] for models with auxiliary statistics. A link between ABC and II is provided by [131], in which the authors compare a collection of parametric Bayesian indirect inference (pBII) methods. The first class combines ABC and II computing the summary statistic on the basis of the auxiliary model, using ideas from II. The second approach, called parametric Bayesian indirect likelihood (pBIL), uses the auxiliary likelihood as a replacement to the intractable likelihood. The authors show that pBIL is a fundamentally different approach to ABC II and devise theoretical results for pBIL to give extra insights into its behavior and its differences with ABC II. Moreover, they investigate the assumptions required to use each pBII method. [131, p. 94] conclude that pBIL based on full data avoids the choice of the ABC discrepancy function and the ABC tolerance. On the other hand, ABC II is able to incorporate additional summary statistics outside the set formed by the auxiliary model and to provide better approximations when the auxiliary model is a simplified version of the generative model.[5]

We end this section with a comparison of the different estimation methods. As pointed out by [465, p. 2032] and [339, pp. 264-265], the researcher has to consider several criteria when choosing among different estimation methods; the most important are: (i) statistical performance and (ii) ease of use and computational speed. In this respect, according to [339, p. 264], the MSM tends to have larger biases and variances but is easier to implement than II, while II is more robust but can be computationally more costly than MSM when the auxiliary model is difficult to estimate. GMM (and hence MSM and II) can be included in the broad class of minimum-distance estimators as it involves the minimization of the distance of the empirical moments from the theoretical ones (see [355, p. 2118]). However, minimum-distance estimation allows to bypass the drawback related to the choice of moments to match. The simulated maximum likelihood method is a good candidate for the estimation of models requiring the simulation of the choice probability corresponding to the observed alternative, as for the case of multinomial logit (see [465, p. 2032]). However, as argued by [475, p. 239], among others, the main weakness of the simulated maximum likelihood is that if the number of draws used in the simulation, say R, is fixed, the estimator does not converge to the true parameters. In particular, the simulation bias disappears as the sample size N (and hence R) goes to infinity. Furthermore, the simulated maximum likelihood estimator is asymptotically equivalent to the maximum likelihood estimator only if R grows faster than $\sqrt{N}$. Finally, as well as MSM, ABC methods overcome the intractability of the likelihood function comparing simulated and observed summary statistics. The main issues of ABC methods concern the necessity to find a vector of low-dimensional summary statistics from the observed data with minimal loss of information and the computational challenge of achieving stringent matching between the observed and simulated summary statistics (see [131, p. 72]). Many reduction techniques are possible, the most important are summarized in [61].

[5] An exhaustive treatment of the different techniques used in ABC can be found in [458].

2.3 A review of estimation, calibration and validation techniques for agent-based models

Since agent-based models are able to describe the general dynamics of a system starting from the interactions between individuals, in the last years these models have become a very useful tool in different fields of research. Due to the expansion in ABM applications, estimation, calibration and validation have also broadly developed. The majority of the contributions apply simulation-based econometric methods to find the parameters that best accommodate the model (see [444, p. 3]).

A wide part of literature exploits indirect inference for the estimation of agent-based models. Starting from the early 2000s, different authors have brought interesting contributions in this direction. The first attempt at estimating the parameters of an agent-based model by considering indirect-based inference can be found in [189]. In their paper, the authors use the characteristic moments of the USD/DM exchange rate to evaluate the similarity of simulation outcomes. ABM parameters are then estimated optimizing that similarity. In order to overcome some inefficiencies due to the Monte Carlo variance in the objective function (e.g., local minima), the authors apply the *threshold-accepting optimization heuristic* ([132, 10]) instead of standard numerical optimization tools. This specific solution is explained by [190] in a subsequent study. Here, the same authors propose a continuous global optimization heuristic for a stochastic approximation of the objective function, using simulation-based indirect estimation of the parameters. The two authors describe an algorithm which combines the Nelder-Mead simplex with a local search heuristic (the threshold accepting named above). That procedure allows to obtain global minima also for non-globally convex objective functions. An extension of this model is then presented in [55] and in [150]. The first exploits the algorithm by [190] in order to provide an ex-post validation of the simulated parameters of the complex adaptive trivial system (CATS) model, in relation to actual data. In the second article, Fabretti compares the algorithm of [189] and [190] with a genetic algorithm implemented to calibrate the model presented in [155].

Despite II produces accurate estimation results, there is evidence that this method may suffer in terms of simulation times (see, e.g., [465, p. 2032] and [339, pp. 264-265]). This has spurred some research in different, less time-intensive methods of estimation.

Another technique broadly diffused in the ABM estimation and calibration literature is the method of simulated moments. A first example is provided in [499]. In this case, the authors compute an objective function based on some robust, i.e. not sensitive to outliers, statistics of the time series. The objective function can identify the presence of autoregressive conditional heteroskedasticity (i.e. changes of the variance of the series over time), and long memory (i.e. long-range dependence between values of the series). Subsequently, they obtain an estimate of the variance-covariance matrix of the different moments included in the objective function using a bootstrap procedure. The properties of the objective function are then analyzed with a GARCH-model and a stochastic volatility model using the USD/DM exchange rate as a benchmark. The results show that the objective function is able to discriminate between different parameter settings and models, and the optimal values are obtained for parameter values close to those resulting from a direct estimation approach.

A second example of the use of the method of simulated moments has been introduced for ABM in [206] in order to estimate ergodic and stationary models. Stationarity and ergodicity ensure that simulated moments converge to the theoretical ones.[6] While in GMM one compares the statistics estimated on real data with the theoretical statistics expressed as functions of the parameters, when no analytical expression is available for the latter, they are replaced by their simulated counterparts. The final estimate is the value of the parameters such that the simulated moments are closer to the ones based on the observed values.

During the last decade, different estimation techniques based on simulated-minimum distance[7] have been developed in order to overcome analytical difficulties in ABM estimation and to solve the issues related to computational costs. As stated before, the method of moments is also a minimum-distance technique (see [355]), in which the distance to be minimized is the one between the moments. However, minimum-distance estimators are often based on more refined definitions of distance between time series. This establishes a link with the stream of the literature concerning time series validation (see [498, 151]).

In a study by [329], three types of similarity measures between simulated and historical time series are identified. The first one is the classic Kullback-Leibler information criterion of [278], a statistical discrepancy based on the concept of relative entropy developed by Shannon in [447] which measures the loss of information deriving by using a model to approximate reality. As argued in [152], the problem of the Kullback-Leiber divergence is that this criterion weights too much the events for which the two models under comparison differ the most in terms of predictions, even if these differences concern aspects of data behavior that are not interesting for the analysis. The second one is the state similarity measure (SSM), which is the sum of the absolute difference in the frequency of every possible state in each of the time series. In particular, each time series/vectors belonging to the set of vectors $\mathbf{P}$ ($\mathbf{Q}$) is partitioned and the lags of the series are computed. Then, the number of occurrence of each state in a given period are computed, grouped in a n-dimensional vector $\mathbf{p}$ ($\mathbf{q}$), and the number of observations in $\mathbf{Q}$ are subtracted from $\mathbf{P}$. Finally, the third one, is the Generalized Hartley Metric (GHM). This distance represents a weighted average of the Hartley Metric (see [226]) and measures the amount of uncertainty associated with a finite set of possible alternatives deriving from the lack of specificity. The larger is the set of possible alternatives, the less specific is the identification of any desired alternative of the set. The possibility is estimated in terms of frequency.

The comparison between SSM and GHM shows the closeness in measuring similarity between the two metrics, but SSM could be preferred to GHM for its simplicity and transparency. Nevertheless, these two metrics can be used only when the simulated and the historical time series have the same length.

[401] have examined a distance given by the sum of the squares of the difference between the observed and simulated market price at a given date/time. The authors minimize this function

[6] For general results on ergodic processes see [233], while for an analysis of stationarity and ergodicity conditions through nonparametric Wald-Wolfowitz tests see [205].

[7] For a thorough review of metrics for statistical data process see [32]. In this paper, the author identifies six different classes of divergence measures: f-divergences, Bregman divergences, α-divergence, divergences between more than two distributions, inference based on entropy and divergence criteria and spectral divergence measures.

through gradient-based calibration and illustrate the procedure using the simple [72] model.

An interesting contribution is provided by [282]. The author proposes the use of an information theoretic criterion called "Generalized Subtracted L-divergence (GSL-div)" based on the work of [302], who developed a symmetric measure for relative entropy (L-div). [282] first exploits the metric developed by Lin to recursively estimate patterns with different lengths. Then, he combines all the L-divergences into a unique information criterion, the GSL-div. The advantage is that this method makes possible to measure how simulated time series replicate the properties of observable outputs without imposing stationarity requirements or defining any likelihood function. In a subsequent paper, [281] proves the discrimination ability of the GSL-div using the model proposed by [72]. Findings show that the calibrated model tracks well the evolution of two real stock indexes.

As anticipated in Section 2.2, an important field of study linked to the minimum-distance approach is SA (see [424, 34, 66, 472, 65, 469]). This branch of research has been successfully applied in the ABM literature. Here we quote two contributions, [472] and [469]. In [469], the authors compare three different methodologies used in SA: sensitivities based on one-factor-at-a-time (OFAT), sensitivities based on model-free output variance decomposition and sensitivities based on model-based output variance decomposition. When analyzing an ABM, instead of the Sobol method suggested by [423], they recommend to use OFAT as a starting point. Indeed, OFAT appears to be able to explore the patterns and to capture the nonlinear interactions and the emergent properties of the model better than other methodologies. [472], instead, provides a practical guide to link the simulation with the calibration of an ABM.

Some techniques based on nonparametric likelihood functions are investigated, among others, in the works of [277] and [314]. In order to approximate the conditional density of the data-generating process from the simulated observations, [277] exploit a nonparametric simulated maximum likelihood based on kernel methods (see also [272]). The unknown likelihood function is then substituted by the simulated likelihood in the estimation. Instead, [314] exploits Sequential Monte Carlo (SMC) estimation based on a particle filter to numerically approximate the conditional densities that enter into the likelihood function. This approximation is then used to simultaneously obtain parameter estimates and filtered state probabilities for the unobservable variables that drive the dynamics of the observable time series.

Something different is proposed by [207]. The authors develop an ABC methodology for likelihood-free estimation. They first provide a kernel density estimation of the likelihood together with Markov chain Monte Carlo sampling schemes. Then, they switch to parametric approximations of the likelihood by assuming a specific form for the distribution of external deviations in the data and, at the end, they introduce Approximate Bayesian Computation.

Despite nonparametric simulated likelihood is widely applied, the kernel smoothing step often leads to the so-called "curse of dimensionality" of the summary statistics (i.e., the summary statistics grow exponentially as the dimension of the parameter space increase, see, e.g., [261] for a general definition). Different "dimension reduction" techniques are possible, the most promising are exposed by [61].

Sometimes the performance of these estimation methods can be improved using *surrogate metamodels* ([424, 265, 65]), often in the form of *kriging*. As the ABM generating the data is generally very slow, the researcher can supplement the simulations from the ABM using a different, generally much faster, data-generating mechanism. The researcher identifies several parameters, simulates the output for each one of these values and estimates the relation between the parameters and the output. This new statistical model is called a surrogate meta-model and can be used to approximate the output of the ABM, especially for new parameter values. This implies that at the end of the procedure, the researcher will have some simulations from the ABM and a set of other simulations from the meta-model. The latter are generally noisier, as they contain some uncertainty arising from model estimation and some bias from model choice, but they are available in a much larger number than the former. This leads to a sensitive reduction of simulation time. The difference between kriging and surrogate meta-models relies on the fact that kriging considers Gaussian process interpolation, while surrogate meta-models are more general (see, e.g., [387]). Among the most recent research on this topic, we quote a paper by [422], who first apply kriging meta-models to computational economics, and a work by [283], in which the authors investigate the estimation of ABM meta-models through machine learning. In their contribution, [283] develop a two-steps procedure. In the first round, an initial pool of parameter combinations is drawn (e.g., training set) using a standard sampling routine. Subsequently, a random subset of combinations (e.g., test set) is drawn without replacement from the pool to initialize the learning procedure and the ABM is evaluated for each of these combinations receiving a specific label ("positive"/"negative") according to a pre-specified calibration criterion. Finally, a surrogate is learned over the combinations of the parameters. Then, the probabilities that the unlabeled combinations in the pool belong to the "positive" category are forecast.. In the second round, given the predicted positive probability, the algorithm draws a small sample from the pool. These drawings are then evaluated in the ABM and aggregated to the set of combinations sampled in the first step. The procedure is repeated until the algorithm reaches a pre-specified level of accomplishment.

Another promising method is the mean-field approximation (MFA). This technique was developed by [167] and [17] who exploited a method used in statistical mechanics (see also [123] for an application to financial markets). As explained by [125], the MFA focuses on the fraction of agents occupying a certain state of a state-space at a certain time. To this end, the agents are divided into sub-groups according to a particular feature. These clusters determine the evolution of the whole system. The result of the functional-inferential method identifies the most probable path of the system dynamics. This method is interesting because leads to an analytical solution of the ABM. In particular, the direct interactions among agents are replaced by indirect mean-field interaction between sub-groups expressed in terms of the transition rates of the master equations. From these equations, it is then possible to compute the statistical stationary equilibrium of the system.

The literature on ABM estimation is characterized by a trade-off between the availability of a complete statistical theory and computational feasibility. On the one end of the spectrum, when estimating a model through indirect inference, one can use most statistical tools that are available in other applications (tests, confidence intervals, etc.) but this comes at the cost of extremely lengthy estimation procedures. On the other end of the spectrum, surrogate meta-models can significantly speed up computations but no specific theory is available for the construction of otherwise simple inferential tools (tests, confidence intervals, etc.).

We can identify at least three problems related with the previous estimation methods. In the following we outline the main weaknesses.

First of all, they involve the minimization of an objective function. As explained in Section 2.1, ABM do not allow to use the same set of random numbers for different values of $\boldsymbol{\theta}$. This leads to a rugged objective function. The roughness of the objective function requires the use of specific optimization algorithms that are often computational expensive (see, e.g., the routine developed by [189, 190]). Other problems are related to the presence of multiple local minima and identification issues leading to large standard deviations. These phenomena are outlined by [277], who examine the graph of the objective function (see [277, pp. 23, 36, 39]).

Second, these estimation methods rely on the stochastic equicontinuity hypothesis (see [335, 372, 355, pp. 2136-2137]). This drawback is related to the first one. Again, as new random numbers are drawn for different values of $\boldsymbol{\theta}$, the objective function is not only rugged, but also discontinuous and no differentiable at all. This implies that the asymptotic properties of the estimator, whose derivation relies on the differentiability of the objective function, may not hold and the associated inferential tools may not be available.

Finally, they assume stationarity and ergodicity of the statistical process. However, as pointed out by [213, p. 49], if we consider long periods, the micro-rules governing the behavior of an ABM are often unstable and can change over time. Hence, it is sometimes useless to study the stationary state of the model. Moreover, the process could be non-ergodic (i.e., the dynamic of the model could exhibit transients, see also the conclusion of [153]).

To overcome some of these issues, recently, different frameworks based on regression have been developed (see [403, 266]). Most of them are based on nonparametric econometrics and machine learning techniques ([449, 110, 400, 396]). [449] exploit deep learning to develop a likelihood-free inference framework. In particular, they use multilayer neural networks to learn a feature-based function from the input (many correlated summary statistics of data) to the output (parameters of interest). [110] proposes to use neural networks with simulated data to learn the limited information posterior mean or a good approximation to it. The inputs to the net are the simulated statistics and the net is trained to fit the parameter values as well as possible according to a pre-specified criterion. [396] provide a method based on convolutional neural networks for ABC allowing to derive simultaneously the mean and the variance of multidimensional posterior distributions directly from simulated data. Once trained on simulated data, the convolutional neural network maps real data samples of variable size to the first two posterior moments of the relevant parameters distributions. A different approach for ABC is presented by [400], who conduct likelihood-free Bayesian inferences on the model parameters with no prior selection of the relevant components of the summary statistics. The parameters are selected applying a random forest to a nonparametric regression setting.

However, the techniques outlined above seem to focus on prediction rather than estimation. In a related work, instead, [86] consider a simpler method. The authors estimate the parameters of an ABM using regularized linear regression. The approach of [86] consists of four main steps. First of all, they identify a set of parameter values in the parameter space. Second, they run the simulation model "many" times computing a vector of statistics for each parameter value. Then, they estimate a regression where the parameters are the dependent variable and the corresponding simulated summary statistics are the independent ones. Finally, the real-data summary statistics are used as input in the regression to obtain the estimates of the parameters. As pointed out by the authors, this approach allows to overcome some issues related to the selection of the summary statistics (see [179], [499], [111] and [75]), the definition of the distance function and the numerical minimization of the objective function, typically met in standard calibration approach. However, the method of [86] presents some disadvantages. Indeed, they assume a linear relation between the statistics and the parameters to estimate, and they do not provide any asymptotic theory for the estimator. A solution to these drawbacks is offered in Chapter 6 of this thesis, in which we develop an estimation technique exploiting nonparametric least absolute shrinkage and selection operator (Lasso) regression and we rigorously characterize its asymptotic properties.

Finally, one of the most recent contribution in the field of calibration can be attributed to [444]. Given a statistical distance, the authors build Model Confidence Sets (MCS) to rank simulated models based on their closeness to the benchmark, and to construct a plausible set of alternative specifications of parameters that cannot be distinguished from the chosen one from a statistical point of view. Therefore, this procedure is twofold: (i) it can be used in a first calibration step, to select the vectors of parameters that are closer to the benchmark data, and, (ii) it can also be used for validation to compare the output of a calibrated model with an external source of data in a second step (see, e.g., [151]). MCS has been developed by [223] and consists of a subset of the set of models containing the best model with a given level of confidence. A first attempt in calibrating ABM via MCS can be found in [26]. In his paper, the author exploits the technique developed in [27] to select among different models, scoring each one of them at the level of individual empirical observations (see also [28] for a multivariate extension of the Markov Information Criterion, MIC, of [27]). Differently from [27], [444] rigorously study the construction of a MCS and provide the statistical properties of the rate function, that is a measure of improbability of the configuration of parameters minimizing the distance between simulated and benchmark data.

The last phase of calibration and estimation process concerns validation (see [63, 152, 498] for a discussion on the issues of calibration and [315, 151] for complete surveys of validation methods). An innovative work in this field is due to [215]. In their paper, the authors propose a method to empirically validate simulation models comparing structures of vector autoregression models (VAR), estimated from both artificial and real-world data, via causal search algorithms. Differently from the information criteria of [26, 27] and [282], [215] focus on the similarity of the dynamic causal structures of the variables of interest and consider a multivariate time-series comparison. The authors identify five steps in their validation procedure:

1. Dataset uniformity: a monotonic transformation is performed in order to allow the matching between the real-world and simulated time series;

2. Analysis of ABM properties: the authors verify that the time series is time-independent and ergodic (see [233] and [205]);

3. VAR estimation: the lag of the model is selected according to some information criteria and the VAR is estimated via ordinary least squares (OLS) and maximum likelihood estimation (MLE);

4. SVAR identification: the vector of residuals are extracted from the estimated VAR and their statistical properties are tested. The researchers choose among two algorithms, according to the results of the tests ("PC" for the Gaussian case and "VAR-LiNGAM" for the non-Gaussian case);

5. Validation assessment: the estimated causal structures are compared according to three distance measures ($\mathbf{\Omega^{sign}}$, $\mathbf{\Omega^{size}}$ and $\mathbf{\Omega^{conj}}$). $\mathbf{\Omega^{sign}}$ considers the directions of the causal relations entailed by the SVAR; $\mathbf{\Omega^{size}}$ compares only the size of the causal effects; $\mathbf{\Omega^{conj}}$ accounts for both the directions and the size.

Another interesting contribution in terms of validation can be found in [29], in which the authors develop a three-phase validation protocol. They first generate a set of 513 parameter combinations using Nearly-Orthogonal Latin Hypercube sampling. Then, they score simulated data against real-world observations via the Markov Information Criterion of [27]. Finally, they build a surrogate model of the MIC response surface exploiting stochastic kriging. The local minima of the interpolated MIC response surface are the parameters combinations that best fit to the data. MCS is then performed to restrict the set of model calibrations to those models that are statistically indistinguishable. At the end, the surrogate model is validated repeating step two of the validation protocol on the restricted set and verifying that the new scores are equal to the scores forecast by the surrogate model.

Chapter 3

Model Calibration and Validation via Confidence Sets[1]

In this chapter, we consider the issues of calibrating and validating a theoretical model, when researchers wish to select the parameters that better approximate the data among a finite number of alternatives. Based on a user-defined loss function, we propose to construct Model Confidence Sets to restrict the number of plausible alternatives, and measure the uncertainty associated to the preferred model. We further suggest an asymptotically exact logarithmic approximation of the probability of choosing a model via a multivariate rate function. We outline a simple numerical procedure for the computation of the latter and we show that it yields results consistent with Model Confidence Sets. We finally showcase the implementation of our approach in a model of inquisitiveness in ad hoc teams, relevant for bounded rationality and organizational research.

3.1 Introduction

In the social sciences, there are usually two methods for determining the parameters of a specific model: calibration and estimation. While the latter is more appealing in a number of cases, it usually requires making additional, and sometimes arbitrary, assumptions about the way data were generated. Exogenous shocks may be added to the model in an ad hoc fashion, and point identification and estimation of the model's parameters are often an issue for complicated, nonlinear and non-separable structural equations. Examples of the latter instances are applied general equilibrium analysis, in which hypotheses for identification and estimation of parameters are often difficult (see [84] for a discussion on the issues of identification in dynamic general equilibrium models), and agent-based models, that are often very hard to describe within a unified statistical framework (see, e.g., [187, 152, 185, 498, 125, 215, 151]).

Therefore, calibration has evolved to be a very popular method for parameter determination. Calibration allows one to set model parameters to certain values that are chosen according to prevailing

[1] This chapter is jointly written with Raffaello Seri, Davide Secchi and Samuele Centorrino, and is in press on Econometrics and Statistics (see [444]).

theoretical and/or empirical evidence (see [70, 112]). This definition of calibration is consistent with most of the economic literature and goes back at least to the seminal business cycle paper of [280] (see also [104] and [414, Sec. 5.8, p. 217]). Calibration is followed by a step that is referred to as model validation (or verification, see [368], and [63]), in which moments of random sequences generated by the model are compared (or eye-balled) with sample moments to show that, given the parameters' value, the economic model provides outcomes consistent with observable data.

From a statistical perspective, drawbacks of this particular definition of calibration and validation are that it does not usually provide a quantitative measure of distance of the chosen model to the data; and that it does not take into account the uncertainty related to the choice of a specific vector of parameters.

This definition has been challenged by [222]. The authors have noticed that the distinction between calibration and estimation is "artificial at best" (see also [406]). They argue that "[calibration] looks remarkably like a way of doing estimation without accounting for sampling errors in the sample mean". They further claim that the validation step used by [280] is tantamount to testing. They argue that both steps are coming from the minimization of an implicit loss function which, if made explicit, would make the principle by which a particular model is chosen easier to understand.

While there is little agreement as to what "calibration" is, there is even more confusion about what are "verification" and "validation". Some authors ([368, pp. 642-643] or [222, pp. 91-92]) use "verification" for the comparison of the output of a calibrated model with the real world. Other authors ([365, pp. 30-31], or [112, p. 419]) use "verification" as the process of making sure that the implementation is coherent with the structure of a model and "validation" as the process of making sure that the implementation is coherent with the real world, thus leading to an overlap between "validation" and "calibration" [112, p. 419]. In [406], "validation" is used as a check of the adequacy of the simulated model with another model it is intended to portray. At last, [63, Sec. 2.1-2.4] use "verification" as in [365, 112], "calibration" as in [368, 222] and "validation" for the comparison of the output of a calibrated model with an external source of data (see also [151]).

In this paper, we follow the approach suggested in [222] and consider the issue of calibrating and validating an economic model by minimization of a user-specified loss function, when researchers need to choose among several (albeit finite; see also [63, Sec. 2]) combinations of the parameters. The underlying economic model does not have to admit a closed-form solution, or to be point identified in a statistical sense. It should, however, be possible to generate series of simulated data from the model itself, for specific configurations of parameters. While the restriction to a finite number of configurations of the parameters may appear to be a limitation in some cases, it is, in our opinion, a minor one. First, the number of configurations may be increased at a relatively low computational burden. Second, most optimization algorithms are based on a finite number of evaluations and on stopping rules that are not guaranteed to reach the true optimum. Possible extensions to the infinite case are deferred to future work.

We thus consider the issue of this theoretical model matching some benchmark data. It is important to note that the benchmark is not necessarily composed of real-world observations: it could also be a function, or results from another simulation, for example. In some cases, one looks for values of parameters that provide simulations that replicate a deterministic target or achieve a fixed goal (see, e.g., [112]). What is more relevant instead is that the real data should contain patterns, i.e. “observations of any kind showing nonrandom structure and therefore containing information on the mechanisms from which they emerge” (see [210, p. 991]). The match between the benchmark and the simulated samples is assessed through a distance. In the following it will be clear that our developments do not rely on any specific property of a distance. However, we prefer to use this name as most applications will deal with this case. Given a distance, we propose to construct Model Confidence Sets in the spirit of [223]. This technique not only allows us to rank models based on their closeness to the benchmark, but also to construct a plausible set of alternative specifications of parameters that cannot be distinguished from the chosen one, at least not in a statistical sense. In this respect our procedure can be used in a first calibration step, to select the vectors of parameters that are closer to the benchmark data; and, being related to a testing procedure through the duality between confidence sets and tests (see, e.g., [56, Sec. 4.5, p. 241]), it can also be used for validation in a second step. Notice that we do not see the requirement that parameters have discrete support as a drawback. As a matter of fact, most simulation-based estimators are obtained by sampling from conveniently selected points in the sample space (as sampling from the entire space is often impossible in practice). Moreover, when used for calibration, our procedure can help identify multiple local optima of the objective function, when direct minimization would only identify one of those local optima. This feature is especially useful for highly parametrized models where point identification is often impossible to show (see [112, p. 419]). Along the lines of [223], we show that MCS are valid confidence sets when the number of simulated samples, n, diverges.

We complement this analysis by characterizing the model’s choice as an estimation problem in which the parameters have discrete support (see [93]). We show that, despite the computation of the probability of choosing a combination of parameters over the others is probably out of reach, one can still provide, using large deviation theory, a measure of the rate of decrease of this probability with n through a multivariate rate function, which depends on the vector of observables. The finite sample approximation of this rate function is interesting for at least two reasons. On the one hand, it can be used empirically as an alternative metric to establish the plausibility of our set of models. However, its computation is not straightforward. On the other hand, its theoretical properties have not been explored thus far, to the best of our knowledge, as published papers only consider rate functions in the univariate context (see, e.g., [134, 135, 136, 413, 137]). In the univariate case, the rate function needs to be calculated at one point only and this does not involve particular computational issues. In our case, however, the measure of the decrease rate of the probability arises from the optimization of a function over a multivariate domain and this requires high accuracy in the computation of the whole function to be minimized.

First, we propose a method based on the *linear-time Legendre Transform* (LLT) to numerically compute the rate function. As the name suggests, the numerical complexity of this method only grows linearly with the number of data points, and it thus provides a computationally efficient method, even when the number of observations is large. Second, we show that this rate function approximation is consistent, and we obtain its rate of convergence. We then perform a short simulation study to demonstrate the validity of this theoretical result in finite samples.

We conclude the paper with an application of this methodology to an ABM of inquisitiveness (henceforth, *inquisitiveness ABM*) (see [30]). Inquisitiveness is a development of “docility”, as a strategy that decision makers use to cope with their bounded rationality [454, 455] and applied to agent-based simulations recently for its cognitive links to behavior [433, 435, 345]. A docile individual is someone who leans on other people’s advice to make decisions. In particular, when docile individuals operate inside a team (or any community of reference) to solve a problem, trust is placed on the advice of team members, while data coming from outside the team are mistrusted. An inquisitive individual, instead, does not operate on a team basis but s/he is rather a problem solver, who may look for help outside the team. We show that both Model Confidence Sets and the approach based on a rate function approximation yield similar results. This ABM has been selected for three reasons: (a) theoretical relevance, (b) applicability to other economic domains, and (c) computational simulation robustness. The first criterion entails selecting a model that is anchored to a solid theoretical background while, at the same time, it should not be a mere replication of other models. The inquisitiveness ABM is based on Simon’s bounded rationality [452, 453, 456]. A theory with high relevance, given the value that its theoretical and empirical applications have had in economics in the recent decades (e.g., [252, 471, 83]). Inquisitiveness exemplifies the tension toward finding more suitable theoretical conceptualizations of bounded rationality. At the same time, it has wide applicability to organizations of various sizes as well as wider economic systems. Finally, to make sure that we selected a code that had been openly peer reviewed in a certifiable and documentable way, we looked onto those ABM uploaded into OpenABM. This platform is supported by the Network for Computational Modeling in Social and Ecological Sciences (CoMSES) and these scholars offer peer review specifically on the code, when modelers ask for it. Among the (very few) models that had been peer reviewed, the one on inquisitiveness ticked also all of our other criteria for selection.

The inquisitiveness model shows high parametrization and nonlinearities in both variables and parameters, and calls for a rigorous calibration procedure, as it is widely done in ABM research. Alternative procedures have also been extensively explored (see [190, 498, 499, 150, 206, 401, 207, 215, 277, 283, 151]). The most recent trend in calibration of ABM is to apply sound econometric techniques to select appropriate parameter values (see [87]). Methods such as Indirect Inference [189, 190], Simulated Method of Moments [499], Simulated Maximum Likelihood [277], Simulated Minimum-Distance [206, 401, 282, 281], and Approximate Bayesian Computation [207] have emerged as calibration techniques in ABM. We believe there are three main drawbacks of performing calibration in this way. First, its results may vary widely depending on sample selection, as models are not necessarily identified, and therefore uniqueness of the optimum is not guaranteed (not even asymptotically). Similarly, there has been little effort to formally characterize the uncertainty associated to a specific choice of parameters [445]. Finally, given the high parametrization of ABM, computational costs can be extremely high (e.g., [472, 307]). Our technique is instead computationally fast; and effectively uses the sample to validate one or more choices of parameters.

Our procedure bears a resemblance with other ones introduced in the literature. For instance, [283] use Surrogate Meta Models and Machine Learning to search the parameter space. Their iterative technique is based on a surrogate model which learns from an initial random subset of parameter combinations. We perceive that our method is similar in spirit, although one learns from a finite number of combinations of parameters that are directly specified by users, instead of being determined iteratively from a (finite-dimensional) subset of the parameters' space. Similarly, [26] applies the methodology developed by [27] to select among different models, scoring each one of them at the level of individual empirical observations. Differently from [26], we study rigorously the construction of a Model Confidence Set and we provide the statistical properties of the rate function.

Our approach can complement existing ones, and be used in a validation step to compare estimates obtained by any of the aforementioned estimation methods. By appropriately splitting the data into a fitting sample and a training sample, one can obtain several plausible parameter values from an initial estimation step, and then compare them in a testing (or validation) step using MCS.

The paper is structured as follows. The mathematical notation that is needed for the discussion of the model is introduced in Section 3.2. Section 3.3 introduces the statistical framework and provides the asymptotic relations between the probabilities of choosing a model and the rate functions. Section 3.4 discusses the construction of Model Confidence Sets for a problem of calibration. Section 3.5 outlines the numerical approximation of the rate function and provides its statistical properties. Readers more interested in the application of our procedure can directly move to Section 3.6. We provide an outline of our findings in Section 3.7. All technical proofs are relegated to the Appendix.

3.2 Notation

This section introduces the notation that will be used throughout the paper. Capital bold letters, such as $\mathbf{A}$, denote matrices while lowercase bold letters, such as $\mathbf{a}$, usually denote vectors. The $i-$th element of vector $\mathbf{a}$ is generally denoted a_i. $\mathbf{u}_n$ is a n-vector composed of ones. $\mathbf{I}_n$ is the $(n \times n)$-identity matrix. $\mathbf{U}_n$ is a $(n \times n)$-matrix composed of ones. $\mathbf{e}_{i,n}$ is a n-vector of zeros with a one in the i-th position; when the length is clear from the context we simply use $\mathbf{e}_i$. $\mathbf{0}_{m\times n}$ is a $(m \times n)$-matrix composed of zeros. We do not indicate the dimensions when they are clear from the context. $\operatorname{diag}(\mathbf{a})$ is a diagonal matrix with $\mathbf{a}$ on its diagonal. $\mathbf{A}'$ and $\mathbf{A}^{-1}$ are respectively the transpose and the classical inverse of the matrix $\mathbf{A}$, provided they exist.

For a set A, the notations $\operatorname{int}A$, $\overline{A}$, ∂A, $\operatorname{co}A$ and $\overline{\operatorname{co}}A$ respectively denote the interior, the closure, the boundary, the convex hull and the closed convex hull of A. The positive half-line is $\mathbb{R}_+ := \{x \in \mathbb{R} : x \geq 0\}$. The positive orthant is $\mathbb{R}^d_+ := (\mathbb{R}_+)^d$. Hyperplanes are indicated with the point-normal notation, i.e. as $\{\mathbf{x} \in \mathbb{R}^d : \mathbf{x}'\boldsymbol{\alpha} = \mathbf{v}'\boldsymbol{\alpha}\}$ where $\boldsymbol{\alpha}$ is a normal vector and $\mathbf{v}$ is a point of the hyperplane. The same notation is used for a half-space as $H^+(\boldsymbol{\alpha}, \mathbf{v}) := \{\mathbf{x} \in \mathbb{R}^d : \mathbf{x}'\boldsymbol{\alpha} \geq \mathbf{v}'\boldsymbol{\alpha}\}$. For a scalar β, $\beta A := \{\beta\mathbf{x} : \mathbf{x} \in A\}$.

We introduce the definition of the *effective domain* of a function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ as $\mathcal{D}(f) := \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) < \infty\}$ (see [124, p. 4]). The *Legendre* or *Legendre-Fenchel transform* of f is the function $f^\star : \mathbb{R}^d \to \overline{\mathbb{R}}$ defined by the variational formula $f^\star(\mathbf{y}) := \sup_{\mathbf{x}\in\mathbb{R}^d} \{\mathbf{y}'\mathbf{x} - f(\mathbf{x})\}$. $f^\star$ is also said to be the *convex conjugate* of f. χ_A is the (convex analysis) *characteristic function* taking the value 0 in A and $+\infty$ outside.

Expectations are denoted as $\mathbb{E}$. When the integration variable is not clear from the context, we indicate it explicitly as a subscript as in $\mathbb{E}_x$. We will do the same for the variance $\mathbb{V}_x$. For a random vector $\mathbf{X}$ taking values in $\mathbb{R}^d$, we define the *moment generating function* (*MGF*) of $\mathbf{X}$ as

$M(\mathbf{u}) := \mathbb{E}\exp\{\mathbf{u}'\mathbf{X}\}$ and the *cumulant generating function* (*CGF*) as $\Lambda(\mathbf{u}) := \ln M(\mathbf{u})$.

3.3 Framework and Preliminary Results

Let $\mathbf{y}$ be a vector of benchmark observations, that can contain individual level data, a time series, etc. It can be determined by a deterministic or a stochastic mechanism, but in any case it is supposed to be fixed in the subsequent analysis. Fixing the outputs $\mathbf{y}$ may come from the difficulty in modeling the real-world phenomenon leading to the data, or from an interest in the real-world data rather than in the data generating process (DGP) leading to them.

Example 3.1. [Multiple outputs] Suppose that the simulation model describes a situation characterized by several outputs $\mathbf{y} = (y_1, \ldots, y_p)$ measured at the same time. The researcher may be interested in looking for the parameters of the simulation model yielding outputs that are as similar as possible to these observations, without the need to model the DGP leading to these data.

Example 3.2. [Time trend] Consider a simulation model describing the evolution of a system over time. In this case $\mathbf{y} = (y_1, \ldots, y_t)$ may be a time series.

Example 3.3. [Spatial patterns] In some cases, the aim of the simulation is to simulate the spatial patterns arising from a complex system. This situation takes place in the study of animal diffusions, geographical plant distributions, etc. In this case $\mathbf{y}$ is a curve in a space or a spatial distribution of points.

We suppose to have m_0 configurations of parameters, say θ_i for $i = 1, \ldots, m_0$. We will not need any special structure on the set of possible parameters. The set of all parameters is denoted by $\mathcal{M}^0 := \{1, \ldots, m_0\}$. For each one of them, we simulate n realizations $\mathbf{z}_j(\theta_i)$, for $j = 1, \ldots, n$. We compute the distances $d(\mathbf{y}, \mathbf{z}_j(\theta_i))$ for $i \in \mathcal{M}^0$ and $j = 1, \ldots, n$. Here and in the following we will use the term distance loosely; as an example, we will never use any specific property of a distance, such as non-negativity or the triangle inequality.

Example 3.4. [Multiple outputs - Example 3.1 continued] A reasonable choice of a distance is:

$$d(\mathbf{y}, \mathbf{z}_j(\theta_i)) = \sum_{h=1}^{p} a_h |y_h - z_{jh}(\theta_i)|,$$

where the a_h's, for $h = 1, \ldots, p$, are constants keeping into account the different contributions of the observations to the overall distance.

Example 3.5. [Time trend - Example 3.2 continued] The time series case admits several distances, with different interpretations. In the ergodic case, when a single instance of a time series is sufficient to estimate probabilities of events, one can consider distances between the probability measures that generated the data. In this case it is quite natural to keep into account the DGP of $\mathbf{y}$ when drawing inferences. For a review of metrics or similarity measures for statistical data processes see [32] and [329]. Two further interesting contributions are provided by [27] and [282]. Other distances cannot be cast as metrics between the probability measures that generated the data, but only as distances between the trajectories over time. These distances are probably more interesting for the following as they apply to trending non-ergodic time series (e.g., [300, 176, 495]).

Example 3.6. [Spatial patterns - Example 3.3 continued] If one considers an animal roaming in a certain area, several distances can be considered, such as the Fréchet distance, used for dynamic time warping (see [273, 47, 214]).

We introduce the following notations. The expected distance relative to the i-th combination of parameters is $\overline{D}_i := \mathbb{E}_{\mathbf{z}} d\left(\mathbf{y}, \mathbf{z}\left(\theta_i\right)\right)$. We denote as $\overline{\mathbf{D}} := \left(\overline{D}_1, \dots, \overline{D}_{m_0}\right)'$ the vector containing the average distances. The distance corresponding to the j-th realization and the i-th combination of parameters is $D_{j,i} := d\left(\mathbf{y}, \mathbf{z}_j\left(\theta_i\right)\right)$, and the sample average distance is $\overline{D}_{n,i} := \frac{1}{n}\sum_{j=1}^{n} D_{j,i}$. Let us define the m_0-vector:

$$\overline{\mathbf{D}}_n := \left(\overline{D}_{n,1}, \dots, \overline{D}_{n,m_0}\right)'.$$

If we denote:

$$\mathbf{D}_k := \left(D_{k,1}, \dots, D_{k,m_0}\right)'$$

for $k = 1, \dots, n$, we can write $\overline{\mathbf{D}}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{D}_k$.

We want to identify the values i corresponding to the smallest expected distance $\overline{D}_i$. Other criterion choices may be more relevant in some situations but in this case, we consider only the expected distance. The set of parameters achieving the minimal distance is denoted as $\mathcal{M}^\star := \left\{j \in \mathcal{M}^0 : \overline{D}_j = \min_{i\in\mathcal{M}^0} \overline{D}_i\right\}$. It is clear that a choice would be to take the value $\widehat{\imath}_n$ minimizing $\overline{D}_{n,i}$ for $i \in \mathcal{M}^0$. We suppose that $\widehat{\imath}_n$ is always single valued while $\mathcal{M}^\star$ may not be a singleton.

The choice of a value from a finite set can be characterized as a discrete-parameter estimation problem (see [93]) or as a classification problem. In the following we characterize the behavior of $\widehat{\imath}_n$ as in [93]. We can write:

$$\begin{aligned}\mathbb{P}\left\{\widehat{\imath}_n = j\right\} &= \mathbb{P}\left\{\widehat{\theta}_n = \theta_j\right\} = \mathbb{P}\left\{\overline{D}_{n,j} \le \overline{D}_{n,i}, 1 \le i \le m_0, i \ne j\right\} \\ &= \mathbb{P}\left\{\overline{\mathbf{D}}_n \in \mathcal{P}_j\right\}\end{aligned}$$

where $\mathcal{P}_j$ is the (unbounded) polytope characterized by the inequalities:

$$\mathcal{P}_j := \left\{\mathbf{x} \in \mathbb{R}^{m_0} : x_j \le \min_{1\le\ell\le m_0, \ell\ne j} x_\ell\right\}.$$

Figure 3.3.1 provides a graphical representation of $\mathcal{P}_3 = \left\{\mathbf{x} \in \mathbb{R}^3 : x_3 \le \min\left\{x_1, x_2\right\}\right\}$ in $\mathbb{R}^3$.

Note that, in the previous definition, there is no special reason to use strict inequalities instead of non-strict ones. As the polytopes are a partition of the whole space, $\overline{\mathbf{D}}$ can belong to one and only one of the polytopes. But in some cases, $\overline{\mathbf{D}}$ may be on the face between two or more polytopes and, as the attribution of the faces to one polytope or another is arbitrary, we prefer to explicitly allow $\mathcal{M}^\star$ not to be a singleton. For any $j \notin \mathcal{M}^\star$, our assumptions guarantee that the probability $\mathbb{P}\left\{\overline{\mathbf{D}}_n \in \partial\mathcal{P}_j\right\}$ is asymptotically negligible, i.e. that $\mathbb{P}\left\{\overline{\mathbf{D}}_n \in \text{int}\mathcal{P}_j\right\}$ and $\mathbb{P}\left\{\overline{\mathbf{D}}_n \in \overline{\mathcal{P}}_j\right\}$ behave similarly for any $j \notin \mathcal{M}^\star$. This makes immaterial whether the faces of each $\mathcal{P}_j$ are attributed to a polytope or the

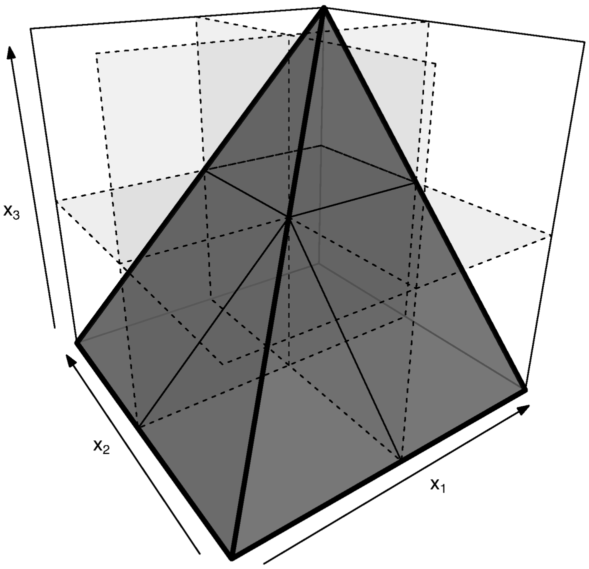

Figure 3.3.1: Polytope $\mathcal{P}_3$ in $\mathbb{R}^3$: the dark grey surfaces delimit $\mathcal{P}_3$ from above; the light grey rectangles represent the three planes $x_1 = 0$, $x_2 = 0$ and $x_3 = 0$; the dashed lines are the intersections of the planes $x_1 = 0$, $x_2 = 0$ and $x_3 = 0$; the solid thin lines are the intersections of the previous planes with the surfaces delimiting $\mathcal{P}_3$.

other.

Once the attribution of the faces is solved, the polytopes $\left\{\mathcal{P}_j, j \in \mathcal{M}^0\right\}$ form a partition of $\mathbb{R}^{m_0}$. Under appropriate assumptions (see A1 below), Khintchin's WLLN shows that $\overline{\mathbf{D}}_n$ converges in probability to $\overline{\mathbf{D}}$. This implies that:

$$\mathbb{P}\left\{\widehat{i}_n \in \mathcal{M}\right\} = \mathbb{P}\left\{\overline{\mathbf{D}}_n \in \bigcup_{j \in \mathcal{M}} \mathcal{P}_j\right\} \to \mathbb{P}\left\{\overline{\mathbf{D}} \in \bigcup_{j \in \mathcal{M}} \mathcal{P}_j\right\}$$

for any $\mathcal{M} \subset \mathcal{M}^0$. It is clear that:

$$\mathbb{P}\left\{\widehat{i}_n \in \mathcal{M}^\star\right\} = \mathbb{P}\left\{\overline{\mathbf{D}}_n \in \bigcup_{j \in \mathcal{M}^\star} \mathcal{P}_j\right\} \to 1$$

while $\mathbb{P}\left\{\widehat{i}_n = j\right\} = \mathbb{P}\left\{\overline{\mathbf{D}}_n \in \mathcal{P}_j\right\} \to 0$ for any $j \notin \mathcal{M}^\star$. In the following we show that the probabilities like $\mathbb{P}\left\{\widehat{i}_n = j\right\}$ for any $j \notin \mathcal{M}^\star$ converge to 0 exponentially fast and characterize the rate of exponential convergence of these probabilities.

This shows that our estimator $(\widehat{i}_n)$ is consistent, i.e. it takes a value in the set of minimizers of the average distance $(\mathcal{M}^\star)$ with probability converging to 1 as the number of replications (n) diverges. Moreover, the probability of the estimator $(\widehat{i}_n)$ taking any value outside the set of minimizers of the average distance $(\mathcal{M}^\star)$ converges to 0 exponentially. As in [93], the estimator has an exponential convergence rate. In the following, we try to characterize, and then estimate, the rate of convergence to 0 of this probability. This is done through large deviations principles (LDP).

As they do not belong to the toolbox of the average econometrician, it is useful to provide a brief explanation of LDP in the case that is most relevant for our purposes (Corollary 6.1.6 in [124, p. 253]). Let $\{\mathbf{X}_1, \dots\}$ be a sequence of independent and identically distributed random vectors. If $\mathbb{E}\mathbf{X}$ exists, $\overline{\mathbf{X}}_n := \frac{1}{n}\sum_{k=1}^n \mathbf{X}_k$ converges almost surely to $\mathbb{E}\mathbf{X}$. For a given set $A \subset \mathbb{R}^d$, $\mathbb{E}\mathbf{X} \in A$ implies that $\mathbb{P}\left\{\overline{\mathbf{X}}_n \in A\right\} \to 1$. Instead, if $\mathbb{E}\mathbf{X} \notin A$, $\mathbb{P}\left\{\overline{\mathbf{X}}_n \in A\right\} \to 0$ exponentially fast. In particular, $\frac{1}{n}\ln \mathbb{P}\left\{\overline{\mathbf{X}}_n \in A\right\}$ is asymptotically bracketed between two bounds depending on the distribution of $\mathbf{X}_1$ and on the set A. These bounds can respectively be expressed as the infimum over $\mathrm{int}A$ and $\overline{A}$ of a so-called rate function $\Lambda^\star$ that is the Legendre-Fenchel transform of the CGF Λ of $\mathbf{X}_1$. When the set A and the random vector $\mathbf{X}_1$ are sufficiently well behaved, the infima over $\mathrm{int}A$ and $\overline{A}$ coincide and one can identify a limit of $\frac{1}{n}\ln \mathbb{P}\left\{\overline{\mathbf{X}}_n \in A\right\}$. This limit can be expressed as the infimum of the rate function over the set A, where the infimum is generally located on the boundary of A. The point in which the infimum is reached is called a dominating point (see [358, 359]) and it can be used to express the large deviations, result in an alternative way.

The role played by the CGF and by its Cramér transform in these bounds can be justified considering the upper bound. The lower bound is too complicated to be explained here. We consider the scalar case with $A = [x, \infty)$. For every $\lambda \geq 0$, the Chernoff bound yields:

$$\begin{aligned}\mathbb{P}\left\{\overline{X}_n \in [x, \infty)\right\} =& \mathbb{E}1\left\{\overline{X}_n - x \geq 0\right\} \leq \mathbb{E}\exp\left\{\lambda\left(\overline{X}_n - x\right)\right\} \\ =& \exp\{-\lambda x\} \prod_{k=1}^{n} \mathbb{E}\exp\left\{\frac{\lambda}{n} X_k\right\} = \exp\{-\lambda x\}\left[M\left(\frac{\lambda}{n}\right)\right]^n \\ =& \exp\left\{-\lambda x + n\Lambda\left(\frac{\lambda}{n}\right)\right\}\end{aligned}$$

where we have used the inequality $1\{x \geq 0\} \leq \exp\{\lambda x\}$ for $\lambda \geq 0$. As $\lambda \geq 0$ is arbitrary, we have:

$$\begin{aligned}\mathbb{P}\left\{\overline{X}_n \in [x, \infty)\right\} \leq& \exp\left\{\inf_{\lambda \geq 0}\left[-\lambda x + n\Lambda\left(\frac{\lambda}{n}\right)\right]\right\} = \exp\left\{-\sup_{\lambda \geq 0}\left[\lambda x - n\Lambda\left(\frac{\lambda}{n}\right)\right]\right\} \\ =& \exp\left\{-n\sup_{\eta \geq 0}\left[\eta x - \Lambda(\eta)\right]\right\} = \exp\left\{-n\Lambda^\star(x)\right\}\end{aligned}$$

or $\frac{1}{n}\ln\mathbb{P}\left\{\overline{X}_n \in [x, \infty)\right\} \leq -\Lambda^\star(x)$. If $\eta^\star$ is the dominating point, i.e. the point $\eta^\star$ at which $\eta^\star x - \Lambda(\eta^\star) = \sup_{\eta \geq 0}[\eta x - \Lambda(\eta)]$, we have:

$$\mathbb{P}\left\{\overline{X}_n \in [x, \infty)\right\} \leq \exp\left\{-n\left[\eta^\star x - \Lambda(\eta^\star)\right]\right\}.$$

This justifies the logarithmic asymptotics of the probabilities and explains the rationale behind the assumptions we are going to put forward. Indeed, they guarantee that a dominating point exists or, equivalently, that $\Lambda(\cdot)$ is sufficiently well behaved that $\sup_{\eta \geq 0}[\eta x - \Lambda(\eta)]$ takes a finite value.

Now we turn to the assumptions. The following one contains the basic properties of the distances. It may be useful to explain its rationale. Each simulation run $\mathbf{z}_j(\theta_i)$ is independent of all the other runs for each $j \in \{1, \ldots, n\}$ and $i \in \mathcal{M}_0$. For fixed $i \in \mathcal{M}_0$, the simulation runs $\mathbf{z}_j(\theta_i)$ are also identically distributed across $j \in \{1, \ldots, n\}$. As a result, the distances $D_{j,i}$ share the same independence properties. This is sufficient to yield consistency and measurability of the estimator $\widehat{i}_n$ (see [93, Proposition 1, p. 280] for a slightly different result).

A1 For $k = 1, \ldots, n$ the vectors $\mathbf{D}_k$ are independent and identically distributed. For each $j \in \mathcal{M}^0$, the distances $D_{k,j}$ are independent. The mean $\mathbb{E}D_{k,j}$ exists and is finite for each $j \in \mathcal{M}^0$.

As briefly explained above, in order to obtain a LDP we need some functions related to the vector $\mathbf{D}_k$. We define the MGF and CGF of $\mathbf{D}_k$:

$$\begin{aligned}M(\mathbf{u}) = \mathbb{E}\exp\left\{\mathbf{u}'\mathbf{D}_k\right\} =& \mathbb{E}\exp\left\{\sum_{\ell=1}^{m_0} u_\ell D_{k,\ell}\right\} \\ =& \prod_{\ell=1}^{m_0} \mathbb{E}\exp\left\{u_\ell D_{k,\ell}\right\} = \prod_{\ell=1}^{m_0} M_\ell(u_\ell)\end{aligned}$$

and:

$$\Lambda(\mathbf{u}) = \sum_{\ell=1}^{m_0} \ln M_\ell(u_\ell) = \sum_{\ell=1}^{m_0} \Lambda_\ell(u_\ell).$$

We introduce the Legendre-Fenchel or Cramér transform of Λ as:

$$\begin{aligned}\Lambda^{\star}(\mathbf{y}) &:= \sup_{\mathbf{u}\in\mathbb{R}^{m_0}} \left\{\mathbf{y}'\mathbf{u} - \Lambda(\mathbf{u})\right\} = \sup_{\mathbf{u}\in\mathbb{R}^{m_0}} \sum_{\ell=1}^{m_0} \left\{y_\ell u_\ell - \Lambda_\ell(u_\ell)\right\} \\ &= \sum_{\ell=1}^{m_0} \sup_{u_\ell\in\mathbb{R}} \left\{y_\ell u_\ell - \Lambda_\ell(u_\ell)\right\} = \sum_{\ell=1}^{m_0} \Lambda_\ell^{\star}(y_\ell).\end{aligned}$$

We now state three technical assumptions. The first one (A2) guarantees that the level sets of $\Lambda^{\star}$, i.e. the sets $\left\{\mathbf{u} \in \mathbb{R}^d : \Lambda^{\star}(\mathbf{u}) \leq \alpha\right\}$, are compact, a property that is useful in order to compute the infimum of the rate function over a set (see Theorem 3.1). The second one (A3) forces the upper and the lower bounds to be equal. The third one (A4) guarantees that, for any two parameter values θ_j and θ_i, the supports of the distances $D_{k,j}$ and $D_{k,i}$ intersect, otherwise the probabilities involving them may end up to be identically equal to 0 or 1. More complete discussions of these assumptions are contained in the remarks following them.

A2 There exists $\delta > 0$ such that, for any $\eta \in (-\delta, +\delta)$, $\Lambda_\ell(\eta) < \infty$ for any $\ell \in \mathcal{M}^0$.

It may be interesting to discuss this assumption in greater detail. It is well known that the CGF at the origin is 0, as the MGF is equal to 1. This assumption requires something more, i.e. that the MGF exists in a neighborhood of the origin, and can be shown (see, e.g., [93, Lemma 1, p. 281]) to be equivalent to the requirement that the so-called Cramér condition $\mathbf{0} \in \mathrm{int}\mathcal{D}(\Lambda)$ holds true. This ensures that the level sets of $\Lambda^{\star}$, i.e. the sets $\left\{\mathbf{u} \in \mathbb{R}^d : \Lambda^{\star}(\mathbf{u}) \leq \alpha\right\}$, are compact. See [360] for large deviations principles without this assumption.

A3 Let $\mathcal{D}_\ell := \mathrm{int}\left\{u \in \mathbb{R} : \Lambda_\ell(u) < \infty\right\}$. For any $\ell \in \mathcal{M}^0$, the function Λ_ℓ is *steep*, i.e. for any sequence $\{u_m\}$ in $\mathrm{int}\mathcal{D}_\ell$ converging to a boundary point of $\mathcal{D}_\ell$:

$$\lim_m \left|\Lambda_\ell'(u_m)\right| = \infty.$$

If the aim of the analysis is to bracket asymptotically the quantity $\frac{1}{n}\ln\mathbb{P}\left\{\widehat{\theta} = \theta_j\right\}$ between two constants, one can dispense with the previous assumption. Otherwise, if the aim is to obtain a more precise limit, this assumption is not necessary but very comfortable, as the results that do not use it are generally quite complex (see, e.g., [100]). This can be linked to another problem arising in large deviations. Consider the scalar case and suppose that Λ is not steep, i.e. $\lim_m |\Lambda'(u_m)| = \lambda < \infty$ for $u_m \to u_\infty$ with $u_\infty \in \partial\mathcal{D}(\Lambda)$. Convex conjugacy implies that $\Lambda'(u) = y$ is equivalent to $u = \Lambda^{\star,\prime}(y)$. This implies $\Lambda^{\star}$ is linear with slope u_∞ for $y > \lambda$, if u_∞ is the right endpoint of $\mathcal{D}(\Lambda)$, or for $y < -\lambda$, if u_∞ is the left endpoint of $\mathcal{D}(\Lambda)$. Therefore, the rate function $\Lambda^{\star}$ is not strictly convex. In this case, large deviations principles are not guaranteed to hold (see, e.g., [120]).

A4 Let $[L_h, U_h]$ be the closure of the convex hull of the support of the law of $D_{k,h}$ for $h \in \mathcal{M}^0$. Let j be the index of the parameter under scrutiny, i.e. $\mathbb{P}\left\{\widehat{\theta} = \theta_j\right\}$. Then $U_\ell > L_j$ for any $\ell \in \mathcal{M}^0$.

In order to see what can go wrong when this assumption is not verified, consider the case when

$\mathcal{M}^0 = \{1, 2\}$. Suppose that $L_2 > U_1$ or, equivalently, that $[U_2, L_2] \cap [U_1, L_1] = \emptyset$. Then, for any n:

$$\mathbb{P}\left\{\widehat{i}_n = 1\right\} = \mathbb{P}\left\{\widehat{\theta}_n = \theta_1\right\} = \mathbb{P}\left\{\overline{D}_{n,1} \leq \overline{D}_{n,2}\right\} = 1$$

and $\mathbb{P}\left\{\widehat{i}_n = 2\right\} = 0$. This implies that $\frac{1}{n} \ln \mathbb{P}\left\{\widehat{i}_n = 2\right\} = -\infty$ and the LDP does not apply. This also suggests that combinations of parameters for which the hypothesis is not verified can be safely removed from $\mathcal{M}^0$ as they dominate or are dominated by the other ones.

Theorem 3.1. *Suppose that $j \notin \mathcal{M}^\star$. Then, under A1:*

$$\liminf \frac{1}{n} \ln \mathbb{P}\left\{\widehat{\theta} = \theta_j\right\} \geq - \inf_{\mathbf{y} \in \operatorname{int} \mathcal{P}_j} \Lambda^\star(\mathbf{y}).$$

Under A1-A2:

$$\limsup \frac{1}{n} \ln \mathbb{P}\left\{\widehat{\theta} = \theta_j\right\} \leq - \inf_{\mathbf{y} \in \overline{\mathcal{P}}_j} \Lambda^\star(\mathbf{y}).$$

Under A1-A4:

$$\lim \frac{1}{n} \ln \mathbb{P}\left\{\widehat{\theta} = \theta_j\right\} = - \inf_{\mathbf{y} \in \operatorname{int} \mathcal{P}_j} \Lambda^\star(\mathbf{y}) = - \inf_{\mathbf{y} \in \overline{\mathcal{P}}_j} \Lambda^\star(\mathbf{y}).$$

Under A1-A4:

$$\inf_{\mathbf{y} \in \overline{\mathcal{P}}_j} \Lambda^\star(\mathbf{y}) = \widetilde{\mathbf{u}}'\widetilde{\mathbf{y}} - \Lambda(\widetilde{\mathbf{u}})$$

where:

- $\widetilde{\mathbf{y}} \in \partial \mathcal{P}_j$*;*
- *the equation* $(\Lambda')^{-1}(\widetilde{\mathbf{y}}) = \widetilde{\mathbf{u}}$ *has a unique solution* $\widetilde{\mathbf{u}}$*;*
- $\mathcal{P}_j \subset H^+(\widetilde{\mathbf{u}}, \widetilde{\mathbf{y}})$.

Remark 3.1. (i) The requirement that $j \notin \mathcal{M}^\star$ is quite natural. Indeed, if $j \in \mathcal{M}^\star$, $\overline{D}_j = \min_{i \in \mathcal{M}^0} \overline{D}_i$ and $\overline{\mathbf{D}} \in \overline{\mathcal{P}}_j$. Now, $\Lambda^\star\left(\overline{\mathbf{D}}\right) = 0$ and $\Lambda^\star(\mathbf{y}) > 0$ for any $\mathbf{y} \neq \overline{\mathbf{D}}$. Therefore, if $j \in \mathcal{M}^\star$:

$$\lim \frac{1}{n} \ln \mathbb{P}\left\{\widehat{\theta} = \theta_j\right\} = - \inf_{\mathbf{y} \in \overline{\mathcal{P}}_j} \Lambda^\star(\mathbf{y}) = -\Lambda^\star\left(\overline{\mathbf{D}}\right) = 0.$$

(ii) It is easy to see why $\inf_{\mathbf{y} \in \operatorname{int} \mathcal{P}_j} \Lambda^\star(\mathbf{y}) = \inf_{\mathbf{y} \in \overline{\mathcal{P}}_j} \Lambda^\star(\mathbf{y})$ provides a measure of improbability. Each probability $\mathbb{P}\left\{\widehat{\theta} = \theta_j\right\}$ is associated with the infimum of the same function $\Lambda^\star(\cdot)$ over a different set $\mathcal{P}_j$. Now, the function $\Lambda^\star(\cdot)$ is strictly convex, positive and has a single zero in $\overline{\mathbf{D}}$. When j is such that $\mathcal{P}_j$ is far away from $\overline{\mathbf{D}}$, $\inf_{\mathbf{y} \in \mathcal{P}_j} \Lambda^\star(\mathbf{y})$ will be larger than 0. The farther away from $\overline{\mathbf{D}}$ is $\mathcal{P}_j$, the larger is $\inf_{\mathbf{y} \in \mathcal{P}_j} \Lambda^\star(\mathbf{y})$, and the faster is the convergence of $\mathbb{P}\left\{\widehat{\theta} = \theta_j\right\}$ to 0. This establishes a relation connecting the distance between $\mathcal{P}_j$ and $\overline{\mathbf{D}}$, on the one hand, and the rate of decrease of $\mathbb{P}\left\{\widehat{\theta} = \theta_j\right\}$.

(iii) This kind of large deviations result provides a formula for the logarithm of the probability. In a similar context, it was shown by [93] that exact or sharp large deviations results and saddlepoint approximations were able to provide valid approximations for the probabilities (and not only for their logarithms). However, in this paper we do not investigate exact large deviations or saddlepoint approximations. The reason is that we aim at using them for the computation of some measures of precision, and exact large deviations are not suitable for this purpose (see the end of Section 3.5).

Example 3.7. In order to illustrate the principle behind the formulas, we consider the following example. We suppose that $D_{k,1}$ is a sample of n exponential random variables with parameter 2 (and mean 1/2) and $D_{k,2}$ is a sample (independent of the previous one) of n exponential random variables with parameter 1 (and mean 1). We want to study:

$$\mathbb{P}\left\{\overline{D}_{n,1} \geq \overline{D}_{n,2}\right\} = \mathbb{P}\left\{\sum_{k=1}^{n} D_{k,1} \geq \sum_{k=1}^{n} D_{k,2}\right\} = \mathbb{P}\left\{\sum_{k=1}^{n} \mathbf{D}_k \in \mathcal{P}_2\right\},$$

where $\mathbf{D}_k := (D_{k,1}, D_{k,2})'$. We have:

$$\begin{aligned} M(\mathbf{u}) &= \frac{2}{2-u_1} \cdot \frac{1}{1-u_2}, \\ \Lambda(\mathbf{u}) &= \ln 2 - \ln(2-u_1) - \ln(1-u_2) \end{aligned}$$

and

$$\Lambda^\star(\mathbf{y}) = 2y_1 - \ln y_1 + y_2 - \ln y_2 - \ln 2e^2.$$

The equation $\widetilde{\mathbf{y}} = \Lambda'(\widetilde{\mathbf{u}})$ is:

$$\begin{cases} \widetilde{y}_1 = \frac{1}{2-\widetilde{u}_1}, \\ \widetilde{y}_2 = \frac{1}{1-\widetilde{u}_2}. \end{cases}$$

As $\widetilde{\mathbf{y}} \in \partial\mathcal{P}_2$, $\widetilde{y}_1 \equiv \widetilde{y}_2$ and $\widetilde{u}_2 = \widetilde{u}_1 - 1$. As $\mathcal{P}_2 \subset H^+(\widetilde{\mathbf{u}}, \widetilde{\mathbf{y}})$, the vector $\widetilde{\mathbf{u}}$ must be normal to the line $\widetilde{y}_1 \equiv \widetilde{y}_2$, or $\widetilde{u}_1 = -\widetilde{u}_2$. The final solution is $\widetilde{u}_1 = 1/2$, $\widetilde{u}_2 = -1/2$, and $\widetilde{y}_1 = \widetilde{y}_2 = \frac{2}{3}$. As a result:

$$\inf_{\mathbf{y}\in\overline{\mathcal{P}}_2} \Lambda^\star(\mathbf{y}) = \Lambda^\star(\widetilde{\mathbf{y}}) = \widetilde{\mathbf{u}}'\widetilde{\mathbf{y}} - \Lambda(\widetilde{\mathbf{u}}) = \ln(9/8) \doteq 0.1177830357.$$

This means that:

$$\lim \frac{1}{n} \ln \mathbb{P}\left\{\overline{D}_{n,1} \geq \overline{D}_{n,2}\right\} = -\ln(9/8).$$

Figure 3.3.2 provides a graphical representation of the function $\Lambda^\star$.

3.4 The Model Confidence Set

In this section we investigate the construction of a Model Confidence Set, i.e. a subset of $\mathcal{M}^0$ containing the models minimizing the average distance with prescribed probability. We will investigate the properties of the procedure under the assumption that n, the number of runs per each configuration of parameters, diverges.

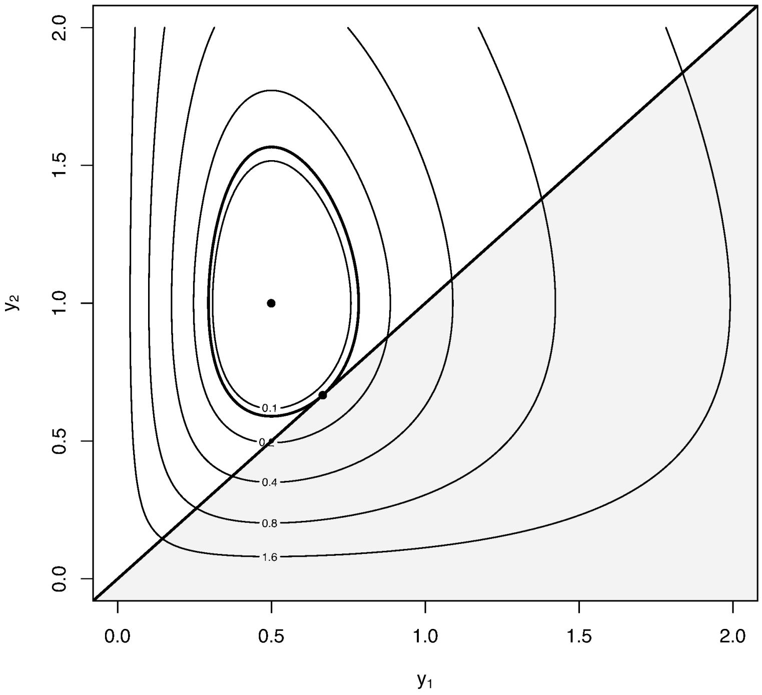

Figure 3.3.2: Example 3.7: level curves of the function $\Lambda^\star$ (in thin solid lines), level curve corresponding to $\Lambda^\star(\cdot) = \ln(9/8)$ (in thick solid line), points $\widetilde{\mathbf{y}} = (3/2, 3/2)'$ and $\overline{\mathbf{D}} = (1/2, 1)'$, half-plane $\mathcal{P}_2$ (in shaded grey area).

3.4.1 The General Procedure

The method starts from a set $\mathcal{M}^0 := \{1, \ldots, m_0\}$. In the following, $\mathcal{M} \subset \mathcal{M}^0$ denotes a generic set of discrete parameters. The iterative procedure is based on an equivalence test $\delta_{\mathcal{M}}$ and a selection rule $e_{\mathcal{M}}$, that are associated to the set $\mathcal{M}$.

The test $\delta_{\mathcal{M}}$ is used to test the null hypothesis:

$$\mathsf{H}_{0,\mathcal{M}} : \overline{D}_i = \overline{D}_j, \forall i, j \in \mathcal{M}.$$

We defined above $\mathcal{M}^\star := \left\{ j \in \mathcal{M}^0 : \overline{D}_j = \min_{i \in \mathcal{M}^0} \overline{D}_i \right\}$. Note that $\mathsf{H}_{0,\mathcal{M}^\star}$ is true while $\mathsf{H}_{0,\mathcal{M}}$ is false whenever $\mathcal{M} \neq \mathcal{M}^\star$. We also introduce the alternative hypothesis:

$$\mathsf{H}_{A,\mathcal{M}} : \exists i, j \in \mathcal{M} \text{ such that } \overline{D}_i \neq \overline{D}_j.$$

We say that $\delta_{\mathcal{M}} = 1$ when the test rejects the null hypothesis and $\delta_{\mathcal{M}} = 0$ when it does not reject it.

The elimination rule $e_{\mathcal{M}}$ is used to delete an element from $\mathcal{M}$ when $\delta_{\mathcal{M}} = 1$. We suppose that $e_{\mathcal{M}}$ takes its values in $\mathcal{M}$, so that the result of the elimination rule is to pass from $\mathcal{M}$ to $\mathcal{M} \backslash e_{\mathcal{M}}$.

One starts from the set $\mathcal{M} = \mathcal{M}^0$ and performs the test $\delta_{\mathcal{M}} = \delta_{\mathcal{M}^0}$. If the test is rejected, then an elimination step $e_{\mathcal{M}} = e_{\mathcal{M}^0}$ is performed to get a new set $\mathcal{M}_1$. The process is repeated until the test $\delta_{\mathcal{M}}$ does not reject the null hypothesis. The final set of models is called $\widehat{\mathcal{M}}^\star$. If all the tests are performed at the same significance level α, one can explicitly write $\widehat{\mathcal{M}}^\star = \widehat{\mathcal{M}}^\star_{1-\alpha}$.

3.4.2 The Implementation

In order to build a MCS, we have to choose a test procedure $\delta_{\mathcal{M}}$ and an elimination procedure $e_{\mathcal{M}}$.

As a test procedure $\delta_{\mathcal{M}}$, we suppose to estimate the mean $\overline{D}_i := \mathbb{E}_{\mathbf{z}} d\left(\mathbf{y}, \mathbf{z}\left(\theta_i\right)\right)$ and the variance $\sigma_i^2 := \mathbb{V}_{\mathbf{z}}\left[d\left(\mathbf{y}, \mathbf{z}\left(\theta_i\right)\right)\right]$ corresponding to each value of θ_i, through Gaussian quasi-likelihood. We will need the following assumptions.

A5 The variances σ_i^2 are finite for any $i \in \mathcal{M}^0$.

Consider the estimators $\overline{D}_{n,i} := \frac{1}{n} \sum_{j=1}^n d\left(y, z_j\left(\theta_i\right)\right)$ and $\widehat{\sigma}_i^2 := \frac{1}{n} \sum_{j=1}^n d^2\left(y, z_j\left(\theta_i\right)\right) - \overline{D}_{n,i}^2$. Suppose that we want to test that all $\overline{D}_i$'s are equal. We write $\overline{\mathbf{D}} = \left(\overline{D}_1, \ldots, \overline{D}_m\right)'$ and

$$\mathbf{\Sigma} := \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_m^2 \end{bmatrix}.$$

Let $\mathbf{\Sigma}_{-1}$ be the $(m-1, m-1)$-matrix obtained from $\mathbf{\Sigma}$ removing the first line and column. In the following, an estimator is indicated adding a hat to the same symbol used for the corresponding parameter. Consider the matrix $\mathbf{A}$ defined by:

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & -1 \end{bmatrix} = \begin{bmatrix} \mathbf{u}_{m-1} & -\mathbf{I}_{m-1} \end{bmatrix}.$$

The null hypothesis $\mathsf{H}_{0,\mathcal{M}}$ is that $\mathbf{A}\overline{\mathbf{D}} = \mathbf{0}_{1,m-1}$. The test statistic is:

$$W_{\mathcal{M}} = n\left(\mathbf{A}\overline{\mathbf{D}}_n\right)' \left[\widehat{\sigma}_1^2 \mathbf{U}_{m-1} + \widehat{\mathbf{\Sigma}}_{-1}\right]^{-1} \left(\mathbf{A}\overline{\mathbf{D}}_n\right).$$

As an elimination procedure $e_{\mathcal{M}}$, we choose the index $j \in \mathcal{M}$ with the largest value $\overline{D}_{n,j}$, i.e. we identify $e_{\mathcal{M}} := \arg\max_{j\in\mathcal{M}} \overline{D}_{n,j}$.

Theorem 3.2. *Let the test procedure $\delta_{\mathcal{M}}$ be based on the test $W_{\mathcal{M}}$, with asymptotic distribution $W_{\mathcal{M}} \overset{\mathcal{D}}{\to} \chi^2_{m-1}$, and let the elimination procedure $e_{\mathcal{M}}$ be based on the elimination from $\mathcal{M}$ of the index $j \in \mathcal{M}$ with the largest value $\overline{D}_{n,j}$. Then, under A1 and A5, we have:*

- $\lim_{n\to\infty} \mathbb{P}\left\{\mathcal{M}^\star \subset \widehat{\mathcal{M}}^\star_{1-\alpha}\right\} \geq 1-\alpha$,
- $\lim_{n\to\infty} \mathbb{P}\left\{i \in \widehat{\mathcal{M}}^\star_{1-\alpha}\right\} = 0 \quad i \notin \mathcal{M}^\star$.

Remark 3.2. The confidence interval obtained in this way will likely be conservative. Indeed, in Corollary 1 in [223, p. 460] it is shown that, when $\mathcal{M}^\star$ is a singleton, $\lim_{n\to\infty} \mathbb{P}\left\{\mathcal{M}^\star = \widehat{\mathcal{M}}^\star_{1-\alpha}\right\} = 1$.

There is an alternative way to see the procedure. We set $\mathcal{M}_1 := \mathcal{M}^0$. Then, let us define a sequence of subsets of $\mathcal{M}^0$ through the elimination rule as:

$$\mathcal{M}_{i+1} = \mathcal{M}_i \backslash e_{\mathcal{M}_i} \qquad i = 1, \ldots, m_0 - 1$$

or:

$$\mathcal{M}_i = \left\{e_{\mathcal{M}_i}, e_{\mathcal{M}_{i+1}}, \ldots, e_{\mathcal{M}_{m_0}}\right\}.$$

In our case, this amounts to ordering the elements $\mathcal{M}^0$ according to the value of $\overline{D}_{n,j}$, from the largest to the smallest.

To each element $e_{\mathcal{M}_i}$, we can associate the p-value of the test procedure $\delta_{\mathcal{M}_i}$ to test the null hypothesis $\mathsf{H}_{0,\mathcal{M}_i}$. We call this p-value $p_{\mathsf{H}_{0,\mathcal{M}_i}}$, with the convention that $p_{\mathsf{H}_{0,\mathcal{M}_{m_0}}} \equiv 1$. These p-values are not necessarily decreasing in i. However, it is possible to define an MCS p-value as:

$$\widehat{p}_{e_{\mathcal{M}_j}} := \max_{i\leq j} p_{\mathsf{H}_{0,\mathcal{M}_i}}.$$

The interest of the MCS p-values $\widehat{p}_{e_{\mathcal{M}_j}}$, for $j = 1, \ldots, m_0$, is that $i \in \widehat{\mathcal{M}}^\star_{1-\alpha}$ if and only if $\widehat{p}_i \geq \alpha$. This allows us to compute the MCS over a range of values α. In this case, the MCS can be used to assess the stability of the optimal solution.

3.5 Estimation of Rate Functions

The asymptotic behavior of the probability is dictated by the infimum of the rate function $\Lambda^{\star}(\cdot)$ over the polytope $\mathcal{P}_j$. This infimum can be approximated as in [134, 135, 136]: in the following we will provide a method of approximation.

3.5.1 Approximating the Rate Function

Consider the empirical moment generating function defined by $\widehat{M}(\mathbf{u}) := \prod_{\ell=1}^{m_0} \widehat{M}_{\ell}(u_{\ell})$, where $\widehat{M}_{\ell}(u_{\ell}) := \frac{1}{n}\sum_{k=1}^{n} \exp\{u_{\ell} D_{k,\ell}\}$. Let $\widehat{\Lambda}_{\ell}(u_{\ell}) := \ln \widehat{M}_{\ell}(u_{\ell})$ and $\widehat{\Lambda}(\mathbf{u}) := \sum_{\ell=1}^{m_0} \widehat{\Lambda}_{\ell}(u_{\ell})$. We define:

$$\widehat{\Lambda}^{\star}(\mathbf{y}) := \sup_{\mathbf{u}\in\mathbb{R}^{m_0}} \left[\mathbf{u}'\mathbf{y} - \widehat{\Lambda}(\mathbf{u})\right] = \sum_{\ell=1}^{m_0} \widehat{\Lambda}_{\ell}^{\star}(y_{\ell})$$

where $\widehat{\Lambda}_{\ell}^{\star}(y_{\ell}) := \sup_{u_{\ell}\in\mathbb{R}} \left\{ y_{\ell} u_{\ell} - \widehat{\Lambda}_{\ell}(u_{\ell}) \right\}$.

We will need the following assumption.

A6 Both L_i and U_i are finite for $i \in \mathcal{M}^0$.

The assumption that the support of the law of $D_{k,i}$ is bounded can be replaced by an assumption of equi-coercivity on $\widehat{\Lambda}_i^{\star}$ (see [117, Chapter 7]). As equi-coercivity is rather specialized and difficult to check, we prefer to use the simpler and more manageable assumption of boundedness, that is customary in this literature [135, 136].

Theorem 3.3. *Suppose that $j \notin \mathcal{M}^{\star}$. Under A1-A4 and A6:*

$$\lim_{n\to\infty} \inf_{\mathbf{y}\in\mathcal{P}_j} \widehat{\Lambda}^{\star}(\mathbf{y}) = \inf_{\mathbf{y}\in\mathcal{P}_j} \Lambda^{\star}(\mathbf{y})$$

and, for n large enough:

$$\inf_{\mathbf{y}\in\overline{\mathcal{P}_j}} \widehat{\Lambda}^{\star}(\mathbf{y}) = \widehat{\mathbf{u}}'\widehat{\mathbf{y}} - \widehat{\Lambda}(\widehat{\mathbf{u}})$$

where:

- $\widehat{\mathbf{y}} \in \partial\mathcal{P}_j$;
- *the equation* $\left(\widehat{\Lambda}'\right)^{-1}(\widehat{\mathbf{y}}) = \widehat{\mathbf{u}}$ *has a unique solution* $\widehat{\mathbf{u}}$;
- $\mathcal{P}_j \subset H^{+}(\widehat{\mathbf{u}}, \widehat{\mathbf{y}})$.

This approximation can be studied from two further points of view. First of all, computational issues arise in the identification of the rate function and of the dominating point. Second, one may be interested in its statistical properties and, in particular, its rate of convergence. We will investigate these points in the following sections.

3.5.2 Computation of the Rate Function

In this paper, we will estimate $\widehat{\Lambda}_\ell^\star$ through one of the algorithms in [312] and [313], in particular the LLT algorithm of [312]. The name indicates that the algorithmic complexity of the algorithm, as measured by the number of operations, is a linear function of that number. In particular, if a grid of q points is used to approximate $\widehat{\Lambda}_\ell$ and if one needs the values of $\widehat{\Lambda}_\ell^\star$ on a grid of p points, the worst-case time complexity of the LLT algorithm is $O\left(q+p\right)$. A graphical representation of the algorithm is in Figure 3.5.1 (see below for details). The figure represents the case in which $\Lambda^\star$ is estimated using $n=10$ observations extracted from an exponential distribution with parameter 1. Note that, when referring to the figure, we remove the index ℓ from functions and scalars.

Let us start with $\widehat{\Lambda}_\ell\left(u\right)=\ln\widehat{M}_\ell\left(u\right)=\ln\frac{1}{n}\sum_{k=1}^n\exp\left\{uD_{k,\ell}\right\}$ (here and in the following we drop the index ℓ from the arguments u_ℓ and y_ℓ). It is clear that:

$$\ln\frac{1}{n}\exp\left\{u\max_k D_{k,\ell}\right\}\leq\widehat{\Lambda}_\ell\left(u\right)\leq\ln\frac{1}{n}\sum_{k=1}^n\exp\left\{u\max_k D_{k,\ell}\right\}$$

and

$$-\ln n+u\max_k D_{k,\ell}\leq\widehat{\Lambda}_\ell\left(u\right)\leq u\max_k D_{k,\ell},$$

so that, for $u\to+\infty$:

$$\widehat{\Lambda}_\ell\left(u\right)\sim u\max_k D_{k,\ell}.$$

In the same way, for $u\to-\infty$:

$$\widehat{\Lambda}_\ell\left(u\right)\sim u\min_k D_{k,\ell}.$$

It can also be shown that:

$$\lim_{u\to+\infty}\widehat{\Lambda}'_\ell\left(u\right)=\max_k D_{k,\ell}$$

and

$$\lim_{u\to-\infty}\widehat{\Lambda}'_\ell\left(u\right)=\min_k D_{k,\ell}.$$

Indeed, $\widehat{\Lambda}'_\ell\left(u\right)=\frac{\sum_{k=1}^n D_{k,\ell}\exp\{uD_{k,\ell}\}}{\sum_{k=1}^n\exp\{uD_{k,\ell}\}}$ and:

$$\min_k D_{k,\ell}\leq\widehat{\Lambda}'_\ell\left(u\right)=\frac{\sum_{k=1}^n D_{k,\ell}\exp\left\{uD_{k,\ell}\right\}}{\sum_{k=1}^n\exp\left\{uD_{k,\ell}\right\}}\leq\max_k D_{k,\ell}. \tag{3.5.1}$$

We will use this fact later. The upper-left quadrant of Figure 3.5.1 represents, in black, the cumulant generating function in the exponential case $\Lambda\left(u\right)=-\ln\left(1-u\right)$, and, in grey, the estimated CGF $\widehat{\Lambda}\left(u\right)$ and the two asymptotes with slope $\max_k D_k$ and $\min_k D_k$.

Then we discuss the approximation of $\widehat{\Lambda}_\ell^\star\left(y\right)$. We recall that:

$$\widehat{\Lambda}_\ell^\star\left(y\right)=\sup_{u\in\mathbb{R}}\left\{yu-\widehat{\Lambda}_\ell\left(u\right)\right\}. \tag{3.5.2}$$

The main problem is that the functions $\widehat{\Lambda}_\ell^\star$ involve an optimization step that can turn out to be computationally intensive, especially when it has to be repeated for several values of y.

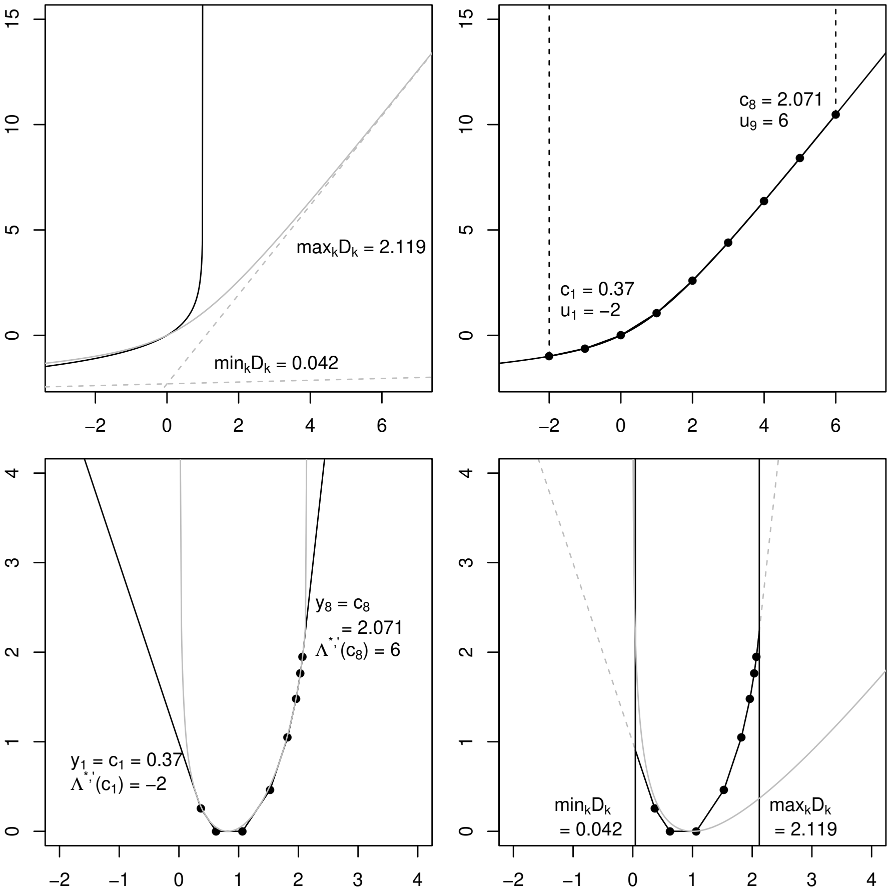

Figure 3.5.1: Illustration of the LLT algorithm with $q = 9$ for $n = 10$ observations extracted from an exponential distribution with parameter 1.
Upper-left quadrant: CGF $\Lambda(u) = -\ln(1-u)$ (in black), estimated cgf $\widehat{\Lambda}(u)$ (in grey), two asymptotes with slope $\max_k D_k$ and $\min_k D_k$.
Upper-right quadrant: points $\left\{\left(u_i, \widehat{\Lambda}(u_i)\right), i = 1, \ldots, 9\right\}$, piecewise linear approximation $\widetilde{\widehat{\Lambda}}$ to the empirical cgf $\widehat{\Lambda}$, four values c_1, u_1, c_8 and u_9.
Lower-left quadrant: points $\left\{\left(c_i, \widetilde{\widehat{\Lambda}}^{\star}(c_i)\right), i = 1, \ldots, 9\right\}$, piecewise linear approximation $\widetilde{\widehat{\Lambda}}^{\star}$ (in black), empirical Legendre transform $\widehat{\Lambda}^{\star}$ (in grey), four values $y_1 := c_1$, $u_1 = \widetilde{\widehat{\Lambda}}^{\star,\prime}(c_1)$, $y_8 := c_8$ and $u_9 = \widetilde{\widehat{\Lambda}}^{\star,\prime}(c_8)$.
Lower-right quadrant: modified version of $\widetilde{\widehat{\Lambda}}^{\star}$ with effective domain equal to $[\min_k D_k, \max_k D_k]$ (in black), true rate function $\Lambda^{\star}(y) = y - 1 - \ln y$ for the exponential case (in grey).

Most algorithms use therefore a property of the functions to avoid the optimization step. Indeed, convex conjugacy implies that $\widehat{\Lambda}'_\ell(u) = y$ is equivalent to $u = \widehat{\Lambda}^{\star,\prime}_\ell(y)$. This means that, if for a fixed u we are able to find $y := \widehat{\Lambda}'_\ell(u)$, we can identify:

$$\widehat{\Lambda}^\star_\ell(y) = u\widehat{\Lambda}'_\ell(u) - \widehat{\Lambda}_\ell(u).$$

The LLT belongs to a whole class of methods based on the *discrete Legendre transform* (DLT), i.e. the replacement in (3.5.2) of the maximum over $\mathbb{R}$ with the maximum over a finite set $\{u_1, \ldots, u_q\}$ (see Step 0 in [312, p. 176]):

$$\widehat{\Lambda}^\star_\ell(y) \simeq \sup_{u\in\{u_1,\ldots,u_q\}} \left\{yu - \widehat{\Lambda}_\ell(u)\right\}.$$

In the following, we suppose that $u_i < u_{i+1}$ for all i.

Step 1 in [312, p. 176], i.e. the convexification of the function $\widehat{\Lambda}_\ell$ through the construction of the convex hull of the point-set $\left\{\left(u_i, \widehat{\Lambda}_\ell(u_i)\right), i = 1, \ldots, q\right\}$, can be skipped as the function $\widehat{\Lambda}_\ell$ is already convex. Therefore, the function $\widehat{\Lambda}_\ell$ can be approximated on the interval $[u_1, u_q]$ by a piecewise linear function $\widetilde{\widehat{\Lambda}}_\ell$ passing through the points $\left\{\left(u_i, \widehat{\Lambda}_\ell(u_i)\right), i = 1, \ldots, q\right\}$. As our aim is to identify $y := \widehat{\Lambda}'_\ell(u)$, we introduce the slopes of the approximated function $\widetilde{\widehat{\Lambda}}_\ell$ between u_i and u_{i+1}:

$$c_i := \frac{\widehat{\Lambda}_\ell(u_{i+1}) - \widehat{\Lambda}_\ell(u_i)}{u_{i+1} - u_i}, \quad i = 1, \ldots, q-1.$$

At the points u_i, for $i = 1, \ldots, q$, the function $\widetilde{\widehat{\Lambda}}_\ell$ is not differentiable. The upper-right quadrant of Figure 3.5.1 represents the points $\left\{\left(u_i, \widehat{\Lambda}(u_i)\right), i = 1, \ldots, 9\right\}$, the piecewise linear approximation $\widetilde{\widehat{\Lambda}}$ to the empirical CGF $\widehat{\Lambda}$, and the four values c_1, u_1, c_8 and u_9.

Now (see Step 2 in [312, p. 176]), for any $c_{i-1} < y < c_i$:

$$\widetilde{\widehat{\Lambda}}^\star_\ell(y) = yu_i - \widehat{\Lambda}_\ell(u_i)$$

so that the function $\widehat{\Lambda}^\star_\ell(y)$ is approximated by a piecewise linear function with intercept $-\widehat{\Lambda}_\ell(u_i)$ and slope u_i over (c_{i-1}, c_i). When $y = c_i$, using the definition of c_i:

$$\widetilde{\widehat{\Lambda}}^\star_\ell(y) = yu_i - \widehat{\Lambda}_\ell(u_i) = yu_{i+1} - \widehat{\Lambda}_\ell(u_{i+1})$$

so that it is immaterial whether we use u_i or u_{i+1} when computing the function. The lower-left quadrant of Figure 3.5.1 represents the points $\left\{\left(c_i, \widetilde{\widehat{\Lambda}}^\star(c_i)\right), i = 1, \ldots, 9\right\}$, the piecewise linear approximation $\widetilde{\widehat{\Lambda}}^\star$ (in black), the true empirical Legendre transform $\widehat{\Lambda}^\star$ (in grey) and the four values $y_1 := c_1$, $u_1 = \widetilde{\widehat{\Lambda}}^{\star,\prime}_\ell(c_1)$, $y_8 := c_8$ and $u_9 = \widetilde{\widehat{\Lambda}}^{\star,\prime}_\ell(c_8)$.

Now we see what happens at the boundary of the function $\widetilde{\widehat{\Lambda}}^\star_\ell$ (this topic does not seem to be covered in detail in [312]).

The original algorithm does not directly approximate the function $\widehat{\Lambda}_\ell$ but rather the function $\widehat{\Lambda}_\ell + \chi_{[u_1,u_q]}$. Thus, $\widehat{\Lambda}_\ell + \chi_{[u_1,u_q]}$ is a version of $\widehat{\Lambda}_\ell$ taking the value $+\infty$ outside the interval $[u_1, u_q]$. The slopes of $\widehat{\Lambda}_\ell$ at u_1 and u_q are respectively $\widehat{\Lambda}'_\ell(u_1)$ and $\widehat{\Lambda}'_\ell(u_q)$, but are approximated by the values c_1 and c_{q-1}. As the function $\widehat{\Lambda}_\ell$ is convex, $\widehat{\Lambda}'_\ell(u_q) \geq c_{q-1}$ and $\widehat{\Lambda}'_\ell(u_1) \leq c_1$. The property of convex conjugacy implies that the function approximating $\widehat{\Lambda}^\star_\ell(y)$ on the basis of $\widehat{\Lambda}_\ell + \chi_{[u_1,u_q]}$ has slope u_1 over $(-\infty, c_1)$ and slope u_q over $(c_{q-1}, +\infty)$. As $u_1 \to -\infty$ and $u_q \to +\infty$, the slopes diverge but this can take place quite slowly. This means that the effective domain of $\widetilde{\widehat{\Lambda}}^\star_\ell$ will be $\mathbb{R}$. However, we know (see [22, Proposition 9.7]) that the effective domain of $\widehat{\Lambda}^\star_\ell$ is $\operatorname{co}\{D_{k,\ell}, k = 1, \ldots, n\} = [\min_k D_{k,\ell}, \max_k D_{k,\ell}]$, the convex hull of the points $\{D_{k,\ell}, k = 1, \ldots, n\}$.

A solution is to suppose that, outside $[u_1, u_q]$, the function $\widetilde{\widehat{\Lambda}}_\ell$ has slopes $\min_k D_{k,\ell}$, over $(-\infty, u_1)$, and $\max_k D_{k,\ell}$, over $(u_q, +\infty)$. Note that, from (3.5.1), this does not alter the convexity of $\widetilde{\widehat{\Lambda}}_\ell$, as it will remain true that $\min_k D_{k,\ell} \leq c_1 \leq \cdots \leq c_{q-1} \leq \max_k D_{k,\ell}$. This corresponds to setting the effective domain of $\widetilde{\widehat{\Lambda}}^\star_\ell$ to $[\min_k D_{k,\ell}, \max_k D_{k,\ell}]$, so that both $\widetilde{\widehat{\Lambda}}^\star_\ell$ and $\widehat{\Lambda}^\star_\ell$ will have the same effective domain. The lower-right quadrant of Figure 3.5.1 represents, in black, the modified version of $\widetilde{\widehat{\Lambda}}^\star$ as well as, in grey, the true rate function $\Lambda^\star(y) = y - 1 - \ln y$ for the exponential case.

This can also be used as a check for the choice of u_1 and u_q. Indeed, if u_1 and u_q are large enough, $c_1 - \min_k D_{k,\ell}$ and $\max_k D_{k,\ell} - c_{q-1}$ should be small. If this is not the case, one should move u_1 and u_q to achieve smaller values of these two quantities.

3.5.3 Computation of the Constrained Minimum

Once each $\widehat{\Lambda}^\star_\ell$ has been computed, we need to obtain $\inf_{\mathbf{y} \in \mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y})$, i.e. to optimize it under a constraint.

In order to avoid the inequality constraints, we use the technique of [67, pp. 72-73]. The replacements:

$$\begin{cases} y_j \mapsto x_j \\ y_\ell \mapsto x_j + x_\ell^2 \end{cases}$$

transform the computation of $\inf_{\mathbf{y} \in \mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y})$ into an unconstrained optimization problem:

$$\inf_{\mathbf{x} \in \mathbb{R}^{m_0}} \sum_{1 \leq \ell \leq m_0, \ell \neq j} \widehat{\Lambda}^\star_\ell\left(x_j + x_\ell^2\right) + \widehat{\Lambda}^\star_j(x_j).$$

Note that the objective of this technique is not to identify on which face of $\mathcal{P}_j$ the infimum is achieved, but to get a reasonably accurate value of $\inf_{\mathbf{y} \in \mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y})$ without using inequality-constrained algorithms.

Example 3.8. We continue Example 3.7. The replacement above provides:

$$\widehat{\Lambda}^\star\left(\begin{bmatrix} x_2 + x_1^2 \\ x_2 \end{bmatrix}\right) = 2\left(x_2 + x_1^2\right) - \ln\left(x_2 + x_1^2\right) + x_2 - \ln x_2 - \ln 2e^2.$$

Note that the non-negativity constraint $x_2 > 0$ does not pose any problem, as the function is not

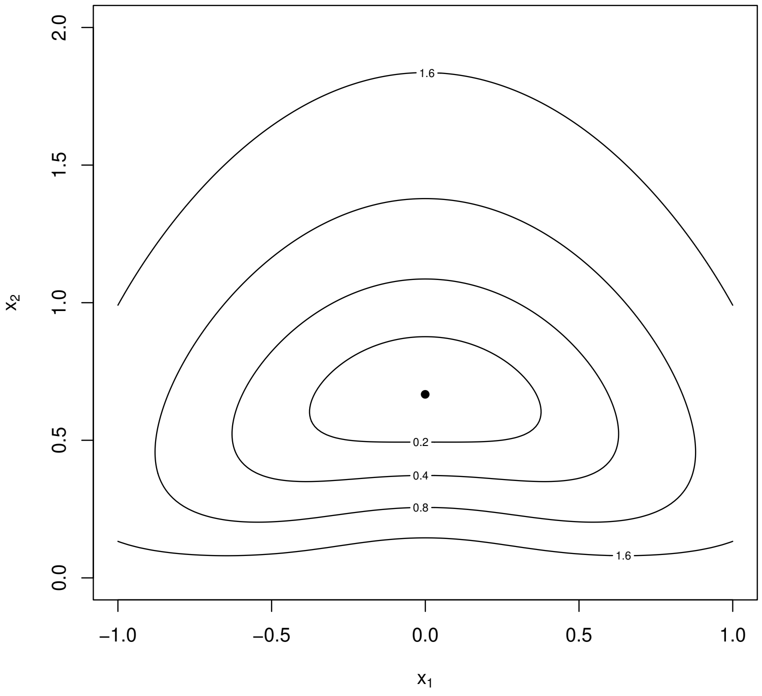

Figure 3.5.2: Example 3.8: level curves of the function $\Lambda^\star$ in the new coordinate system (in thin solid lines), point $\widetilde{\mathbf{x}} = (0, 3/2)'$.

defined for $x_2 \leq 0$. The level curves of the function are represented in Figure 3.5.2. The point $\widetilde{\mathbf{x}} = (0, 3/2)'$ corresponds to $\widetilde{\mathbf{y}} = (3/2, 3/2)'$ in the new coordinates.

3.5.4 Asymptotic Error of the Rate Function

In this paper we do not provide a full statistical analysis of this method. However, it is not difficult to obtain some hints about the behavior of the solution. The minimum $\inf_{\mathbf{y}\in\mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y})$ is reached in a point $\widehat{\mathbf{y}}$ converging to the point $\widetilde{\mathbf{y}}$ at which $\inf_{\mathbf{y}\in\overline{\mathcal{P}}_j} \Lambda^\star(\mathbf{y})$ is reached. Therefore:

$$\left| \inf_{\mathbf{y}\in\mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y}) - \inf_{\mathbf{y}\in\mathcal{P}_j} \Lambda^\star(\mathbf{y}) \right| \leq \sup_{\mathbf{y}\in\mathcal{N}(\widetilde{\mathbf{y}})} \left| \widehat{\Lambda}^\star(\mathbf{y}) - \Lambda^\star(\mathbf{y}) \right|$$

where $\mathcal{N}(\widetilde{\mathbf{y}})$ is a neighborhood of $\widetilde{\mathbf{y}}$ contained in the interior of the effective domain of $\Lambda^\star$. Using the reasoning in [398, Section 4.3], $\sup_{\mathbf{y}\in\mathcal{N}(\widetilde{\mathbf{y}})} \left| \widehat{\Lambda}^\star(\mathbf{y}) - \Lambda^\star(\mathbf{y}) \right|$ is of the same order as:

$$\sup_{\mathbf{u}\in\mathcal{N}(\widetilde{\mathbf{u}})} \left| \widehat{\Lambda}(\mathbf{u}) - \Lambda(\mathbf{u}) \right| = \sup_{\mathbf{u}\in\mathcal{N}(\widetilde{\mathbf{u}})} \left| \sum_{\ell=1}^{m_0} \widehat{\Lambda}_\ell(u_\ell) - \sum_{\ell=1}^{m_0} \Lambda_\ell(u_\ell) \right| \leq \sup_{\mathbf{u}\in\mathcal{N}(\widetilde{\mathbf{u}})} \sum_{\ell=1}^{m_0} \left| \widehat{\Lambda}_\ell(u_\ell) - \Lambda_\ell(u_\ell) \right|.$$

The behavior of $\widehat{\Lambda}_\ell - \Lambda_\ell$ has been considered by [158] (see also [41]). We first note that, provided $\widehat{M}_\ell$ is a consistent estimator of M_ℓ, the behavior of $\widehat{\Lambda}_\ell - \Lambda_\ell$ is similar to $\widehat{M}_\ell - M_\ell$:

$$\widehat{\Lambda}_\ell - \Lambda_\ell = \ln\left(1 + \frac{\widehat{M}_\ell - M_\ell}{M_\ell}\right) = \frac{\widehat{M}_\ell - M_\ell}{M_\ell} + O\left(\left(\widehat{M}_\ell - M_\ell\right)^2\right).$$

Then, we note that $\mathbb{V}\left(\widehat{M}_\ell\left(u_\ell\right)\right) = M_\ell\left(2u_\ell\right) - M_\ell^2\left(u_\ell\right)$ so that, if $M_\ell\left(2u_\ell\right)$ is infinite, $\mathbb{V}\left(\widehat{M}_\ell\left(u_\ell\right)\right)$ will be infinite too. This justifies what we say below.

Let I_ℓ be the effective domain of M_ℓ. Theorem 2.3 in [158] shows that, for any $u_\ell \in \frac{1}{2} \cdot I_\ell$, $\sqrt{n} \cdot \left(\widehat{M}_\ell - M_\ell\right)$ converges weakly to a Gaussian process. This means that $\widehat{M}_\ell - M_\ell = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$. The convergence is uniform over compact subsets of $\frac{1}{2} \cdot I_\ell$ and a similar result applies to derivatives. Theorem 2.4 in [158] extends the result to $\widehat{\Lambda}_\ell - \Lambda_\ell$, that is $\widehat{\Lambda}_\ell - \Lambda_\ell = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$. In [158, p. 460], the interval $\frac{1}{2} \cdot I_\ell$ is called zone of normal convergence.

Consider now any $u_\ell \in I_\ell \backslash \left(\frac{1}{2} \cdot I_\ell\right)$. Let $\alpha := C_\ell / u_\ell$ and C_ℓ is the extreme of I_ℓ having the same sign as u_ℓ, i.e. the (left or right) abscissa of convergence of the Laplace-Stieltjes transform. Now, under mild assumptions (see Theorem 1 in [430]), $e^{u_\ell D_{k,\ell}}$ has a tail that is regularly varying with index $-\alpha$. This means that $e^{u_\ell D_{k,\ell}}$ is the domain of attraction of a stable law with exponent α and $n^\delta\left(\widehat{M}_\ell - M_\ell\right) \to 0$ and $n^\delta\left(\widehat{\Lambda}_\ell - \Lambda_\ell\right) \to 0$ almost surely for any $\delta < \delta_0 = \alpha^{-1} - 1 = -\frac{C_\ell - u_\ell}{C_\ell}$, while the same quantities almost surely diverge for $\delta > \delta_0$ (see [158, Theorem 2.5]). The result extends to derivatives and holds uniformly over intervals in I_ℓ. Therefore, $\widehat{M}_\ell - M_\ell = o\left(n^{-\delta}\right)$ and $\widehat{\Lambda}_\ell - \Lambda_\ell = o\left(n^{-\delta}\right)$ for any $\delta < \delta_0$, where δ_0 is interpreted as the minimum value of $\max\left\{-\frac{C_\ell - u_\ell}{C_\ell}, -\frac{1}{2}\right\}$ over the interval.

Now, A6 above implies that the effective domain of the CGF (and of the MGF) is the whole real line. Therefore, also the zone of normal convergence is $\mathbb{R}$, and $\widehat{\Lambda}_\ell - \Lambda_\ell = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$ uniformly over compact subsets of $\mathbb{R}$. This means that the mean squared error of $\inf_{\mathbf{y} \in \mathcal{P}_j} \widehat{\Lambda}^\star\left(\mathbf{y}\right)$, i.e. the quantity:

$$\mathsf{MSE}\left(\inf_{\mathbf{y} \in \mathcal{P}_j} \widehat{\Lambda}^\star\left(\mathbf{y}\right)\right) = \mathbb{E}\left(\inf_{\mathbf{y} \in \mathcal{P}_j} \widehat{\Lambda}^\star\left(\mathbf{y}\right) - \inf_{\mathbf{y} \in \mathcal{P}_j} \Lambda^\star\left(\mathbf{y}\right)\right)^2$$

is expected to decrease as $O\left(n^{-1}\right)$. The following example shows that this may take place also when A6 does not hold but the value y corresponds to a value of u that is in the zone of normal convergence.

Example 3.9. In order to prove that this is the case, we have run a small simulation study. Consider the behavior of the mean of n exponential random variables X_i, $i = 1, \ldots, n$, with parameter 1. We want to study the behavior of the probability $\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n X_i \leq \frac{1}{2}\right\}$ for large n. In this case an exact characterization is possible. Indeed, $\sum_{i=1}^n X_i$ is Gamma distributed with shape n and scale 1. Therefore, the CDF of $\sum_{i=1}^n X_i$ is $\gamma\left(n, x\right) / \Gamma\left(n\right)$ and:

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n X_i \leq \frac{1}{2}\right\} = \frac{\gamma\left(n, \frac{n}{2}\right)}{\Gamma\left(n\right)}.$$

From 8.11(iii) in [366, p. 180], $\gamma\left(n, \frac{n}{2}\right) \sim \left(\frac{n}{2}\right)^{n-1} e^{-\frac{n}{2}}$ and, from 5.11.3 in [366, p. 140], $\Gamma\left(n\right) \sim$

Table 3.5.1: Example 3.9: bias, variance and MSE of the estimator $\widehat{\Lambda}^{\star}\left(\frac{1}{2}\right)$ of $\Lambda^{\star}\left(\frac{1}{2}\right)=\ln 2-\frac{1}{2}\doteq 0.1931471806$, i.e the rate function associated with exponential random variables evaluated at $\frac{1}{2}$, on the basis of $10,000$ replications.

n	bias	variance	MSE
10	0.08326409	0.09852722	0.1054601
20	0.03118596	0.02266419	0.02363675
40	0.0135553	0.009384405	0.009568151
80	0.007289302	0.004495908	0.004549042
160	0.003312765	0.002183584	0.002194558
320	0.001757503	0.001064703	0.001067792
640	0.0008518621	0.0005252574	0.000525983

$e^{-n}n^{n}\left(\frac{2\pi}{n}\right)^{\frac{1}{2}}$. At last:

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}X_i\leq\frac{1}{2}\right\}\sim\left(\frac{2}{\pi n}\right)^{\frac{1}{2}}e^{-n\left(\ln 2-\frac{1}{2}\right)}$$

where $\ln 2-\frac{1}{2}\doteq 0.1931471806$. Large deviations principles give the same solution, but in logarithmic form. Indeed:

$$\lim\frac{1}{n}\ln\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}X_i\leq\frac{1}{2}\right\}=\inf_{y\in\left(-\infty,\frac{1}{2}\right]}\Lambda^{\star}(y)=\Lambda^{\star}\left(\frac{1}{2}\right)$$

where $\Lambda^{\star}(y)=y-1-\ln y$. Now, $\Lambda^{\star}\left(\frac{1}{2}\right)=\ln 2-\frac{1}{2}\doteq 0.1931471806$. In Table 3.5.1, on the basis of $10,000$ replications, we compute the bias, variance and MSE of $\widehat{\Lambda}^{\star}\left(\frac{1}{2}\right)$ when $\widehat{\Lambda}^{\star}$ is based on a sample of n exponential random variables. Apart from an initial transient, each doubling of the sample size leads roughly to a halving of the MSE (and of the variance), thus suggesting that the rate of convergence is indeed $O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$. Indeed, the value $y=\frac{1}{2}$ corresponds to $\Lambda^{\star,\prime}\left(\frac{1}{2}\right)=-1$ and this means that $u=-1$ too, that is in the zone of normal convergence $\left(-\infty,\frac{1}{2}\right)$.

This explains why it is hopeless to use $\inf_{\mathbf{y}\in\mathcal{P}_j}\widehat{\Lambda}^{\star}(\mathbf{y})$ to reconstruct the probability $\mathbb{P}\left\{\widehat{i}_n=j\right\}=\mathbb{P}\left\{\widehat{\theta}=\theta_j\right\}$. Indeed, Theorem 3.1 can be restated as:

$$\frac{1}{n}\ln\mathbb{P}\left\{\widehat{\theta}=\theta_j\right\}=-\inf_{\mathbf{y}\in\mathcal{P}_j}\Lambda^{\star}(\mathbf{y})+o(1)$$

but more precise estimates can be obtained (see [358, 359, 245, 93]) as:

$$\frac{c}{n^{\frac{m_0}{2}}}e^{-n\inf_{\mathbf{y}\in\mathcal{P}_j}\Lambda^{\star}(\mathbf{y})}\leq\mathbb{P}\left\{\widehat{\theta}=\theta_j\right\}\leq\frac{C}{n^{\frac{1}{2}}}e^{-n\inf_{\mathbf{y}\in\mathcal{P}_j}\Lambda^{\star}(\mathbf{y})}.$$

Note that this is compatible with:

$$\frac{1}{n}\ln\mathbb{P}\left\{\widehat{\theta}=\theta_j\right\}=-\inf_{\mathbf{y}\in\mathcal{P}_j}\Lambda^{\star}(\mathbf{y})+O\left(\frac{\ln n}{n}\right).$$

Replacing $\inf_{\mathbf{y}\in\mathcal{P}_j}\Lambda^\star\left(\mathbf{y}\right)$ with $\inf_{\mathbf{y}\in\mathcal{P}_j}\widehat{\Lambda}^\star\left(\mathbf{y}\right)+O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$ yields:

$$\frac{c}{n^{\frac{m_0}{2}}}e^{-n\inf_{\mathbf{y}\in\mathcal{P}_j}\widehat{\Lambda}^\star(\mathbf{y})+O_{\mathbb{P}}\left(n^{\frac{1}{2}}\right)}\le\mathbb{P}\left\{\widehat{\theta}=\theta_j\right\}\le\frac{C}{n^{\frac{1}{2}}}e^{-n\inf_{\mathbf{y}\in\mathcal{P}_j}\widehat{\Lambda}^\star(\mathbf{y})+O_{\mathbb{P}}\left(n^{\frac{1}{2}}\right)}$$

that is not accurate enough. Nevertheless, the formulas can be used to approximate some useful quantities, such as ratios of logarithms of probabilities that measure their relative rates of decrease. For $i,j\notin\mathcal{M}^\star$:

$$\frac{\ln\mathbb{P}\left\{\widehat{\theta}=\theta_j\right\}}{\ln\mathbb{P}\left\{\widehat{\theta}=\theta_i\right\}}=\frac{\inf_{\mathbf{y}\in\mathcal{P}_j}\Lambda^\star\left(\mathbf{y}\right)}{\inf_{\mathbf{y}\in\mathcal{P}_i}\Lambda^\star\left(\mathbf{y}\right)}\left(1+O\left(\frac{\ln n}{n}\right)\right)=\frac{\inf_{\mathbf{y}\in\mathcal{P}_j}\widehat{\Lambda}^\star\left(\mathbf{y}\right)}{\inf_{\mathbf{y}\in\mathcal{P}_i}\widehat{\Lambda}^\star\left(\mathbf{y}\right)}\left(1+O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)\right).$$

3.6 Application

In this section, we consider a published model (see [30]) available on the online repository OpenABM at `https://www.comses.net/codebases/4749/releases/1.0.0/`. This is a model of "inquisitiveness in ad hoc teams" and, as already mentioned in the introduction, it has been selected because of (a) theoretical relevance, (b) applicability to economics at large, and (c) computational simulation robustness.

3.6.1 Short theoretical background

According to the authors of the model, inquisitiveness is a development of "docility", a concept introduced by Herbert A. [454, 455] to indicate individuals that lean on information, advice, recommendations, and suggestions from others to make decisions. As [434] and [433] show, this attitude requires some conditions in place before it is activated. One such conditions is the presence of a *community* of reference—i.e. the idea that individuals must feel like they belong to a group of like-minded others. This could be, for example, the community of mathematical sociologists, or that of a small organisation. Or, again, among the many examples, this "community" could also be the work team one belongs to. "Docile" individuals would fit in the team in a way such that there is exchange of information and cognition is very much distributed in an ecological, enacted, embedded sense (see [108]). The idea of inquisitiveness comes up when [30] attempt to break free from the limits of docility. While it can be a good proxy in explaining the dynamics of effective teams, docility could also point at the potentially closed loops in which some may find themselves trapped. This is due to the lack of trust in data coming from outside of the team. When a team member connects and establishes a dialogue with people in other teams then he/she is starting an enquiry. The idea is that such individuals focus on the problem at hand rather than on what can be achieved within the team per se; hence they are driven by a quest for competences, skills, and help that may not be available in just one team. As stated in [30, p. 68], "[w]e use the word 'inquisitiveness' to refer to an agent who mostly relies on *learning by inquiry* and *open explorations* of his or her own environment, including social channels" (italics in the original text).

3.6.2 ABM characteristics

The model is based on the theory of docility and attempts to expand it by considering docile (i.e. team-bound), non-docile (i.e. lone wolves), and inquisitive (i.e. docile *sans frontières*) individuals. The aim of the simulation is to understand under which circumstances the inquisitive team member adds something useful to the team. To study this aspect, the model presents problems to individuals and teams and calculates the efficiency with which they are solved. A full description of the model can be found in the paper by [30] and in the supplementary materials file, available online at the link indicated above. The inquisitiveness ABM can be classified as a highly stochastic simulation and the version run for this exercise maintains many of them. In the following, we have decided to succinctly recall only those features that are relevant to our calculations. The purpose of this exercise is to demonstrate how the Model Confidence Sets technique works, not to fully introduce and discuss assumptions and results of a computational simulation model.

Table 3.6.1 presents key parameters and describes them shortly. The organisational space—i.e. the simulation environment—features `problems` P that can take any number, but that have been set at $[100, 300]$ for this simulation. Each problem is attributed with a `difficulty` level d, distributed normally at random as indicated in Table 3.6.1. Very difficult problems (with d larger than 95% the maximum level of d) could, based on a random algorithm, `spin-off` up to 4 other (smaller) problems at each simulation step. Also, a random number of up to 3 very difficult problems increase their `difficulty` level over time at a 2% rate.

The model also features agents, called `decision makers` dm; they take two values, $[100, 300]$. Each dm has a level of `competence`—distributed normally at random (see Table 3.6.1)—that uses to solve the problems that it is set to deal with. When a problem is solved, there is an increase of competence of up to 0.30, when the problem is abandoned (not solved), then there is a decrease in competence of 0.05. In addition to this, `decision makers` are attributed a level of `docility` $sodm$ that determines their propensity towards working with others, and a level of `enquiry` that is used to cooperate with coworkers outside of one's team (i.e. an extension of docility, as described above). This latter characteristic is enabled by a binomial parameter `inquisitiveness`, when set to 'on' (or 'true').

3.6.3 Rules of interaction

The simulation works with agents moving around in the environment as they attempt to reach the target problem (selected at random). Once agents connect to a problem, they also start a number of interactions with neighbouring agents, forming ad hoc teams. All agents appear in random positions on the environment every time the simulation is ready to start. This allows to generate a random allocation of `decision makers` on a so-called organizational "problem solving space" where `problems` also appear and are found at random—i.e. not as a function of agent's `competence`. When the simulation is performed the appropriate number of times (see below), this stochasticity guarantees a variety of combinations of d, on the one hand, and $sodm$ and c on the other.

While `problems` do not move from their position, `decision makers` do so in a way that sets them to look for problems and move towards them. If they find other problems in their way—i.e.

Table 3.6.1: Parameter Notations and Values

Parameter	**Values**	**Description**
problems, $N_{P,0}$	$100, 300$	The number of problems P at time $t = 0$, at the start of the simulation.
problem spin-off, *pso*	4	The maximum rate at which problems can multiply—i.e. spin-off simpler problems.
difficulty, d	$\sim \mathcal{N}(3, 1)$	The level of difficulty each problem is associated with.
decision makers, $N_{dm,0}$	$100, 300$	Number of agents dm at time $t = 0$ in the organisation
competence, c	$\sim \mathcal{N}(1, 1.5)$	The level of knowledge that each dm carries and that can be applied to any P in order to solve it.
docility, *sodm*	$\sim \mathcal{N}(0, 1)$	Socially-oriented decision making (or, simply, docility) that is associated to each dm.
enquiry, e	$\sim \mathcal{N}(0, 1)$	Another characteristic of the dm agent, that indicates the extent to which curiosity lead it to explore knowledge of team members other than those of its own team.
inquisitiveness	true/false	A binomial parameter that enables or disables the use of enquiry in dm.
range	$6, 9$	The extent to which dm explore the environment around them to seek problems P and/or other dm to cooperate with.

problems that are not their main "goal"—they stop and attempt to solve them as well. They will resume the movement after a solution or abandonment, if that problem is still there. If not, they will look for another to deal with.

As a decision maker deals with a problem, it establishes a link to it. While this happens, it can also establish cooperation links with other decision makers in the radius indicated in Table 3.6.1 under parameter `range`. The amount of `competence` that these agents share is proportional to their level of `docility`, such that this latter parameter gives an indication of how much knowledge is actually shared. It is fair to assume that one team member does not transfer its knowledge completely to the rest of the team. Some is transferred while some is kept, for various reasons, tacit and inaccessible to others. Only agents with $sodm > \mu_{sodm} + 0.75 \cdot \sigma_{sodm}$, where μ_{sodm} is the mean docility in the system, and σ_{sodm} is its standard deviation, are sharing significant amounts of their `competence` c. Contrary to the average docile, the way in which competence is used by highly docile agents is incremental and not simply additive. That is, there is an upgrade of the knowledge gained thanks to one's own competence. Agents can make full use of this incremental add-on when the switch `inquisitiveness` is set to true, and for agents mating `docility` with high levels of `enquiry` (see Table 3.6.1).

A problem is solved and disappears from the space, when the combination of competencies in the team is more than the problem's difficulty. When a problem is too difficult for a team, then each team member evaluates its own contribution and may leave the problem after the efforts have been infused for 20 steps of the simulation.

3.6.4 Running the simulation model

In Table 3.6.2 we describe each one of the 16 configurations of parameters.

Table 3.6.2: Definition of the different configurations of parameters.

cop	**inquisitiveness**	$N_{P,0}$	$N_{dm,0}$	**range**
1	false	100	100	9
2	false	100	100	6
3	false	100	300	9
4	false	100	300	6
5	false	300	100	9
6	false	300	100	6
7	false	300	300	9
8	false	300	300	6
9	true	100	100	9
10	true	100	100	6
11	true	100	300	9
12	true	100	300	6
13	true	300	100	9
14	true	300	100	6
15	true	300	300	9
16	true	300	300	6

Note. *cop*: configuration of parameters.

The simulation model runs for 300 steps—a 'step' could be thought of as an opportunity each agent has to interact with another agent and/or with a problem—and the configurations of parameters (as per Table 3.6.2) are $2 \times 2 \times 2 \times 2 = 16$. This factorial design derives from the results shown by the proponents of this model, and have been selected to increase variability in the outcome variable as well as introduce some novelty in the understanding of how this inquisitiveness ABM works. To determine how many times an ABM with a highly stochastic component should be performed, we followed [437] and [445], and calculated power analysis for ANOVA for $\alpha = 0.01, 1 - \beta = 0.95$ and effect size of 0.1, consistent with what found in [30]. As a result, the simulation was performed 200 times per each configuration of parameters, for a total of 3200 runs.

3.6.5 Analyzing the data

In the following, we illustrate the technique outlined above. Since our aim is purely expository, we will consider what happens when the h-th element y_h of $\mathbf{y}$ is $y_h = 300 + 4.9 \cdot h$. This is motivated by the aim to establish an ideal benchmark for the output variable. In so doing, the equation represents an optimal solution threshold that is set at 90% of all problems solved at time 300. Figure 3.6.1 represents the trajectories $\mathbf{z}_j(\theta_i)$ for any $j = 1, \ldots, 200$ and $i = 1, \ldots, 16$, as well as the trajectory $\mathbf{y} = (y_1, \ldots, y_p)$. As a distance, we choose the square of the Euclidean distance, normalized dividing it by $301 \cdot 1,000,000$, where 301 is the length of the series and $1,000,000$ is a normalizing factor. Therefore:

$$d(\mathbf{y}, \mathbf{z}_j(\theta_i)) = \frac{\sum_{h=1}^{301} |y_h - z_{jh}(\theta_i)|^2}{301 \cdot 1,000,000}.$$

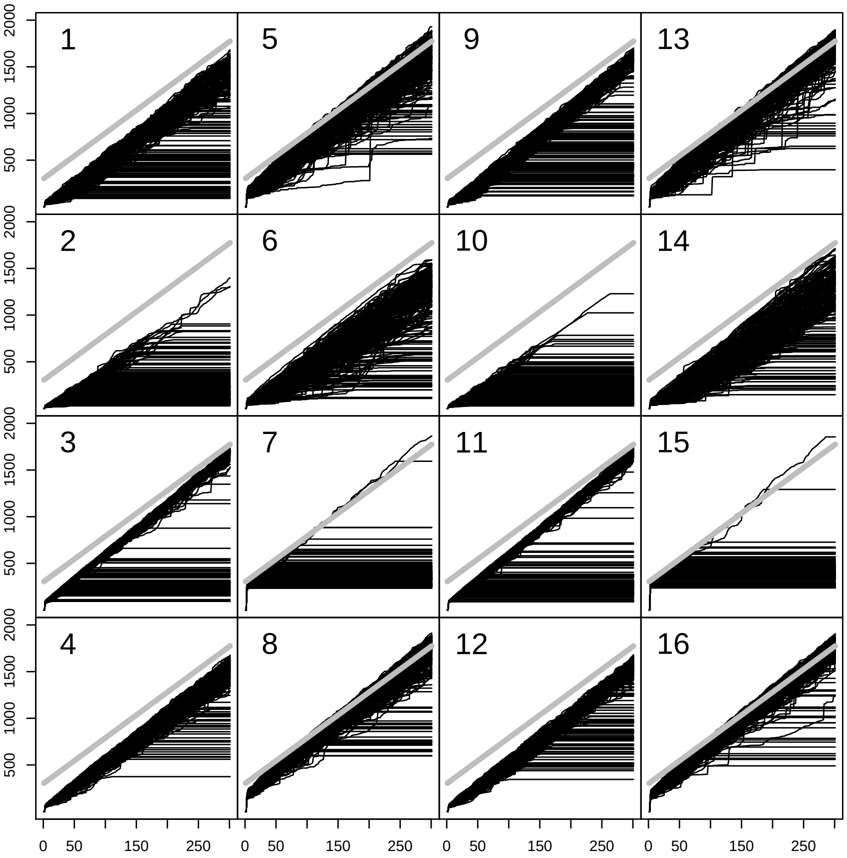

Figure 3.6.1: Trajectories $\mathbf{z}_j(\theta_i)$ for any $j = 1, \ldots, 200$ and $i = 1, \ldots, 16$ (in solid lines) and trajectory $\mathbf{y} = (y_1, \ldots, y_p)$ when $y_h = 300 + 4.9 \cdot h$ (in solid grey line).

These choices respect all the assumptions. A preliminary consideration, that will be used in the following, is that the distances in our example are bounded: indeed, the number of problems to be solved in a finite horizon is bounded and so is the distance between the benchmark and the simulated data. This automatically implies that A2, A3, A5 and A6 are verified. A1 is respected by construction. A4 is verified as shown by an analysis of the data (see also Figure 3.6.3).

In this case we have $\widehat{i}_n = 16$. The construction of the MCS is illustrated in Table 3.6.3. The MCS p-values are represented graphically in Figure 3.6.2. The Model Confidence Sets at 95% and at 99% are $\{8, 13, 16\}$.

Now we consider the large deviations rate functions associated with each one of the configurations of parameters (apart from $\widehat{i}_n$). We have approximated each $\widehat{\Lambda}_\ell$ on a grid of mesh 0.0005 from -200 to 400. This means that $q = 200,001$. The largest value taken by $c_1 - \min_k D_{k,\ell}$ and $\max_k D_{k,\ell} - c_{q-1}$ over $\ell = 1, \ldots, 16$ is 0.005270962, that seems to be small enough. The functions $\widehat{\Lambda}_\ell^\star$ are approximated each one on a grid of length $10,000$ going from $\min_k D_{k,\ell}$ to $\max_k D_{k,\ell}$. The final results are

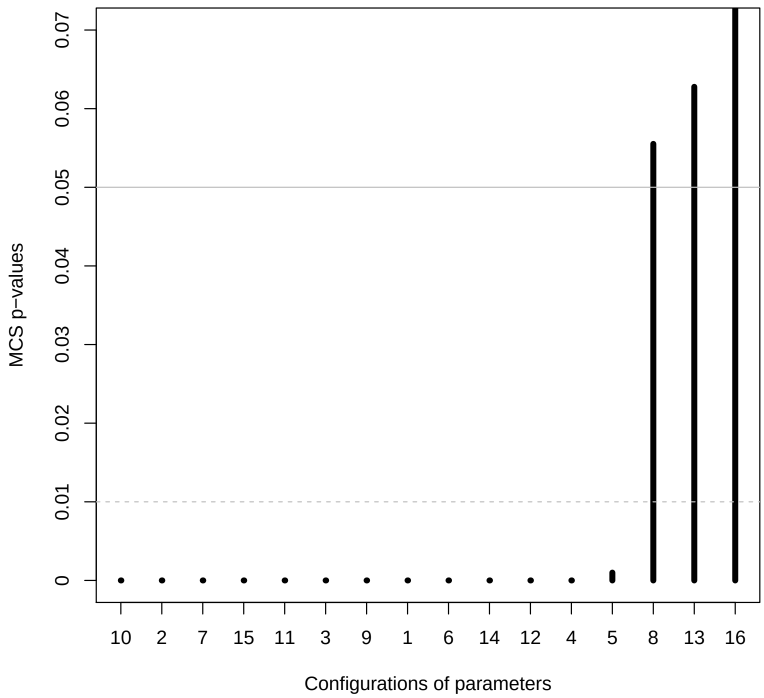

Figure 3.6.2: Graphical representation of the MCS p-values, $\widehat{p}_{e_{\mathcal{M}_k}}$, as vertical black lines and of the thresholds as horizontal grey lines.

Table 3.6.3: Order of elimination of the different configurations of parameters with means and p-values for $y_h = 300 + 4.9 \cdot h$.

k	$e_{\mathcal{M}_k}$	**mean of** $e_{\mathcal{M}_k}$	p-value of $\delta_{\mathcal{M}_k}$ ($p_{\mathsf{H}_0,\mathcal{M}_k}$)	MCS p-value ($\widehat{p}_{e_{\mathcal{M}_k}}$)
1	10	0.87555	0.00000	0.00000
2	2	0.84660	0.00000	0.00000
3	7	0.63282	0.00000	0.00000
4	15	0.62762	0.00000	0.00000
5	11	0.45591	0.00000	0.00000
6	3	0.37997	0.00000	0.00000
7	9	0.31235	0.00000	0.00000
8	1	0.30822	0.00000	0.00000
9	6	0.26487	0.00000	0.00000
10	14	0.25420	0.00000	0.00000
11	12	0.14426	0.00000	0.00000
12	4	0.10884	0.00000	0.00000
13	5	0.05956	0.00099	0.00099
14	8	0.04722	0.05552	0.05552
15	13	0.04353	0.06277	0.06277
16	16	0.02995	1.00000	1.00000

Note. *cop*: configuration of parameters.

illustrated in Figure 3.6.3. Note that the arg min of each $\widehat{\Lambda}^\star_\ell$ coincides with $\overline{D}_{n,\ell}$.

Table 3.6.4 reports the values of $\inf_{\mathbf{y}\in\mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y})$, i.e. the estimated rate of decrease of the probabilities. It can be seen that the accordance with the MCS p-values is extremely good.

Table 3.6.4: Value of the rate function $\inf_{\mathbf{y}\in\mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y})$ for all $j \neq \widehat{\imath}_n$ ordered from largest to smallest.

j	$\inf_{\mathbf{y}\in\mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y})$
10	6.93468540
2	6.14584947
15	3.96367376
7	3.70826307
14	1.35071156
6	1.35043742
1	1.10375307
9	0.91656526
11	0.71091121
3	0.54869452
12	0.49722953
4	0.36842910
5	0.03924210
8	0.01211586
13	0.00860211

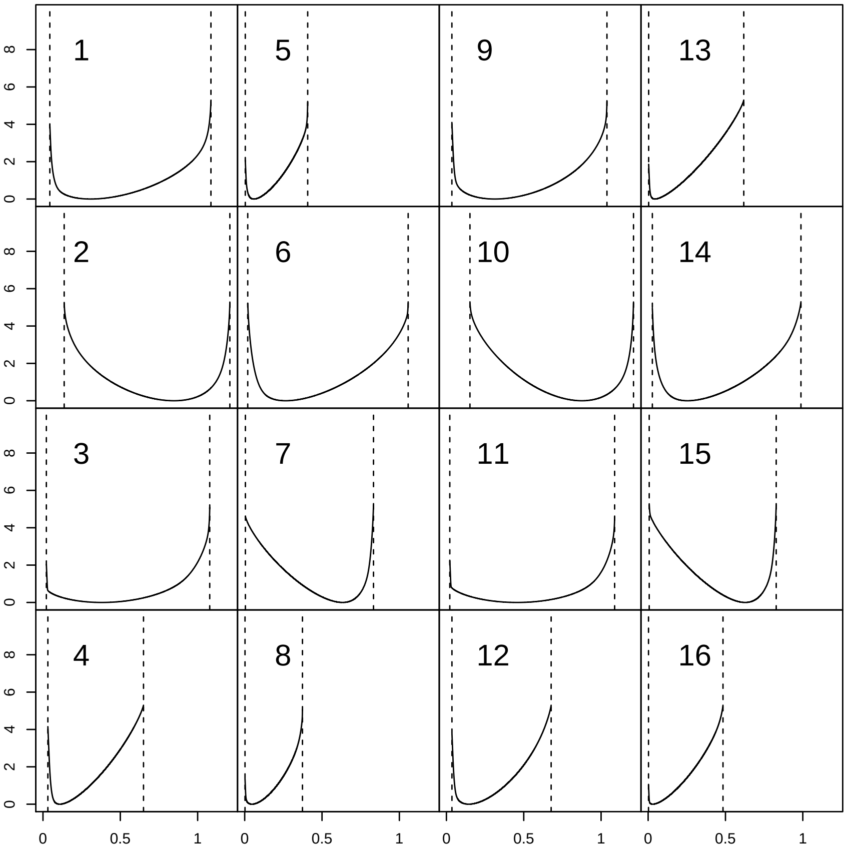

Figure 3.6.3: Rate functions (in solid line) $\widehat{\Lambda}_\ell^\star$ for all $j = 1, \ldots, 16$; boundaries of the effective domain of $\widehat{\Lambda}_\ell^\star$ (in dashed vertical lines).

3.7 Conclusions

The standard approach in calibration is to select a unique vector of parameters based on previously available estimations or on the basis of the *wisdom-of-the-crowd*. Assessing the validity of such calibration exercises is daunting as no robustness or sensitivity checks are performed. In this paper, we consider one of these possible checks, based on the comparison of several finite combinations of the model parameters. We propose to use a measure of distance to rank models from the most plausible to the least plausible. Model Confidence Sets can be employed to further restrict the number of plausible alternatives and provide a sensitivity check for the preferred specification. We also discuss a complementary analysis based on rate functions. The estimation of the latter allows the researcher to assign to all models, except the best one, an estimated rate of decrease of the probability of being the correct model. This approach has comparable interpretation to the Model Confidence Sets, and can be equivalently used to capture the distance between the chosen model and its alternatives, as our empirical application shows.

Acknowledgements

The authors are grateful to the editor, the associate editor, and three reviewers, for insightful comments which improved the original manuscript. Moreover, they would like to thank the participants of the CFE'18, EURAM 2019 and CFE'19 conferences, the ABMO4 workshop and the first meeting of the NetCIEx (Network on Counterfactual Impact Evaluation) for helpful discussions.

3.8 Appendix

Proof of Theorem 3.1. Let $\mathcal{S}$ be the closure of the convex hull of the support of the law of $\mathbf{D}_k$.

Cramér's Theorem in $\mathbb{R}^d$ (see Corollary 6.1.6 in [124, p. 253]) allows us to derive the first two statements. In particular, the lower bound holds with no supplementary assumption. The upper bound requires the Cramér condition $\mathbf{0} \in \mathrm{int}\mathcal{D}(\Lambda)$ that is verified under A2.

For the third statement, we apply part (ii) in Lemma on page 903 of [359]. Let us start from (I). Under A2, $\mathrm{int}\mathcal{D}(\Lambda)$ is non-empty. On $\mathrm{int}\mathcal{D}(\Lambda)$ the function Λ is differentiable (see [22, Proposition 9.7, p. 49]). Under A3, Λ is steep (see [124, p. 44] for a definition). Therefore, Λ is *essentially smooth* and (I) is verified. According to Corollary 6.1.6 in [124, p. 253], condition (II) is verified if $\mathbf{0} \in \mathrm{int}\mathcal{D}(\Lambda)$, that is verified under A2. As far as (III) is concerned, we first take $B \equiv \mathrm{int}\mathcal{P}_j$. From [22, Proposition 9.7, p. 49], $\mathrm{int}\mathcal{S} = \mathrm{int}\mathcal{D}(\Lambda^\star) \subset \mathcal{D}(\Lambda^\star) \subset \overline{\mathcal{S}} = \mathcal{S}$. Therefore, the condition $\mathrm{int}\left(\mathrm{int}\mathcal{P}_j \cap \mathcal{D}(\Lambda^\star)\right) \neq \emptyset$ is equivalent to:

$$\begin{aligned}\mathrm{int}\left(\mathrm{int}\mathcal{P}_j \cap \mathcal{D}(\Lambda^\star)\right) &= \mathrm{int}\mathcal{P}_j \cap \mathrm{int}\mathcal{D}(\Lambda^\star)\\ =&\mathrm{int}\mathcal{P}_j \cap \mathrm{int}\mathcal{S} = \mathrm{int}\left(\mathcal{P}_j \cap \mathcal{S}\right) \neq \emptyset.\end{aligned}$$

Under A4, the closure of the convex hull of the support of the law of $\mathbf{D}_k$ is $\prod_{1\leq \ell \leq m_0}[L_\ell, U_\ell]$. Therefore, $\mathrm{int}\mathcal{S} = \prod_{1\leq \ell \leq m_0}(L_\ell, U_\ell)$ and A4 implies $\mathrm{int}\left(\mathcal{P}_j \cap \mathcal{S}\right) \neq \emptyset$. This implies that $\inf_{\mathbf{y}\in\mathrm{int}\mathcal{P}_j}\Lambda^\star(\mathbf{y})$ is achieved at a unique point belonging to $\overline{\mathcal{P}_j} \cap \mathrm{int}\mathcal{D}(\Lambda^\star) = \overline{\mathcal{P}_j} \cap \mathrm{int}\mathcal{S}$ (see [22, Proposition 9.7, p. 49]). The same is true if we take $B \equiv \overline{\mathcal{P}_j}$, thus implying that $\inf_{\mathbf{y}\in\mathrm{int}\mathcal{P}_j}\Lambda^\star(\mathbf{y}) = \inf_{\mathbf{y}\in\overline{\mathcal{P}_j}}\Lambda^\star(\mathbf{y})$.

For the fourth statement, we use the theorem in [359, p. 904]. The conditions are easily verified. Under A2, $\mathcal{D}(\Lambda)$ contains a neighborhood of the origin. The proof that Λ is essentially smooth is provided above. The set $B = \mathcal{P}_j$ is clearly convex. The condition $\mathrm{int}\left(\mathcal{P}_j \cap \mathcal{S}\right) \neq \emptyset$ is verified above. At last, it is also clear that $\mathbb{E}\mathbf{D}_k \notin \mathcal{P}_j$. Therefore, there is a dominating point respecting the conditions in the statement (we use here the definition in [358] instead of the one in [359]). QED

Proof of Theorem 3.2. We briefly recall the framework of [223]. For any $\mathcal{M} \subset \mathcal{M}^0$, we need the following assumptions:

B1 $\limsup_{n\to\infty}\mathbb{P}\left\{\delta_{\mathcal{M}} = 1 \,|\mathsf{H}_{0,\mathcal{M}}\right\} \leq \alpha$.

B2 $\lim_{n\to\infty}\mathbb{P}\left\{\delta_{\mathcal{M}} = 1 \,|\mathsf{H}_{A,\mathcal{M}}\right\} = 1$.

B3 $\lim_{n\to\infty}\mathbb{P}\left\{e_{\mathcal{M}} \in \mathcal{M}^\star \,|\mathsf{H}_{A,\mathcal{M}}\right\} = 0$.

Under these conditions, the following results hold (see Theorem 1 in [223, p. 459]):

- $\lim_{n\to\infty}\mathbb{P}\left\{\mathcal{M}^\star \subset \widehat{\mathcal{M}}^\star_{1-\alpha}\right\} \geq 1-\alpha$,
- $\lim_{n\to\infty}\mathbb{P}\left\{i \in \widehat{\mathcal{M}}^\star_{1-\alpha}\right\} = 0 \quad i \notin \mathcal{M}^\star$.

Now we turn to the proof in our case. Under A1 and A2 it is trivial to verify that:

$$\sqrt{n}\left(\overline{\mathbf{D}}_n - \overline{\mathbf{D}}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}_{1,m}, \mathbf{\Sigma}\right).$$

Under $\mathsf{H}_{0,\mathcal{M}} : \mathbf{A}\overline{\mathbf{D}} = \mathbf{0}_{m-1}$, we have:

$$\sqrt{n}\left(\mathbf{A}\overline{\mathbf{D}}_n - \mathbf{A}\overline{\mathbf{D}}\right) = \sqrt{n}\mathbf{A}\overline{\mathbf{D}}_n \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}_{1,m-1}, \mathbf{A\Sigma A}'\right)$$

where:

$$\begin{aligned}\mathbf{A\Sigma A}' &= \left[\begin{array}{cc} \mathbf{u}_{m-1} & -\mathbf{I}_{m-1} \end{array}\right] \left[\begin{array}{cc} \sigma_1^2 & \mathbf{0}_{1,m-1} \\ \mathbf{0}_{m-1,1} & \mathbf{\Sigma}_{-1} \end{array}\right] \left[\begin{array}{c} \mathbf{u}'_{m-1} \\ -\mathbf{I}_{m-1} \end{array}\right] \\ &= \sigma_1^2 \mathbf{U}_{m-1} + \mathbf{\Sigma}_{-1}.\end{aligned}$$

Therefore, under the null hypothesis $\mathsf{H}_{0,\mathcal{M}}$, the asymptotic distribution of this test is $W_{\mathcal{M}} \overset{\mathcal{D}}{\to} \chi^2_{m-1}$. This does not change if the σ_i^2's are replaced by estimators. Therefore, Assumption B1 is verified.

Under the alternative $\mathsf{H}_{A,\mathcal{M}}$, i.e. when $\mathbf{A}\overline{\mathbf{D}} \neq \mathbf{0}_{m-1}$, the quantity $W_{\mathcal{M}}/n$ converges in probability to $\left(\mathbf{A}\overline{\mathbf{D}}\right)' \left[\sigma_1^2 \mathbf{U}_{m-1} + \mathbf{\Sigma}_{-1}\right]^{-1} \left(\mathbf{A}\overline{\mathbf{D}}\right) > 0$ and $W_{\mathcal{M}}$ diverges to infinity. Therefore, Assumption B2 is verified.

Under $\mathsf{H}_{A,\mathcal{M}}$, $\mathcal{M} \neq \mathcal{M}^\star$ and at least an element of $\mathcal{M}$ does not belong to $\mathcal{M}^\star$. It is clear that this element will have an average distance that is larger than the one observed in $\mathcal{M}^\star$. Therefore, selecting the index $j \in \mathcal{M}$ with the largest value $\widehat{\mu}_j$ will eventually lead to eliminate an element in $\mathcal{M}\backslash\mathcal{M}^\star$, as requested by Assumption B3. QED

Proof of Theorem 3.3. As seen above, the moment generating function $M(\mathbf{u})$ can be approximated through $\widehat{M}(\mathbf{u})$. We will define $\widehat{M_\ell}(u_\ell) := \frac{1}{n}\sum_{k=1}^n \exp\{u_\ell D_{k,\ell}\}$. We will introduce the notation $\widehat{\mathbb{P}}_n(\cdot) = \frac{1}{n}\sum_{k=1}^n \delta_{\mathbf{D}_k}(\cdot)$, where δ_x is the Dirac measure in x, and use $\widehat{\mathbb{E}}_n$ for the expectation under $\widehat{\mathbb{P}}_n$.

Using [233, Section 3], it is simple to see that $\widehat{M_\ell}(u_\ell)$ converges almost surely pointwise to $M_\ell(u_\ell)$. As $\widehat{M_\ell}$ and M_ℓ have in general different effective domains, it is clear that the convergence cannot be uniform. For this reason, we will use epigrahical convergence or, for short, epi-convergence (see, e.g., [117, 411]) that is especially suitable to analyse the convergence of infima and minimizers of sequences of functions, especially when they have different effective domains (see, e.g., [385, 443, 234]). Using Theorem 3.1 in [232] (or Theorem 2.3 in [92]), it is trivial to verify that the approximation is not only pointwise but also epigraphically convergent, i.e. $\text{epi} - \lim_n \widehat{M_\ell}(\cdot) = M_\ell(\cdot)$. Now, the very definition of epi-convergence and the strict positivity of $\widehat{M_\ell}(\cdot)$ imply that $\text{epi} - \lim_n \widehat{\Lambda}_\ell(\cdot) = \Lambda_\ell(\cdot)$ where $\widehat{\Lambda}_\ell(\cdot) := \ln \widehat{M_\ell}(\cdot)$. It is in general false that sums of epi-convergent functions are epi-convergent, but exploiting Proposition 6.25 in [117, p. 64] it is easy to see that $\text{epi} - \lim_n \widehat{\Lambda}(\mathbf{u}) = \Lambda(\mathbf{u})$ where $\Lambda(\mathbf{u}) := \sum_{\ell=1}^{m_0} \Lambda_\ell(u_\ell)$ and $\widehat{\Lambda}(\mathbf{u}) := \sum_{\ell=1}^{m_0} \widehat{\Lambda}_\ell(u_\ell)$. Using the continuity of the Fenchel transform with respect to epi-convergence (Wijsman's theorem, see Theorem 11.34 in [411, p. 500]), this shows that also $\Lambda^\star(\cdot)$ has an epigraphically convergent approximation, in the form:

$$\widehat{\Lambda}^\star(\mathbf{y}) := \sup_{\mathbf{u}\in\mathbb{R}^{m_0}} \left[\mathbf{u}'\mathbf{y} - \widehat{\Lambda}(\mathbf{u})\right].$$

This can at last be written as $\widehat{\Lambda}^\star(\mathbf{y}) := \sum_{\ell=1}^{m_0} \widehat{\Lambda}_\ell^\star(y_\ell)$ where $\widehat{\Lambda}_\ell^\star(y_\ell) := \sup_{u_\ell\in\mathbb{R}} \left\{y_\ell u_\ell - \widehat{\Lambda}_\ell(u_\ell)\right\}$.

Now, convergence of the minima holds true only under an equi-coercivity condition [117, Chapter

7]. To avoid messier assumptions, we assume that $\mathcal{S} = \prod_{1 \leq \ell \leq m_0} [L_\ell, U_\ell]$ is compact (see A5; this assumption seems to be common in this literature, see [134, 135]). From [22, Proposition 9.7] we know that $\mathcal{D}(\Lambda^\star) \subset \mathcal{S}$. We also know, applying the same result to the function $\widehat{\Lambda}^\star$, that $\mathcal{D}\left(\widehat{\Lambda}^\star\right) \subset \overline{\text{co}}\{\mathbf{D}_k, 1 \leq k \leq n\} \subset \mathcal{S}$. Therefore:

$$\inf_{\mathbf{y} \in \mathcal{P}_j \cap \mathcal{S}} \widehat{\Lambda}^\star(\mathbf{y}) = \inf_{\mathbf{y} \in \mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y}).$$

Now we characterize $\inf_{\mathbf{y} \in \mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y})$ in terms of its dominating point. As $\widehat{\Lambda}$ is everywhere finite for finite n, $\mathcal{D}\left(\widehat{\Lambda}\right) = \mathbb{R}^{m_0}$ contains a neighborhood of the origin. For the same reason $\widehat{\Lambda}$ is essentially smooth. The set $B = \mathcal{P}_j$ is convex. The condition $\text{int}(\mathcal{P}_j \cap \overline{\text{co}}\{\mathbf{D}_k, 1 \leq k \leq n\}) \neq \emptyset$ is justified for large n. At last, it is also clear that $\widehat{\mathbb{E}}_n \mathbf{D}_k \notin \mathcal{P}_j$ for large enough n. QED

Chapter 4

Asymptotic Properties of the Plug-in Estimator of the Discrete Entropy under Dependence[1]

This chapter is devoted to the estimation of the entropy of a discretely-supported time series through a plug-in estimator. We provide a correction of the bias and we study the asymptotic properties of the estimator. We show that the widely-used correction proposed by [418] is incorrect as it does not remove the $O\left(N^{-1}\right)$ part of the bias while ours does. We provide the asymptotic distribution and we show that it differs when the values taken by the marginal distribution of the process are equiprobable (a situation that we call *degeneracy*) and when they are not. We introduce estimators of the bias, the variance and the distribution under degeneracy and we study the estimation error. Finally, we propose a goodness-of-fit test based on entropy and give two motivations for it. The theoretical results are supported by specific numerical examples.

4.1 Introduction

Since the seminal works of [447] and [278], the entropy plays a central role in information and communication theory. The Shannon entropy and the Kullback-Leibler divergence can be observed from several perspectives but, overall, they can be classified as uncertainty measures.

The Shannon entropy has been applied to many fields of information theory, including the estimation of the entropy rate of information sources (see [378, 254, 379]), the estimation of functionals of probability distributions (see [15, 249]), the analysis of texts and symbol sequences (see [431, 269, 23]) and machine learning research (see, e.g., [466]). Other important branches of application of Shannon entropy are psychology (see [310, 311]), physics (see [199, 64, 7, 502]), and economics and finance (see [509, 332, 282]).

[1] This chapter is jointly written with Raffaello Seri.

It is therefore natural that many efforts have focused on its estimation. Many papers have been devoted to the estimation of the entropy for observations with a continuous distribution function or with a discrete one, as well as for independent and identically distributed (iid) observations or for dependent data.

The main contributions related to the case of data with continuous distributions exploit nonparametric estimation methods, such as kernel or nearest-neighbor estimators. Among these, we quote several papers dealing with iid data (see [4, 218, 303, 49]), some of which investigate the behavior of the bias, and a few works that tackle the case of time series (see, for instance, [198] and [240]). The most important results achieved in the literature are reviewed by [39].

However, the case that attracted most attention, and that we will consider in this paper, is the one of the entropy of data coming from discretely supported distributions, a situation that applies both to genuinely discrete data and to discretized (also called symbolized) ones. In the iid case, the most natural estimator is the so-called maximum likelihood or plug-in estimator, obtained replacing the discrete probability with its maximum likelihood estimator. Two facts about it were early recognized in a 1954 unpublished report by Miller and Madow. As witnessed by a summary of this paper in [310, p. 45], the authors showed that the asymptotic behavior of the estimated entropy depends on whether all values assumed by the discrete process have the same probability or not. In the second case, the statistic is asymptotically normally distributed, while in the first case it is asymptotically distributed as a chi-square. The second discovery (see also [342, 412, 85]) is that the estimator is biased in finite samples.

The case of iid discrete observations has further been explored by several authors (see [224, 381, 16, 15, 201, 378, 508, 507]). This has led to the availability of a large number of alternatives to the plug-in estimator, like the Grassberger ([199, 201]), the best upper bound ([378]), the unseen ([480]) as well as several polynomial approximation estimators ([248, 500, 249]). Many of these papers provided methods to overcome the bias. We refer to [248, 249] for a complete review of the most recent contributions in this field, and to [248, Sec. V] for an extensive simulation study comparing the performance of several estimators.

The extension of the iid case to the one of dependent data has proceeded along two directions, both of them associated with different estimation strategies. To clarify what we mean, we consider a stationary stochastic process $\{\ldots, x_{t-1}, x_t, x_{t+1}, \ldots\}$ and we use the informal notation $p(\cdot)$ $(p(\cdot\,|\cdot))$ to denote the (conditional) probability mass or density associated with its argument. The symbol $\mathbb{E}$ denotes expectation.

The first direction is associated with the *entropy rate* appearing in the Shannon–McMillan–Breiman theorem. This result states that, when $N \to \infty$, minus the normalized logarithm of the density of the time series $\{x_1, \ldots, x_N\}$, i.e. $-\frac{1}{N}\ln p(x_1, \ldots, x_N)$, converges to the entropy rate$-\mathbb{E}\ln p(x_t\,|\ldots, x_{t-1})$, defined as minus the expected value of the logarithm of the conditional probability of x_t given its past $\{\ldots, x_{t-1}\}$ (see, e.g., [8]). The relevance of this result to information theory lies in the *asymptotic equipartition property* (see [8] for some references). Estimation in this direction has been thoroughly investigated in the literature, see [200, 268, 269, 253, 180]. In dynamical systems, a similar quantity to the entropy rate appears in the *metric entropy* defined by Kolmogorov and Sinai (see, e.g., [105]), whose estimation has been considered in [202, 451, 431].

The second direction is connected with quantities appearing in several measures or tests of statistical dependence, both in time series and dynamical systems, like in [174, 173, 251, 376, 391, 373, 374, 375, 473, 494]. In this case, time series data are used to estimate the same formula of the iid case, $-\mathbb{E}\ln p\left(x_t\right)$, where the probability p is defined by the marginal distribution of the process (see, e.g. [418]). This second approach can be extended to compute the *block entropy* $-\mathbb{E}\ln p\left(x_{t-k+1},\ldots,x_{t-1},x_t\right)$, i.e. the entropy of the block $\{x_{t-k},\ldots,x_{t-1},x_t\}$. By using the properties of conditional probabilities, one can then compute the *conditional* or *differential entropy* $-\mathbb{E}\ln p\left(x_t\left|x_{t-k},\ldots,x_{t-1}\right.\right)$, the entropy of x_t conditionally on its recent past $\{x_{t-k},\ldots,x_{t-1}\}$, as a difference of block entropies. This fact seems to draw together the two directions as, letting k diverge, one recovers the entropy rate of the Shannon–McMillan–Breiman theorem. However, the two directions are generally associated with different estimation strategies and problems. Indeed, estimating the differential entropy $-\mathbb{E}\ln p\left(x_t\left|x_{t-k},\ldots,x_{t-1}\right.\right)$ as an approximation of the entropy rate $-\mathbb{E}\ln p\left(x_t\left|\ldots,x_{t-1}\right.\right)$ may introduce a severe bias (see [431, p. 416] and [180, p. 75]), unless the process is a Markov chain (see [248, pp. 2859-2860]). However, even bounding ourselves to the block entropy with small k or $k=1$, the estimation efforts in this direction have been limited. Indeed, most articles study the plug-in estimator or variants thereof and apply to the dependent case formulas derived for iid observations without modification (see [230, p. 102], [431, p. 416] and [418]). Among these, we highlight the paper of [418], who evaluates the bias and the asymptotic distribution of the entropy. However, as we will show below, the bias formulas proposed by [418], and thus his bias correction, are not correct, as are his asymptotic variance formulas.

In this paper we consider the entropy appearing in the second direction outlined above. We analyze in detail the plug-in entropy estimator H_N obtained replacing the probabilities of each value assumed by the process with their natural estimators based on a sequence of dependent observations of length N. Our aim is to fill some gaps in the literature, mainly concerning its consistency and asymptotic distribution, and to correct some incorrect results.

First of all, we show that, under stationarity, the observed entropy H_N converges almost surely to a limit H_∞ which is a random variable. Under stationary ergodicity, this limit H_∞ becomes a fixed value. We characterize the bias of H_N showing that it disappears asymptotically and, if the process is a fortiori α-mixing with $\sum_{n=1}^{\infty}\alpha\left(n\right)<\infty$, H_N has bias $O\left(N^{-1}\right)$. We then propose a bias correction and we compare it with the one proposed by [418]. The evidence shows that the correction in [418] does not remove the $O\left(N^{-1}\right)$ part of the bias while ours does. Despite the wrong correction proposed by [418], during the last twenty years many authors have considered his formulas, fostering the propagation of the error in information theory (see [211, 382, 275, 383, 377, 384, 503]), neurosciences (see [81, 260, 299, 404, 237]), physiology (see [506]), engineering (see [250]) and organizational research (see [58]).

Subsequently, we provide asymptotic distributional results under α-mixing. We show that in general the statistic, when centered and scaled by $\sqrt{N}$, has a normal asymptotic distribution but, under a condition that we call *degeneracy*, it must be scaled by N and it converges in distribution to a weighted sum of chi square random variables. The name "degeneracy" is due both to the fact that the variance of the asymptotic normal distribution is null (or degenerate) and to the fact that the entropy behaves like a degenerate V-statistic (see [441, Chapters 5 and 6]). We then propose some estimators of the bias of the entropy. One of them exploits an autocorrelation-consistent covariance matrix estimator (see [356] and [14]). The second one applies when the process is a Markov chain and features the fundamental matrix of the chain (see, for instance, [258] and [440]). Finally, we give a result on the average error induced by the estimation of bias. Our outcomes demonstrate that the Markov bias correction is more precise than the estimator based on the autocorrelation-consistent covariance matrix estimator, and bias correction slightly increases the variance of the estimator, but the mean squared error is generally improved by the corrections. In the non-degenerate case, we also address estimation of the variance of the entropy. Under degeneracy, the asymptotic distribution depends on some weights that can be estimated. However, this impacts directly on the significance level of tests. Indeed, we show that the Kolmogorov distance between the exact asymptotic distribution and the estimated one is $O_{\mathbb{P}}\left(N^{-1/2}\right)$.

At last, we provide an application of the entropy to a goodness-of-fit test for the marginal distribution of the process and we report the results of a simulation study showing the finite-sample properties of the test.

Throughout the paper, we apply our results to two different examples: a dichotomized first-order autoregressive process and the Gilbert–Shannon–Reeds model (see, e.g., [36]).

The article is organized as follows. Section 4.2 introduces some notations that will be used throughout the paper. Section 4.3 investigates the limiting behavior of the entropy and proposes formulas for its bias. Section 4.4 introduces the estimators of bias, variance and distribution under degeneracy, and provides results on the errors in the estimation. Section 4.5 propose a test of goodness-of-fit based on the entropy. Section 4.6 wraps up the main conclusions. Appendix 4.7 contains the proofs of the results.

4.2 Notation

We introduce some notation.

We write $\mathbb{N}$ for the positive integers, $\mathbb{N}_0$ for the non-negative integers and $\mathbb{R}$ for the real numbers. We follow the convention that $0 \ln 0 = 0$.

For sequences, when $n \to \infty$, we use $a_n \simeq b_n$ when $a_n = b_n \cdot (1 + o(1))$, $a_n \asymp b_n$ when $b_n/C \leq a_n \leq C b_n$ for $\infty > C > 0$ and n large enough, $a_n \ll b_n$ when $a_n = o(b_n)$, $a_n \lesssim b_n$ when $a_n \leq C b_n$ (with a_n and b_n non-negative) for $\infty > C > 0$ and n large enough. We use the same notation when the limit is with respect to a continuous variable.

We use capital bold letters, such as $\mathbf{A}$, to denote matrices and lowercase bold letters, such as $\mathbf{a}$, to denote vectors. Let $\boldsymbol{\iota}$ be a vector of ones, $\mathbf{U}$ a square matrix of ones, $\mathbf{I}$ the identity matrix, $\mathbf{0}$ a matrix or a vector of zeros. If a confusion is possible, the dimension will be indicated through an index, as in $\boldsymbol{\iota}_N$. For a vector $\mathbf{a}$, let $\bar{\mathbf{a}}$ be the vector containing the reciprocals of the elements of $\mathbf{a}$. Let $\mathrm{dg}(\mathbf{a})$ be a diagonal matrix having $\mathbf{a}$ on its diagonal. Let $\mathrm{tr}(\mathbf{A})$ be the trace of $\mathbf{A}$, i.e. the sum of the diagonal elements of a square matrix $\mathbf{A}$. For a suitable matrix $\mathbf{A}$, $\mathbf{A}'$ is its transpose, $\mathbf{A}^{\star}$ its conjugate transpose, $\mathbf{A}^{-1}$ its inverse and $\mathbf{A}^{+}$ its Moore-Penrose pseudoinverse. The element-wise power of a vector or a matrix is denoted by $\mathbf{A}^{\odot b}$ (so that $\bar{\mathbf{a}} = \mathbf{a}^{\odot(-1)}$), while $\mathbf{A}^{b}$ is the usual power obtained multiplying $\mathbf{A}$ by itself b times. The element of $\mathbf{A}$ in position (i, j) is denoted as $\mathbf{A}_{ij}$ or

$[\mathbf{A}]_{ij}$; the matrix with generic element a_{ij} is denoted $[a_{ij}]$.

The notation $\|\cdot\|_p$ indicates the Schatten norm, that is $\|\mathbf{A}\|_p := \left[\sum_i \left(s_i\left(\mathbf{A}\right)\right)^p\right]^{\frac{1}{p}}$ where s_i is the i-th singular value of $\mathbf{A}$, i.e. the square root of the i-th non-negative eigenvalue of $\mathbf{A}^\star \mathbf{A}$. We will use mainly the nuclear norm $\|\cdot\|_1$ and the Frobenius norm $\|\cdot\|_2$, also written $\|\cdot\|_F$. When applied to a vector $\mathbf{a}$, the notation $\|\cdot\|_{L^p}$ denotes the vector norm defined as $\|\mathbf{a}\|_{L^p} := \left(\sum_i |a_i|^p\right)^{\frac{1}{p}}$; when applied to a matrix $\mathbf{A}$, it denotes the matrix norm induced by the vector norm as $\|\mathbf{A}\|_{L^p} := \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{Ax}\|_{L^p}}{\|\mathbf{x}\|_{L^p}}$.

We use $\sim$, as in $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, to denote that X is distributed as the random variable on the right-hand side. The notations $\rightarrow_{\mathbb{P}}$ and $\rightarrow_{\mathcal{D}}$ denote convergence in probability and in distribution respectively. For $\sim$ and $\rightarrow_{\mathcal{D}}$ we sometimes write, with a small abuse of notation, that $X_n \rightarrow_{\mathcal{D}} X$ where X is a random variable with a given distribution. The symbols $\mathbb{E}$ and $\mathbb{V}$ respectively denote the expectation and the variance of a random variable or vector.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $T : \Omega \rightarrow \Omega$ a measurable transformation. T is called *measure-preserving* if $\mathbb{P}(TA) = \mathbb{P}(A)$ for any $A \in \mathcal{A}$. Let us define a trajectory of a stochastic process as $x(\omega) = \{\dots, x_1(\omega), x_2(\omega), \dots\}$ (note that in the following we will systematically neglect the argument ω). We can identify T with the *shift transformation*, i.e. as the function such that $x_t(T\omega) = x_{t+1}(\omega)$, so that $x(T\omega) = \{\dots, x_2(\omega), x_3(\omega), \dots\}$. In this case, a stochastic process $\{\dots, x_1, x_2, \dots\}$ is stationary if the sequences $\{\dots, x_1, x_2, \dots\}$ and $\{\dots, x_{k+1}, x_{k+2}, \dots\}$ have the same distributions, for every $k > 0$. The set A is said to be *invariant* under T if $\mathbb{P}(A \triangle TA) = 0$. The set of invariant sets under T is a σ-algebra denoted $\mathcal{I}$. T is called *ergodic* if $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$ for any $A \in \mathcal{I}$. A process is ergodic iff:

$$\begin{aligned}\lim_{K\rightarrow\infty} \frac{1}{K} \sum_{k=1}^{K-1} & \mathbb{P}\left\{\{\dots, x_{k+1}, \dots\} \in B, \{\dots, x_1, \dots\} \in A\right\} \\ & = \mathbb{P}\left\{\{\dots, x_1, \dots\} \in B\right\} \mathbb{P}\left\{\{\dots, x_1, \dots\} \in A\right\}\end{aligned}$$

for any measurable set A and B. For two sub-σ-fields $\mathcal{G}$ and $\mathcal{H}$ of $\mathcal{A}$, we define the strong and the uniform mixing coefficients as:

$$\begin{aligned}\alpha(\mathcal{G}, \mathcal{H}) &= \sup_{G\in\mathcal{G}, H\in\mathcal{H}} \left|\mathbb{P}(G \cap H) - \mathbb{P}(G)\mathbb{P}(H)\right|, \\ \varphi(\mathcal{G}, \mathcal{H}) &= \sup_{G\in\mathcal{G}, H\in\mathcal{H}} \left|\mathbb{P}(H|G) - \mathbb{P}(H)\right|.\end{aligned}$$

Let us define $\mathcal{F}_{-\infty}^t = \sigma(\dots, x_{t-1}, x_t)$ and $\mathcal{F}_{t+m}^\infty = \sigma(x_{t+m}, x_{t+m+1}, \dots)$ the σ-algebras generated by the random variables inside the parentheses. We define:

$$\begin{aligned}\alpha(m) &= \sup_t \alpha\left(\mathcal{F}_{-\infty}^t, \mathcal{F}_{t+m}^\infty\right), \\ \varphi(m) &= \sup_t \varphi\left(\mathcal{F}_{-\infty}^t, \mathcal{F}_{t+m}^\infty\right).\end{aligned}$$

We say that the process is *strong* or *α-mixing* if $\lim_{m\rightarrow\infty} \alpha(m) = 0$ and *uniform* or *φ-mixing* if $\lim_{m\rightarrow\infty} \varphi(m) = 0$.

4.3 Main Results

Consider a stochastic process $\{x_1, \dots\}$ with finite support, i.e. such that each $x_i \in \{1, 2, \dots, B\}$. This process can be genuinely discrete or can come from the symbolization of a stochastic process $\{\widetilde{x}_1, \dots\}$ whose support is divided into B intervals. The choice of a finite and bounded B may seem restrictive at first, as several recent papers consider in detail what happens when B is infinite or diverges with N (see [15, 248, 500, 249]). However, this assumption will turn out to be natural in Section 4.4 as the estimation procedures we use require finite B and, moreover, it allows us to concentrate on the main aim of this paper, ruling out several interesting but unexpected behaviors (see, in particular, [15]).

We suppose that the process $\{x_1, \dots\}$ is stationary. Note that under this assumption the probability on $\{1, 2, \dots, B\}^{\mathbb{N}}$ can be extended to a probability on $\{1, 2, \dots, B\}^{\mathbb{Z}}$ and this will allow us to refer interchangeably to one-sided $\{x_1, \dots\}$ or two-sided processes $\{\dots, x_1, \dots\}$. The hypothesis of stationarity can be generalized using the concept of asymptotic mean stationarity (see [204, 92, 203, 233]), but we will not pursue this improvement here.

The results we are going to prove are stated for the estimation of the entropy computed on the marginal distribution of the process. They can be easily adapted to the computation of the block entropy for blocks of length k. Indeed, let us consider a process $\{y_1, \dots\}$. We take a process $\{x_1, \dots\}$ where we identify $x_i := (y_i, \dots y_{i+k-1})$. If $y_i \in \{1, 2, \dots, b\}$, $x_i \in \{1, 2, \dots, b\}^k$ and it is easy to reorder the elements of this set in such a way that $x_i \in \{1, 2, \dots, B\}$ where $B = b^k$. If $\{y_1, \dots\}$ is stationary, ergodic and mixing with mixing coefficients $\alpha(m)$ $(\varphi(m))$ for $m \in \mathbb{N}$, then the process $\{x_1, \dots\}$ is respectively stationary, ergodic and mixing with mixing coefficients $\alpha(m-k+1)$ $(\varphi(m-k+1))$ for $m \in \mathbb{N}$. However, this estimator of the block entropy may suffer from some drawbacks: as the number of cells whose probability is small increases with k, the bias tends to increase and the bias corrections are less reliable.

The proportions of values equal to i is:

$$q_i = \frac{n_i}{N} = \frac{\sum_{j=1}^{N} 1\{x_j = i\}}{N}.$$

The observed entropy is therefore:

$$H_N = -\sum_{i=1}^{B} \frac{n_i}{N} \ln \frac{n_i}{N} = -\sum_{i=1}^{B} q_i \ln q_i.$$

With respect to the case in which the observations are from a sequence of iid random variables, the asymptotic theory is quite different.

We will need the following quantities, characterizing the distribution of the process:

$$\begin{aligned} p_i &= \mathbb{P}\{x_1 = i\} \quad i = 1, \dots, B \\ p_{ij}^{(h)} &= \mathbb{P}\{x_1 = i, x_{h+1} = j\} \quad i, j = 1, \dots, B,\, h \in \mathbb{N}_0. \end{aligned}$$

It is clear that $p_{ii}^{(0)} \equiv p_i$ and that $p_{ij}^{(0)} \equiv 0$ if $i \neq j$. Moreover, we will use the notation $p_i^{(h)} \equiv p_{ii}^{(h)}$.

Stationarity allows us to extend $p_{ij}^{(h)}$ to $h \in \mathbb{Z}$, in which case $p_{ij}^{(h)} = p_{ji}^{(-h)}$. We also define the vector of dichotomic variables $\mathbf{x}_j = (1\{x_j = 1\}, 1\{x_j = 2\}, \ldots, 1\{x_j = B\})'$.

In the following, we outline two examples that will be used throughout the paper to show and support our main results. The two examples concern a dichotomized first-order autoregressive process and the Gilbert–Shannon–Reed model whose behavior follows a Markov chain.

Example 4.1. [Dichotomized $\mathsf{AR}(1)$ process] Let us consider a process $\{\widetilde{x}_1, \ldots\}$ defined by the first-order autoregressive (i.e. $\mathsf{AR}(1)$) equation:

$$\widetilde{x}_i = \alpha \cdot \widetilde{x}_{i-1} + \varepsilon_i \quad i = 2, \ldots$$

where $\{\varepsilon_1, \ldots\}$ is an iid process of normally distributed random variables with mean 0 and variance $1-\alpha^2$. The initial value has the distribution $\widetilde{x}_1 \sim \mathcal{N}(0,1)$ that guarantees that the process is strictly stationary. A symbolized process requires the choice of a partition of the real line $\{\mathcal{I}_1, \ldots, \mathcal{I}_B\}$. Then:

$$\begin{aligned} p_i^{(h)} =& \mathbb{P}\{x_1 = i, x_{h+1} = i\} = \mathbb{P}\{\widetilde{x}_1 \in \mathcal{I}_i, \widetilde{x}_{h+1} \in \mathcal{I}_i\} \\ =& \mathbb{P}\left\{\mathcal{N}\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left[\begin{array}{cc} 1 & \alpha^h \\ \alpha^h & 1 \end{array}\right]\right) \in \mathcal{I}_i \times \mathcal{I}_i\right\}. \end{aligned}$$

Here we consider only a dichotomized process, i.e. a symbolized process with $B = 2$, $\mathcal{I}_1 = (-\infty, 0)$ and $\mathcal{I}_2 = [0, +\infty)$. This process retains only the signs of the original process, i.e.:

$$x_i = 1 + 1\{\widetilde{x}_i \geq 0\} \quad i \in \mathbb{N}.$$

It is clear that:

$$\begin{aligned} p_1 &= \mathbb{P}\{x_i = 1\} = \mathbb{P}\{\widetilde{x}_i \geq 0\} = {}^1\!/_2 \\ p_2 &= 1 - p_1 = {}^1\!/_2. \end{aligned}$$

As to the probabilities of couples separated by h time periods, we first derive the expressions for $\widetilde{x}_i$ as a function of $\widetilde{x}_{i-h}$:

$$\widetilde{x}_i = \alpha^h \cdot \widetilde{x}_{i-h} + \sum_{j=0}^{h-1} \alpha^j \cdot \varepsilon_{i-j}$$

or:

$$\widetilde{x}_{h+1} = \alpha^h \cdot \widetilde{x}_1 + \sum_{\ell=1}^{h} \alpha^{h-\ell} \cdot \varepsilon_{\ell+1}.$$

This implies that $\mathrm{Cov}(\widetilde{x}_1, \widetilde{x}_{h+1}) = \alpha^h \cdot \mathbb{V}(\widetilde{x}_1) = {}^{(\alpha^h \sigma^2)}\!/_{(1-\alpha^2)}$ and the correlation is α^h. From [474,

p. 189], we have:

$$\begin{aligned}
p_{22}^{(h)} &= \mathbb{P}\left\{x_1 = 2, x_{h+1} = 2\right\} = \mathbb{P}\left\{\widetilde{x}_1 \geq 0, \widetilde{x}_{h+1} \geq 0\right\} \\
&= \mathbb{P}\left\{\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \alpha^h \\ \alpha^h & 1 \end{bmatrix}\right) \in \mathbb{R}_+^2\right\} \\
&= \pi_h = {}^1\!/_4 + {}^1\!/_{2\pi} \arcsin\left(\alpha^h\right) \\
p_{11}^{(h)} &= \pi_h \\
p_{12}^{(h)} &= p_{21}^{(h)} = {}^1\!/_2 - \pi_h.
\end{aligned}$$

It is clear that $\pi_h \to {}^1\!/_4$, $|\pi_h - {}^1\!/_4| \leq {}^{\alpha^h}\!/_4$ and, for large h, $\pi_h \sim {}^1\!/_4 + {}^1\!/_{2\pi}\alpha^h$.

Example 4.2. [Gilbert–Shannon–Reeds (GSR) model] We consider the Gilbert–Shannon–Reeds model of shuffles (see, e.g., [36]), but in the following we only need a short description. Let B the number of cards in a deck. First, a number C is chosen from $\{0, 1, \ldots, B\}$ according to the binomial distribution with probabilities $\binom{B}{C} / (2B)$. Second, the first C cards are held in the left hand and the remaining $B - C$ cards in the right. Third, cards are dropped from a given hand with probability proportional to packet size. Thus, the first card is dropped from the left hand packet with probability C/B and from the right hand packet with probability $(B - C)/B$. If the first card is dropped from the left packet, the next card is dropped from the left packet with probability $(C - 1)/(B - 1)$ and from the right packet with probability $(B - C)/(B - 1)$. The process continues until there is no card left. This describes a Markov chain whose state space is the set of all possible permutations of the deck of cards, but we will not focus on this process. We will instead consider what happens to a single randomly selected card when the deck is repeatedly shuffled. Even if there is no guarantee that aggregating a Markov chain will result in a Markov chain of the same order (see, e.g., [166]), it is easy to convince oneself that what matters for the position of the card after a shuffle is the position of that same card before the shuffle, the positions of the other cards being irrelevant. The transition matrix of this Markov chain is called *position matrix* in [95]. From Lemma 2.1 in [95] or Proposition 2.1 in [19], the probability of going from state i to state j is:

$$\pi_{ij} = \begin{cases} 2^{-j} + 2^{j-1-B} & \text{if } i = j \\ 2^{j-1-B}\binom{B-j}{i-j} & \text{if } i > j \\ 2^{-j}\binom{j-1}{i-1} & \text{if } j > i \end{cases}$$

The position matrices $\mathbf{P} = [\pi_{ij}]$ are therefore given by the following formulas, valid respectively for

$B = 2, 3, 4, 5, 6$:

$$\mathbf{P} = 4^{-1} \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$

$$\mathbf{P} = 8^{-1} \begin{bmatrix} 5 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 5 \end{bmatrix}$$

$$\mathbf{P} = 16^{-1} \begin{bmatrix} 9 & 4 & 2 & 1 \\ 3 & 6 & 4 & 3 \\ 3 & 4 & 6 & 3 \\ 1 & 2 & 4 & 9 \end{bmatrix}$$

$$\mathbf{P} = 32^{-1} \begin{bmatrix} 17 & 8 & 4 & 2 & 1 \\ 4 & 10 & 8 & 6 & 4 \\ 6 & 6 & 8 & 6 & 6 \\ 4 & 6 & 8 & 10 & 4 \\ 1 & 2 & 4 & 8 & 17 \end{bmatrix}$$

$$\mathbf{P} = 64^{-1} \begin{bmatrix} 33 & 16 & 8 & 4 & 2 & 1 \\ 5 & 18 & 16 & 12 & 8 & 5 \\ 10 & 8 & 12 & 12 & 12 & 10 \\ 10 & 12 & 12 & 12 & 8 & 10 \\ 5 & 8 & 12 & 16 & 18 & 5 \\ 1 & 2 & 4 & 8 & 16 & 33 \end{bmatrix}.$$

The stationary probability of these Markov chains is a uniform distribution on the states, so that $p_i = B^{-1}$ for $i = 1, \ldots, B$. The probabilities $p_{ij}^{(h)}$ can then be easily obtained multiplying the position matrices and the stationary probability.

4.3.1 Limiting behavior

A trivial application of the Birkhoff Ergodic Theorem and of, e.g., Corollary 6.3.1 in [405, p. 174] yields almost sure convergence of H_N to a limiting random variable measurable with respect to the invariant σ-algebra $\mathcal{I}$.

Proposition 4.1. *Under stationarity:*

$$\lim_{N\to\infty} H_N = H_\infty = -\sum_{i=1}^{B} \mathbb{P}\left\{x_1 = i \,|\mathcal{I}\right\} \ln \mathbb{P}\left\{x_1 = i \,|\mathcal{I}\right\} \quad \mathbb{P}-\mathsf{as}$$

where H_∞ is an invariant random variable. Under stationary ergodicity, $H_\infty = -\sum_{i=1}^{B} p_i \ln p_i$ is a constant.

4.3.2 Bias

The nonlinearity introduced by the logarithm implies that H_N will not be an unbiased estimator of H_∞. Moreover, it is well known that the bias is always negative (see, e.g., [378, Proposition 1] and note that the proof does not depend on the iid assumption). The next result characterizes the bias of H_N.

Proposition 4.2. *If the process $\{x_1, \dots\}$ is stationary ergodic:*

$$\mathbb{E}\left[H_N\right] - H_\infty = -\frac{B-1}{2N} - \frac{1}{N}\sum_{i=1}^{B} \frac{\sum_{h=1}^{N-1}\left(p_i^{(h)} - p_i^2\right)}{p_i} + o\left(1\right)$$

where $-\frac{B-1}{2N} - \frac{1}{N}\sum_{i=1}^{B} \frac{\sum_{h=1}^{N-1}\left(p_i^{(h)} - p_i^2\right)}{p_i} \leq 0$ and the right-hand side is $o\left(1\right)$. If the process $\{x_1, \dots\}$ is α-mixing with $\sum_{n=1}^{\infty} \alpha\left(n\right) < \infty$, then:

$$\mathbb{E}\left[H_N\right] - H_\infty = -\frac{B-1}{2N} - \frac{1}{N}\sum_{i=1}^{B} \frac{\sum_{h=1}^{\infty}\left(p_i^{(h)} - p_i^2\right)}{p_i} + o\left(N^{-1}\right)$$

where the right-hand side is indeed $O\left(N^{-1}\right)$.

Remark 4.1. (i) Positive values of the covariance $\mathrm{Cov}\left(1\left\{x_j = i\right\}, 1\left\{x_\ell = i\right\}\right) = p_i^{(j-\ell)} - p_i^2$, for $i = 1, \dots, B$ and $j, \ell \in \mathbb{N}$, for most values of the indices are sometimes used as an indicator of persistence of the stochastic process$\{x_1, \dots\}$, often defined as the tendency to assume in a time period values that are near to the ones of previous time periods. Persistent stochastic processes will usually have $\sum_{i=1}^{B} \frac{\sum_{j=1}^{N-1}\left(p_i^{(j)} - p_i^2\right)}{p_i N} > 0$. This implies not only that the observed entropy is systematically biased downwards from the true entropy, but that this effect is stronger for the case of stochastic processes with persistence. Antipersistence can instead reduce the bias.
(ii) Under ergodic stationarity, we have:

$$\frac{1}{N-1}\sum_{h=1}^{N-1}\left(p_i^{(h)} - p_i^2\right) \to 0$$

but this term is not necessarily $O\left(N^{-1}\right)$ and so isn't the bias.
(iii) In the rest of the paper, and especially in Section 4.4, we will use the following definition, valid under α-mixing with $\sum_{n=1}^{\infty} \alpha\left(n\right) < \infty$:

$$\mathrm{bias}\left(H_N\right) := -\frac{B-1}{2N} - \frac{1}{N}\sum_{i=1}^{B} \frac{\sum_{h=1}^{\infty}\left(p_i^{(h)} - p_i^2\right)}{p_i}.$$

The reason is that most results on which we will rely for the estimation of $\mathrm{bias}\left(H_N\right)$ require conditions stronger than $\sum_{n=1}^{\infty} \alpha\left(n\right) < \infty$ (see, e.g., [14]).
(iv) It is interesting to see what this result implies for the stationary not necessarily ergodic case. For a stationary ergodic process, the time average and the ensemble average coincide and the bias

correction is quite simple to understand and, as we will see below, implement. However, in the general case of a stationary process, the limit of H_N is the time average H_∞, that is an $\mathcal{I}$-measurable random variable, and a bias correction for the time average should be an $\mathcal{I}$-measurable random variable too. Nevertheless, in most applications one observes a single time series. It is well known that a stationary process can be written as a mixing of ergodic processes with respect to a measure that, e.g., is called *contingency law* in [479]. This means that each single realization of a stationary process is obtained, first, extracting a random value from the contingency law and, second, extracting a realization from the ergodic process associated with the previous random value. This implies that each time series is extracted from an ergodic process whose properties can be inferred using the Ergodic Theorem, but nothing can be inferred about the contingency law. This point of view is made very clear in [316, p. 202] with reference to prediction. As a result, the bias correction applies to the entropy computed on the single trajectory, that can be supposed to be extracted from an ergodic law.

Example 4.3. [Dichotomized $\mathsf{AR}(1)$ process - Example 4.1 continued] The process $\{x_1, \dots\}$ is ergodic and mixing with $\alpha(h) \leq {}^1\!/_{2\pi} \arcsin\left(\alpha^h\right) \leq {}^{\alpha^h}\!/_{2\pi}$. Therefore, we have:

$$\begin{aligned}
\mathbb{E}[H_N] &= H_\infty - \frac{1}{2N} - \frac{2}{\pi}\left(\frac{\sum_{j=1}^{N-1}\left(1-\frac{j}{N}\right)\arcsin\left(\alpha^j\right)}{N}\right) + o\left(N^{-1}\right) \\
&= H_\infty - \frac{1}{2N} - \frac{2}{\pi N}\left(\sum_{j=1}^{\infty}\arcsin\left(\alpha^j\right)\right) + o\left(N^{-1}\right).
\end{aligned}$$

In Figure 4.3.1 we show the performance of the bias correction. For the moment, as we have not yet considered estimation, we correct H_∞ to approximate $\mathbb{E}[H_N]$.[2] The trajectories of H_N are represented by the grey jigsaw lines that oscillate around the dark grey line representing $\mathbb{E}H_N$. They converge from below towards the fixed limiting value H_∞. In finite samples, H_N is a badly biased estimator of H_∞. We show two corrections to H_∞, i.e. in black dotted line $H_\infty - {}^1\!/_{2N}$, the correction for the iid case (the one proposed in [418] for the time-series case), and in black dashed line $H_\infty - {}^1\!/_{2N} - {}^2\!/_{\pi N}\left(\sum_{j=1}^{\infty}\arcsin\left(\alpha^j\right)\right)$, our correction. It is clear that our correction is much better than the one in [418]. On the right plot, we display the empirical cdf of H_N with $N = 25$ (black dashed line), $N = 50$ (black dotted line), $N = 100$ (black dash-dot line), $N = 200$ (black solid line). This shows that in the ergodic case H_N converges (almost surely) to H_∞.

4.3.3 Central Limit Theorem

The following proposition provides an asymptotic distributional result.

Proposition 4.3. *If the process $\{x_1, \dots\}$ is α-mixing with $\sum_{n=1}^{\infty}\alpha(n) < \infty$, we have:*

$$\sqrt{N}\left(H_N - H_\infty\right) \to_{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\right)$$

[2] In general, the bias correction is applied to H_N in order to reduce its distance with respect to the value H_∞ that is being estimated (see Example 4.7).

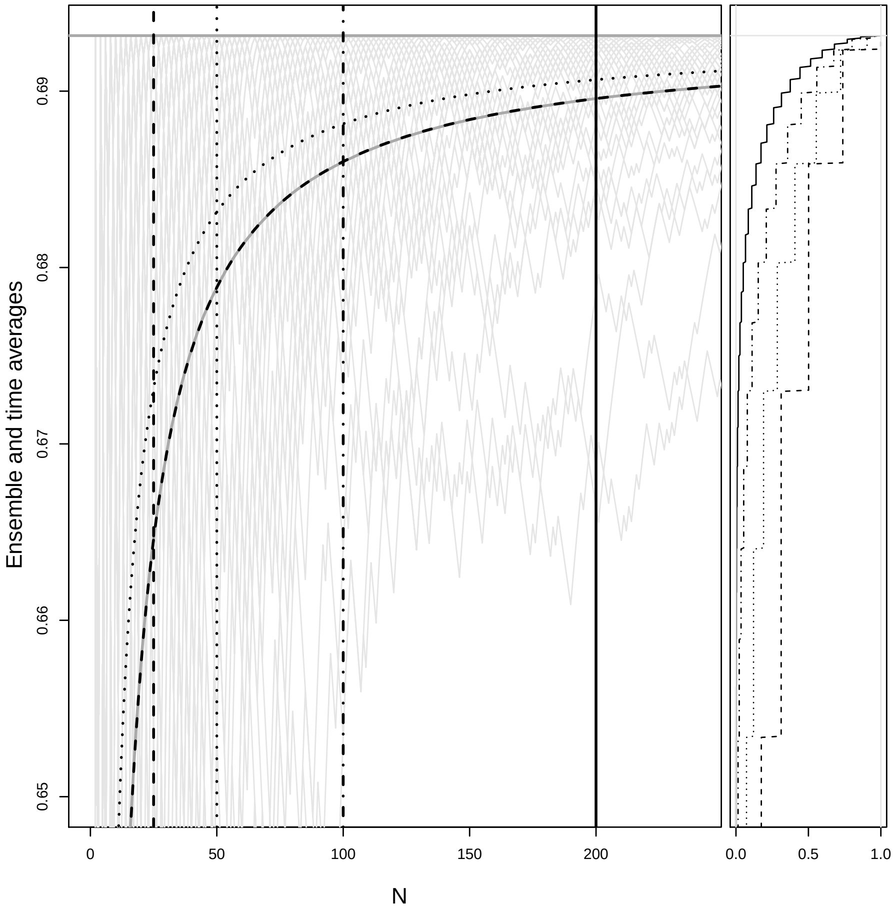

Figure 4.3.1: Ensemble and time averages of the entropy in the ergodic (dichotomized $\mathsf{AR}\,(1)$) case: on the left plot, 50 trajectories of H_N as a function of N (light grey jigsaw lines), H_∞ (dark grey horizontal line), $\mathbb{E}H_N$ (dark grey curved line), $H_\infty - {}^1\!/_{2N}$ (black dotted line), $H_\infty - {}^1\!/_{2N} - {}^2\!/_{\pi N}\left(\sum_{j=1}^{\infty} \arcsin\left(\alpha^j\right)\right)$ (black dashed line), vertical lines at $N \in \{25, 50, 100, 200\}$ (respectively black dashed, dotted, dash-dot, solid lines); on the right plot, empirical cdf of H_N with $N = 25$ (black dashed line), $N = 50$ (black dotted line), $N = 100$ (black dash-dot line), $N = 200$ (black solid line).

where

$$\sigma^2 := \sum_{i=1}^{B} p_i \ln^2 p_i - \left(\sum_{i=1}^{B} p_i \ln p_i \right)^2 + 2 \sum_{i=1}^{B} \sum_{i'=1}^{B} \sum_{h=1}^{\infty} \left(\frac{p_{ii'}^{(h)} + p_{i'i}^{(h)}}{2} - p_i p_{i'} \right) \ln p_i \ln p_{i'}.$$

Provided $\sigma^2 \neq 0$ *and* $\varphi(n) \leq \kappa (n+1)^{-2}$ *for any* n*:*

$$\left\| F_{\sqrt{N}(H_N - H_\infty)/\sigma} - \Phi \right\|_\infty = O\left(N^{-1/2}\right).$$

Remark 4.2. (i) When $p_i = B^{-1}$ for any i, the asymptotic variance annihilates. For the iid case, this was remarked by Miller and Madow in 1954 (see [310, p. 45]) but was later overlooked by [31] (see [224, pp. 326-327]). Here we show that the same result extends to the case in which the observations are dependent. The asymptotic distribution in this case is dealt with in Section 4.3.4.
(ii) The Berry-Esséen bound involves a condition on the uniform mixing coefficients because Berry-Esséen bounds for the strong mixing case are less satisfactory (see [407, Theorem 2, Remark 2, Application 2]).

Example 4.4. [Dichotomized $\mathsf{AR}(1)$ process - Examples 4.1, 4.3 continued] In this case we have $\sigma^2 = 0$.

Combining together Propositions 4.2 and 4.3, we obtain the following trivial result.

Corollary 4.1. *If the process is* α*-mixing with* $\sum_{n=1}^{\infty} \alpha(n) < \infty$*, we have:*

$$\sqrt{N} \left(H_N - \text{bias}(H_N) - H_\infty \right) \to_{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\right).$$

4.3.4 Asymptotic distribution under degeneracy

Now we turn to the properties when $p_i = B^{-1}$ for any i. In the following proposition we need the definition of a matrix $\mathbf{\Omega}$. In Section 4.4 we will show that this is a modification of a covariance matrix $\mathbf{\Sigma}$.

Proposition 4.4. *Suppose that the process* $\{x_1, \dots\}$ *is* α*-mixing with* $\sum_{n=1}^{\infty} \alpha(n) < \infty$*. Consider the matrix* $\mathbf{\Omega}$*, whose elements are given by:*

$$\mathbf{\Omega}_{ii} = \frac{2 \sum_{h=1}^{\infty} \left(p_i^{(h)} - p_i^2 \right) + p_i (1 - p_i)}{2 p_i},$$

$$\mathbf{\Omega}_{ii'} = \frac{2 \sum_{h=1}^{\infty} \left(\frac{p_{ii'}^{(h)} + p_{i'i}^{(h)}}{2} - p_i p_{i'} \right) - p_i p_{i'}}{2 \left(p_i p_{i'} \right)^{1/2}}.$$

Let $(\lambda_1, \dots, \lambda_B)$ be the eigenvalues of the matrix $\mathbf{\Omega}$ arranged in decreasing order. Therefore:

$$N\left(H_N - H_\infty\right) \to_{\mathcal{D}} -\sum_{i=1}^{B} \lambda_i \chi^2_{1,i}.$$

Remark 4.3. (i) The derivation of a precise rate of convergence of the finite-sample distribution to the asymptotic one, namely $\left\| F_{N(H_N-H_\infty)} - F_{-\sum_{i=1}^{B}\lambda_i\chi^2_{1,i}} \right\|_\infty$, seems to be out of reach given the state of the literature. According to the proof of Proposition 4.4, the rate of convergence of $N\left(H_N - H_\infty\right)$ to its asymptotic distribution can be linked to the rate of convergence of the chi-square statistic $-N\sum_{i=1}^{B}\frac{(q_i-p_i)^2}{2p_i}$. In the dependent case there seems to be no available result for the lattice case, but one can consider what happens in the independent case as a benchmark. In that case, [149] showed that the convergence rate is $O\left(N^{-\frac{B-1}{B}}\right)$ (see also [53]), while [194] showed that the rate of convergence is $O\left(N^{-1}\right)$ for $B \geq 6$. We investigate the rate of convergence in Examples 4.5 and 4.6 below.

Example 4.5. [Dichotomized $\mathsf{AR}\,(1)$ process - Examples 4.1, 4.3, 4.4 continued] We have:

$$\mathbf{\Omega}_{11} = \mathbf{\Omega}_{22} = {}^1\!/\!_4 + {}^1\!/\!_\pi \sum_{h=1}^{\infty} \arcsin\left(\alpha^h\right),$$
$$\mathbf{\Omega}_{12} = \mathbf{\Omega}_{21} = -{}^1\!/\!_4 - {}^1\!/\!_\pi \sum_{h=1}^{\infty} \arcsin\left(\alpha^h\right).$$

Therefore:

$$\mathbf{\Omega} = \left\{ {}^1\!/\!_4 + {}^1\!/\!_\pi \sum_{h=1}^{\infty} \arcsin\left(\alpha^h\right) \right\} \cdot \begin{bmatrix} +1 & -1 \\ -1 & +1 \end{bmatrix}.$$

This matrix is singular and therefore $\lambda_1 = \operatorname{tr}(\mathbf{\Omega}) = {}^1\!/\!_2 + {}^2\!/\!_\pi \sum_{h=1}^{\infty} \arcsin\left(\alpha^h\right)$ and $\lambda_2 = 0$. We have:

$$N\left(H_N - H_\infty\right) \to_{\mathcal{D}} -\left\{ {}^1\!/\!_2 + {}^2\!/\!_\pi \sum_{h=1}^{\infty} \arcsin\left(\alpha^h\right) \right\} \cdot \chi^2_1.$$

In Figure 4.3.2 we show the difference between the cdf of the entropy and its asymptotic approximation for $N \in \{250, 1000, 4000\}$. The finite-sample distribution of the entropy is discrete as q_1 and q_2 can assume only $N+1$ values. These distribution are obtained through 1,000,000 samplings. The Kolmogorov distances between the finite-sample distributions with $N \in \{250, 1000, 4000\}$ and the asymptotic one are respectively 0.04209815, 0.02110595 and 0.01056157, thus suggesting a rate of convergence of $O\left(N^{-1/2}\right)$, that is in line with Remark 4.3 for $B = 2$.

Example 4.6. [GSR model - Example 4.2 continued] For $B \in \{2, 3, 4, 5, 6\}$ we compute the asymptotic distribution and we compare it with the finite-sample distributions for $N \in \{10, 11, \dots, 250\}$. These curves are represented in Figure 4.3.3 and are consistent with an increase in the rate of convergence when B increases. The jigsaw profile of the curves in the figure does not seem to be an artifact of our simulations as it appears consistently across different replications. The curves for B running from 2 to 6 are respectively based on over $4 \cdot 10^7$, $4.5 \cdot 10^7$, $4.5 \cdot 10^7$, $2 \cdot 10^8$ and $5 \cdot 10^8$

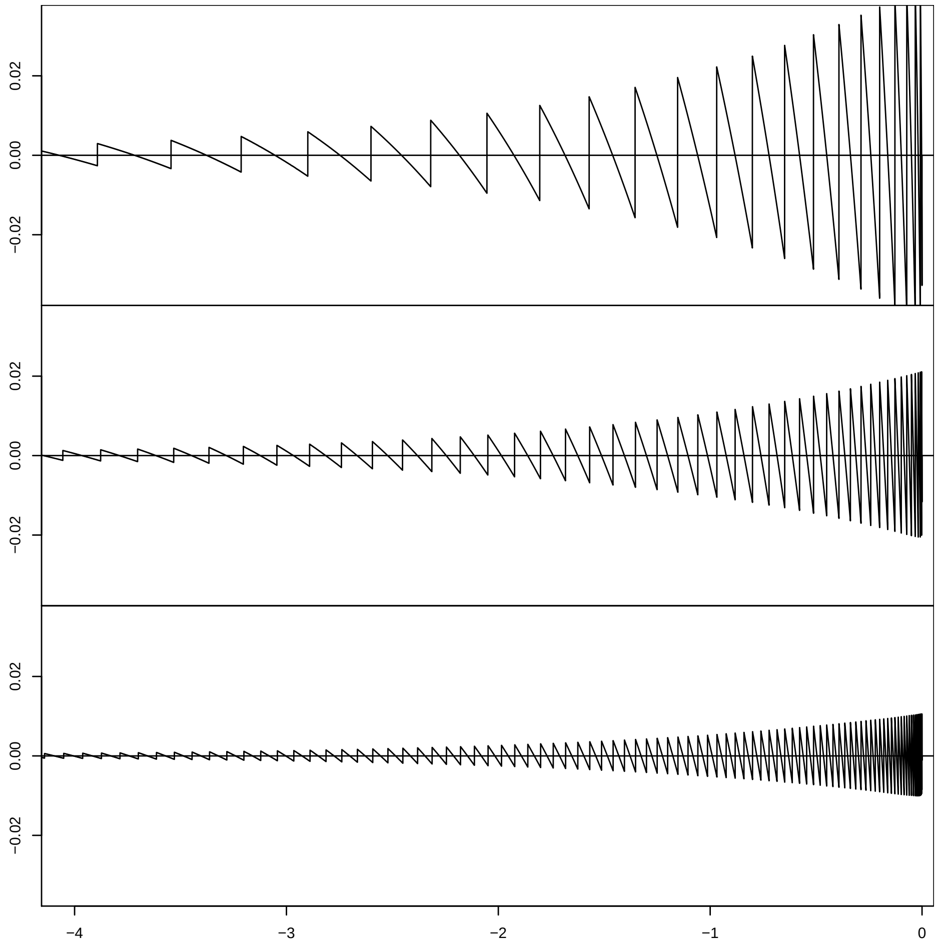

Figure 4.3.2: Difference between the cdf of the entropy and its asymptotic approximation for $N =$ 250, $N = 1000$ and $N = 4000$ (from above to below).

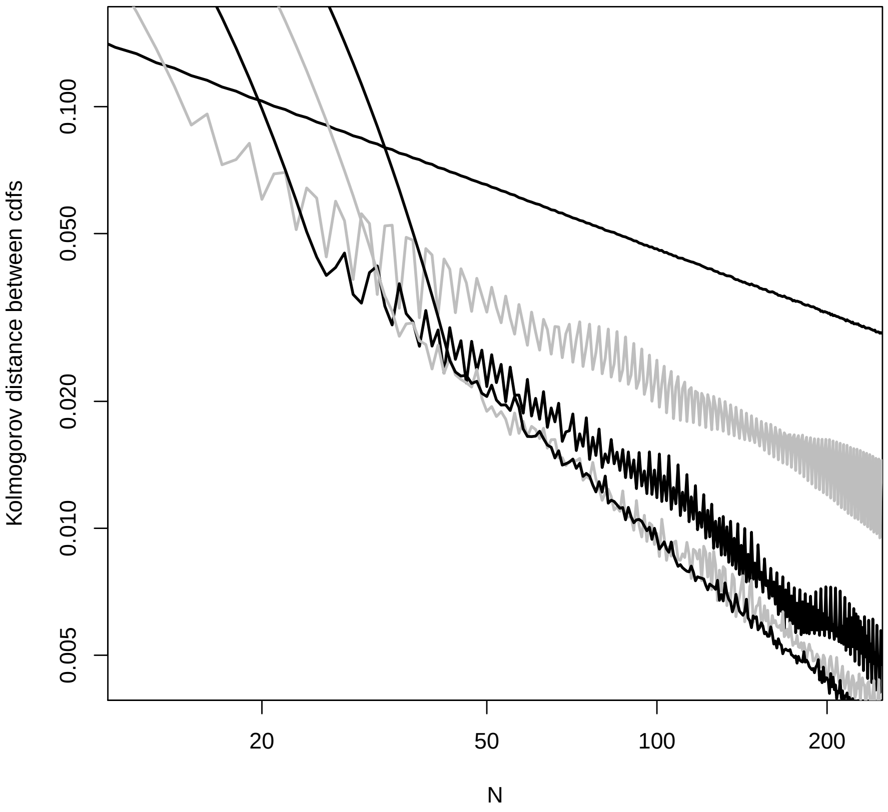

Figure 4.3.3: Rate of convergence to zero of the Kolmogorov distance between the cdf of the entropy and its asymptotic approximation for $N \in \{10, \dots, 250\}$ and $B \in \{2, \dots, 6\}$.

(non-independent) observations.

The combination of Propositions 4.2 and 4.4 gives the following result.

Corollary 4.2. *Suppose that the process* $\{x_1, \dots\}$ *is* α*-mixing with* $\sum_{n=1}^{\infty} \alpha(n) < \infty$*. Consider the matrix* $\mathbf{\Omega}$ *defined in Proposition 4.4. Therefore:*

$$N\left(H_N - \mathrm{bias}\left(H_N\right) - H_\infty\right) \to_{\mathcal{D}} -\sum_{i=1}^{B} \lambda_i \left(\chi^2_{1,i} - 1\right).$$

4.4 Estimation

When correcting for bias or computing the asymptotic variance of the entropy, we need to compute the matrix $\mathbf{\Sigma}$ whose elements are (see Lemma 4.3 in Section 4.7.2):

$$\mathbf{\Sigma}_{ii} = p_i\left(1-p_i\right) + 2\sum_{h=1}^{\infty}\left(p_i^{(h)} - p_i^2\right),$$
$$\mathbf{\Sigma}_{ii'} = 2\sum_{h=1}^{\infty}\left(\frac{p_{ii'}^{(h)} + p_{i'i}^{(h)}}{2} - p_i p_{i'}\right) - p_i p_{i'}.$$

We can rewrite the elements as:

$$\mathbf{\Sigma}_{ii} = p_i\left(1-p_i\right) + 2\sum_{h=1}^{\infty}\left(p_i^{(h)} - p_i^2\right) = \sum_{h=-\infty}^{\infty}\left(p_i^{(h)} - p_i^2\right) \tag{4.4.1}$$
$$\mathbf{\Sigma}_{ii'} = 2\sum_{h=1}^{\infty}\left(\frac{p_{ii'}^{(h)} + p_{i'i}^{(h)}}{2} - p_i p_{i'}\right) - p_i p_{i'} = \sum_{h=-\infty}^{\infty}\left(p_{ii'}^{(h)} - p_i p_{i'}\right) \tag{4.4.2}$$

where we have used the fact that, under stationarity, $p_{i\ell}^{(h)} = p_{\ell i}^{(-h)}$.

The computation of the bias is performed as follows rewriting it as:

$$\begin{aligned}\text{bias}\left(H_N\right) &:= -\frac{B-1}{2N} - \frac{1}{N}\sum_{i=1}^{B}\frac{\sum_{h=1}^{\infty}\left(p_i^{(h)} - p_i^2\right)}{p_i} \\ &= -\frac{1}{2N}\sum_{i=1}^{B}\frac{\mathbf{\Sigma}_{ii}}{p_i} = -\frac{\operatorname{tr}\left(\operatorname{dg}\left(\bar{\mathbf{p}}\right)\mathbf{\Sigma}\right)}{2N}.\end{aligned} \tag{4.4.3}$$

The matrix $\mathbf{\Omega}$, used to define the distribution in the degenerate case (see Proposition 4.4), is defined as:

$$\mathbf{\Omega} = \frac{1}{2}\operatorname{dg}\left(\mathbf{p}^{\odot\left(-\frac{1}{2}\right)}\right)\mathbf{\Sigma}\operatorname{dg}\left(\mathbf{p}^{\odot\left(-\frac{1}{2}\right)}\right). \tag{4.4.4}$$

In the following we propose two methods aimed at correcting the bias of the estimator of the entropy and at computing its variance.

4.4.1 First method

A first method that holds with generality is to estimate the elements of the matrix $\mathbf{\Sigma}$ through an autocorrelation-consistent (AC) covariance-matrix estimator, like the ones considered in [356, 14, 388]. We show their general structure.

We define:

$$\mathbf{\Pi}^{(h)} = \operatorname{Cov}\left(\mathbf{x}_1, \mathbf{x}_{1+h}'\right)$$

whose generic element is $\left[\mathbf{\Pi}^{(h)}\right]_{ii'} = \operatorname{Cov}\left(1\left\{x_1 = i\right\}, 1\left\{x_{1+h} = i'\right\}\right) = p_{ii'}^{(h)} - p_i p_{i'}$. Thus, from

(4.4.1) and (4.4.2):

$$\mathbf{\Sigma} = \sum_{h=-\infty}^{\infty} \mathbf{\Pi}^{(h)}.$$

Estimators take the form:

$$\hat{\mathbf{\Sigma}} = \sum_{h=-N+1}^{N-1} k\left(\frac{h}{S_N}\right)\hat{\mathbf{\Pi}}^{(h)}$$

where k is a kernel function, S_N is a bandwidth parameter and:

$$\hat{\mathbf{\Pi}}^{(h)} = \begin{cases} \frac{1}{N}\sum_{n=h+1}^{N}\left(\mathbf{x}_n - \mathbf{q}\right)\left(\mathbf{x}_{n-h} - \mathbf{q}\right)' & h \geq 0, \\ \frac{1}{N}\sum_{n=-h+1}^{N}\left(\mathbf{x}_{n+h} - \mathbf{q}\right)\left(\mathbf{x}_n - \mathbf{q}\right)' & h < 0. \end{cases}$$

A plug-in estimator of $\mathbf{\Omega}$ is:

$$\hat{\mathbf{\Omega}} = \frac{1}{2}\mathrm{dg}\left(\mathbf{q}^{\odot\left(-\frac{1}{2}\right)}\right)\hat{\mathbf{\Sigma}}\mathrm{dg}\left(\mathbf{q}^{\odot\left(-\frac{1}{2}\right)}\right).$$

Therefore, a plug-in estimator of the bias is:

$$\widehat{\mathrm{bias}\left(H_N\right)} = -\frac{\mathrm{tr}\left(\mathrm{dg}\left(\bar{\mathbf{q}}\right)\hat{\mathbf{\Sigma}}\right)}{2N}.$$

Example 4.7. [Dichotomized $\mathsf{AR}(1)$ process - Examples 4.1, 4.3, 4.4, 4.5 continued] Here we consider bias correction using the estimator of [356] and [14]. The light grey jigsaw lines are the trajectories of $H_N - \widehat{\mathrm{bias}\left(H_N\right)}$ with Newey-West bias correction. They oscillate around a curved solid grey line that is $\mathbb{E}\left(H_N - \widehat{\mathrm{bias}\left(H_N\right)}\right)$ with Newey-West bias correction, a curved dashed grey line that is $\mathbb{E}\left(H_N - \widehat{\mathrm{bias}\left(H_N\right)}\right)$ with Andrews bias correction, an horizontal dark grey line that is H_∞, a black solid curve that is $\mathbb{E}\left(H_N\right)$, and a black dashed curve (almost indistinguishable from H_∞) that is $\mathbb{E}\left(H_N\right) - \mathrm{bias}\left(H_N\right)$. It is apparent from the plot that $H_N - \widehat{\mathrm{bias}\left(H_N\right)}$ is in both cases less biased than H_N (see also Figure 4.3.1). However, the replacement of $\mathrm{bias}\left(H_N\right)$ with $\widehat{\mathrm{bias}\left(H_N\right)}$ is not without consequences. Indeed, the quantity $\mathbb{E}\left(H_N\right) - \mathrm{bias}\left(H_N\right)$ is represented by a black dashed line that is almost undistinguishable from H_∞, while $\mathbb{E}\left(H_N - \widehat{\mathrm{bias}\left(H_N\right)}\right)$ is not. The vertical lines at $N \in \{100, 125, 150, 200\}$ (respectively black dashed, dotted, dash-dot, solid lines) represent the values of $N \in \{100, 125, 150, 200\}$ at which the empirical cdf of $H_N - \widehat{\mathrm{bias}\left(H_N\right)}$ with Newey-West bias correction (black lines) and with Andrews bias correction (grey lines) are computed.

In the following we will prove our results under the following assumption.

AC Let $q > 0$ be such that:

$$\sum_{h=-\infty}^{\infty} |h|^q \left\|\mathbf{\Pi}^{(h)}\right\|_{L^2} < \infty$$

and:

$$\lim_{x \to 0} \frac{1 - k(x)}{|x|^q} = k_q < \infty.$$

Then, the following conditions hold:

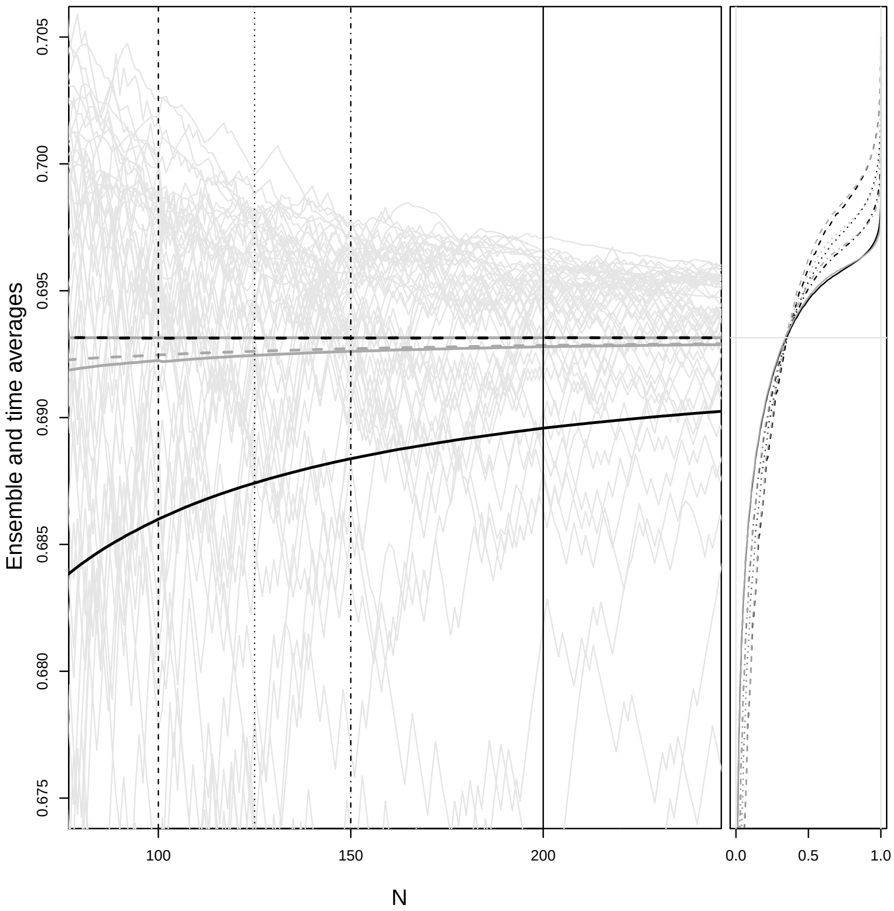

Figure 4.4.1: Ensemble and time averages of the entropy in the ergodic (dichotomized $\mathsf{AR}\,(1)$) case with bias corrections: on the left plot, 50 trajectories of $H_N - \widehat{\text{bias}\,(H_N)}$ with Newey-West bias correction as a function of N (light grey jigsaw lines), H_∞ (dark grey horizontal line), $\mathbb{E}\,(H_N)$ (black solid curve), $\mathbb{E}\left(H_N - \widehat{\text{bias}\,(H_N)}\right)$ with Newey-West bias correction (grey solid curve), $\mathbb{E}\left(H_N - \widehat{\text{bias}\,(H_N)}\right)$ with Andrews bias correction (grey dashed curve), $\mathbb{E}\,(H_N) - \text{bias}\,(H_N)$ (black dashed curve), vertical lines at $N \in \{100, 125, 150, 200\}$ (respectively black dashed, dotted, dash-dot, solid lines); on the right plot, empirical cdf of $H_N - \widehat{\text{bias}\,(H_N)}$ with Newey-West bias correction with $N = 100$ (black dashed line), $N = 125$ (black dotted line), $N = 150$ (black dash-dot line), $N = 200$ (black solid line), with Andrews bias correction with $N = 100$ (grey dashed line), $N = 125$ (grey dotted line), $N = 150$ (grey dash-dot line), $N = 200$ (grey solid line).

1. the process is α-mixing with $\sum_{n=1}^{\infty} n^2 \alpha(n) < \infty$;
2. $S_N/N \to 0$ as $N \to \infty$;
3. $k : \mathbb{R} \to [-1, 1]$ is symmetric, continuous at 0 and for all but a finite number of points, and satisfies $k(0) = 1$ and $\int_{-\infty}^{\infty} k^2(x)\,\mathrm{d}x < \infty$;
4. if $S_N \not\to \infty$, $S_N^{-1} \sum_{j=-N+1}^{N-1} |k(j/S_N)| = O(1)$;
5. if $q < 1/2$, $N^{1/2-q} S_N^{-1/2} = O(1)$;
6. one of the following three sets of conditions hold true:
   (a) if $S_N \to \infty$ and $k_q \neq 0$, $S_N^{-q-1/2} N^{1/2} = O(1)$;
   (b) if $S_N \to \infty$ and $k_q = 0$:
   $$S_N^{-1/2} N^{1/2} \sum_{h=1}^{N-1} \left(1 - k\left(\frac{h}{S_N}\right)\right) \mathbf{\Pi}^{(h)} = O(1)$$
   and this is a fortiori true if $S_N^{-q-1/2} N^{1/2} = O(1)$;
   (c) if $S_N \not\to \infty$:
   $$S_N^{-1/2} N^{1/2} \sum_{h=1}^{N-1} \left(1 - k\left(\frac{h}{S_N}\right)\right) \mathbf{\Pi}^{(h)} = O(1).$$

Remark 4.4. (i) The case in which $S_N \not\to \infty$ is required by some recent results on the estimation of AC covariance matrices (see Theorem 2.1 in [388, p. 707]).
(ii) The case in which $S_N \to \infty$ is surely the most interesting. If $k_q \neq 0$, conditions 2 and 6 imply that $N^{\frac{1}{2q+1}} \lesssim S_N \ll N$. In this case condition 4 is always verified and condition 5 is redundant, as $N^{1-2q} \lesssim N^{\frac{1}{2q+1}}$.
(iii) If the function k is non-negative and non-increasing over $[0, \infty)$, one can adapt the reasoning in Theorem 1 in [18, p. 410] to show that assumption 4 is automatically true:

$$\begin{aligned} S_N^{-1} \sum_{j=-N+1}^{N-1} |k(j/S_N)| =& S_N^{-1} \left\{ 1 + 2 \sum_{j=1}^{N-1} k(j/S_N) \right\} \\ \leq& S_N^{-1} \left\{ 1 + 2 \int_1^N k(x/S_N)\,\mathrm{d}x + 2k(1/S_N) \right\} \\ \leq& S_N^{-1} + S_N^{-1} \int_{-\infty}^{\infty} k(x/S_N)\,\mathrm{d}x + 2 S_N^{-1} k(1/S_N) \\ \leq& \sqrt{\int_{-\infty}^{\infty} k^2(y)\,\mathrm{d}y} + 3 S_N^{-1} = O(1). \end{aligned}$$

This holds irrespective of the fact that $S_N \not\to \infty$ or $S_N \to \infty$.

4.4.2 Second method

Whenever the process is a Markov chain, an alternative is to use the transition matrix in order to compute the probabilities appearing in the formulas above. We suppose below that the Markov chain is ergodic and regular, i.e. irreducible and aperiodic.

We have $\mathbf{p}' = \mathbf{p}'\mathbf{P}$, i.e. $\mathbf{p}$ is a normalized right eigenvector of the stochastic transition matrix $\mathbf{P}$ corresponding to the eigenvalue equal to 1. We define $\mathbf{H} := (\mathbf{I} - \mathbf{P} + \boldsymbol{\iota}\mathbf{p}')^{-1}$, the *fundamental matrix* of [259] (see also [440]).

Proposition 4.5. *For a Markov chain with transition matrix* $\mathbf{P}$ *and ergodic distribution* $\mathbf{p}$*, we have:*

$$\text{bias}\left(H_N\right) = -\frac{2\text{tr}\left(\mathbf{H}\right) - B - 1}{2N}$$

and:

$$\boldsymbol{\Omega} = -\frac{1}{2}\mathbf{I} + \frac{1}{2}\text{dg}\left(\mathbf{p}^{\odot\frac{1}{2}}\right)\left(\mathbf{H}\text{dg}\left(\bar{\mathbf{p}}\right) + \text{dg}\left(\bar{\mathbf{p}}\right)\mathbf{H}' - \mathbf{U}\right)\text{dg}\left(\mathbf{p}^{\odot\frac{1}{2}}\right).$$

Example 4.8. [GSR model - Examples 4.2 and 4.6 continued] We can characterize the quantities appearing in the GSR model. From Theorem 2.2 in [95], the matrix $\mathbf{P}$ has eigenvalues given by 2^{-m} for $0 \leq m \leq B-1$. Using Theorem 1 in [501], the eigenvalues of $\mathbf{H}^{-1}$ are 1 and $1-2^{-m}$ for $1 \leq m \leq B-1$, and the column eigenvector associated with 1 is proportional to $\boldsymbol{\iota}$. Therefore, the eigenvalues of $\mathbf{H}$ are 1 and $\left(1-2^{-m}\right)^{-1}$ for $1 \leq m \leq B-1$. This implies that $\text{tr}\left(\mathbf{H}\right) = B + \sum_{m=1}^{B-1}\frac{1}{2^m-1}$. Now, $\mathbf{p} = B^{-1}\boldsymbol{\iota}$ from which the matrix $\boldsymbol{\Omega}$ in Proposition 4.5 becomes:

$$\boldsymbol{\Omega} = \frac{1}{2}\left(\mathbf{H} + \mathbf{H}' - \mathbf{I} - B^{-1}\mathbf{U}\right).$$

We suppose to estimate $\mathbf{P}$ through $\hat{\mathbf{P}}$ defined as:

$$\left[\hat{\mathbf{P}}\right]_{ii'} = \frac{\sum_{j=1}^{N-1}\mathbb{1}\left\{x_j = i, x_{j+1} = i'\right\}}{\sum_{i'=1}^{B}\sum_{j=1}^{N-1}\mathbb{1}\left\{x_j = i, x_{j+1} = i'\right\}}$$

and $\mathbf{p}$ through $\hat{\mathbf{p}}$, the normalized left eigenvector of $\hat{\mathbf{P}}$, i.e. $\hat{\mathbf{p}}'\hat{\mathbf{P}} = \hat{\mathbf{p}}'$ (in general $\hat{\mathbf{p}}$ does not coincide with $\mathbf{q}$). Moreover we define $\hat{\mathbf{H}} := \left(\mathbf{I} - \hat{\mathbf{P}} + \boldsymbol{\iota}\hat{\mathbf{p}}'\right)^{-1}$.

4.4.3 Error in the estimation of bias

We provide a result on the average error induced by the estimation of the bias.

Proposition 4.6. *For the method in Section 4.4.1, under AC:*

$$\widehat{\text{bias}\left(H_N\right)} = \text{bias}\left(H_N\right) + O_{\mathbb{P}}\left(S_N^{1/2}N^{-3/2}\right).$$

For the method in Section 4.4.2:

$$\widehat{\text{bias}\left(H_N\right)} = \text{bias}\left(H_N\right) + O_{\mathbb{P}}\left(N^{-3/2}\right).$$

Remark 4.5. (i) As $S_N = o\left(N\right)$ in AC, for the method in Section 4.4.1, we get that the error is $o_{\mathbb{P}}\left(N^{-1}\right)$.
(ii) The optimal rate of divergence of S_N for the Newey-West estimator in [356] is $S_N \asymp N^{1/3}$, and for

the second-order kernels in [14] it is $S_N \asymp N^{1/5}$. The rate of error in the bias decreases respectively as $O_{\mathbb{P}}\left(N^{-4/3}\right)$ and $O_{\mathbb{P}}\left(N^{-7/5}\right)$.

Corollary 4.3. *Under the conditions of Propositions 4.3 and 4.6, if $S_N = o(N)$, we have:*

$$\sqrt{N}\left(H_N - \widehat{\operatorname{bias}(H_N)} - H_\infty\right) \to_{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\right)$$

and

$$N\left(H_N - \widehat{\operatorname{bias}(H_N)} - H_\infty\right) \to_{\mathcal{D}} -\sum_{i=1}^{B} \lambda_i \left(\chi_{1,i}^2 - 1\right).$$

Using the previous results we can derive the following corollary concerning the MSE.

Corollary 4.4. *Under the hypotheses of Propositions 4.3, 4.4 and 4.6:*

$$\begin{aligned}
&\operatorname{MSE}(H_N) = \begin{cases} O\left(N^{-1}\right) & \text{if } \sigma^2 > 0 \\ O\left(N^{-2}\right) & \text{if } \sigma^2 = 0 \end{cases} \\
&\operatorname{MSE}(H_N) - \operatorname{MSE}(H_N - \operatorname{bias}(H_N)) = O\left(N^{-2}\right) \\
&\operatorname{MSE}\left(H_N - \widehat{\operatorname{bias}(H_N)}\right) - \operatorname{MSE}(H_N - \operatorname{bias}(H_N)) \\
&\qquad = \begin{cases} O\left(S_N^{1/2} N^{-2}\right) & \text{if } \sigma^2 > 0 \\ O\left(S_N^{1/2} N^{-5/2}\right) & \text{if } \sigma^2 = 0 \end{cases}
\end{aligned}$$

where the left-hand sides of the first two expressions are always non-negative. For the method in Section 4.4.2, the formulas still hold with $S_N \equiv 1$.

Example 4.9. [GSR model - Examples 4.2, 4.6 and 4.8 continued] We consider a deck of $B = 10$ cards and we shuffle it $N = 1,000$ times. We record the position taken by the card occupying the first position in the original order of the deck. It is expected that, in a series of shuffles, the card will visit each integer number between 1 and B with a probability converging to B^{-1}. Therefore, the limit value of the entropy is $H_\infty = \ln B = \ln 10 \doteq 2.302585$. We have simulated 1,000,000 times the process of shuffling. The average $\mathbb{E}H_N$ without bias correction is 2.296466. We have then computed the Newey-West, Andrews and Markov bias corrections using each time series of N observations. One should note that the matrix $\hat{\mathbf{P}}$ for the Markov bias correction is estimated using only $N - 1$ observations. The values of $\mathbb{E}\left(H_N - \widehat{\operatorname{bias}(H_N)}\right)$ with Newey-West, Andrews and Markov bias corrections are respectively 2.302135, 2.302413 and 2.302571, thus confirming the order suggested by Proposition 4.6. The empirical cdfs of H_N and $H_N - \widehat{\operatorname{bias}(H_N)}$ with Newey-West, Andrews and Markov bias corrections confirm these findings. Note that the Markov bias correction is more precise than the other two. It is also possible to estimate the variance $\mathbb{V}(H_N)$ as $1.121541 \cdot 10^{-5}$, as well as the variances $\mathbb{V}\left(H_N - \widehat{\operatorname{bias}(H_N)}\right)$ with Newey-West, Andrews and Markov bias corrections respectively as $1.143687 \cdot 10^{-5}$, $1.128285 \cdot 10^{-5}$ and $1.126361 \cdot 10^{-5}$. This means that the bias correction slightly increases the variance of the estimator, but the MSE is still improved by the corrections; indeed, the MSE is respectively $4.865827 \cdot 10^{-5}$, $1.163983 \cdot 10^{-5}$,

$1.131258 \cdot 10^{-5}$ and $1.126382 \cdot 10^{-5}$ for H_N and $H_N - \widehat{\text{bias}\,(H_N)}$ with Newey-West, Andrews and Markov bias corrections.

4.4.4 Error in the estimation of the distribution under degeneracy

One of the problems raised by the previous result is to determine what is the effect of estimating the weights on the significance level of tests.

Proposition 4.7. *Let $\left(\hat{\lambda}_1, \ldots, \hat{\lambda}_B\right)$ be the eigenvalues of the matrix $\hat{\mathbf{\Omega}}$ defined in Sections 4.4.1 and 4.4.2. The following bound holds true:*

$$\left\| F_{-\sum_{i=1}^{B} \hat{\lambda}_i \chi^2_{1,i}} - F_{-\sum_{i=1}^{B} \lambda_i \chi^2_{1,i}} \right\|_\infty = O\left(\left\| \mathbf{\Omega} - \hat{\mathbf{\Omega}} \right\|_1 \right).$$

For the method in Section 4.4.1, under AC the bound is $O_{\mathbb{P}}\left(\left({}^{s_N}/_{N} \right)^{1/2} \right)$. For the method in Section 4.4.2, the bound is $O_{\mathbb{P}}\left(N^{-1/2}\right)$.

Remark 4.6. (i) This result can be used as follows. Suppose that we determine the quantile q_α of a test of level α using $-\sum_{i=1}^{B} \hat{\lambda}_i \chi^2_{1,i}$. Then:

$$F_{-\sum_{i=1}^{B} \lambda_i \chi^2_{1,i}} (q_\alpha) = \alpha + O\left(\left\| \mathbf{\Omega} - \hat{\mathbf{\Omega}} \right\|_1 \right).$$

(ii) The distance $\left\| F_{-\sum_{i=1}^{B} \hat{\lambda}_i \chi^2_{1,i}} - F_{-\sum_{i=1}^{B} \lambda_i \chi^2_{1,i}} \right\|_\infty$ is $O_{\mathbb{P}}\left(N^{-1/3}\right)$ for the Newey-West estimator in [356], $O_{\mathbb{P}}\left(N^{-2/5}\right)$ for the second-order kernels in [14], and $O_{\mathbb{P}}\left(N^{-1/2}\right)$ for the flat-top kernels in [388].

Example 4.10. [GSR model - Examples 4.2, 4.6, 4.8 and 4.9 continued] We compute the asymptotic distribution $F_{-\sum_{i=1}^{B} \lambda_i \chi^2_{1,i}}$ and we compare it with the distributions using matrices $\hat{\mathbf{\Sigma}}$ based on time series of length N for several values B, as described in Example 4.9. In Table 4.4.1 we compute the quantity:

$$\mathbb{E} \left\| F_{-\sum_{i=1}^{B} \hat{\lambda}_i \chi^2_{1,i}} - F_{-\sum_{i=1}^{B} \lambda_i \chi^2_{1,i}} \right\|_\infty$$

where the expectation is computed over the distribution of the weights based on a series of length N. It is apparent that when the number of observations N is multiplied by 4 there is a division by 2 of the average Kolmogorov distance, coherently with the $O_{\mathbb{P}}\left(N^{-1/2}\right)$ rate predicted by Proposition 4.7.

4.5 A goodness-of-fit test

In this section, we propose a test of goodness-of-fit based on the entropy.

We first describe the setup, then we give two different interpretations of the test procedure. Suppose to observe a stationary time series $\{\widetilde{x}_1, \ldots, \widetilde{x}_N\}$ with $\widetilde{x}_1$ taking its values in $\mathbb{R}$. Suppose that the process is α-mixing with $\sum_{n=1}^{\infty} \alpha(n) < \infty$. Its marginal distribution has a density f with respect to a measure σ. We want to test the null hypothesis $\mathsf{H}_0 : f \equiv f_0$, where f_0 is a completely

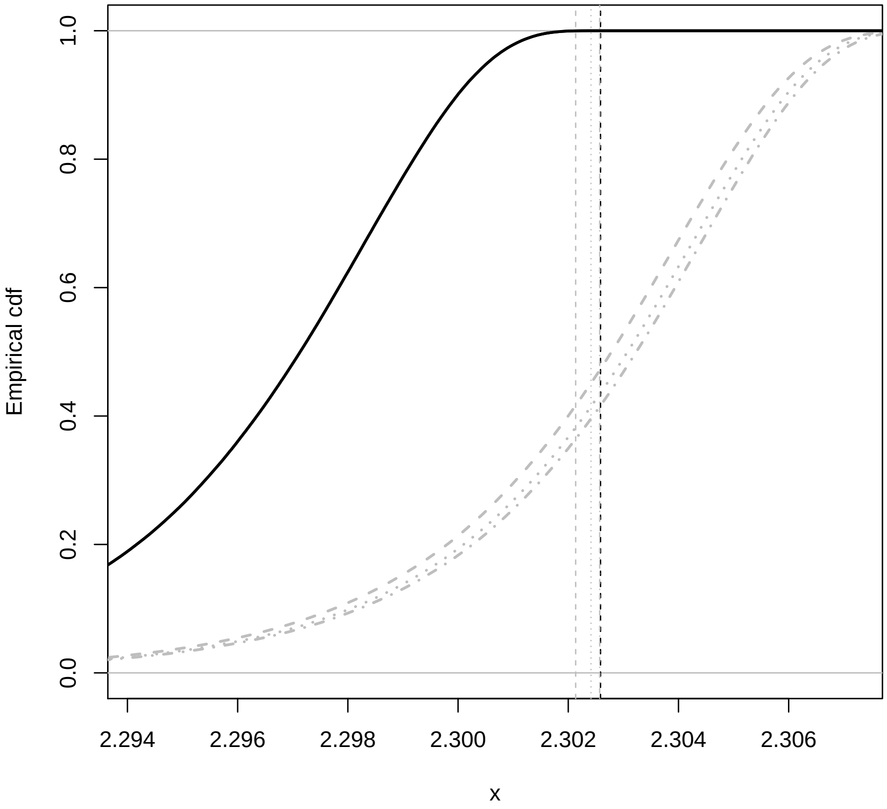

Figure 4.4.2: Empirical cumulative distribution function (cdf) of H_N (black line), $H_N - \widehat{\text{bias}\,(H_N)}$ with Newey-West bias correction (grey dashed line), $H_N - \widehat{\text{bias}\,(H_N)}$ with Andrews bias correction (grey dotted line), $H_N - \widehat{\text{bias}\,(H_N)}$ with Markov bias correction (grey dash-dot line), in comparison with the expected values $\mathbb{E}H_N$ (vertical black solid line), $\mathbb{E}\left(H_N - \widehat{\text{bias}\,(H_N)}\right)$ with Newey-West (vertical grey dashed line), Andrews (vertical grey dotted line) and Markov (vertical grey dash-dot line) bias corrections, and the true value H_∞ (vertical black dashed line).

		N					
		125	250	500	1000	2000	4000
B	2	0.04074774	0.02851654	0.02004530	0.01416706	0.009979191	0.007065414
	3	0.04793574	0.03318252	0.02348522	0.01646886	0.011574740	0.008097934
	4	0.05325378	0.03648672	0.02570183	0.01806233	0.012672534	0.008901423
	5	0.05832340	0.03949900	0.02735191	0.01931552	0.013527968	0.009557864
	6	0.06242307	0.04239093	0.02925993	0.02067011	0.014549110	0.010190659

Table 4.4.1: Average Kolmogorov distance between the exact asymptotic distribution and the one obtained estimating $\hat{\lambda}_i$, for $i = 1, \ldots, B$ through N observations.

specified density function with respect to σ. We identify a partition of the real line $\{\mathcal{I}_1, \ldots, \mathcal{I}_B\}$ such that:

$$\int_{\mathcal{I}_b} f_0(x)\, \sigma(\mathrm{d}x) = B^{-1}, \quad b = 1, \ldots, B.$$

We introduce the symbolized time series $\{\widetilde{x}_1, \ldots, \widetilde{x}_N\}$ defined by:

$$x_i = \sum_{b=1}^{B} b \cdot 1\{\widetilde{x}_i \in \mathcal{I}_b\}.$$

The symbolization is clearly much simpler when the measure σ is the Lebesgue measure and the density with respect to σ is a classical probability density function. Note that, by the very definition of α-mixing, the mixing coefficients of the symbolized process are majorized by the ones of the original process.

The first justification of the test uses the different behavior of the entropy under the null and the alternative hypotheses. Under the null hypothesis, the entropy computed on the time series $\{\widetilde{x}_1, \ldots, \widetilde{x}_N\}$ is degenerate as in Proposition 4.4. The asymptotic distribution of the entropy based on $\{\widetilde{x}_1, \ldots, \widetilde{x}_N\}$ satisfies:

$$N(H_N - H_\infty) \to_{\mathcal{D}} -\sum_{i=1}^{B} \lambda_i \chi^2_{1,i}$$

where $H_\infty = \ln B$. Therefore, under H_0, an acceptance region $\mathcal{A} = [q_\alpha, 0]$ corresponding to a significance level α for $N(H_N - \ln B)$ can be built using the quantile q_α of $-\sum_{i=1}^{B} \lambda_i \chi^2_{1,i}$ such that $F_{-\sum_{i=1}^{B} \lambda_i \chi^2_{1,i}}(q_\alpha) = \alpha$.

Now, we investigate what happens under the alternative hypothesis $\mathsf{H}_1 : f \equiv f_1 \neq f_0$. The test appears to have no power against any f_1 such that:

$$\int_{\mathcal{I}_b} f_1(x)\, \sigma(\mathrm{d}x) = B^{-1}, \quad b = 1, \ldots, B. \tag{4.5.1}$$

If, however, this does not hold true, Proposition 4.3 implies that σ is strictly positive and that:

$$\begin{aligned}
\mathbb{P}\{N(H_N - \ln B) \in \mathcal{A}\} =& \mathbb{P}\left\{H_N \geq \ln B + \frac{q_\alpha}{N}\right\} \\
=& \mathbb{P}\left\{\sqrt{N}\frac{H_N - H_\infty}{\sigma} \geq \sqrt{N}\frac{\ln B - H_\infty}{\sigma} + \frac{q_\alpha}{\sqrt{N}\sigma}\right\} \\
\leq& \left\| F_{\sqrt{N}\frac{H_N - H_\infty}{\sigma}} - \Phi \right\|_\infty + \Phi\left(-\sqrt{N}\frac{\ln B - H_\infty}{\sigma} - \frac{q_\alpha}{\sqrt{N}\sigma}\right) \\
=& O\left(N^{-1/2}\right) + \Phi\left(\sqrt{N}\frac{H_\infty - \ln B}{\sigma} - \frac{q_\alpha}{\sqrt{N}\sigma}\right).
\end{aligned}$$

Now, $H_\infty \leq \ln B$ with equality if and only if $\int_{\mathcal{I}_b} f_1(x)\, \sigma(\mathrm{d}x) = B^{-1}$ for $b = 1, \ldots, B$ (see, e.g., [310, p. 27]). Therefore, $\Phi\left(\sqrt{N}\frac{H_\infty - \ln B}{\sigma} - \frac{q_\alpha}{\sqrt{N}\sigma}\right) \downarrow 0$, $\mathbb{P}\{N(H_N - \ln B) \in \mathcal{A}\} \downarrow 0$ and the power of the test converges to 1.

Now we come to the second justification. We build the likelihood of the symbolized time series$\{\widetilde{x}_1, \ldots, \widetilde{x}_N\}$ neglecting the dependence between the values, i.e. supposing that they are independent. This object is sometimes called a pseudolikelihood (see, e.g., [92, Section 2.5] for a general result and [410, 197] for earlier examples). Despite the data are dependent, it is still possible to formulate a LR test of $\mathsf{H}_0 : f \equiv f_0$ that takes the form (see, e.g., [483, p. 252]):

$$\mathsf{LR} = \sum_{i=1}^{B} q_i \ln \frac{q_i}{B^{-1}} = \sum_{i=1}^{B} q_i \ln q_i + \ln B = H_\infty - H_N.$$

In the context of [355, Section 9], this is called a *distance metric statistic*. This test will not have the usual asymptotic distribution of LR tests but its distribution can be obtained from the one of the entropy. Linking this goodness-of-fit test with a LR test also shows that the test enjoys some optimality properties in the case of independent data and outperforms commonly used tests such as the chi-square test (see, e.g., [483, Section 17.6]).

In the following we provide two examples showing the finite-sample properties of the test.

Example 4.11. [Iid process] We consider the previous procedure when applied to an iid standard Gaussian sample. We symbolize the process in $B = 4$ equally probable intervals. In Figure 4.5.1, we depict the deviation between the actual and the nominal significance level:

$$\begin{aligned}\alpha \mapsto & \mathbb{P}\left\{N\left(H_N - \ln B\right) \notin \mathcal{A}\right\} - \mathbb{P}\left\{-\sum_{i=1}^{B} \lambda_i \chi^2_{1,i} \notin \mathcal{A}\right\} \\ = & \mathbb{P}\left\{N\left(H_N - \ln B\right) < q_\alpha\right\} - \mathbb{P}\left\{-\sum_{i=1}^{B} \lambda_i \chi^2_{1,i} < q_\alpha\right\} \\ = & \mathbb{P}\left\{N\left(H_N - \ln B\right) < q_\alpha\right\} - \alpha\end{aligned}$$

under the null hypothesis, for $B = 4$, $N \in \{50, 100, 200, 400\}$ and α ranging from 0.01 to 0.1. As the curves are based on $5 \cdot 10^7$ replications, for both plots the irregular profile of the curves is not an artifact of the simulations. In Figure 4.5.2 we depict the statistical power function:

$$\alpha \mapsto \pi = \mathbb{P}\left\{N\left(H_N - \ln B\right) \notin \mathcal{A}\right\} = \mathbb{P}\left\{N\left(H_N - \ln B\right) < q_\alpha\right\}$$

under some alternative hypotheses, i.e. when the data are from a sample of iid Gaussian random variables with mean $c \in \{0.1, 0.2, 0.3\}$ and variance 1. These curves are based on 10^7 replications.

Example 4.12. [Symbolized $\mathsf{AR}\,(1)$ process] We consider the process described in Example 4.1. We want to test that its marginal distribution is standard Gaussian. In order to do so, we symbolize the process as explained above. We apply the Newey-West variance estimator with bandwidth equal to $S_N = \lceil N^{1/3} \rceil$ and the Andrews quadratic spectral variance estimator with bandwidth equal to $S_N = \lceil N^{1/5} \rceil$. The second choice is advocated in [177, pp. 551, 573] and criticized in [12, p. 17]. The first choice is rather similar to other ones proposed in the literature, such as the commonly used $S_N = \lceil 0.75 \cdot N^{1/3} \rceil$, but we have chosen the present one for simplicity. We have considered the adaptive procedures of [14] and [357], but in a small percentage of cases they fail to deliver reliable results, and this can be a problem in large simulations. Figure 4.5.3 represents the deviation between the actual and the nominal significance level for $B = 4$, $N \in \{50, 100, 200, 400\}$, α ranging

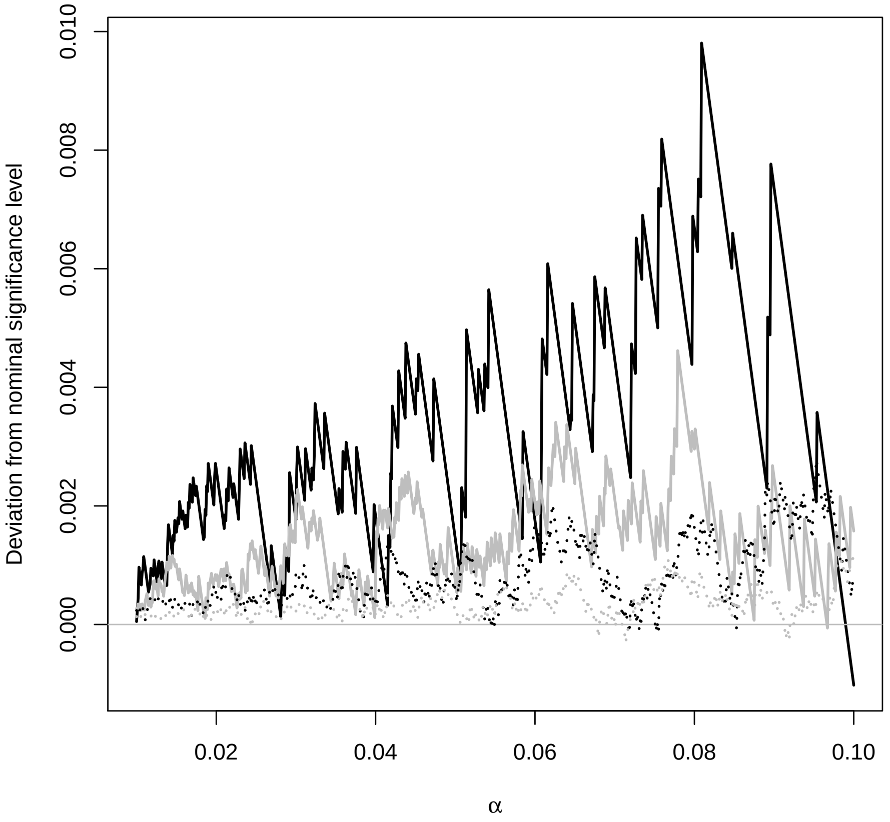

Figure 4.5.1: Difference between the actual and the nominal significance level in the independent case, for $\alpha \in [0.01, 0.1]$, $B = 4$ and $N \in \{50, 100, 200, 400\}$ (continuous black line for $N = 50$, continuous grey line for $N = 100$, dotted black line for $N = 200$, dotted grey line for $N = 400$).

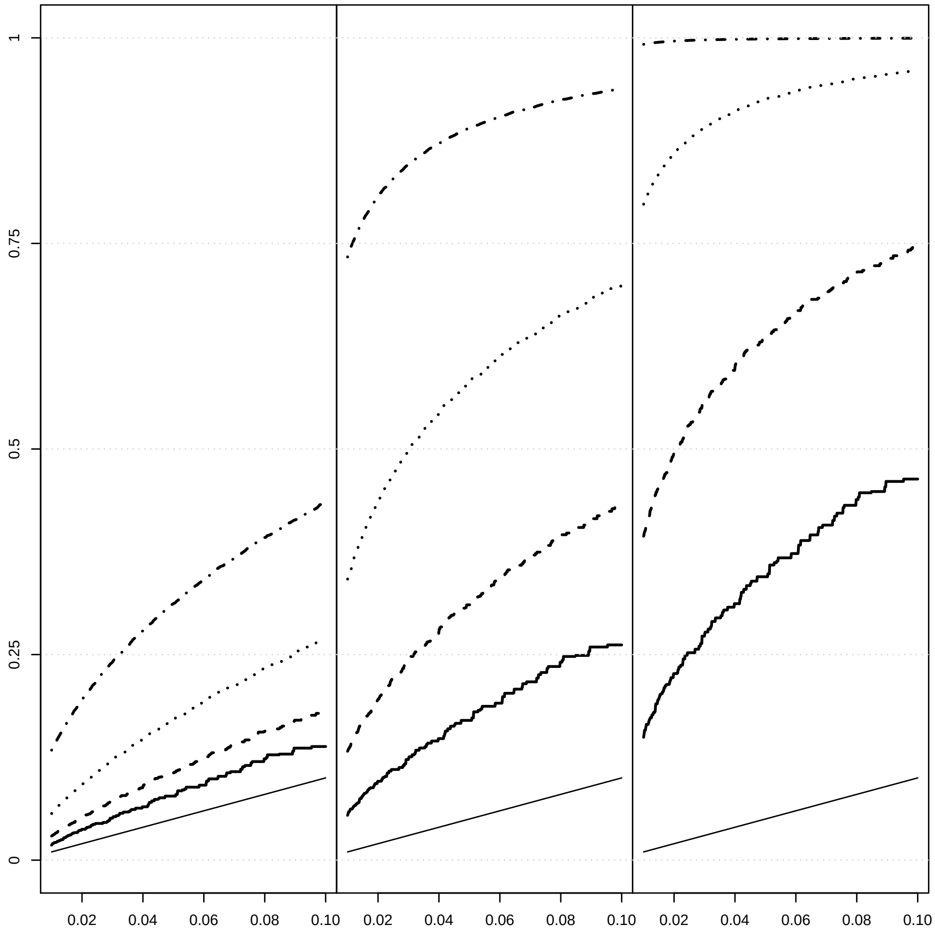

Figure 4.5.2: Power function in the independent case, for $\alpha \in [0.01, 0.1]$, $B = 4$, $N \in \{50, 100, 200, 400\}$, $c \in \{0.1, 0.2, 0.3\}$ (thin line for power equal to α, continuous line for $N = 50$, dashed line for $N = 100$, dotted line for $N = 200$, dash-dot line for $N = 400$; left column for $c = 0.1$, central column for $c = 0.2$, right column for $c = 0.3$).

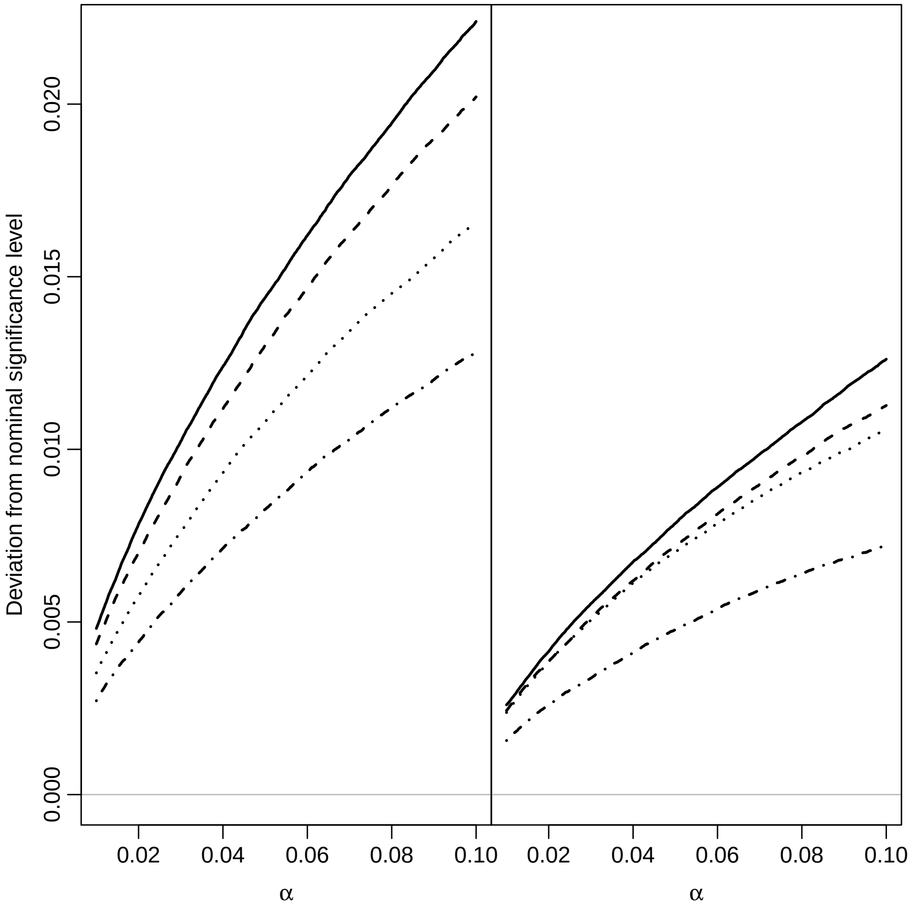

Figure 4.5.3: Difference between the actual and the nominal significance level in the dependent case with Newey-West estimator (on the left) and Andrews estimator (on the right), for $\alpha \in [0.01, 0.1]$, $B = 4$ and $N \in \{50, 100, 200, 400\}$ (continuous line for $N = 50$, dashed line for $N = 100$, dotted line for $N = 200$, dash-dot line for $N = 400$).

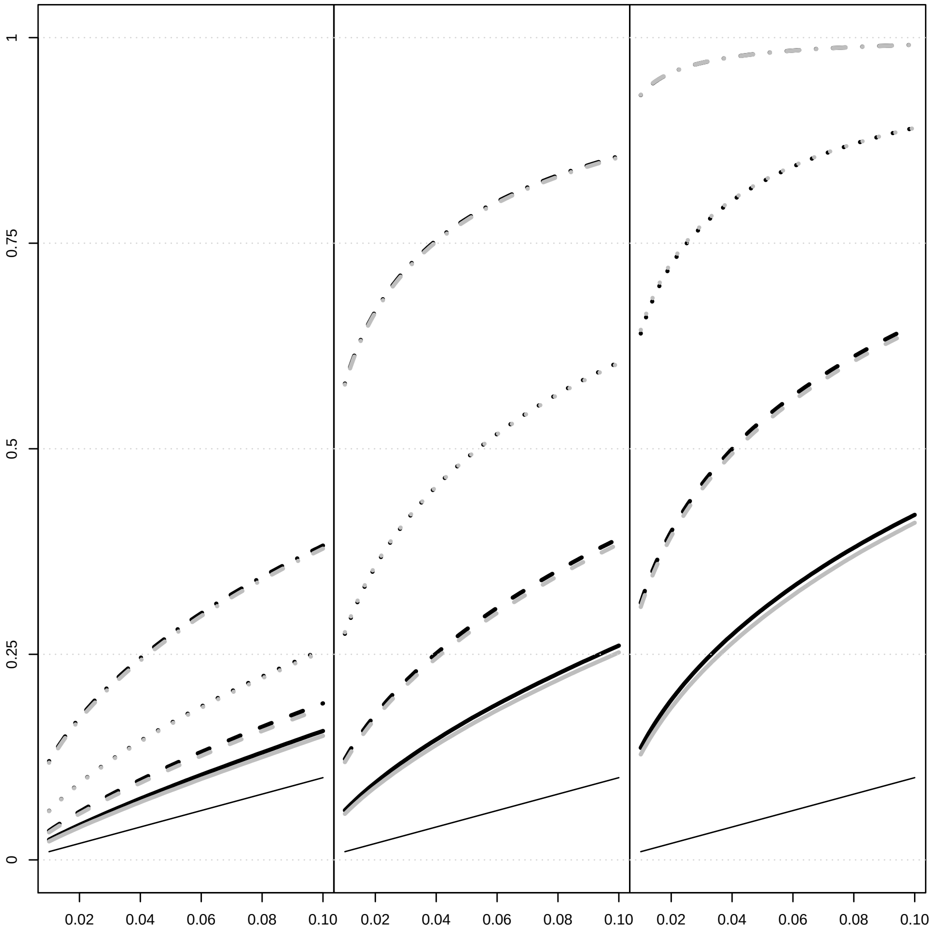

Figure 4.5.4: Power function in the dependent case with Newey-West (in black) and Andrews (in grey) estimators, for $\alpha \in [0.01, 0.1]$, $B = 4$, $N \in \{50, 100, 200, 400\}$, $c \in \{0.1, 0.2, 0.3\}$ (thin line for power equal to α, continuous line for $N = 50$, dashed line for $N = 100$, dotted line for $N = 200$, dash-dot line for $N = 400$; left column for $c = 0.1$, central column for $c = 0.2$, right column for $c = 0.3$).

from 0.01 to 0.1 with Newey-West (on the left) and Andrews (on the right) estimators. The curves look smoother than the ones for the independent case because the weights of the distribution are computed on the basis of the data and differ for each replication. Even for $N = 50$, the error in the significance level is rather small. Figure 4.5.4 represents the statistical power functions for Newey-West (in black) and Andrews (in grey) estimators. In this case the plots are more similar to the ones for the independent case.

4.6 Conclusions

In this contribution we consider the estimation of the entropy of data coming from a discretely supported stochastic process. In order to do so, we use the plug-in estimator of the entropy, in which the probabilities of the different values are replaced by their empirical estimators. With respect to the state of the art, we provide new results concerning asymptotic normality and bias, and we fix an error about formulas for bias correction and variance that, started in [418], has propagated through the literature. We demonstrate that our correction of the bias removes the $O\left(N^{-1}\right)$ part of the bias of the observed entropy H_N. One of the central outcomes of the paper is represented by the behavior of the distribution under degeneracy, i.e. when the marginal distribution of the process assume equal probabilities for each value of the time series. Indeed, at odds with the general case, under degeneracy the statistic–with a different scaling–converges in distribution to a weighted sum of chi square random variables. We finally introduce some estimators of the distribution under degeneracy and we provide results on the error in the estimation. To complete our analysis, we showcase an application of the entropy to a goodness-of-fit test for the marginal distribution of the process. The simulation studies performed throughout the paper to investigate the finite-sample properties of the estimators enhance the theoretical conclusions.

The present study has some limitations. First of all, we only consider strictly stationary stochastic processes under ergodicity and, for some results, mixing. However, some processes used in signal processing and information theory exhibit limited amounts of nonstationarity such as cyclostationarity (see [182]). Further generalizations of the properties could be obtained through asymptotic mean stationarity, a more general concept (see, e.g., [233]) than stationarity, encompassing cyclostationarity. Second, we consider the properties of the plug-in estimator of the entropy in the discrete or discretized case. This has the consequence, among other things, that the goodness-of-fit test that we propose has no power against some alternative hypotheses (see (4.5.1)). It could be possible to circumvent this problem by letting the number of classes B to diverge together with the number of observations N.

4.7 Proofs

4.7.1 Preliminary results

Here we collect two results for future reference.

The first result (Lemma 4.1) is a general expansion of H_N already introduced in [224]. To keep the paper self-contained, we reproduce the proof here. The second result (Lemma 4.2) is a multivariate second-order delta method that will be used in the proof of the asymptotic distribution under degeneracy.

Lemma 4.1. *We have:*

$$H_N = H_\infty - \sum_{i=1}^{B} (q_i - p_i) \ln p_i + \sum_{m=2}^{r} \frac{(-1)^{m-1}}{m(m-1)} \sum_{i=1}^{B} \frac{(q_i - p_i)^m}{p_i^{m-1}} + R_{r+1}$$

where $|R_{r+1}| \leq \frac{1}{r(r+1)} \sum_{i=1}^{B} \frac{|q_i - p_i|^{r+1}}{(\lambda^\star p_i)^r}$ *for* $\lambda^\star > 0$ *independent of any* q_i.

Proof. We take a limited development of $-q_i \ln q_i$ for q_i around p_i with Lagrange remainder. We have:

$$\begin{aligned} -q_i \ln q_i &= -p_i \left(1 + \frac{q_i - p_i}{p_i}\right) \ln \left[p_i \left(1 + \frac{q_i - p_i}{p_i}\right)\right] \\ &= -p_i \ln p_i - (q_i - p_i) \ln p_i + \sum_{m=2}^{r} \frac{(-1)^{m-1}}{m(m-1)} \frac{(q_i - p_i)^m}{p_i^{m-1}} + R_{r+1,i} \end{aligned}$$

where:

$$\begin{aligned} R_{r+1,i} =& \frac{(-1)^r}{r(r+1)} \frac{(q_i - p_i)^{r+1}}{\xi_i^r}, \\ \xi_i =& \lambda_i p_i + (1 - \lambda_i) q_i, \quad 0 < \lambda_i < 1. \end{aligned}$$

This implies that:

$$H_N = H_\infty - \sum_{i=1}^{B} (q_i - p_i) \ln p_i + \sum_{m=2}^{r} \frac{(-1)^{m-1}}{m(m-1)} \sum_{i=1}^{B} \frac{(q_i - p_i)^m}{p_i^{m-1}} + R_{r+1}$$

where $R_{r+1} = \sum_{i=1}^{B} R_{r+1,i}$. Now we majorize R_{r+1}.

Let us first suppose that $|q_i - p_i| < (1 - \varepsilon) p_i$ for $0 < \varepsilon < 1$. This implies that $\left|\frac{q_i - p_i}{p_i}\right| < 1 - \varepsilon$ and $-q_i \ln q_i$ can be expanded in an infinite series:

$$\begin{aligned} -q_i \ln q_i =& - p_i \ln p_i - (q_i - p_i) \ln p_i + \sum_{m=2}^{r} \frac{(-1)^{m-1}}{m(m-1)} \frac{(q_i - p_i)^m}{p_i^{m-1}} \\ &+ \sum_{m=r+1}^{\infty} \frac{(-1)^{m-1}}{m(m-1)} \frac{(q_i - p_i)^m}{p_i^{m-1}} \end{aligned}$$

so that $R_{r+1,i} = \sum_{m=r+1}^{\infty} \frac{(-1)^{m-1}}{m(m-1)} \frac{(q_i - p_i)^m}{p_i^{m-1}}$. Now:

$$\begin{aligned} |R_{r+1,i}| \leq& \sum_{m=r+1}^{\infty} \frac{1}{m(m-1)} \frac{|q_i - p_i|^m}{p_i^{m-1}} \leq \sum_{m=r+1}^{\infty} \frac{|q_i - p_i|^m}{p_i^{m-1}} \\ =& \frac{|q_i - p_i|^{r+1}}{p_i^r} \sum_{j=0}^{\infty} \frac{|q_i - p_i|^j}{p_i^j} \leq \frac{|q_i - p_i|^{r+1}}{\varepsilon p_i^r} \end{aligned}$$

where we have used the fact that, as $\left|\frac{q_i - p_i}{p_i}\right| < 1 - \varepsilon$, $\sum_{j=0}^{\infty} \frac{|q_i - p_i|^j}{p_i^j} = \left(1 - \frac{|q_i - p_i|}{p_i}\right)^{-1} \leq \varepsilon^{-1}$.

Now, let us consider the case $|q_i - p_i| \geq (1 - \varepsilon) p_i$ for $0 < \varepsilon < 1$. Let

$$A_\varepsilon(p_i) := \{q_i \in [0,1] : |q_i - p_i| \geq (1 - \varepsilon) p_i\}.$$

[3] Then, $R_{r+1,i} = \frac{(-1)^r}{r(r+1)} \frac{(q_i - p_i)^{r+1}}{\xi_i^r}$ will not be zero on $A_\varepsilon(p_i)$. We have:

$$|R_{r+1,i}| = \frac{1}{r(r+1)} \frac{|q_i - p_i|^{r+1}}{(\lambda_i p_i + (1-\lambda_i) q_i)^r}$$

or:

$$\lambda_i = \frac{\left(\frac{|q_i - p_i|^{r+1}}{r(r+1)|R_{r+1,i}|} \right)^{\frac{1}{r}} - q_i}{p_i - q_i}.$$

The set $A_\varepsilon(p_i)$ is compact and $q_i \mapsto \lambda_i$ is a continuous positive function, therefore it attains its minimum on that set and the minimum must be positive. We define:

$$\lambda_B^\star := \min_{1 \le i \le B} \min_{q_i \in A_\varepsilon(p_i)} \lambda_i(q_i) > 0$$

and we note that this is independent of N.

We define $\lambda^\star := \min \left\{ \lambda_B^\star, \varepsilon^{\frac{1}{r}} \right\} > 0$ and we note it is independent of N. The final formula is easily obtained.

Lemma 4.2. *Let $\{\mathbf{X}_n\}$ be a sequence of vectors in $\mathbb{R}^k$. Assume that $\tau_n(\mathbf{X}_n - \boldsymbol{\mu}) \to_{\mathcal{D}} \mathbf{X}$ where $\boldsymbol{\mu}$ is a constant vector and $\{\tau_n\}$ is a sequence of constants such that $\tau_n \to \infty$. Let $g : \mathbb{R}^k \to \mathbb{R}$ be twice differentiable at $\boldsymbol{\mu}$ with continuous derivatives and suppose that $\left.\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}'}\right|_{\mathbf{x}=\boldsymbol{\mu}} (\mathbf{x} - \boldsymbol{\mu}) \equiv 0$ for $\mathbf{x}$ in a neighborhood of $\boldsymbol{\mu}$. Then:*

$$\tau_n^2 (g(\mathbf{X}_n) - g(\boldsymbol{\mu})) \to_{\mathcal{D}} \frac{1}{2} \mathbf{X}' \left. \frac{\partial^2 g(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \right|_{\mathbf{x}=\boldsymbol{\mu}} \mathbf{X}.$$

Proof. The proof follows the one of [294, Theorem 11.2.14 (i), p. 436]. A limited development gives the following formula:

$$g(\mathbf{x}) = g(\boldsymbol{\mu}) + \left. \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}'} \right|_{\mathbf{x}=\boldsymbol{\mu}} (\mathbf{x} - \boldsymbol{\mu}) + \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \left. \frac{\partial^2 g(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \right|_{\mathbf{x}=\boldsymbol{\mu}} (\mathbf{x} - \boldsymbol{\mu}) + R(\mathbf{x} - \boldsymbol{\mu})$$

where $R(\mathbf{y}) = o\left(\|\mathbf{y}\|_{L^2}^2 \right)$ as $\|\mathbf{y}\|_{L^2} \downarrow 0$. Now, from $\left.\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}'}\right|_{\mathbf{x}=\boldsymbol{\mu}} (\mathbf{x} - \boldsymbol{\mu}) \equiv 0$:

$$\tau_n^2 (g(\mathbf{X}_n) - g(\boldsymbol{\mu})) = \frac{1}{2} \tau_n^2 (\mathbf{X}_n - \boldsymbol{\mu})' \left. \frac{\partial^2 g(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \right|_{\mathbf{x}=\boldsymbol{\mu}} (\mathbf{X}_n - \boldsymbol{\mu}) + \tau_n^2 R(\mathbf{X}_n - \boldsymbol{\mu}).$$

By the Continuous Mapping Theorem, the first term on the right-hand side yields:

$$\frac{1}{2} \tau_n^2 (\mathbf{X}_n - \boldsymbol{\mu})' \left. \frac{\partial^2 g(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \right|_{\mathbf{x}=\boldsymbol{\mu}} (\mathbf{X}_n - \boldsymbol{\mu}) \to_{\mathcal{D}} \frac{1}{2} \mathbf{X}' \left. \frac{\partial^2 g(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \right|_{\mathbf{x}=\boldsymbol{\mu}} \mathbf{X}.$$

We then show that $\tau_n^2 R(\mathbf{X}_n - \boldsymbol{\mu}) = o_{\mathbb{P}}(1)$. We define the function $h(\mathbf{y}) := R(\mathbf{y}) / \|\mathbf{y}\|_{L^2}^2$ for $\mathbf{y} \neq \mathbf{0}$

[3] We amend the notation used by [224], $A_\varepsilon(p_i, q_i)$, as the set does not depend on q_i.

and $h(\mathbf{0}) := 0$. This function is continuous at $\mathbf{0}$ and, therefore:

$$\tau_n^2 R(\mathbf{X}_n - \boldsymbol{\mu}) = \tau_n^2 \|\mathbf{X}_n - \boldsymbol{\mu}\|_{L^2}^2 h(\mathbf{X}_n - \boldsymbol{\mu})$$

where $\tau_n^2 \|\mathbf{X}_n - \boldsymbol{\mu}\|_{L^2}^2 = O_{\mathbb{P}}(1)$, by definition, and $h(\mathbf{X}_n - \boldsymbol{\mu}) = o_{\mathbb{P}}(1)$, by the fact that $\tau_n(\mathbf{X}_n - \boldsymbol{\mu}) \to_{\mathcal{D}} \mathbf{X}$ implies that $\mathbf{X}_n \to_{\mathbb{P}} \boldsymbol{\mu}$ and by the Continuous Mapping Theorem. By Slutsky's theorem, we get the final result.

4.7.2 Variances

The following lemma contains some formulas for the variances and covariances of $\mathbf{q}$ and a central limit theorem for $\mathbf{q}$.

Lemma 4.3. *Under stationarity:*

$$\begin{aligned}
\mathbb{V}\left[\sqrt{N}(q_i - p_i)\right] =& p_i(1-p_i) + 2\sum_{h=1}^{N-1}\left(1 - \frac{h}{N}\right)\left(p_i^{(h)} - p_i^2\right), \\
\operatorname{Cov}\left[\sqrt{N}(q_i - p_i), \sqrt{N}(q_{i'} - p_{i'})\right] =& 2\sum_{h=1}^{N-1}\left(1 - \frac{h}{N}\right)\left(\frac{p_{ii'}^{(h)} + p_{i'i}^{(h)}}{2} - p_i p_{i'}\right) - p_i p_{i'}.
\end{aligned}$$

Under α-mixing, if $\sum_{n=1}^{\infty} \alpha(n) < \infty$, $\sqrt{N}(\mathbf{q} - \mathbf{p}) \to_{\mathcal{D}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ where:

$$\begin{aligned}
\boldsymbol{\Sigma}_{ii} =& \lim_{N\to\infty} \mathbb{V}\left[\sqrt{N}(q_i - p_i)\right] = p_i(1-p_i) + 2\sum_{h=1}^{\infty}\left(p_i^{(h)} - p_i^2\right), \\
\boldsymbol{\Sigma}_{ii'} =& \lim_{N\to\infty} \operatorname{Cov}\left[\sqrt{N}(q_i - p_i), \sqrt{N}(q_{i'} - p_{i'})\right] \\
=& 2\sum_{h=1}^{\infty}\left(\frac{p_{ii'}^{(h)} + p_{i'i}^{(h)}}{2} - p_i p_{i'}\right) - p_i p_{i'}.
\end{aligned}$$

Proof. In the following, we will frequently use the rewriting:

$$\sum_{k=1}^{N}\sum_{\ell=1}^{N} p_{ij}^{(k-\ell)} = N p_i \cdot \mathbb{1}\{i = j\} + \sum_{h=1}^{N-1}(N-h)\left(p_{ij}^{(h)} + p_{ji}^{(h)}\right) \tag{4.7.1}$$

where we have used the equality $p_{i\ell}^{(h)} = p_{\ell i}^{(-h)}$, valid under stationarity.

We have:

$$\begin{aligned}\mathbb{V}\left[\sqrt{N}\left(q_i-p_i\right)\right]=N\mathbb{V}\left(q_i\right)&=\frac{1}{N}\mathbb{V}\left(\sum_{j=1}^{N}1\left\{x_j=i\right\}\right)\\&=\frac{1}{N}\sum_{j=1}^{N}\sum_{j'=1}^{N}\mathrm{Cov}\left(1\left\{x_j=i\right\},1\left\{x_{j'}=i\right\}\right)\\&=\frac{1}{N}\sum_{j=1}^{N}\sum_{j'=1}^{N}\left\{\mathbb{E}\left(1\left\{x_j=i\right\}1\left\{x_{j'}=i\right\}\right)-p_i^2\right\}\\&=\frac{1}{N}\sum_{j=1}^{N}\sum_{j'=1}^{N}\left(p_i^{(j-j')}-p_i^2\right)\\&=p_i\left(1-p_i\right)+2\sum_{h=1}^{N-1}\left(1-\frac{h}{N}\right)\left(p_i^{(h)}-p_i^2\right)\end{aligned}\tag{4.7.2}$$

and:

$$\begin{aligned}\mathrm{Cov}\left[\sqrt{N}\left(q_i-p_i\right),\sqrt{N}\left(q_{i'}-p_{i'}\right)\right]=&N\mathrm{Cov}\left(q_i,q_{i'}\right)\\=&\frac{1}{N}\mathrm{Cov}\left(\sum_{j=1}^{N}1\left\{x_j=i\right\},\sum_{j'=1}^{N}1\left\{x_{j'}=i'\right\}\right)\\=&\frac{1}{N}\sum_{j=1}^{N}\sum_{j'=1}^{N}\left(p_{ii'}^{(j-j')}-p_ip_{i'}\right)\\=&2\sum_{h=1}^{N-1}\left(1-\frac{h}{N}\right)\left(\frac{p_{ii'}^{(h)}+p_{i'i}^{(h)}}{2}-p_ip_{i'}\right)-p_ip_{i'}\end{aligned}\tag{4.7.3}$$

where we have used repeatedly (4.7.1) and $2\sum_{h=1}^{N}\left(1-\frac{h}{N}\right)=N-1$.

Now, we can apply Lemma 1.1 in [409, p. 2] to $\mathbb{V}\left[\sqrt{N}\left(q_i-p_i\right)\right]$. If $\lim_{\ell\to\infty}p_i\left(1-p_i\right)+2\sum_{h=1}^{\ell}\left(p_i^{(h)}-p_i^2\right)$ exists, then $\mathbb{V}\left[\sqrt{N}\left(q_i-p_i\right)\right]$ converges to the same limit. Now, it is clear that $\left|p_i^{(h)}-p_i^2\right|\leq\alpha\left(h\right)$ for any i. If the process is α-mixing with $\sum_{n=1}^{\infty}\alpha\left(n\right)<\infty$, we can apply Lemma 1.2 in [409, p. 3]. Then $\lim_{N\to\infty}\sum_{h=1}^{N-1}\left(p_i^{(h)}-p_i^2\right)$ exists and can be written as $\sum_{h=1}^{\infty}\left(p_i^{(h)}-p_i^2\right)$. Therefore:

$$\lim_{N\to\infty}\mathbb{V}\left[\sqrt{N}\left(q_i-p_i\right)\right]=p_i\left(1-p_i\right)+2\sum_{h=1}^{\infty}\left(p_i^{(h)}-p_i^2\right).\tag{4.7.4}$$

The same reasoning allows us to write:

$$\lim_{N\to\infty}\mathrm{Cov}\left[\sqrt{N}\left(q_i-p_i\right),\sqrt{N}\left(q_{i'}-p_{i'}\right)\right]=2\sum_{h=1}^{\infty}\left(\frac{p_{ii'}^{(h)}+p_{i'i}^{(h)}}{2}-p_ip_{i'}\right)-p_ip_{i'}.$$

To prove the asymptotic normality of $\sqrt{N}\left(\mathbf{q}-\mathbf{p}\right)$, it is enough to apply the CLT in Theorem 18.5.4 of [244]. A vector version of the same result is in [129, p. 67].

4.7.3 Bias

Proof of Proposition 4.2. From Lemma 4.1 in Section 4.7.1 with $r = 2$, we have:

$$\mathbb{E}H_N = H_\infty - \frac{1}{2}\sum_{i=1}^{B} \frac{\mathbb{E}\left(q_i - p_i\right)^2}{p_i} + \mathbb{E}R_3$$

where:

$$\mathbb{E}\left|R_3\right| \leq \frac{1}{6}\sum_{i=1}^{B} \frac{\mathbb{E}\left|q_i - p_i\right|^3}{\left(\lambda^\star p_i\right)^2}.$$

Under stationarity:

$$\begin{aligned}\sum_{i=1}^{B} \frac{\mathbb{E}\left(q_i - p_i\right)^2}{2p_i} &= \frac{1}{2N}\sum_{i=1}^{B}\left\{\left(1 - p_i\right) + \frac{2\sum_{h=1}^{N-1}\left(1 - \frac{h}{N}\right)\left(p_i^{(h)} - p_i^2\right)}{p_i}\right\}\\ &= \frac{B-1}{2N} + \frac{1}{N}\sum_{i=1}^{B} \frac{\sum_{h=1}^{N-1}\left(1 - \frac{h}{N}\right)\left(p_i^{(h)} - p_i^2\right)}{p_i}\\ &= \frac{B-1}{2N} + \frac{1}{N}\sum_{i=1}^{B} \frac{\sum_{h=1}^{N-1}\left(p_i^{(h)} - p_i^2\right)}{p_i} - \frac{1}{N^2}\sum_{i=1}^{B} \frac{\sum_{h=1}^{N-1} h\left(p_i^{(h)} - p_i^2\right)}{p_i}.\end{aligned}$$

If the process is ergodic stationary, then:

$$\frac{1}{N-1}\sum_{h=1}^{N-1}\left(p_i^{(h)} - p_i^2\right) \to 0$$

(see, e.g., [118, Theorem 13.13]). Now, we show that $\frac{1}{N^2}\sum_{h=1}^{N-1} h\left(p_i^{(h)} - p_i^2\right) \to 0$. Note that most majorizations do not work here as they would involve taking the absolute value of $p_i^{(h)} - p_i^2$. Let us define $s_n := p_i^{(n)} - p_i^2$ for $n \in \mathbb{N}_0$ and $s_0 := 0$, and define $\{a_n\}$ as the sequence whose partial sums are given by $\{s_n\}$, i.e. $s_n = \sum_{j=0}^{n} a_j$ or $a_n = s_n - s_{n-1}$. Now we define the Cesàro averaging methods (C, α) for $\alpha \geq 0$ (see [448, Section 2.2]). Using Definitions 2.9 and 2.10 and Lemma 2.11 in [448], we are led to consider:

$$\frac{A_n^\alpha}{E_n^\alpha} = \frac{\sum_{k=0}^{n}\binom{n-k+\alpha}{\alpha} a_k}{\binom{n+\alpha}{\alpha}}.$$

If $\lim_{n\to\infty} A_n^\alpha / E_n^\alpha$ converges to a limit, we say that $\{a_n\}$ is summable (C, α), where summability (C, k) implies summability $(C, k+1)$ to the same limit (see [510, Vol. I, p. 76]). Now:

$$\begin{aligned}\frac{A_n^0}{E_n^0} &= \sum_{k=0}^{n} a_k = s_n\\ \frac{A_n^1}{E_n^1} &= \frac{\sum_{k=0}^{n}\left(n - k + 1\right)a_k}{n+1} = \frac{\sum_{k=0}^{n} s_k}{n+1}\\ \frac{A_n^2}{E_n^2} &= \frac{\sum_{k=0}^{n}\left(n - k + 1\right)\left(n - k + 2\right)a_k}{\left(n+1\right)\left(n+2\right)} = 2\left\{\frac{\sum_{k=0}^{n} s_k}{n+2} - \frac{\sum_{k=0}^{n} k s_k}{\left(n+1\right)\left(n+2\right)}\right\}.\end{aligned}$$

The sequence $\{a_n\}$ is summable $(C,1)$ with limit 0 as $\lim_{n\rightarrow\infty} {}^{A_n^1}/_{E_n^1} = 0$. Therefore it is also summable $(C,2)$, i.e.:

$$\lim_{n\rightarrow\infty} 2\left\{\frac{\sum_{k=0}^{n} s_k}{n+2} - \frac{\sum_{k=0}^{n} k s_k}{(n+1)(n+2)}\right\} = 0$$

and, as a result, $\lim_{n\rightarrow\infty} n^{-2}\sum_{k=0}^{n} k s_k = 0$. In our case, $\frac{1}{N^2}\sum_{h=1}^{N-1} h\left(p_i^{(h)} - p_i^2\right) \rightarrow 0$.

As to the remainder term:

$$\mathbb{E}\left|R_3\right| \leq \frac{1}{6}\sum_{i=1}^{B}\frac{\mathbb{E}\left|q_i - p_i\right|^3}{\left(\lambda^\star p_i\right)^3} \leq \frac{1}{6}\sum_{i=1}^{B}\frac{\mathbb{E}\left|q_i - p_i\right|^2}{\left(\lambda^\star p_i\right)^3} = o\left(1\right).$$

This can also be proved in a different way as in [378, Section 4]. Let us start from the formula:

$$H_N = H_\infty - \sum_{i=1}^{B}\left(q_i - p_i\right)\ln p_i - D_{KL}\left(\mathbf{q};\mathbf{p}\right)$$

where $D_{KL}\left(\mathbf{q};\mathbf{p}\right) := \sum_{i=1}^{B} q_i \ln\frac{q_i}{p_i}$ is the Kullback-Leibler divergence. Therefore, $\mathbb{E}D_{KL}\left(\mathbf{q};\mathbf{p}\right) = H_\infty - \mathbb{E}H_N$ and, as $D_{KL}\left(\mathbf{q};\mathbf{p}\right) \geq 0$ (see [184, p. 422]), $\mathbb{E}H_N \leq H_\infty$ and the bias of H_N is always negative. Now, from [184, Theorem 5]:

$$D_{KL}\left(\mathbf{q};\mathbf{p}\right) \leq \ln\left(1 + \sum_{i=1}^{B}\frac{\left(q_i - p_i\right)^2}{p_i}\right)$$

and, through Jensen inequality:

$$\mathbb{E}D_{KL}\left(\mathbf{q};\mathbf{p}\right) \leq \mathbb{E}\ln\left(1 + \sum_{i=1}^{B}\frac{\left(q_i - p_i\right)^2}{p_i}\right) \leq \ln\left(1 + \sum_{i=1}^{B}\frac{\mathbb{E}\left(q_i - p_i\right)^2}{p_i}\right).$$

At last:

$$0 \leq H_\infty - \mathbb{E}H_N \leq \ln\left(1 + \mathbb{E}\sum_{i=1}^{B}\frac{\left(q_i - p_i\right)^2}{p_i}\right) \leq \sum_{i=1}^{B}\frac{\mathbb{E}\left(q_i - p_i\right)^2}{p_i}$$

and:

$$\left|H_\infty - \mathbb{E}H_N - \frac{1}{2}\sum_{i=1}^{B}\frac{\mathbb{E}\left(q_i - p_i\right)^2}{p_i}\right| \leq \frac{1}{2}\sum_{i=1}^{B}\frac{\mathbb{E}\left(q_i - p_i\right)^2}{p_i}.$$

Now we turn to the mixing case. We can apply the reasoning leading to (4.7.4) in Lemma 4.3 in Section 4.7.2 to show that

$$\lim_{N\rightarrow\infty} N\sum_{i=1}^{B}\frac{\mathbb{E}\left(q_i - p_i\right)^2}{2p_i} = \frac{B-1}{2} + \sum_{i=1}^{B}\frac{\sum_{h=1}^{\infty}\left(p_i^{(h)} - p_i^2\right)}{p_i}.$$

As far as $\mathbb{E}\left|R_3\right|$ is concerned, we use Theorem 6.3 in [409]:

$$N^3\mathbb{E}\left|q_i - p_i\right|^3 = \mathbb{E}\left|N\left(q_i - p_i\right)\right|^3 \leq 2^{11}3\left\{s_N^3 + N\int_0^1\left[\alpha^{-1}\left(u\right)\wedge N\right]^2 Q^3\left(u\right)\mathrm{d}u\right\}$$

where $s_N^2 := \sum_{j=1}^N \sum_{\ell=1}^N \left|\operatorname{Cov}\left(\mathbf{1}\left\{x_j = i\right\}, \mathbf{1}\left\{x_\ell = i\right\}\right)\right|$ and $Q\left(\cdot\right)$ is the quantile function of the random variable $\mathbf{1}\left\{x_j = i\right\}$. Now:

$$\begin{aligned} s_N^2 :=& \sum_{j=1}^N \sum_{\ell=1}^N \left| p_i^{(j-\ell)} - p_i^2 \right| \leq N p_i \left(1 - p_i\right) + 2N \sum_{h=1}^{N-1} \alpha\left(h\right) - 2 \sum_{h=1}^{N-1} h\alpha\left(h\right) \\ \leq& N p_i \left(1 - p_i\right) + 2N \sum_{h=1}^{N-1} \alpha\left(h\right). \end{aligned}$$

As $\sum_{h=1}^\infty \alpha\left(h\right) < \infty$, $s_N^2 = O\left(N\right)$. From (C.10) in [409], using the fact that $Q\left(\cdot\right) \leq 1$:

$$M_{p,\alpha,N}\left(Q\right) = \int_0^1 \left[\alpha^{-1}\left(u\right) \wedge N\right]^{p-1} Q^p\left(u\right) \mathrm{d}u \leq \max\left(1, p-1\right) \sum_{h=1}^{N-1} \left(h+1\right)^{p-2} \alpha\left(h\right)$$

and $\mathbb{E}\left|q_i - p_i\right|^3 \lesssim N^{-\frac{3}{2}} + N^{-2} \sum_{h=1}^{N-1} h\alpha\left(h\right)$. From $\sum_{h=1}^\infty \alpha\left(h\right) < \infty$ we get $\alpha\left(h\right) = o\left(h^{-1}\right)$ and $\mathbb{E}\left|R_3\right| = o\left(N^{-1}\right)$.

4.7.4 Asymptotic normality and Berry-Esséen bound

Proof of Proposition 4.3. Asymptotic normality of H_N follows from asymptotic normality of $\sqrt{N}\left(\mathbf{q} - \mathbf{p}\right)$ (see Lemma 4.3 in Section 4.7.2) and the delta method (see, e.g., Example 6.1 (b) in [450, p. 279]). The quantity $\sqrt{N}\left(H_N - H_\infty\right)$ is asymptotically equivalent to:

$$\sum_{i=1}^B \frac{\partial H_\infty}{\partial p_i} \cdot \sqrt{N}\left(q_i - p_i\right) = -\sum_{i=1}^B \left(1 + \ln p_i\right) \cdot \sqrt{N}\left(q_i - p_i\right) = -\sum_{i=1}^B \ln p_i \cdot \sqrt{N}\left(q_i - p_i\right).$$

This is asymptotically normal with variance:

$$\begin{aligned} &\sum_{i=1}^B \sum_{i'=1}^B \mathbf{\Sigma}_{ii'} \ln p_i \ln p_{i'} \\ &\quad= \sum_{i=1}^B p_i \ln^2 p_i + \sum_{i=1}^B \sum_{i'=1}^B \left\{ 2 \sum_{h=1}^\infty \left(\frac{p_{ii'}^{(h)} + p_{i'i}^{(h)}}{2} - p_i p_{i'} \right) - p_i p_{i'} \right\} \cdot \ln p_i \ln p_{i'} \\ &\quad= \sum_{i=1}^B p_i \ln^2 p_i - \left(\sum_{i=1}^B p_i \ln p_i \right)^2 + 2 \sum_{i=1}^B \sum_{i'=1}^B \sum_{h=1}^\infty \left(\frac{p_{ii'}^{(h)} + p_{i'i}^{(h)}}{2} - p_i p_{i'} \right) \ln p_i \ln p_{i'} \end{aligned}$$

where $\mathbf{\Sigma}_{ii}$ and $\mathbf{\Sigma}_{ii'}$ are defined in Lemma 4.3 in Section 4.7.2, and the first equality uses the fact that $\mathbf{\Sigma}_{ii}$ is identical to p_i plus the expression for $\mathbf{\Sigma}_{ii'}$ in which i' is formally replaced by i.

Now we turn to the Berry-Esséen bound. We will apply Lemma 1.3 in [450, p. 261], i.e. the inequality:

$$\begin{aligned} \left\|F_{W+\Delta} - \Phi\right\|_\infty &\leq \left\|F_W - \Phi\right\|_\infty + 4\mathbb{E}\left|W\Delta\right| + 4\mathbb{E}\left|\Delta\right| \\ &\leq \left\|F_W - \Phi\right\|_\infty + 4\sqrt{\mathbb{E}W^2\mathbb{E}\Delta^2} + 4\sqrt{\mathbb{E}\Delta^2}, \end{aligned}$$

valid for any random variables W and Δ. We identify $W+\Delta$ with $\sqrt{N}\left(H_N - H_\infty\right)/\sigma$ and W with $-\sigma^{-1}\sum_{i=1}^{B}\sqrt{N}\left(q_i - p_i\right)\cdot \ln p_i$. As far as $\left\|F_W - \Phi\right\|_\infty$ is concerned, if $\sum_{j=1}^{\infty} j\varphi\left(j\right) < \infty$, it is shown to be $O\left(N^{-1/2}\right)$ in [408, Théorème 1]. Now we turn to the second term, and we remark that $\mathbb{E}W^2 = 1$. Therefore:

$$\left\|F_{W+\Delta} - \Phi\right\|_\infty \leq O\left(N^{-1/2}\right) + 8\sqrt{\mathbb{E}\Delta^2}.$$

From Lemma 4.1 in Section 4.7.1, we have:

$$H_N = H_\infty - \sum_{i=1}^{B}\left(q_i - p_i\right)\ln p_i + R_2$$

where $\left|R_2\right| \leq \frac{1}{2}\sum_{i=1}^{B}\frac{\left(q_i-p_i\right)^2}{\lambda^\star p_i}$. Therefore, $\Delta = -\frac{\sqrt{N}}{\sigma}R_2$ and:

$$\Delta^2 \leq \frac{N}{4\lambda^{\star,2}\sigma^2}\left(\sum_{i=1}^{B}\frac{\left(q_i-p_i\right)^2}{p_i}\right)^2 \leq \frac{NB}{4\lambda^{\star,2}\sigma^2}\sum_{i=1}^{B}\frac{\left(q_i-p_i\right)^4}{p_i^2}.$$

Using the fact that $\alpha\left(n\right) \leq \varphi\left(n\right)$ and $\varphi\left(n\right) \leq \kappa\left(n+1\right)^{-2}$ for any n, by Remark 6.3 in [409] or Eq. (2.10) in [409, p. 36], $\mathbb{E}\left(q_i - p_i\right)^4 = O\left(N^{-2}\right)$ and $\mathbb{E}\Delta^2 = O\left(N^{-1}\right)$. As a consequence the whole bound is $O\left(N^{-1/2}\right)$.

4.7.5 Distribution under degeneracy

Proof of Proposition 4.4. A seldom observed fact is that, if $p_i \equiv 1/B$ for any i, the first-order term in Lemma 4.1 in Section 4.7.1 is:

$$\sum_{i=1}^{B}\left(q_i-p_i\right)\ln p_i = -\ln B\cdot\sum_{i=1}^{B}\left(q_i - \frac{1}{B}\right) = -\ln B\cdot\left(\sum_{i=1}^{B}q_i - 1\right) = 0.$$

In this case the asymptotic distribution is a degenerate normal random variable with null variance. Therefore, we apply Lemma 4.2 in Section 4.7.1 identifying $k=B$, $\tau_n = \sqrt{N}$, $\mathbf{X}_n = \mathbf{q}$, $\boldsymbol{\mu} = \mathbf{p}$ and $g\left(\mathbf{x}\right) = -\sum_{i=1}^{B}x_i \ln x_i$. The convergence $\sqrt{N}\left(\mathbf{q}-\mathbf{p}\right) \to_{\mathcal{D}} \mathcal{N}\left(\mathbf{0},\boldsymbol{\Sigma}\right)$ (that is $\tau_n\left(\mathbf{X}_n - \boldsymbol{\mu}\right) \to_{\mathcal{D}} \mathbf{X}$) is proved in Lemma 4.3 in Section 4.7.2. We need to compute $\frac{\partial^2 g(\mathbf{x})}{\partial\mathbf{x}\partial\mathbf{x}'}$ that is the diagonal matrix with $\left[\frac{\partial^2 g(\mathbf{x})}{\partial\mathbf{x}\partial\mathbf{x}'}\right]_{ii} = -\frac{1}{x_i}$, so that $\left.\frac{\partial^2 g(\mathbf{x})}{\partial\mathbf{x}\partial\mathbf{x}'}\right|_{\mathbf{x}=\boldsymbol{\mu}}$ is $-\mathrm{dg}\left(\overline{\mathbf{p}}\right)$. If we write $\mathbf{G}$ for a standard normal vector, we have:

$$N\left(H_N - H_\infty\right) \to_{\mathcal{D}} -\frac{1}{2}\mathbf{G}'\boldsymbol{\Sigma}^{\frac{1}{2}}\mathrm{dg}\left(\overline{\mathbf{p}}\right)\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{G}.$$

It can be shown (see, e.g., [467]) that the asymptotic distribution of $N\left(H_N - H_\infty\right)$ is minus a weighted sum of chi square random variables whose weights are the eigenvalues, arranged in decreasing order, $\left(\lambda_1,\ldots,\lambda_B\right)$ of the matrix $\boldsymbol{\Omega}$ where:

$$\boldsymbol{\Omega}_{ii} = \frac{1}{2p_i}\boldsymbol{\Sigma}_{ii},$$

$$\boldsymbol{\Omega}_{ij} = \frac{1}{2\left(p_ip_j\right)^{1/2}}\boldsymbol{\Sigma}_{ij},$$

where $\mathbf{\Sigma}_{ii}$ and $\mathbf{\Sigma}_{ii'}$ are defined in Lemma 4.3 in Section 4.7.2. At last:

$$N\left(H_N - H_\infty\right) \to_{\mathcal{D}} -\sum_{i=1}^{B} \lambda_i \chi^2_{1,i}.$$

Proof of Corollary 4.2. We have:

$$N\left(H_N - \text{bias}\left(H_N\right) - H_\infty\right) = N\left(H_N - H_\infty\right) - N\text{bias}\left(H_N\right).$$

From (4.4.3) and (4.4.4):

$$\text{bias}\left(H_N\right) = -\frac{\text{tr}\left(\text{dg}\left(\bar{\mathbf{p}}\right)\mathbf{\Sigma}\right)}{2N} = -\frac{\text{tr}\left(\mathbf{\Omega}\right)}{N} = -\frac{\sum_{i=1}^{B}\lambda_i}{N},$$

from which we get the result.

4.7.6 Preliminary results on Markov chain estimation

Proof of Proposition 4.5. We can write:

$$\begin{aligned}
&\left[p_{ii'}^{(h)} - p_i p_{i'}\right] \\
&\quad = \text{dg}\left(\mathbf{p}\right)\mathbf{P}^h - \mathbf{p}\mathbf{p}' \\
&\qquad \cdot \left[\sum_{h=1}^{\infty}\left(p_{ii'}^{(h)} - p_i p_{i'}\right) + \sum_{h=1}^{\infty}\left(p_{i'i}^{(h)} - p_i p_{i'}\right) + p_i \mathbf{1}\left\{i = i'\right\} - p_i p_{i'}\right] \\
&\quad = \sum_{h=1}^{\infty}\left(\text{dg}\left(\mathbf{p}\right)\mathbf{P}^h - \mathbf{p}\mathbf{p}'\right) + \sum_{h=1}^{\infty}\left(\text{dg}\left(\mathbf{p}\right)\mathbf{P}^h - \mathbf{p}\mathbf{p}'\right)' + \text{dg}\left(\mathbf{p}\right) - \mathbf{p}\mathbf{p}'.
\end{aligned}$$

We then note that $\mathbf{p}\mathbf{p}' = \text{dg}\left(\mathbf{p}\right)\boldsymbol{\iota}\mathbf{p}'$ allows us to write:

$$\sum_{h=1}^{\infty}\left(\text{dg}\left(\mathbf{p}\right)\mathbf{P}^h - \mathbf{p}\mathbf{p}'\right) = \text{dg}\left(\mathbf{p}\right)\sum_{h=1}^{\infty}\left(\mathbf{P}^h - \boldsymbol{\iota}\mathbf{p}'\right).$$

It is well known that:

$$\sum_{h=0}^{\infty}\left(\mathbf{P}^h - \boldsymbol{\iota}\mathbf{p}'\right) = \left(\mathbf{I} - \mathbf{P} + \boldsymbol{\iota}\mathbf{p}'\right)^{-1} = \mathbf{H}$$

from which:

$$\sum_{h=1}^{\infty}\left(\mathbf{P}^h - \boldsymbol{\iota}\mathbf{p}'\right) = \mathbf{H} - \mathbf{I}. \tag{4.7.5}$$

Therefore:

$$\begin{aligned}\boldsymbol{\Sigma} &= \sum_{h=1}^{\infty}\left(\mathrm{dg}\left(\mathbf{p}\right)\mathbf{P}^{h}-\mathbf{p}\mathbf{p}'\right)+\sum_{h=1}^{\infty}\left(\mathrm{dg}\left(\mathbf{p}\right)\mathbf{P}^{h}-\mathbf{p}\mathbf{p}'\right)'+\mathrm{dg}\left(\mathbf{p}\right)-\mathbf{p}\mathbf{p}' \\ &= \mathrm{dg}\left(\mathbf{p}\right)\sum_{h=1}^{\infty}\left(\mathbf{P}^{h}-\boldsymbol{\iota}\mathbf{p}'\right)+\sum_{h=1}^{\infty}\left(\mathbf{P}^{h}-\boldsymbol{\iota}\mathbf{p}'\right)'\mathrm{dg}\left(\mathbf{p}\right)+\mathrm{dg}\left(\mathbf{p}\right)-\mathbf{p}\mathbf{p}' \\ &= \mathrm{dg}\left(\mathbf{p}\right)\left(\mathbf{H}-\mathbf{I}\right)+\left(\mathbf{H}'-\mathbf{I}\right)\mathrm{dg}\left(\mathbf{p}\right)+\mathrm{dg}\left(\mathbf{p}\right)-\mathbf{p}\mathbf{p}' \\ &= \mathrm{dg}\left(\mathbf{p}\right)\mathbf{H}+\mathbf{H}'\mathrm{dg}\left(\mathbf{p}\right)-\mathrm{dg}\left(\mathbf{p}\right)-\mathbf{p}\mathbf{p}'.\end{aligned}$$

The bias can be computed as:

$$\begin{aligned}\mathrm{bias}\left(H_N\right) &= -\frac{\mathrm{tr}\left(\mathrm{dg}\left(\bar{\mathbf{p}}\right)\boldsymbol{\Sigma}\right)}{2N} \\ &= -\frac{2\mathrm{tr}\mathbf{H}-B-1}{2N}\end{aligned}$$

where we have used the equalities $\mathrm{tr}\left(\mathbf{AB}\right)=\mathrm{tr}\left(\mathbf{BA}\right)$, $\mathrm{dg}\left(\mathbf{a}\right)\mathrm{dg}\left(\bar{\mathbf{a}}\right)=\mathbf{I}$, $\mathrm{tr}\left(\mathbf{A}\right)=\mathrm{tr}\left(\mathbf{A}'\right)$, $\mathbf{p}\mathbf{p}'=\mathbf{p}\boldsymbol{\iota}'\mathrm{dg}\left(\mathbf{p}\right)$ and $\mathrm{tr}\left(\mathbf{p}\boldsymbol{\iota}'\right)=\mathrm{tr}\left(\boldsymbol{\iota}'\mathbf{p}\right)=1$.

The matrix $\boldsymbol{\Omega}$ used to obtain the distribution in the degenerate case is then defined as:

$$\boldsymbol{\Omega}=-\frac{1}{2}\mathbf{I}+\frac{1}{2}\mathrm{dg}\left(\mathbf{p}^{\odot\frac{1}{2}}\right)\left(\mathbf{H}\mathrm{dg}\left(\bar{\mathbf{p}}\right)+\mathrm{dg}\left(\bar{\mathbf{p}}\right)\mathbf{H}'-\mathbf{U}\right)\mathrm{dg}\left(\mathbf{p}^{\odot\frac{1}{2}}\right).$$

Lemma 4.4. *For the method in Section 4.4.2:*

$$\widehat{\mathrm{bias}\left(H_N\right)}\simeq\mathrm{bias}\left(H_N\right)+O_{\mathbb{P}}\left(N^{-3/2}\right).$$

Proof. Using the matrix differential notation (see [320, 235, 46]), we write $\mathrm{d}\mathbf{P}:=\hat{\mathbf{P}}-\mathbf{P}$, where $\mathrm{d}\mathbf{P}$ is asymptotically negligible of order $O_{\mathbb{P}}\left(N^{-1/2}\right)$. It is easy to see that:

$$\left(\hat{\mathbf{p}}-\mathbf{p}\right)'\left(\mathbf{I}-\mathbf{P}+\boldsymbol{\iota}\mathbf{p}'\right)=\hat{\mathbf{p}}'\left(\hat{\mathbf{P}}-\mathbf{P}\right).$$

We have:

$$\begin{aligned}\left(\hat{\mathbf{p}}-\mathbf{p}\right)'\left(\mathbf{I}-\mathbf{P}+\boldsymbol{\iota}\mathbf{p}'\right) &= \hat{\mathbf{p}}'\mathrm{d}\mathbf{P} \\ \left(\hat{\mathbf{p}}-\mathbf{p}\right)' &= \hat{\mathbf{p}}'\mathrm{d}\mathbf{P}\left(\mathbf{I}-\mathbf{P}+\boldsymbol{\iota}\mathbf{p}'\right)^{-1}=\hat{\mathbf{p}}'\mathrm{d}\mathbf{P}\mathbf{H} \\ \hat{\mathbf{p}}' &= \mathbf{p}'+\hat{\mathbf{p}}'\mathrm{d}\mathbf{P}\mathbf{H} \\ \hat{\mathbf{p}} &= \mathbf{p}+\mathbf{H}'\mathrm{d}\mathbf{P}'\hat{\mathbf{p}}.\end{aligned}$$

Replacing the expression for $\hat{\mathbf{p}}$ in the last formula we get:

$$\hat{\mathbf{p}}=\mathbf{p}+\mathbf{H}'\mathrm{d}\mathbf{P}'\hat{\mathbf{p}}\simeq\mathbf{p}+\mathbf{H}'\mathrm{d}\mathbf{P}'\mathbf{p}.$$

From this:

$$\begin{aligned}\hat{\mathbf{H}} &= \left(\mathbf{I}-\hat{\mathbf{P}}+\boldsymbol{\iota}\hat{\mathbf{p}}'\right)^{-1}\\ &= \left(\mathbf{I}-\mathbf{P}-\mathrm{d}\mathbf{P}+\boldsymbol{\iota}\left(\mathbf{p}'+\hat{\mathbf{p}}'\mathrm{d}\mathbf{P}\mathbf{H}\right)\right)^{-1}\\ &= \left(\mathbf{H}^{-1}+\left(\boldsymbol{\iota}\hat{\mathbf{p}}'\mathrm{d}\mathbf{P}\mathbf{H}-\mathrm{d}\mathbf{P}\right)\right)^{-1}\\ &= \mathbf{H}\left(\mathbf{I}+\mathbf{H}\left(\boldsymbol{\iota}\hat{\mathbf{p}}'\mathrm{d}\mathbf{P}\mathbf{H}-\mathrm{d}\mathbf{P}\right)\right)^{-1}\\ &\simeq \mathbf{H}\left(\mathbf{I}-\mathbf{H}\left(\boldsymbol{\iota}\hat{\mathbf{p}}'\mathrm{d}\mathbf{P}\mathbf{H}-\mathrm{d}\mathbf{P}\right)\right)\\ &\simeq \mathbf{H}\left(\mathbf{I}-\mathbf{H}\left(\boldsymbol{\iota}\mathbf{p}'\mathrm{d}\mathbf{P}\mathbf{H}-\mathrm{d}\mathbf{P}\right)\right)\end{aligned}$$

Then:

$$\begin{aligned}\widehat{\mathrm{bias}\left(H_N\right)} &= -\frac{2\mathrm{tr}\hat{\mathbf{H}}-B-1}{2N}\\ &\simeq -\frac{2\mathrm{tr}\mathbf{H}-B-1}{2N}+\frac{\mathrm{tr}\left[\mathbf{H}^2\left(\boldsymbol{\iota}\mathbf{p}'\mathrm{d}\mathbf{P}\mathbf{H}-\mathrm{d}\mathbf{P}\right)\right]}{N}\\ &= \mathrm{bias}\left(H_N\right)+\frac{\mathrm{tr}\left[\mathbf{H}^2\left(\boldsymbol{\iota}\mathbf{p}'\mathrm{d}\mathbf{P}\mathbf{H}-\mathrm{d}\mathbf{P}\right)\right]}{N}.\end{aligned}$$

This implies that $\widehat{\mathrm{bias}\left(H_N\right)}-\mathrm{bias}\left(H_N\right)=O_{\mathbb{P}}\left(N^{-3/2}\right)$.

4.7.7 Error in the estimation of bias

We first adapt Lemma A.1 in [388, p. 739] to our case. This result is a version of Theorem 10 in [220, p. 283], Lemma 2 in [12, p. 15], and Proposition 1 in [14, p. 825]. It generalizes these results as it allows for a bandwidth not diverging to ∞, as required by some recent results (see Theorem 2.1 in [388, p. 707]). We define:

$$\boldsymbol{F}^{(q)} := \frac{1}{2\pi}\sum_{h=-\infty}^{\infty}|h|^q\,\boldsymbol{\Pi}^{(h)}.$$

Lemma 4.5. *Assume AC. Then,* $\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}=O_{\mathbb{P}}\left(S_N^{1/2}N^{-1/2}\right)$.

Proof. We first restate Assumptions A, B and C in [14]. We identify T with N, t with n, θ with $\mathbf{p}$, $\widehat{\theta}$ with $\mathbf{q}$, $V_t\left(\theta\right)$ with $\mathbf{x}_n-\mathbf{p}$, $V_t\left(\widehat{\theta}\right)$ with $\mathbf{x}_n-\mathbf{q}$.

In order to verify Assumption A in [14, p. 823], we use his Lemma 1. Our process $\left\{\mathbf{x}_1-\mathbf{p},\mathbf{x}_2-\mathbf{p},\dots\right\}$ is zero-mean, fourth-order stationary and bounded. Therefore, we can take $\nu=\infty$ and it's enough to require $\sum_{n=1}^{\infty}n^2\alpha\left(n\right)<\infty$, as in our condition 1 of assumption AC.

Now we consider Assumption B in [14, p. 825]. Assumption B (i) is verified by an application of Theorem 18.5.4 of [244] or of [129, p. 67] under $\sum_{n=1}^{\infty}\alpha\left(n\right)<\infty$. Assumption B (ii) is trivially true because of the boundedness of $\mathbf{x}_n$. For Assumption B (iii) it is enough to identify $\left(\partial/\partial\theta'\right)V_t\left(\theta\right)$ with $\left(\partial/\partial\mathbf{p}'\right)\left(\mathbf{x}_n-\mathbf{p}\right)=-\mathbf{I}$. Assumption B (iv) is true under condition 3 of assumption AC.

As far as Assumption C (i) in [14, p. 826] is concerned, this is true by the reasoning reported just after the statement of his Assumption C. Part (ii) of the assumption is true as $\left(\partial^2/\partial\theta\partial\theta'\right)V_{ta}\left(\theta\right)$ is

0.

Now we turn to the requirements stated in [388, p. 739]. Conditions 2, 3 and 4 in our result come from Lemma A.1 in [388, p. 739], respectively, as a statement in the text, as Eq. (A.1), and as condition (i). We just note that condition (i) is not needed explicitly in [14] because, when $S_N \to \infty$, $S_N^{-1} \sum_{j=-N+1}^{N-1} |k(j/S_N)| \to \int |k(x)| \mathrm{d}x$ (see [14, p. 852]). This is finite by our condition 3 of Assumption AC. Instead, [388] requires his condition (i) (see [388, p. 744]) because S_N may not diverge. As we allow S_N to be bounded, we require it but change its statement. Condition (iii) of Lemma A.1 in [388, p. 739] is automatically verified as $\mathbb{E}V_t(\theta)(\partial/\partial\theta')V_{t-j}(\theta) = -\mathbb{E}V_t(\theta)$ and this is a zero vector. Condition (ii) of Lemma A.1 in [388, p. 739] is trickier. We first note that the matrix denoted $\hat{\Omega}$ in that source corresponds to the matrix $\tilde{\boldsymbol{\Sigma}}$ defined as:

$$\tilde{\boldsymbol{\Sigma}} = \sum_{h=-N+1}^{N-1} k\left(\frac{h}{S_N}\right) \tilde{\boldsymbol{\Pi}}^{(h)}$$

where:

$$\tilde{\boldsymbol{\Pi}}^{(h)} = \begin{cases} \frac{1}{N}\sum_{n=h+1}^{N} (\mathbf{x}_n - \mathbf{p})(\mathbf{x}_{n-h} - \mathbf{p})' & h \geq 0, \\ \frac{1}{N}\sum_{n=-h+1}^{N} (\mathbf{x}_{n+h} - \mathbf{p})(\mathbf{x}_n - \mathbf{p})' & h < 0. \end{cases}$$

(Note that $\tilde{\boldsymbol{\Pi}}^{(h)}$ is similar to $\boldsymbol{\Pi}^{(h)}$, but the centering is different.) Now, we have $\mathbb{E}\tilde{\boldsymbol{\Pi}}^{(h)} = \frac{N-h}{N}\boldsymbol{\Pi}^{(h)}$ and:

$$\mathbb{E}\tilde{\boldsymbol{\Sigma}} = \mathbb{E}\tilde{\boldsymbol{\Pi}}^{(0)} + 2\sum_{h=1}^{N-1} k\left(\frac{h}{S_N}\right)\left(1 - \frac{h}{N}\right)\boldsymbol{\Pi}^{(h)}.$$

This means that:

$$\mathbb{E}\tilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} = -2\sum_{h=1}^{N-1}\left(1 - k\left(\frac{h}{S_N}\right)\right)\boldsymbol{\Pi}^{(h)} - \frac{2}{N}\sum_{h=1}^{N-1} k\left(\frac{h}{S_N}\right) h\boldsymbol{\Pi}^{(h)} - 2\sum_{h=N}^{\infty}\boldsymbol{\Pi}^{(h)}.$$

The requirement in [388, Lemma A.1] is that $\mathbb{E}\tilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} = O\left(S_N^{1/2} N^{-1/2}\right)$. Let $\boldsymbol{\gamma}$ be a vector with $\|\boldsymbol{\gamma}\|_{L^2} = 1$.

Let us start from the second term. For $q \geq 1$, we have:

$$\begin{aligned} \frac{1}{N}\left|\boldsymbol{\gamma}'\left\{\sum_{h=1}^{N-1} k\left(\frac{h}{S_N}\right) h\boldsymbol{\Pi}^{(h)}\right\}\boldsymbol{\gamma}\right| \leq & \frac{1}{N}\sum_{h=1}^{N-1}\left|k\left(\frac{h}{S_N}\right)\right| h \left\|\boldsymbol{\Pi}^{(h)}\right\|_{L^2} \\ \leq & \frac{1}{N}\sum_{h=1}^{N-1} h\left\|\boldsymbol{\Pi}^{(h)}\right\|_{L^2} \leq \frac{1}{N}\sum_{h=1}^{\infty} h\left\|\boldsymbol{\Pi}^{(h)}\right\|_{L^2}. \end{aligned}$$

This is automatically $O\left(S_N^{1/2}N^{-1/2}\right)$. For $q<1$, we use the fact that $|h/N|\leq|h/N|^q$ for $h<N$:

$$\begin{aligned}\frac{1}{N}\left|\gamma'\left\{\sum_{h=1}^{N-1}k\left(\frac{h}{S_N}\right)h\mathbf{\Pi}^{(h)}\right\}\gamma\right|\leq&\frac{1}{N}\sum_{h=1}^{N-1}\left|k\left(\frac{h}{S_N}\right)\right|h\left\|\mathbf{\Pi}^{(h)}\right\|_{L^2}\\ \leq&N^{-q}\sum_{h=1}^{N-1}h^q\left\|\mathbf{\Pi}^{(h)}\right\|_{L^2}.\end{aligned}$$

When $q\geq 1/2$, this is automatically $O\left(S_N^{1/2}N^{-1/2}\right)$. When $q<1/2$, it requires $N^{1/2-q}S_N^{-1/2}=O(1)$.

The last term can be majorized as:

$$\begin{aligned}\left|\gamma'\left\{\sum_{h=N}^{\infty}\mathbf{\Pi}^{(h)}\right\}\gamma\right|\leq&\sum_{h=N}^{\infty}\left\|\mathbf{\Pi}^{(h)}\right\|_{L^2}\\ \leq&N^{-q}\sum_{h=N}^{\infty}h^q\left\|\mathbf{\Pi}^{(h)}\right\|_{L^2}.\end{aligned}$$

We need $N^{-q}\sum_{h=N}^{\infty}h^q\left\|\mathbf{\Pi}^{(h)}\right\|_{L^2}=O\left(S_N^{1/2}N^{-1/2}\right)$. For $q\geq 1/2$, this is automatically verified. For $q<1/2$, it is true under $N^{1/2-q}S_N^{-1/2}=O(1)$.

Now we turn to the first term. If $S_N\not\to\infty$ as $N\to\infty$, we just require it to be $O\left(S_N^{1/2}N^{-1/2}\right)$. If $S_N\to\infty$ as $N\to\infty$, we can reason as in [220, p. 284] and in [12, pp. A4-A5]. We have:

$$S_N^q\sum_{h=1}^{N-1}\left(1-k\left(\frac{h}{S_N}\right)\right)\mathbf{\Pi}^{(h)}=\sum_{h=1}^{N-1}\left(\frac{1-k(h/S_N)}{(h/S_N)^q}-k_q\right)h^q\mathbf{\Pi}^{(h)}+k_q\sum_{h=1}^{N-1}h^q\mathbf{\Pi}^{(h)}.$$

Now, the function defined by $\frac{1-k(x)}{|x|^q}$ for $x\neq 0$ and by k_q for $x=0$ is non-negative and bounded by a constant M. Hence, $\frac{1-k(x)}{|x|^q}\leq M$. Let us choose a fixed N_0 such that $\sum_{h=N_0}^{\infty}h^q\left\|\mathbf{\Pi}^{(h)}\right\|_{L^2}\leq\varepsilon/(2M)$ for $\varepsilon>0$. Then:

$$\begin{aligned}&\left\|\sum_{h=1}^{N-1}\left(\frac{1-k(h/S_N)}{(h/S_N)^q}-k_q\right)h^q\mathbf{\Pi}^{(h)}\right\|_{L^2}\\ &=\left\|\sum_{h=1}^{N_0-1}\left(\frac{1-k(h/S_N)}{(h/S_N)^q}-k_q\right)h^q\mathbf{\Pi}^{(h)}+\sum_{h=N_0}^{N-1}\left(\frac{1-k(h/S_N)}{(h/S_N)^q}-k_q\right)h^q\mathbf{\Pi}^{(h)}\right\|_{L^2}\\ &\leq\sum_{h=1}^{N_0-1}\left|\frac{1-k(h/S_N)}{(h/S_N)^q}-k_q\right|h^q\left\|\mathbf{\Pi}^{(h)}\right\|_{L^2}+\sum_{h=N_0}^{N-1}\left|\frac{1-k(h/S_N)}{(h/S_N)^q}-k_q\right|h^q\left\|\mathbf{\Pi}^{(h)}\right\|_{L^2}\\ &\leq\sum_{h=1}^{N_0-1}\left|\frac{1-k(h/S_N)}{(h/S_N)^q}-k_q\right|h^q\left\|\mathbf{\Pi}^{(h)}\right\|_{L^2}+2M\sum_{h=N_0}^{N-1}h^q\left\|\mathbf{\Pi}^{(h)}\right\|_{L^2}\\ &\lesssim o(1)+\varepsilon=o(1)\end{aligned}$$

where the first term is $o(1)$ due to the bounded convergence of $\frac{1-k(x)}{|x|^q}-k_q$ to 0 and the second is

$o\,(1)$ due to the arbitrariness of ε. When $N \to \infty$, $\sum_{h=1}^{N-1} h^q \mathbf{\Pi}^{(h)}$ converges to $\mathbf{F}^{(q)}$. As a result:

$$\begin{aligned}
&-2\sum_{h=1}^{N-1}\left(1-k\left(\frac{h}{S_N}\right)\right)\mathbf{\Pi}^{(h)}\\
&= -2S_N^{-q}\sum_{h=1}^{N-1}\left(\frac{1-k\,(h/S_N)}{(h/S_N)^q}-k_q\right)h^q\mathbf{\Pi}^{(h)} - 2k_qS_N^{-q}\sum_{h=1}^{N-1}h^q\mathbf{\Pi}^{(h)}\\
&= o\left(S_N^{-q}\right) - 2k_qS_N^{-q}\mathbf{F}^{(q)}.
\end{aligned}$$

If $k_q \neq 0$, the second term dominates and we need $-2k_qS_N^{-q}\mathbf{F}^{(q)} = O\left(S_N^{1/2}N^{-1/2}\right)$. If $k_q = 0$, the condition boils down to the one for $S_N \not\to \infty$.

Proof of Proposition 4.6. We consider a limited development of $\widehat{\text{bias}\,(H_N)}$ with respect to $\widehat{\mathbf{\Sigma}}_{ii}$ and q_i respectively around $\mathbf{\Sigma}_{ii}$ and p_i:

$$\begin{aligned}
\widehat{\text{bias}\,(H_N)} =& \frac{1}{2N}\sum_{i=1}^{B}\frac{\widehat{\mathbf{\Sigma}}_{ii}}{q_i} = \frac{1}{2N}\sum_{i=1}^{B}\frac{\mathbf{\Sigma}_{ii}+\left(\widehat{\mathbf{\Sigma}}_{ii}-\mathbf{\Sigma}_{ii}\right)}{p_i\left(1+\frac{q_i-p_i}{p_i}\right)}\\
\simeq& \frac{1}{2N}\sum_{i=1}^{B}\frac{\mathbf{\Sigma}_{ii}+\left(\widehat{\mathbf{\Sigma}}_{ii}-\mathbf{\Sigma}_{ii}\right)}{p_i}\left(1-\frac{q_i-p_i}{p_i}+\left(\frac{q_i-p_i}{p_i}\right)^2\right)\\
=& \frac{1}{2N}\sum_{i=1}^{B}\left(\frac{\mathbf{\Sigma}_{ii}}{p_i}+\frac{\widehat{\mathbf{\Sigma}}_{ii}-\mathbf{\Sigma}_{ii}}{p_i}-\frac{\mathbf{\Sigma}_{ii}\,(q_i-p_i)}{p_i^2}\right.\\
&\left.-\frac{\left(\widehat{\mathbf{\Sigma}}_{ii}-\mathbf{\Sigma}_{ii}\right)(q_i-p_i)}{p_i^2}+\frac{\mathbf{\Sigma}_{ii}\,(q_i-p_i)^2}{p_i^3}+\frac{\left(\widehat{\mathbf{\Sigma}}_{ii}-\mathbf{\Sigma}_{ii}\right)(q_i-p_i)^2}{p_i^3}\right).
\end{aligned}$$

Now, $q_i - p_i = O_{\mathbb{P}}\left(N^{-1/2}\right)$ and, under AC, Lemma 4.5 implies that $\widehat{\mathbf{\Sigma}}_{ii} - \mathbf{\Sigma}_{ii} = O_{\mathbb{P}}\left(\left(S_N/N\right)^{1/2}\right)$. At last we get:

$$\widehat{\text{bias}\,(H_N)} \simeq \text{bias}\,(H_N) + O_{\mathbb{P}}\left(S_N^{1/2}N^{-3/2}\right). \tag{4.7.6}$$

For the Markov case, we refer to Lemma 4.4 in Section 4.7.6.

Proof of Corollary 4.3. We only consider the case of Section 4.4.1. It is simple to see that:

$$\begin{aligned}
&\sqrt{N}\left(H_N - \widehat{\text{bias}\,(H_N)} - H_\infty\right)\\
&\quad= \sqrt{N}\,(H_N - \text{bias}\,(H_N) - H_\infty)\\
&\qquad+ \sqrt{N}\left(\text{bias}\,(H_N) - \widehat{\text{bias}\,(H_N)}\right)\\
&\quad= \sqrt{N}\,(H_N - \text{bias}\,(H_N) - H_\infty) + O_{\mathbb{P}}\left(S_N^{1/2}N^{-1}\right)
\end{aligned}$$

and:

$$\begin{aligned}
&N\left(H_N - \widehat{\mathrm{bias}\left(H_N\right)} - H_\infty\right)\\
&\quad = N\left(H_N - \mathrm{bias}\left(H_N\right) - H_\infty\right)\\
&\qquad + N\left(\mathrm{bias}\left(H_N\right) - \widehat{\mathrm{bias}\left(H_N\right)}\right)\\
&\quad = N\left(H_N - \mathrm{bias}\left(H_N\right) - H_\infty\right) + O_{\mathbb{P}}\left(S_N^{1/2} N^{-1/2}\right)
\end{aligned}$$

and, provided $S_N = o\left(N\right)$, the results of Corollaries 4.1 and 4.2.

Proof of Corollary 4.4. Let us first consider the case when $\sigma^2 > 0$. Then, from Proposition 4.3 it is trivial to see that:

$$\mathrm{MSE}\left(H_N\right) = \mathbb{E}\left(H_N - H_\infty\right)^2 = O\left(N^{-1}\right).$$

We have:

$$\begin{aligned}
&\mathrm{MSE}\left(H_N\right) - \mathrm{MSE}\left(H_N - \mathrm{bias}\left(H_N\right)\right)\\
&\quad = \mathbb{E}\left(H_N - H_\infty\right)^2 - \mathbb{E}\left(H_N - \mathrm{bias}\left(H_N\right) - H_\infty\right)^2\\
&\quad = \mathrm{bias}\left(H_N\right)\left(2\mathbb{E}H_N - \mathrm{bias}\left(H_N\right) - 2H_\infty\right)\\
&\quad \simeq \left[\mathrm{bias}\left(H_N\right)\right]^2 = O\left(N^{-2}\right).
\end{aligned}$$

At last:

$$\begin{aligned}
&\mathrm{MSE}\left(H_N - \widehat{\mathrm{bias}\left(H_N\right)}\right) - \mathrm{MSE}\left(H_N - \mathrm{bias}\left(H_N\right)\right)\\
&\quad = \mathbb{E}\left(H_N - \widehat{\mathrm{bias}\left(H_N\right)} - H_\infty\right)^2 - \mathbb{E}\left(H_N - \mathrm{bias}\left(H_N\right) - H_\infty\right)^2\\
&\quad = \mathbb{E}\left(\mathrm{bias}\left(H_N\right) - \widehat{\mathrm{bias}\left(H_N\right)}\right)\left(2H_N - \widehat{\mathrm{bias}\left(H_N\right)} - \mathrm{bias}\left(H_N\right) - 2H_\infty\right)\\
&\quad = O\left(\left[\mathbb{E}\left(\mathrm{bias}\left(H_N\right) - \widehat{\mathrm{bias}\left(H_N\right)}\right)^2\right]^{1/2}\right.\\
&\qquad \left.\cdot\left[\mathbb{E}\left(\left(H_N - \widehat{\mathrm{bias}\left(H_N\right)} - H_\infty\right) + \left(H_N - \mathrm{bias}\left(H_N\right) - H_\infty\right)\right)^2\right]^{1/2}\right)\\
&\quad = O\left(\left[\mathbb{E}\left(\mathrm{bias}\left(H_N\right) - \widehat{\mathrm{bias}\left(H_N\right)}\right)^2\right]^{1/2}\right.\\
&\qquad \left.\cdot\left[2\left(\mathbb{E}\left(H_N - \widehat{\mathrm{bias}\left(H_N\right)} - H_\infty\right)^2 + \mathbb{E}\left(H_N - \mathrm{bias}\left(H_N\right) - H_\infty\right)^2\right)\right]^{1/2}\right)\\
&\quad = O\left(S_N^{1/2} N^{-2}\right)
\end{aligned}$$

where the third step comes from Cauchy-Schwarz inequality, the fourth from the inequality $\left(a+b\right)^2 \leq 2\left(a^2+b^2\right)$, and the fifth from Proposition 4.6 and Corollaries 4.1 and 4.3.

When $\sigma^2 = 0$, from Proposition 4.4:

$$\mathrm{MSE}\left(H_N\right) = \mathbb{E}\left(H_N - H_\infty\right)^2 = O\left(N^{-2}\right).$$

The formula for $\mathrm{MSE}\left(H_N\right) - \mathrm{MSE}\left(H_N - \mathrm{bias}\left(H_N\right)\right)$ remains unchanged. The formula for $\mathrm{MSE}\left(H_N - \widehat{\mathrm{bias}\left(H_N\right)}\right) -$ $\mathrm{MSE}\left(H_N - \mathrm{bias}\left(H_N\right)\right)$ becomes:

$$\begin{aligned}
&\mathrm{MSE}\left(H_N - \widehat{\mathrm{bias}\left(H_N\right)}\right) - \mathrm{MSE}\left(H_N - \mathrm{bias}\left(H_N\right)\right)\\
&= O\left(\left[\mathbb{E}\left(\mathrm{bias}\left(H_N\right) - \widehat{\mathrm{bias}\left(H_N\right)}\right)^2\right]^{1/2}\right.\\
&\quad \left.\cdot \left[\mathbb{E}\left(\left(H_N - \widehat{\mathrm{bias}\left(H_N\right)} - H_\infty\right) + \left(H_N - \mathrm{bias}\left(H_N\right) - H_\infty\right)\right)^2\right]^{1/2}\right)\\
&= O\left(S_N^{1/2} N^{-5/2}\right)
\end{aligned}$$

if Corollary 4.1 is replaced by Corollary 4.2.

4.7.8 Error in the estimation of the distribution under degeneracy

Proof of Proposition 4.7. Suppose that the matrix $\mathbf{\Omega}$ is estimated through $\hat{\mathbf{\Omega}}$, where $\hat{\mathbf{\Omega}} - \mathbf{\Omega} = o_{\mathbb{P}}(1)$. This means that we will replace the distribution of $\sum_{i=1}^{B} \lambda_i \chi^2_{1,i}$ by the distribution of $\sum_{i=1}^{B} \hat{\lambda}_i \chi^2_{1,i}$. We would like to characterize the error in this replacement through a bound on:

$$\left\| F_{\sum_{i=1}^{B} \hat{\lambda}_i \chi^2_{1,i}} - F_{\sum_{i=1}^{B} \lambda_i \chi^2_{1,i}} \right\|_\infty .$$

The eigenvalues of both $\mathbf{\Omega}$ and $\hat{\mathbf{\Omega}}$ are non-negative as both of them are variance matrices. Let $B^\star$ be the number of non-zero eigenvalues of $\mathbf{\Sigma}^\star$, so that $\lambda_{B^\star} > 0$. We need to differentiate the case $B^\star = 1$ from the case $B^\star > 1$.

By the Wielandt-Hoffman inequality (see [255, p. 126]), we have:

$$\left|\lambda_i - \hat{\lambda}_i\right| \leq \left\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\right\|_1, \quad i = 1, \ldots, B$$

and therefore $\left|\lambda_{B^\star} - \hat{\lambda}_{B^\star}\right| \leq \left\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\right\|_1$ or $\lambda_{B^\star} - \left\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\right\|_1 \leq \hat{\lambda}_{B^\star}$. As $\left\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\right\|_1 = o_{\mathbb{P}}(1)$, for N large enough, $\hat{\lambda}_{B^\star} > 0$, so that the two matrices have ultimately the same rank.

We start from the case $B^\star > 1$. We define $\boldsymbol{\lambda} = \left(\lambda_1, \ldots, \lambda_B\right)$, $\hat{\boldsymbol{\lambda}} = \left(\hat{\lambda}_1, \ldots, \hat{\lambda}_B\right)$, $\mathbf{\Lambda} = \mathrm{dg}\left(\boldsymbol{\lambda}\right)$ and $\hat{\mathbf{\Lambda}} = \mathrm{dg}\left(\hat{\boldsymbol{\lambda}}\right)$. The techniques in [94, 442] do not work directly here, as they require part of the eigenvalues to coincide. However, we can write:

$$\begin{aligned}
&\left\| F_{\sum_{i=1}^{B} \hat{\lambda}_i \chi_{1,i}^2} - F_{\sum_{i=1}^{B} \lambda_i \chi_{1,i}^2} \right\|_{\infty} \\
&\quad = \sup_{x \geq 0} \left| \mathbb{P}\left\{ \sum_{i=1}^{B} \hat{\lambda}_i \chi_{1,i}^2 \leq x \right\} - \mathbb{P}\left\{ \sum_{i=1}^{B} \lambda_i \chi_{1,i}^2 \leq x \right\} \right| \\
&\quad = \sup_{x \geq 0} \left| \mathbb{P}\left\{ \left\| \mathcal{N}\left(\mathbf{0}, \hat{\boldsymbol{\Lambda}}\right) \right\|_{L_2} \leq \sqrt{x} \right\} - \mathbb{P}\left\{ \left\| \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Lambda}\right) \right\|_{L_2} \leq \sqrt{x} \right\} \right|.
\end{aligned}$$

We use Theorem 1 in [352]:

$$\begin{aligned}
&\sup_{x \geq 0} \left| \mathbb{P}\left\{ \left\| \mathcal{N}\left(\mathbf{0}, \hat{\boldsymbol{\Lambda}}\right) \right\|_{L_2} \leq x \right\} - \mathbb{P}\left\{ \left\| \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Lambda}\right) \right\|_{L_2} \leq x \right\} \right| \\
&\quad \leq C \cdot \left\{ \left(\sum_{i=1}^{B} \lambda_i^2 \cdot \sum_{i=2}^{B} \lambda_i^2 \right)^{-1/4} + \left(\sum_{i=1}^{B} \hat{\lambda}_i^2 \cdot \sum_{i=2}^{B} \hat{\lambda}_i^2 \right)^{-1/4} \right\} \cdot \sum_{i=1}^{B} \left| \lambda_i - \hat{\lambda}_i \right|
\end{aligned}$$

for an absolute constant $C > 0$. By the Wielandt-Hoffman inequality (see [255, p. 126]), $\sum_{i=1}^{B} \left| \lambda_i - \hat{\lambda}_i \right| \leq \left\| \hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega} \right\|_1$. Now, $\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega} = o_{\mathbb{P}}(1)$ implies that also $\hat{\lambda}_i - \lambda_i = o_{\mathbb{P}}(1)$, so that, for N large enough:

$$\left\| F_{\sum_{i=1}^{B} \hat{\lambda}_i \chi_{1,i}^2} - F_{\sum_{i=1}^{B} \lambda_i \chi_{1,i}^2} \right\|_{\infty} \leq C \frac{\left\| \hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega} \right\|_1}{\left(\sum_{i=1}^{B} \lambda_i^2 \cdot \sum_{i=2}^{B} \lambda_i^2 \right)^{1/4}}$$

where the constant C is here different from the one seen above.

When $B^{\star} = 1$, we have:

$$\begin{aligned}
\left\| F_{\hat{\lambda}_1 \chi_1^2} - F_{\lambda_1 \chi_1^2} \right\|_{\infty} = & \sup_{x} \left| \mathbb{P}\left\{ \hat{\lambda}_1 \chi_1^2 \leq x \right\} - \mathbb{P}\left\{ \lambda_1 \chi_1^2 \leq x \right\} \right| \\
= & \sup_{x} \left| \mathbb{P}\left\{ \left| \hat{\lambda}_1^{1/2} Z \right| \leq x \right\} - \mathbb{P}\left\{ \left| \lambda_1^{1/2} Z \right| \leq x \right\} \right| \\
= & \sup_{x} \left| \mathbb{P}\left\{ \hat{\lambda}_1^{1/2} Z \leq x \right\} - \mathbb{P}\left\{ \hat{\lambda}_1^{1/2} Z \leq -x \right\} \right. \\
& \left. - \mathbb{P}\left\{ \lambda_1^{1/2} Z \leq x \right\} + \mathbb{P}\left\{ \lambda_1^{1/2} Z \leq -x \right\} \right| \\
\leq & 2 \sup_{x > 0} \left| \Phi(x) - \Phi\left(\hat{\lambda}_1^{1/2} \lambda_1^{-1/2} x \right) \right|.
\end{aligned}$$

We are only interested in the case in which $\hat{\lambda}_1^{1/2} \lambda_1^{-1/2} \simeq 1$, therefore we write $\hat{\lambda}_1^{1/2} \lambda_1^{-1/2} = 1 + \varepsilon$ and we get:

$$\Phi(x) - \Phi\left(\hat{\lambda}_1^{1/2} \lambda_1^{-1/2} x \right) = \Phi(x) - \Phi\left((1 + \varepsilon) x \right) \simeq \phi(x) x \varepsilon.$$

This implies that $\left| \Phi(x) - \Phi\left(\hat{\lambda}_1^{1/2} \lambda_1^{-1/2} x \right) \right| \lesssim \sup_{x \in \mathbb{R}} \left| \phi(x) x \right| \cdot \left| \varepsilon \right| \leq C \left| \varepsilon \right|$ (where C can be taken equal to or larger than $\sup_{x\in\mathbb{R}} |\phi(x)\, x| = 1/\sqrt{2\pi e} \doteq 0.2419707$). Therefore:

$$\left\| F_{\hat{\lambda}_1\chi_1^2} - F_{\lambda_1\chi_1^2} \right\|_\infty \lesssim 2C \left| \frac{\hat{\lambda}_1 - \lambda_1}{\lambda_1^{1/2}\left(\hat{\lambda}_1^{1/2} + \lambda_1^{1/2}\right)} \right| \lesssim C \left| \frac{\hat{\lambda}_1 - \lambda_1}{\lambda_1} \right|.$$

In this case too, $\left\| F_{\sum_{i=1}^B \hat{\lambda}_i\chi_{1,i}^2} - F_{\sum_{i=1}^B \lambda_i\chi_{1,i}^2} \right\|_\infty = O\left(\left\| \hat{\mathbf{\Omega}} - \mathbf{\Omega} \right\|_1\right)$.

The final part of the statement comes from the results in Lemma A.1 in [388] under assumption AC.

4.8 Supplementary results

4.8.1 A Non-ergodic Example

Consider an iid standard Gaussian sequence $\{y_1, y_2, \dots\}$ and a standard Gaussian random variable z independent of the previous sequence. Then we define:

$$\widetilde{x}_i = \beta z + \left(1 - \beta^2\right)^{1/2} y_i.$$

As above, the dichotomized process is based on the signs of the original process:

$$x_i = 1 + \mathbb{1}\left\{\widetilde{x}_i \geq 0\right\}.$$

It is clear that:

$$\begin{aligned} p_1 &= \mathbb{P}\left\{x_i = 1\right\} = \mathbb{P}\left\{\widetilde{x}_i < 0\right\} = 1/2 \\ p_2 &= 1 - p_1 = 1/2. \end{aligned}$$

However, $\mathrm{Cov}\left(\widetilde{x}_1, \widetilde{x}_{h+1}\right) = \mathrm{Cov}\left(\beta z + \left(1-\beta^2\right)^{1/2} y_1, \beta z + \left(1-\beta^2\right)^{1/2} y_{h+1}\right) = \beta^2$. From [474, p. 189], we have:

$$\begin{aligned} p_{22}^{(h)} &= p_{11}^{(h)} = 1/4 + 1/2\pi \arcsin\left(\beta^2\right) \\ p_{12}^{(h)} &= p_{21}^{(h)} = 1/2 - p_{22}^{(h)}. \end{aligned}$$

In this case, the process $\{\widetilde{x}_1, \widetilde{x}_2, \dots\}$ is stationary but non-ergodic (see, e.g., Example 13.9 in [118, p. 196]), and so is $\{x_1, x_2, \dots\}$. Therefore:

$$\begin{aligned} q_1 &= \frac{n_1}{N} = \frac{\sum_{j=1}^N \mathbb{1}\left\{x_j = 1\right\}}{N} = \frac{\sum_{j=1}^N \mathbb{1}\left\{\widetilde{x}_j < 0\right\}}{N} \\ &= \frac{\sum_{j=1}^N \mathbb{1}\left\{y_i < -\frac{\beta}{\left(1-\beta^2\right)^{1/2}} z\right\}}{N} \to \Phi\left(-\frac{\beta}{\left(1-\beta^2\right)^{1/2}} z\right) \quad \mathbb{P}-\text{as}. \end{aligned}$$

Note that z is an invariant random variable. At last the limit of the observed entropy is:

$$\begin{aligned} H_\infty =& -\Phi\left(-\frac{\beta}{\left(1-\beta^2\right)^{1/2}}z\right)\ln\Phi\left(-\frac{\beta}{\left(1-\beta^2\right)^{1/2}}z\right) \\ & -\Phi\left(\frac{\beta}{\left(1-\beta^2\right)^{1/2}}z\right)\ln\Phi\left(\frac{\beta}{\left(1-\beta^2\right)^{1/2}}z\right). \end{aligned}$$

In Figure 4.8.1 we show what is the behavior of the statistic in this case. The grey jigsaw lines represent some trajectories of H_N, while the dark grey curved line represents $\mathbb{E}H_N$ (based on 250,000 samples) and the dark grey horizontal line represents $\mathbb{E}H_\infty$. On the right plot, we display the empirical cdf of H_N with $N = 25$ (black dashed line), $N = 50$ (black dotted line), $N = 100$ (black dash-dot line), $N = 200$ (black long dash line). The black solid line represents the empirical cdf of H_∞. In the non-ergodic case H_N converges (almost surely) to the random variable H_∞.

In this case, $H_\infty \neq \mathbb{E}H_\infty$. The first quantity has been obtained above. The second one is given by:

$$\begin{aligned} \mathbb{E}H_\infty =& -\mathbb{E}\left\{\Phi\left(-\frac{\beta}{\left(1-\beta^2\right)^{1/2}}z\right)\ln\Phi\left(-\frac{\beta}{\left(1-\beta^2\right)^{1/2}}z\right)\right\} \\ & +\mathbb{E}\left\{\Phi\left(\frac{\beta}{\left(1-\beta^2\right)^{1/2}}z\right)\ln\Phi\left(\frac{\beta}{\left(1-\beta^2\right)^{1/2}}z\right)\right\} \\ =& -2\mathbb{E}\left\{\Phi\left(\frac{\beta}{\left(1-\beta^2\right)^{1/2}}z\right)\ln\Phi\left(\frac{\beta}{\left(1-\beta^2\right)^{1/2}}z\right)\right\}. \end{aligned}$$

The first-order bias correction of $\mathbb{E}H_N$ takes the form:

$$\begin{aligned} & -\frac{B-1}{2N}-\frac{1}{N}\sum_{i=1}^{B}\frac{\sum_{h=1}^{N-1}\left(p_i^{(h)}-p_i^2\right)}{p_i} \\ =& -\frac{1}{2N}-\frac{2\left(N-1\right)}{\pi N}\arcsin\left(\beta^2\right)=O\left(1\right). \end{aligned}$$

Note that, while the first-order bias correction reduces the bias in the estimation of $\mathbb{E}H_\infty$ through $\mathbb{E}H_N$, nothing guarantees that this reduces the bias in the estimation of H_∞ through H_N.

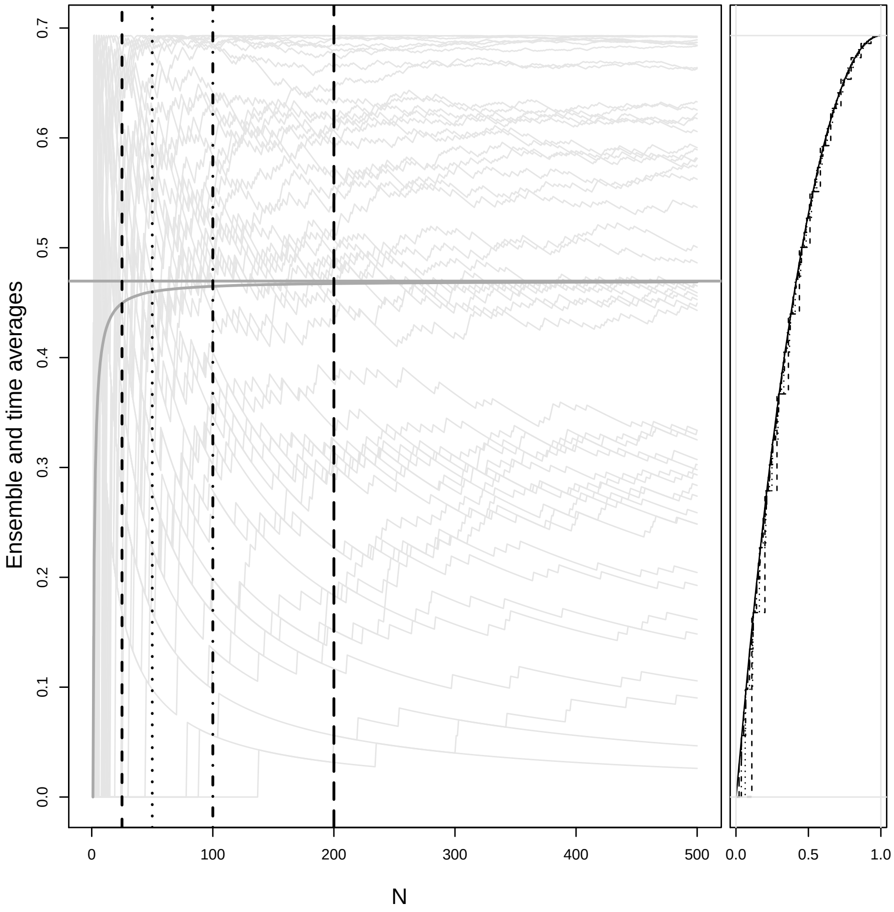

Figure 4.8.1: Ensemble and time averages of the entropy in the non-ergodic case: on the left plot, 50 trajectories of H_N as a function of N (light grey jigsaw lines), $\mathbb{E}H_\infty$ (dark grey horizontal line), $\mathbb{E}H_N$ (dark grey curved line), vertical lines at $N \in \{25, 50, 100, 200\}$ (respectively black dashed, dotted, dash-dot, long dashed lines); on the right plot, empirical cdf of H_N with $N = 25$ (black dashed line), $N = 50$ (black dotted line), $N = 100$ (black dash-dot line), $N = 200$ (black long dashed line) and $N = \infty$ (black solid line).

Chapter 5

Nonparametric Moment-based Estimation of Simulated Models without Optimization[1]

In this chapter, a new method for the estimation of simulated models is presented. It exploits a nonparametric sieve regression estimated through OLS to find the parameters of a simulation model producing statistics that are close to the ones obtained in real-world data. The simulation model is run for several values of the parameters, statistics are computed on each run, and the function linking the generated statistics and the associated parameters is estimated nonparametrically. Estimates of the parameters are then obtained through the previous nonparametric estimate using the real-world statistics as explanatory variables. At odds with simulated minimum-distance techniques (e.g., indirect inference and simulated method of moments), our framework does not involve any objective function and no optimization algorithm is required. The full asymptotic theory of the estimator is explicitly and rigorously characterized, including the order of the bias, confidence intervals and hypotheses tests. The approach is evaluated through a small simulation study.

5.1 Introduction

In this paper, we propose a new method for the estimation of simulation-based models. Our aim is to find the parameters of a simulated model that produce statistics that are close to the ones obtained in real-world/benchmark data. This generally requires the minimization of a distance between simulated and real-world data, but our setup does not require any optimization algorithm typically used in simulated minimum-distance techniques (e.g., indirect inference and method of simulated moments, see below). Instead, the method relies upon ordinary least squares (OLS) nonparametric regression.

[1] This chapter is co-authored with Raffaello Seri.

Classical frameworks for the estimation, calibration and validation of statistical models through simulation involve indirect inference (II, [196, 460, 195]), method of simulated moments (MSM or SMM, [335, 372, 133]), simulated minimum-distance (SMD, [217, 463, 191, 48]), approximated ([292, 293, 157, 2]) and nonparametric simulated maximum likelihood ([156, 272]), and approximate Bayesian computations (ABC, [148, 25, 131, 175, 168, 458]).

As agent-based models (ABM) produce simulated observations, their estimation and calibration has been generally performed adapting the previous econometric techniques in order to select the parameter values that best fit the model (see [444, p. 3]). Many examples can be found in the literature, among these we quote [189, 190, 54, 150] who estimate the parameters of an ABM via a version of II, [499, 172, 206, 90] who exploit MSM, [401, 282] who develop calibration techniques based on SMD, [277, 314] who apply simulated maximum likelihood methods, and [207] who rely on ABC techniques. The computational burden of ABM simulations has spurred the development of algorithms, relying upon kriging (see [271]) or machine learning techniques, improving the performances of these estimators through surrogate meta-models (see [422, 283]).

The calibration algorithms have inspired efforts in some related statistical problems that are relevant for ABM. First, an important branch of application of the minimum-distance approach is sensitivity analysis (SA). This stream of literature aims at studying the structure and the uncertainty associated to the model output when a change occurs in the model input (see [395] for a broad application of SA to different fields of research). Many approaches are possible, the most recent advances being summarized in [424, 34, 66, 472, 65, 469]. Second, recent years have seen a surge of interest in the validation of ABM, that is the comparison of the output of a calibrated model with an external source of data (see [151] and [444, p. 2] for a critical review of the definitions of validation). This involves both the choice of distances between data to be used in calibration (see [215, 282]) and the development of algorithms for the identification of those combinations of parameters that reproduce features of real data (see [29, 444]). In the latter case, some care has been devoted in the literature to avoid numerical optimization.

It is well known that, under certain assumptions, the estimators proposed above are consistent and asymptotically normal (see, e.g., [475, p. 239]). However, they present some drawbacks. First of all, almost all of them have strong requirements in terms of computational time (see [472, 307]). Second, many of them require the choice of a statistical metric measuring the dissimilarity between simulated and real-world observations (see [184, 301, 402, 32, 329, 426]) and/or the identification of the moments to match (see [179, 499, 111, 75]). A third drawback, that has not been fully recognized in the previous literature, is that several simulation models do not respect some of the conditions for the identification of the parameter $\boldsymbol{\theta}$ in simulation-based estimation methods. These issues are mainly related to the violation of the stochastic equicontinuity condition (see, e.g., [335], [372] and [355, pp. 2136-2137] for a definition). A detailed treatment of the dependence of the simulated data on $\boldsymbol{\theta}$ and the consequences of the violation of the stochastic equicontinuity condition are exposed in Section 5.2.

To cope with these issues, some methods based on regression and its variants have been proposed, but generally for prediction rather than for parameter estimation. These methods often rely on complex machine learning (ML) techniques (see, e.g., [265, 403, 266] for a review and a comparison between regression and kriging meta-modeling in stochastic simulations, [449, 110, 400, 396] for the estimation of parameters via quantile random forest regressions and neural networks). In a technique developed by [86], parameter estimation is obtained via the least absolute shrinkage and selection operator (Lasso). The authors identify a set of parameter values in the parameter space, launch a simulation model generating a vector of statistics for each value of the parameters, then they introduce the simulated summary statistics as independent variables in a regression where the parameters are the dependent variables and, at last, they obtain the estimates as forecasts using the real-world statistics as regressors in the regression estimated above.

Our aim is similar in spirit, but we perform parameter estimation using a linear regression combined with sieve estimation (see [208, 354, 88]) and we explicitly and rigorously characterize the full asymptotic theory of the estimator $\widehat{\theta}^{(j)}$, including the order of the bias, confidence intervals and hypotheses tests. In order to do so, we need to introduce a more formalized setup. More specifically, we consider a simulation model indexed by a vector of parameters $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^K$, where the generic element of $\boldsymbol{\theta}$ is $\theta^{(j)}$ for $j = 1, \ldots, K$. For N parameter values selected in Θ we simulate one or more runs used to compute $M \geq K$ statistics. Let the index j be fixed, and suppose we are interested in the estimation of $\theta^{(j)}$ alone. We estimate an OLS nonparametric regression of the form $\theta_n^{(j)} = f_P^{(j)} \left(\widehat{S}_{1n}, \ldots, \widehat{S}_{Mn} \right) + \eta_n^{(j)}$, where $n = 1, \ldots, N$, $f_P^{(j)}$ is a linear combination of basis functions and $\widehat{S}_{1n}, \ldots, \widehat{S}_{Mn}$ are the statistics obtained from the data simulated by $\theta_n^{(j)}$. If $\widehat{s}_1, \ldots, \widehat{s}_M$ (i.e. $\widehat{\mathbf{s}}$) are the statistics computed on real-world data, we can estimate $\widehat{\theta}^{(j)}$ as $\widehat{\theta}^{(j)} = f_P^{(j)} \left(\widehat{s}_1, \ldots, \widehat{s}_M \right)$. More details on the procedure are given below.

We believe that our setup has several advantages when compared to other estimation methods. In the following we discuss these advantages.

First of all, our technique does not require any optimization. The computation of extremum estimators is notoriously tricky because of the existence of local extrema, the occurrence of boundary solutions, the problems of identification of the extrema, etc., and this is exacerbated in simulation-based estimation methods. Even when stochastic equicontinuity holds true because errors or innovations are recycled (see Section 5.2), it is well known that some simulators do not have good properties (see, as an early example, [296] for a discussion of the so-called crude frequency simulator for choice probabilities) and often slow derivative-free methods of optimization are required (see [75, p. 178] for a recent example). This is amplified when errors or innovations are not recycled because in that case stochastic equicontinuity is violated and the objective function is discontinuous or non-differentiable (see [189, 190, 277]). Moreover, as no optimization algorithm is involved in the estimation process, our technique is less computationally intensive than classical simulation-based estimation methods. As an example, in MSM the number of steps of the minimization procedure is not known in advance and it is often larger when the number of replications used to estimate the statistics is smaller. On the other hand, within our framework we are able to estimate the parameters also when only a few points in the parameter space have been selected and the statistics are estimated using a small number of replications. These estimates may be imprecise but are nevertheless available and can be used as starting points for further refinements.

Second, despite relying on the choice of a finite number of statistics, our method is essentially nonparametric. Indeed, we model the dependence of the parameters on the statistics as a nonparametric function through sieve estimation. This peculiarity allows to leave some liberty on the relation linking the statistics and the parameters to be estimated. When the statistics are moments, the method shares some similarities with MSM and, more generally, the generalized method of moments (GMM). GMM compares the moments based on real data with the theoretical moments expressed as functions, known a priori, of the parameters. When these functions cannot be computed exactly, MSM replaces them with sample expectations based on simulated data. However, while GMM is generally semiparametric as it does not impose a functional form on the statistical model but it only constrains its moments, its simulated counterpart, MSM, requires simulations from completely specified statistical models and is therefore parametric. Our method, instead, is more flexible, as the dependence between moments and parameters is left unspecified and estimated nonparametrically. Our technique shows also some analogies with II, as the function linking the simulated statistics and the related parameters can be thought of as the link function of II. However, in our framework this relation is left unspecified and estimated nonparametrically. Therefore, differently from II, we do not have to rely on the choice of an arbitrary auxiliary model.

Third, despite being nonparametric, our framework relies on OLS regression, a simple technique available in most statistical computer programs. This implies that the algorithm can be implemented in most software with less efforts than methods based on simulated minimum-distance, simulated likelihood or machine learning techniques.

Fourth, our method estimates each parameter separately. On the one hand, this does not allow to capture the constraints among the parameters and, as a result, the estimates can be outside the parameter space. On the other hand, however, this may produce a gain of efficiency in terms of computational time and could reduce the curse of dimensionality, particularly in over-parametrized contexts. The number of parameters of the simulation model will still have an impact on computation: the number of parameter values for which simulations are required in order to estimate the function $f_P^{(j)}$ depends on the number of parameters.

Fifth, our technique applies very much in the same way when the statistics are computed on a time series extracted from a single simulation or on independent replications of the process. This depends on the fact that our method only relies on the availability of a vector of statistics. Whether these statistics are computed exploiting the last R observations of a trajectory or taking the last observation from R distinct trajectories, the estimation method does not change.

Sixth, our estimation method allows to derive a full asymptotic theory for the estimator $\widehat{\theta}^{(j)}$. In particular, we are able to show the rate of convergence of $\widehat{\theta}^{(j)}$ to $\theta^{(j)}$, and to characterize the asymptotic distribution of the estimator. This is useful to obtain confidence intervals and hypotheses tests. Note that most estimation methods described above do not provide inferential tools in the case of ABM, as the derivation of the asymptotic distribution is based on a condition of stochastic equicontinuity that is not respected in this case. However, as pointed out below, the technique proposed here does not assume any continuity condition.

At last, as anticipated, an advantage of our setup is that it does not reckon on stochastic equicontinuity, and no condition on the errors or innovations of the model is required. This allows us to derive the asymptotic behavior of the estimator $\widehat{\theta}^{(j)}$ without assuming any continuity condition of the objective function.

Some weaknesses appear also in our estimation technique. First, the use of linear regression implies that the number of basis function is bounded from above by the number N of points chosen out of the parameter space. A solution to this drawback is analyzed in Chapter 6 exploiting Lasso regression. Second, the convergence rate of $\widehat{\theta}^{(j)}$ to $\theta^{(j)}$ is often worse than in the parametric case. Since we propose a nonparametric estimator, we expect to obtain sub-optimal rates of convergence when compared to classical estimation methods. However, the loss of efficiency in terms of convergence rate is compensated by the flexibility deriving from the nonparametric estimation of the function linking the statistics and the parameters. Third, the advantage coming from the lack of an optimization process may be weakened by two factors: (i) the computational cost deriving from the construction of a high-dimensional grid; (ii) the risk of estimating a raw link function for the region under examination, as the function itself is computed over the whole parameter space. Now, we will show below that, under certain combinations of the asymptotic parameters, the estimation of the link function has no impact on the asymptotic distribution of the statistic. This shows that one could in principle refine the estimate of the function without affecting the properties of the estimator. A first approach to address these issues is to use adaptive mesh refinements (see, e.g., [76]). Another approach could be to apply an iterative procedure. A preliminary estimate of $f_P^{(j)}$ and $\widehat{\theta}^{(j)}$ can be obtained taking few, sparse points in the parameter space. Then, $\widehat{\theta}^{(j)}$ is employed to select a finer grid that is exploited to restrict the parameter space and to estimate the new $f_P^{(j)}$ and $\widehat{\theta}^{(j)}$. The procedure can be repeated until an accurate estimate is reached.

The rest of the work is organized as follows. In Section 5.2 we explain the problems related with the violation of the stochastic equicontinuity condition. In Section 5.3, we provide some notations that will be useful throughout the paper. In Section 5.4, we describe the statistical framework and we give some examples of construction of the function $f_P^{(j)}$. In Section 5.5, we outline the assumptions useful to derive the asymptotic theory. In Section 5.6, we derive the main results of the contribution, i.e., the rate of convergence of the estimator to the true value and the (nonparametric) convergence to a normal distribution in two cases: (i) when the asymptotic theory is determined by $\widehat{\mathbf{s}}$, and (ii) when the asymptotic theory is determined by the simulations. The simulation experiment verifying the finite-sample properties of the estimator is performed in Section 5.7. Section 5.8 sums up the main conclusions. The proofs of the results are deferred to Section 5.9.

5.2 Violation of the Stochastic Equicontinuity Condition

Many simulation-based estimators rely on the optimization of an objective function depending on both real-world and simulated data. Most proofs of the asymptotic properties of these estimators require an assumption called stochastic equicontinuity (see, e.g., [335], [372] and [355, pp. 2136-2137]), a condition of probabilistic continuity of the objective function without which asymptotic properties of simulation-based estimators are not guaranteed.[2]

We try to explain the problem with the following very simple example in which the only parameter

[2] Stochastic equicontinuity is useful when the proofs of the asymptotic properties use uniform convergence of the objective function. When epigraphical convergence is used instead of the uniform one in the proofs of consistency, this condition can be replaced with a one-sided version (see [231, 92, 232]), but we do not pursue the topic here.

is the mean $\mu \in \mathbb{R}$: we have a sample of real data $\{y_1, y_2, \ldots, y_n\}$ and, for a given value of μ, we draw a sample of independent Gaussian random variables $\{y_1(\mu), y_2(\mu), \ldots, y_m(\mu)\}$ with mean μ and variance 1. We compute the estimator μ minimizing the distance between the sample mean of the real-world data and the simulated sample mean (see, in the context of a different model, [195, p. 20]):

$$Q(\mu) := \left(\frac{1}{n} \sum_{i=1}^{n} y_i - \frac{1}{m} \sum_{j=1}^{m} y_j(\mu) \right)^2 .$$

We note that $y_j(\mu) = \mu + \varepsilon_j$, where $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m\}$ are independent standard Gaussian random variables.[3] If the variables $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m\}$ are the same for any value of μ or, as often said, they are recycled for different values of μ, $Q(\mu)$ is a continuous function of μ and stochastic equicontinuity holds true. If, on the other hand, a new sample $\{\varepsilon_1, \varepsilon_2, \ldots\}$ is drawn for any μ the resulting function $Q(\mu)$ will be "rugged" and both the computation and the study of the asymptotic properties of the estimator will be difficult. For this reason, in order to obtain stochastic equicontinuity, the variables $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m\}$ of the model must be recycled for different values of $\boldsymbol{\theta}$, as required in [335, p. 999], [195, p. 16], [272, p. 78] and [143, p. 346].

However, several simulation models, and ABM among them, do not allow for the recycling of errors. This creates two problems. First, as the function to optimize is rugged, optimization routines are often very time-consuming and ad hoc algorithms have to be used. As an example, [189, 190] realize that the objective functions arising in the simulation-based estimation of the parameters of ABM are non-differentiable and devise algorithms combining a simplex search approach with a threshold accepting algorithm in order to identify the optimum. In a different example, [277] outline some problems related to the roughness of the objective function (i.e., multiple local minima, identification issues leading to large standard deviations), and perform a graphical inspection of the simulated log-likelihood function (see [277, pp. 23, 36, 39]). Second, the classical asymptotic properties may not hold and the corresponding inferential tools (tests, confidence intervals) are thus not necessarily available.

Therefore, as our method does not involve any objective function in the estimation process, it can be used to bypass the drawbacks related to the violation of the stochastic equicontinuity hypothesis.

5.3 Notations

This section summarizes the notation used below.

We will use capital bold letters, such as $\mathbf{A}$, to denote matrices and lowercase bold letters, such as $\mathbf{a}$, to indicate vectors. The i-th element of vector $\mathbf{a}$ is generally denoted a_i. $\mathbf{u}_n$ is a n-vector composed of ones. $\mathbf{I}_n$ is the $(n \times n)$-identity matrix. $\mathbf{U}_n$ is a $(n \times n)$-matrix composed of ones. $\mathbf{e}_{i,n}$ is a n-vector of zeros with a one in the i-th position; when the length is clear from the context we simply use $\mathbf{e}_i$. $\mathbf{0}_{m \times n}$ is a $(m \times n)$-matrix composed of zeros. We do not indicate the dimensions when they are clear from the context. $\operatorname{diag}(\mathbf{a})$ is a diagonal matrix with $\mathbf{a}$ on its diagonal. $\mathbf{A}'$ and

[3] The variables $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m\}$ are often called errors in cross-sectional models and innovations in dynamic models.

$\mathbf{A}^{-1}$ are respectively the transpose and the classical inverse of the matrix $\mathbf{A}$, provided they exist. Now, $\|\cdot\|$ denotes the L_2 norm and $\|\cdot\|_F$ denotes the Frobenius norm.

5.4 Framework

Suppose to have a simulation model indexed by some parameters $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^K$. The j-th component of $\boldsymbol{\theta}$ is denoted $\theta^{(j)}$. For each value in Θ, it is possible to simulate one or more runs that are used to compute M statistics, with $M \geq K$. Each component of the K-vector can be expressed as an unknown function of the statistics:

$$\theta^{(j)} = f^{(j)}\left(S_1, S_2, \ldots, S_M\right), \qquad j = 1, \ldots, K,$$

where $f^{(j)} : \mathbb{R}^M \to \mathbb{R}$.

We approximate the function f through a function f_P given by an expansion in a series of basis functions:

$$f_P\left(\mathbf{x}\right) = p'_P\left(\mathbf{x}\right) \cdot \boldsymbol{\gamma}$$

where $p_P\left(\mathbf{x}\right) = \left(p_{1P}\left(\mathbf{x}\right), \ldots, p_{PP}\left(\mathbf{x}\right)\right)'$ is a $P \times 1$ vector of given basis functions. For $f : \mathbb{R}^n \to \mathbb{R}$, we define:

$$\partial^\lambda f\left(\mathbf{x}\right) = \frac{\partial^{|\lambda|} f\left(\mathbf{x}\right)}{\partial x_1^{\lambda_1} \ldots \partial x_n^{\lambda_n}}$$

for a multi-index $\lambda = (\lambda_1, \ldots, \lambda_n)$, where $|\lambda| = \lambda_1 + \cdots + \lambda_n$. We will need the following definitions:

$$\begin{aligned}
|f|_s &:= \max_{|\lambda| \leq s} \sup_{\mathbf{x}} \left|\partial^\lambda f\left(\mathbf{x}\right)\right|, \\
\|f\|_\infty &:= \sup_{\mathbf{x}} \left|f\left(\mathbf{x}\right)\right|, \\
\zeta_s\left(P\right) &:= \max_{|\lambda| \leq s} \sup_{\mathbf{x}} \left\|\partial^\lambda p_P\left(\mathbf{x}\right)\right\|.
\end{aligned}$$

Suppose to extract N configurations of parameters $\left(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\right)' \subset \Theta$ indexed with n. The j-th statistic based on the run (or runs) generated by $\boldsymbol{\theta}_n$ is $\widehat{S}_{jn}$.We estimate a nonparametric regression of the form:

$$\theta_n^{(j)} = f_P^{(j)}\left(\widehat{S}_{1n}, \ldots, \widehat{S}_{Mn}\right) + \eta_n^{(j)}, \qquad n = 1, \ldots, N, j = 1, \ldots, K,$$

where

$$\eta_n^{(j)} := f^{(j)}\left(\widehat{S}_{1n}, \ldots, \widehat{S}_{Mn}\right) - f_P^{(j)}\left(\widehat{S}_{1n}, \ldots, \widehat{S}_{Mn}\right) + \varepsilon_n^{(j)}$$

and $\varepsilon_n^{(j)}$ is the error due to the estimation of the statistics. Note that each parameter $\theta_n^{(j)}$ can be estimated separately.

We then compute the statistics $\widehat{s}_1, \ldots, \widehat{s}_M$ on the basis of some real data. We can forecast the value of $\theta^{(j)}$ as:

$$\widehat{\theta}^{(j)} \simeq f_P^{(j)}\left(\widehat{s}_1, \ldots, \widehat{s}_M\right).$$

The construction of the function f_P is outlined below using some examples. We first consider some cases when $\mathbf{x}$ is a scalar, i.e. $\mathbf{x} = x$. Moreover, we suppose that $x \in [0,1]$.

Example 5.1. [Power series] A solution is to use $p_{jP}(x) = x^{j-1}$. In this case, $\zeta_s(P) \lesssim P^{1+2s}$.

Example 5.2. [Orthogonal polynomial series] An alternative is to use orthogonal polynomials as, e.g., Legendre polynomials, that are orthonormal with respect to the Lebesgue measure on $[0,1]$:

$$p_P'(x) = \left(1, \sqrt{3}x, \sqrt{5/4}\left(3x^2 - 1\right), \dots\right).$$

The value of $\zeta_s(P)$ does not change.

Example 5.3. [Spline series] A spline series of order 1 starts from a finite number of equally spaced knots $\ell_1, \dots, \ell_{k-2}$ in $[0,1]$ and defines:

$$p_P(x) = \left(1, x, (x - \ell_1)_+, \dots, (x - \ell_{k-2})_+\right)'.$$

The cubic splines or spline series of order 3 starts instead from the equally spaced knots $\ell_1, \dots, \ell_{k-4}$ to get:

$$p_P(x) = \left(1, x, x^2, x^3, (x - \ell_1)_+^3, \dots, (x - \ell_{k-4})_+^3\right)'.$$

This can be generalized to an arbitrary order s_0. It is often the case that, instead of splines, B-splines are used. In this case, $\zeta_s(P) \lesssim P^{\frac{1}{2}+s}$.

Other examples are in [109], [13], [241], [88] and [40].

Now we cover the case when $\mathbf{x} = (x_1, \dots, x_M)$ is a vector where each component is supposed to belong to $[0,1]$.

Example 5.4. [Tensor products] In this case, the solution is to take a series $p_{P_i}(x_i)$ for any $i = 1, \dots, M$. The vector $p_P(\mathbf{x})$ is then built as the tensor product of the previous ones, i.e.:

$$p_P(\mathbf{x}) = p_{P_1}(x_1) \otimes \cdots \otimes p_{P_M}(x_M).$$

The number of terms is $P = \prod_{i=1}^M P_i$. The value of $\zeta_s(P)$ is the same of the corresponding method in the scalar case.

Example 5.5. [Total degree space of monomials] The previous solution contains elements of order higher than each P_i. As an example, if each $p_{P_i}(x_i)$ is a polynomial series, we will observe a term of order $P_1 + P_2$ like $x_1^{P_1} x_2^{P_2}$ but we will not observe $x_1^{P_1+P_2}$. In some cases, it is possible to build $p_P(\mathbf{x})$ as a union of forms, where a form is a homogeneous polynomial (as in linear or quadratic form). [340] calls total degree space the set of monomials of degree smaller than a certain value. As an example, if $\mathbf{x} = (x_1, x_2)$ we have:

$$p_P(\mathbf{x}) = \left(1, x_1, x_2, x_1^2, x_1 x_2, x_2^2, \dots\right)'.$$

In the general case, if the degree of each scalar polynomial is p, the number of terms is $P = \frac{(M+p)!}{p!M!} \sim$

$\frac{p^M}{M!}$, where the last relation holds for large p. The corresponding product of tensors is:

$$p_P(\mathbf{x}) = \left(1, x_1, x_2, x_1 x_2, x_1^2, x_2^2, x_1^2 x_2^2, \ldots\right)'.$$

The value of $\zeta_s(P)$ can be bounded from above by $\zeta_s^M(p)$.

5.5 Assumptions

In this section we will provide a more formal derivation of the estimation method that will shed some light on the assumptions we will introduce.

A1 The parameter space $\Theta \subset \mathbb{R}^K$ is supposed to be compact.

For any value of the parameter $\boldsymbol{\theta} \in \Theta$, the model produces a statistic (generally a moment) through a function g:

$$S = g\left(\theta^{(1)}, \ldots, \theta^{(K)}\right) = g(\boldsymbol{\theta}).$$

We suppose that the statistic is a fixed number, not a random variable (this is generally associated with the fact that the model is ergodic, see [233]). Later we will see what happens when the statistic is estimated. We suppose to observe K of these statistics and we index the statistic S and the function g with a progressive subscript:

$$S_j = g^{(j)}\left(\theta^{(1)}, \ldots, \theta^{(K)}\right) = g^{(j)}(\boldsymbol{\theta}), \qquad j = 1, \ldots, K. \tag{5.5.1}$$

In vector coordinates, $\mathbf{S} = \mathbf{g}(\boldsymbol{\theta})$, where $\mathbf{g} = (g_1, \ldots, g_K)' : \mathbb{R}^K \to \mathbb{R}^K$. Under Assumption A2 below, we have $\boldsymbol{\theta} = \mathbf{g}^{-1}(\mathbf{S}) = \mathbf{f}(\mathbf{S})$ or:

$$\theta^{(j)} = f^{(j)}(S_1, S_2, \ldots, S_K), \qquad j = 1, \ldots, K, \tag{5.5.2}$$

where $f^{(j)} : \mathbb{R}^K \to \mathbb{R}$. In general, as some statistics may be unable to discriminate among different values of $\boldsymbol{\theta}$, one would like to consider $M \geq K$ statistics. We are therefore led to the model:

$$\theta^{(j)} = f^{(j)}(S_1, S_2, \ldots, S_M), \qquad j = 1, \ldots, K,$$

where, with an abuse of notation, we keep the notation $f^{(j)}$ for the function with domain in $\mathbb{R}^M$. The following assumption is required to guarantee global invertibility of $\mathbf{g}$.

A2 For a subset of cardinality K of $\{1, \ldots, M\}$, the system (5.5.1) can be globally inverted to give the system (5.5.2), where the functions $f^{(j)}$ are continuous.

Note that we do not use the Inverse Function Theorem as this would only provide local invertibility.

Now we come to estimation. We choose N configurations of parameters indexed by n, say $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)' \subset \Theta$. Then:

$$\theta_n^{(j)} = f^{(j)}(S_{1n}, S_{2n}, \ldots, S_{Mn}) = f^{(j)}\left(\underset{M\times 1}{\mathbf{S}_n}\right), \qquad j = 1, \ldots, K, n = 1, \ldots, N.$$

We replace the function $f^{(j)}$ through a function $f_P^{(j)}$ given by an expansion in a series of basis functions, say $f_P^{(j)}(\cdot) = p'_P(\cdot)\boldsymbol{\gamma}_j$. Therefore:

$$\theta_n^{(j)} = f_P^{(j)}(\mathbf{S}_n) + \left[f^{(j)}(\mathbf{S}_n) - f_P^{(j)}(\mathbf{S}_n)\right], \qquad j = 1,\dots,K, n = 1,\dots,N.$$

If we stack all the N observations:

$$\underset{N\times 1}{\boldsymbol{\theta}^{(j)}} = f_P^{(j)}\left(\underset{M\times N}{\mathbf{S}}\right) + \boldsymbol{\varepsilon}^{(j)}, \qquad j = 1,\dots,K.$$

Note that all variables are fixed and $\boldsymbol{\varepsilon}^{(j)}$ is only an approximation error.

We replace $\mathbf{S}_n$ with $\widehat{\mathbf{S}}_n$, estimated on the basis of R observations:

$$\begin{aligned}\theta_n^{(j)} &= f_P^{(j)}\left(\widehat{\mathbf{S}}_n\right) + \left[f_P^{(j)}(\mathbf{S}_n) - f_P^{(j)}\left(\widehat{\mathbf{S}}_n\right)\right] + \left[f^{(j)}(\mathbf{S}_n) - f_P^{(j)}(\mathbf{S}_n)\right]\\ &= \underset{1\times P}{p'_P\left(\widehat{\mathbf{S}}_n\right)}\underset{P\times 1}{\boldsymbol{\gamma}_{Pj}} + \eta_n^{(j)}, \qquad j = 1,\dots,K, n = 1,\dots,N.\end{aligned}$$

Note that the regressors $p'_P\left(\widehat{\mathbf{S}}_n\right)$ and the error $\eta_n^{(j)}$ are generally correlated. If we stack all the observations:

$$\underset{N\times 1}{\boldsymbol{\theta}^{(j)}} = \underset{N\times P}{p'_P\left(\widehat{\mathbf{S}}\right)}\boldsymbol{\gamma}_{Pj} + \boldsymbol{\eta}^{(j)}, \qquad j = 1,\dots,K.$$

Provided $N \geq P$, and neglecting the dependence between $p'_P\left(\widehat{\mathbf{S}}\right)$ and $\boldsymbol{\eta}^{(j)}$, this leads us to the OLS estimator:

$$\widehat{\boldsymbol{\gamma}}_{Pj} = \left(p_P\left(\widehat{\mathbf{S}}\right)\cdot p'_P\left(\widehat{\mathbf{S}}\right)\right)^{-1}\cdot p_P\left(\widehat{\mathbf{S}}\right)\cdot\boldsymbol{\theta}^{(j)}. \tag{5.5.3}$$

The forecast value of $\boldsymbol{\theta}$ is, for $j = 1,\dots,K$:

$$\widehat{\theta}^{(j)} = p'_P(\widehat{\mathbf{s}})\cdot\widehat{\boldsymbol{\gamma}}_{Pj}. \tag{5.5.4}$$

We define:

$$\boldsymbol{\Pi} := p_P(\mathbf{S})\cdot p'_P(\mathbf{S}).$$

A3 As $N \to \infty$ and any $P \ll N$, the matrix $\frac{1}{N}\boldsymbol{\Pi}$ converges to a given positive definite matrix $\boldsymbol{\Pi}_{0P}$, whose smallest eigenvalue is bounded away from zero.

A4 As $N \to \infty$ and any $P \ll N$, the matrix $\frac{1}{N}\cdot p_P(\mathbf{S})\cdot\boldsymbol{\theta}^{(j)}$ converges to a given vector $\boldsymbol{\pi}_{0P}$.

We define the approximation errors:

$$\begin{aligned}E_{NP}^{(1)} &:= \left\|N^{-1}\boldsymbol{\Pi} - \boldsymbol{\Pi}_{0P}\right\|,\\ E_{NP}^{(2)} &:= \left\|N^{-1}p_P(\mathbf{S})\cdot\boldsymbol{\theta}^{(j)} - \boldsymbol{\pi}_{0P}\right\|.\end{aligned}$$

Moreover, we set:

$$E_{NP} := E_{NP}^{(1)} + E_{NP}^{(2)}.$$

Note that A3 and A4 imply that $E_{NP}^{(1)}$ and $E_{NP}^{(2)}$ converge asymptotically to 0 as $N \to \infty$.

If we knew the values of $\mathbf{S}$, the choice of $\boldsymbol{\gamma}_j$ minimizing the squared loss would be:

$$\boldsymbol{\gamma}_{NPj} := \arg\min_{\boldsymbol{\gamma}_j} \frac{1}{N} \sum_{n=1}^{N} \left[\theta_n^{(j)} - p_P' \left(\mathbf{S}_n\right) \boldsymbol{\gamma}_j\right]^2$$

or:

$$\boldsymbol{\gamma}_{NPj} := \boldsymbol{\Pi}^{-1} \cdot p_P \left(\mathbf{S}\right) \cdot \boldsymbol{\theta}^{(j)}. \tag{5.5.5}$$

Assumptions A3 and A4 imply that this converges, when N diverges for fixed P, to:

$$\boldsymbol{\gamma}_{Pj} := \boldsymbol{\Pi}_{0P}^{-1} \cdot \boldsymbol{\pi}_{0P}. \tag{5.5.6}$$

We define the approximation error:

$$N_P := \sup_{\mathbf{s}} \left| f^{(j)} \left(\mathbf{s}\right) - p_P' \left(\mathbf{s}\right) \boldsymbol{\gamma}_{Pj} \right|.$$

The following are two technical assumptions that are required to simplify the formulas and to guarantee consistency.

A5 $\frac{N^{\frac{1}{2}} M}{R^{\frac{1}{2}}} \cdot \zeta_1 \left(P\right) = o\left(1\right)$.

A6 $N^{-1} \zeta_0 \left(P\right) = O\left(1\right)$.

The following assumptions concern the MSE and the asymptotic distribution of $\widehat{s}_1, \ldots, \widehat{s}_M$. They are both consistent with the fact that $\widehat{s}_k - s_k = O_{\mathbb{P}}\left(S^{-\frac{1}{2}}\right)$.

A7 $\mathbb{E}\left(\widehat{s}_k - s_k\right)^2 \leq \frac{c}{S}$ for $k = 1, \ldots, M$.

A8 $S^{\frac{1}{2}} \left(\widehat{\mathbf{s}} - \mathbf{s}\right) \to_{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$.

The statistics $\widehat{S}_{1n}, \ldots, \widehat{S}_{Mn}$ for $n = 1, \ldots, N$ are such that their MSE converge to 0 at rate R^{-1}. This is compatible with the fact that $\widehat{S}_{kn} - S_{kn} = O_{\mathbb{P}}\left(R^{-\frac{1}{2}}\right)$, but no convergence in distribution is explicitly required.

A9 $\mathbb{E}\left(\widehat{S}_{kn} - S_{kn}\right)^2 \leq \frac{c}{R}$ for $n = 1, \ldots, N, k = 1, \ldots, M$.

A10 $\mathbb{E}\left|\widehat{S}_{kn} - S_{kn}\right|^3 \leq \frac{c}{R^{\frac{3}{2}}}$ for $n = 1, \ldots, N, j = 1, \ldots, M$.

In the following, we analyze the approximation errors $E_{NP}^{(1)}$ and $E_{NP}^{(2)}$ under some appropriate assumptions.

As in [109, p. 714], we define the *design measure*, i.e. the discrete uniform distribution supported by the values $\{\boldsymbol{\theta}_n, n = 1, \ldots, N\}$:

$$\mathbb{P}_N \left(A\right) := N^{-1} \sum_{n=1}^{N} 1\left\{\boldsymbol{\theta}_n \in A\right\}$$

where A is a Borel set in $\mathbb{R}^K$. (This is not properly speaking an empirical distribution as the points are not random.) We suppose that $\mathbb{P}_N$ converges to an asymptotic design measure $\mathbb{P}$.

The first assumption requires that the parameter space can be reduced to a hypercube. The condition that parameters are variation free is not uncommon in econometrics (see, e.g., [147]).

A11 The parameters are variation free, i.e. the parameter space is given by the product of the parameter space of each single component of $\boldsymbol{\theta}$. Moreover, each component of the parameter vector is rescaled to the interval $[0,1]$.

The following assumption states that the limiting matrices in A3 and A4 can be written as expectations with respect to a probability measure $\mathbb{P}$.

A12 There is a probability measure $\mathbb{P}$ such that:

$$\begin{aligned}\left[\mathbf{\Pi}_{0P}\right]_{(i,j)} &= \int_{\mathbb{R}^M} p_{iP}\left(\mathbf{g}\left(\mathbf{x}\right)\right) p_{jP}\left(\mathbf{g}\left(\mathbf{x}\right)\right) \mathbb{P}\left(\mathrm{d}\mathbf{x}\right),\\ \left[\boldsymbol{\pi}_{0P}\right]_{(j)} &= \int_{\mathbb{R}^M} p_{iP}\left(\mathbf{g}\left(\mathbf{x}\right)\right) x_j \mathbb{P}\left(\mathrm{d}\mathbf{x}\right).\end{aligned}$$

The next assumption bounds the derivatives of the functions linking the parameters with the statistics.

A13 For any k and j, $\left|\frac{\partial g_k}{\partial \theta^{(j)}}\right| \leq \mu$ in a neighborhood of the true parameter value.

5.6 Results

In this section, we state our main results.

Theorem 5.1. *Under Assumptions A1-A5, A7 and A9, the rate of convergence is:*

$$\widehat{\theta}^{(j)} - \theta^{(j)} = O_{\mathbb{P}}\left(M \cdot \zeta_1\left(P\right) \cdot \left(\frac{1}{S^{\frac{1}{2}}} + \frac{\zeta_0\left(P\right)}{R^{\frac{1}{2}}}\right)\right) + O\left(N_P + \zeta_0\left(P\right) \cdot E_{NP}\right).$$

The first result covers the case in which the asymptotic theory is determined by $\widehat{\mathbf{s}}$. What is remarkable of this result is that, once it is guaranteed that the bias is negligible, the function f enters the result only through the variance. This implies that one can adjust the function f, e.g., varying the subset of Θ used for the simulations of the training set, without affecting the asymptotic distribution of the estimator, provided this does not increase the bias term.

Theorem 5.2. *Under Assumptions A1-A3, A5 and A7-A9, we have:*

$$S^{\frac{1}{2}}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) - \mathbf{A}_{NP} \to_{\mathcal{D}} \mathcal{N}\left(0, \mathbf{\Omega}\right)$$

where:

$$\left[\mathbf{\Omega}\right]_{ij} = \frac{\partial f^{(i)}\left(\mathbf{s}\right)}{\partial \mathbf{s}'} \cdot \mathbf{\Sigma} \cdot \frac{\partial f^{(j)}\left(\mathbf{s}\right)}{\partial \mathbf{s}}$$

and:

$$[\mathbf{A}_{NP}]_j = O\left(S^{\frac{1}{2}}N_P + S^{\frac{1}{2}}\zeta_0(P)\cdot E_{NP}\right) + O_{\mathbb{P}}\left(\frac{S^{\frac{1}{2}}M}{R^{\frac{1}{2}}}\cdot\zeta_0(P)\zeta_1(P) + \frac{M^2\zeta_2(P)}{S^{\frac{1}{2}}}\right).$$

Corollary 5.1. *Under Assumptions A1-A3, A5 and A7-A9, we have:*

$$S^{\frac{1}{2}}\left(\widehat{\theta}^{(j)} - \theta^{(j)}\right) - A_{NP}^{(j)} \to_{\mathcal{D}} \mathcal{N}\left(0, \Omega_P^{(j)}\right)$$

where:

$$\Omega_P^{(j)} = \frac{\partial f^{(j)}(\mathbf{s})}{\partial \mathbf{s}'}\cdot\mathbf{\Sigma}\cdot\frac{\partial f^{(j)}(\mathbf{s})}{\partial \mathbf{s}}$$

and:

$$A_{NP}^{(j)} = O\left(S^{\frac{1}{2}}N_P + S^{\frac{1}{2}}\zeta_0(P)\cdot E_{NP}\right) + O_{\mathbb{P}}\left(\frac{S^{\frac{1}{2}}M}{R^{\frac{1}{2}}}\cdot\zeta_0(P)\zeta_1(P) + \frac{M^2\zeta_2(P)}{S^{\frac{1}{2}}}\right).$$

The second result covers the case in which the asymptotic theory is determined by the simulations, and not by the real-world data. This case is less desirable, as the distribution is more complicated to characterize.

Theorem 5.3. *Let B_{NP} be a sequence such that $B_{NP} \to +\infty$ and $\frac{|\ln B_{NP}|}{PB_{NP}} \downarrow 0$. Under Assumptions A1-A3, A5-A7 and A9-A10, we have:*

$$\begin{aligned}\widehat{\theta}^{(j)} - \theta^{(j)} =& X_{NP} + o_{\mathbb{P}}\left(B_{NP}^{\frac{1}{3}}N^{\frac{1}{3}}MR^{-\frac{1}{2}}P^{\frac{1}{3}}\zeta_1(P)\cdot\left(N_P + \zeta_0(P)\cdot E_{NP} + N^{-1}\zeta_0^2(P)\right)\right)\\ &+ O_{\mathbb{P}}\left(\left(MR^{-1}\zeta_0(P)\zeta_1(P) + S^{-\frac{1}{2}}\right)\cdot M\zeta_1(P)\right)\\ &+ O\left(\left(1 + NR^{-\frac{1}{2}}M\zeta_1(P)\right)\cdot\left(N_P + \zeta_0(P)\cdot E_{NP}\right) + R^{-\frac{1}{2}}M\zeta_0^2(P)\zeta_1(P)\right)\end{aligned}$$

where:

$$\begin{aligned}X_{NP} \sim& \mathcal{N}\left(0, \sum_{n=1}^{N}\mathbf{c}_n'\cdot\frac{\partial p_P(\mathbf{S}_n)}{\partial \mathbf{s}'}\cdot\mathbb{V}\left(\widehat{\mathbf{S}}_n\right)\cdot\left(\frac{\partial p_P(\mathbf{S}_n)}{\partial \mathbf{s}'}\right)'\cdot\mathbf{c}_n\right),\\ \mathbf{c}_n =& \mathbf{e}_n'\left\{\mathbf{I}_N - p_P'(\mathbf{S})\mathbf{\Pi}^{-1}p_P(\mathbf{S})\right\}\boldsymbol{\theta}^{(j)}\cdot\mathbf{\Pi}^{-1}p_P(\mathbf{s})\\ &- \mathbf{e}_n'p_P'(\mathbf{S})\mathbf{\Pi}^{-1}p_P(\mathbf{s})\cdot\mathbf{\Pi}^{-1}p_P(\mathbf{S})\boldsymbol{\theta}^{(j)}\end{aligned}$$

and:

$$\mathbb{V}(X_{NP}) = O\left(\frac{M^2N}{R}\cdot\zeta_1^2(P)\right).$$

Remark 5.1. The previous result uses Yurisnkii's coupling. The reason to prefer this result to a more classical CLT is that it seems very difficult to find primitive conditions under which the variance of the asymptotic distribution converges to a constant or is bounded away from zero.

The following result characterizes the behavior of $E_{NP}^{(1)}$ and $E_{NP}^{(2)}$. It takes two distinct forms according to the type of points that are used. As in [363, p. 14], the term "sequence" is used to denote an infinite sequence, while the term point-set is used to denote a set of points of cardinality N.

Proposition 5.1. *Under A1-A2 and A11-A13, we have:*

$$\begin{aligned}
E_{NP}^{(1)} \leq & C\left(M,K\right)\cdot\left(\mu\vee\mu^{K}\right)\cdot D_{N,\mathbb{P}} \\
& \cdot\zeta_{K}\left(P\right)\cdot\left\{\left(\mu\vee\mu^{K}\right)\cdot\zeta_{K}\left(P\right)+\left(\sum_{i=1}^{P}\left\|p_{iP}\right\|_{\infty}^{2}\right)^{\frac{1}{2}}\right\} \\
E_{NP}^{(2)} \leq & C\left(M,K\right)\cdot D_{N,\mathbb{P}}\cdot\left\{\left(\mu\vee\mu^{K}\right)\cdot\zeta_{K}\left(P\right)+\left(\sum_{i=1}^{P}\left\|p_{iP}\right\|_{\infty}^{2}\right)^{\frac{1}{2}}\right\}
\end{aligned}$$

where:

$$D_{N,\mathbb{P}}:=\sup_{A\subseteq[0,1]^{K}}\left|\mathbb{P}_{N}\left(A\right)-\mathbb{P}\left(A\right)\right|$$

is the non-uniform unanchored discrepancy and A is any axis-parallel rectangle. For any N, it is possible to find a point-set such that $D_{N,\mathbb{P}}=O\left(\frac{(\ln N)^{d-1}}{N}\right)$ if $\mathbb{P}$ is the uniform measure on Θ and such that $D_{N,\mathbb{P}}=O\left(\frac{(\ln N)^{\frac{3d+1}{2}}}{N}\right)$ if $\mathbb{P}$ is not the uniform measure on Θ; there is a sequence such that $D_{N,\mathbb{P}}=O\left(\frac{(\ln N)^{d}}{N}\right)$ if $\mathbb{P}$ is the uniform measure on Θ and such that $D_{N,\mathbb{P}}=O\left(\frac{(\ln N)^{\frac{3d+4}{2}}}{N}\right)$ if $\mathbb{P}$ is not the uniform measure on Θ.

Remark 5.2. (i) Here $C\left(M,K\right)$ is a constant depending only on K and M that changes from place to place. However, following the proof it is possible to obtain a value for it.
(ii) For the case in which $\mathbb{P}$ is the uniform measure but the functions $f^{(j)}$ and $g^{(k)}$ are not smooth, one could use the results in [69].

5.7 Simulation Experiments

In this section, we report the results of some simulation experiments intended to verify the finite-sample properties of our estimator. We start providing two simple examples, concerning the estimation of the mean μ and the standard deviation σ of a Gaussian random variable. Although we are considering two trivial cases, these two cases are very informative and exemplify well the behavior of the estimator in different frameworks.[4] Indeed, μ represents a case in which the parameter to be estimated can be expressed as a finite linear combination of the basis functions used in the regression; this means that μ is a parametric function of the limiting values of the statistics. For σ, instead, the parameter can be represented as a linear combination of an infinite number of basis functions. The first example covers a parametric relation between parameters and statistics, while the second one involves a nonparametric relation.

The algorithm is composed by the following steps:

1. we select a grid of N points $\left\{\left(\mu_{n},\sigma_{n}\right),n=1,\ldots,N\right\}\subset\Theta=\left(-\infty,+\infty\right)\times\left(0,+\infty\right)$, that will be used to create the training dataset;

[4] We are considering to exploit our technique to estimate an agent-based computational model in which the evolutionary dynamics of the financial market are driven by agents with heterogeneous beliefs.

2. for each point in the grid, we simulate a sample of R independent random variables distributed as $\mathcal{N}\left(\mu_n, \sigma_n^2\right)$, and we compute the first four non-central moments on each of the N samples, i.e. $\left(\widehat{S}_{1n}, \widehat{S}_{2n}, \widehat{S}_{3n}, \widehat{S}_{4n}\right)$;

3. for each point in the grid, we compute the Hermite polynomials up to order p for the first and the third moments and the Laguerre polynomials up to order p for the second and the fourth moments: this choice is dictated by the fact that Hermite polynomials have support $(-\infty, +\infty)$ as odd moments have, while Laguerre polynomials have support $[0, +\infty)$ as even moments have;

4. for each point in the grid, we build the regressors as the tensor product of the previously computed polynomials;

5. we use as regressors the variables computed in item 4 and as dependent variable the corresponding values of μ_n (or σ_n), and we estimate the function $f_P^{(j)}\left(\widehat{S}_{1n}, \widehat{S}_{2n}, \widehat{S}_{3n}, \widehat{S}_{4n}\right)$ with a least-squares regression;

6. we create a test dataset: we repeat item 1 with a different selection of the parameters (see below) and item 2 with a different number of observations (replacing R with S); we get the statistics $(\widehat{s}_1, \widehat{s}_2, \widehat{s}_3, \widehat{s}_4)$ and we repeat items 3 and 4 on these statistics;

7. we use the regressors created in item 6 and the regression function estimated in item 5 to estimate the parameters.

The procedure is replicated a certain number of times using 16 different configurations.[5] The parametrization of the simulation experiment is reported in Table 5.7.1. For the training sample, the values of each one of the two parameters μ and σ are chosen according to an equispaced grid with range and cardinality detailed in the table. The final values of (μ, σ) form a two-dimensional grid in Θ, i.e. the parameter values in item 1 above are chosen according to a full factorial design. This is not necessarily the best choice, but it seems to be a rather simple one. Moreover, we reckon on an equispaced grid for two reasons: (i) it is a "worst-case" scenario since it is well known that other configurations of points of smaller cardinality provide equivalent approximations; (ii) we comply with the design matrix used in so-called computational experiments by the most popular ABM softwares (e.g., NetLogo).[6] For the test sample for μ, we have fixed $\sigma \equiv 1$ and we have taken μ on a grid, ranging in the interval $[-0.76, +0.76]$, of cardinality specified in the table. For the test sample for σ, we have taken $\mu \equiv 0$ and σ belonging to a grid, ranging in the interval $[0.6, 1.5]$, of cardinality reported in the table.

Let us see which variances we should expect from the method. We note that $s_1 = \mu$ and $s_2 = \sigma^2 + \mu^2$.

[5] For the estimator of the mean, we replicate the procedure about 20,000 times. For the estimator of the standard deviation, we replicate the procedure about 80,000 times. A larger number of replications is needed for the standard deviation than for the mean because the estimates appear to be more volatile in some regions of the parameter space.

[6] Many alternatives are possible to explore the parameter space. Among these we highlight Nearly Orthogonal Latin Hypercube sampling (see, e.g., [422]) and Quasi-Monte Carlo with sampling based on Sobol' sequences (see, e.g., [274] for a comparison of the two techniques).

Table 5.7.1: Parametrization of the simulation experiment

			Training sample				Test sample	Test sample
p	$R=S$	N	μ		σ		for μ	for σ
			card.	range	card.	range	card.	card.
2	10	980	49	$[-0.96, +0.96]$	20	$[0.1, 2]$	91	91
2	100	980	49	$[-0.96, +0.96]$	20	$[0.1, 2]$	91	91
2	1,000	980	49	$[-0.96, +0.96]$	20	$[0.1, 2]$	91	91
2	10,000	980	49	$[-0.96, +0.96]$	20	$[0.1, 2]$	91	91
3	10	980	49	$[-0.96, +0.96]$	20	$[0.1, 2]$	91	91
3	100	980	49	$[-0.96, +0.96]$	20	$[0.1, 2]$	91	91
3	1,000	980	49	$[-0.96, +0.96]$	20	$[0.1, 2]$	91	91
3	10,000	980	49	$[-0.96, +0.96]$	20	$[0.1, 2]$	91	91
4	10	980	49	$[-0.96, +0.96]$	20	$[0.1, 2]$	91	91
4	100	980	49	$[-0.96, +0.96]$	20	$[0.1, 2]$	91	91
4	1,000	980	49	$[-0.96, +0.96]$	20	$[0.1, 2]$	91	91
4	10,000	980	49	$[-0.96, +0.96]$	20	$[0.1, 2]$	91	91
5	10	980	49	$[-0.96, +0.96]$	20	$[0.1, 2]$	91	91
5	100	980	49	$[-0.96, +0.96]$	20	$[0.1, 2]$	91	91
5	1,000	980	49	$[-0.96, +0.96]$	20	$[0.1, 2]$	91	91
5	10,000	980	49	$[-0.96, +0.96]$	20	$[0.1, 2]$	91	91

Note: “card.” stands for “cardinality”.

For the estimator of μ, we have $\mu = s_1$, i.e. f is the identity. Therefore, $\lim_{S\to\infty} \mathbb{V}\left(S^{\frac{1}{2}}\left(\widehat{\mu}-\mu\right)\right) = \lim_{S\to\infty} \mathbb{V}\left(S^{\frac{1}{2}}\widehat{s}_1\right) = \sigma^2$. This means that we can compare the variance $\mathbb{V}\left(\widehat{\mu}\right)$ with $\frac{\sigma^2}{S}$. By the way, this is the Cramér-Rao lower bound (CRLB) for the estimation of the mean, therefore the comparison provides a measure of efficiency of the procedure.

For the estimator of σ, $\sigma = \sqrt{s_2 - s_1^2}$. Therefore:

$$\begin{aligned}
&\lim_{S\to\infty} \mathbb{V}\left(S^{\frac{1}{2}}\left(\widehat{\sigma}-\sigma\right)\right)\\
&= \begin{bmatrix} \frac{\partial f}{\partial s_1} & \frac{\partial f}{\partial s_2} \end{bmatrix} \lim_{S\to\infty} \begin{bmatrix} \mathbb{V}\left(S^{\frac{1}{2}}\widehat{s}_1\right) & \mathrm{Cov}\left(S^{\frac{1}{2}}\widehat{s}_1, S^{\frac{1}{2}}\widehat{s}_2\right) \\ \mathrm{Cov}\left(S^{\frac{1}{2}}\widehat{s}_1, S^{\frac{1}{2}}\widehat{s}_2\right) & \mathbb{V}\left(S^{\frac{1}{2}}\widehat{s}_2\right) \end{bmatrix} \begin{bmatrix} \frac{\partial f}{\partial s_1} \\ \frac{\partial f}{\partial s_2} \end{bmatrix}\\
&= \frac{1}{s_2 - s_1^2} \begin{bmatrix} -s_1 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \sigma^2 & 2\mu\sigma^2 \\ 2\mu\sigma^2 & 2\sigma^2\left(2\mu^2+\sigma^2\right) \end{bmatrix} \begin{bmatrix} -s_1 \\ \frac{1}{2} \end{bmatrix} = \frac{\sigma^2}{2}.
\end{aligned}$$

Therefore, we can compare the variance $\mathbb{V}\left(\widehat{\sigma}\right)$ with $\frac{\sigma^2}{2S}$. By the way, this is the CRLB for the estimation of the standard deviation.

5.7.1 Estimation of the mean of a Gaussian random variable

The bias and the variance of the estimator of μ are represented in Figures 5.7.1 and 5.7.2 respectively. In the bias plots, the grey horizontal line represents the value 0. In the variance plots, the CRLB for each value of S is represented by the thin horizontal line with the same style as the corresponding variance of the estimator. Note that in the plots we use the fact that the bias is an odd function

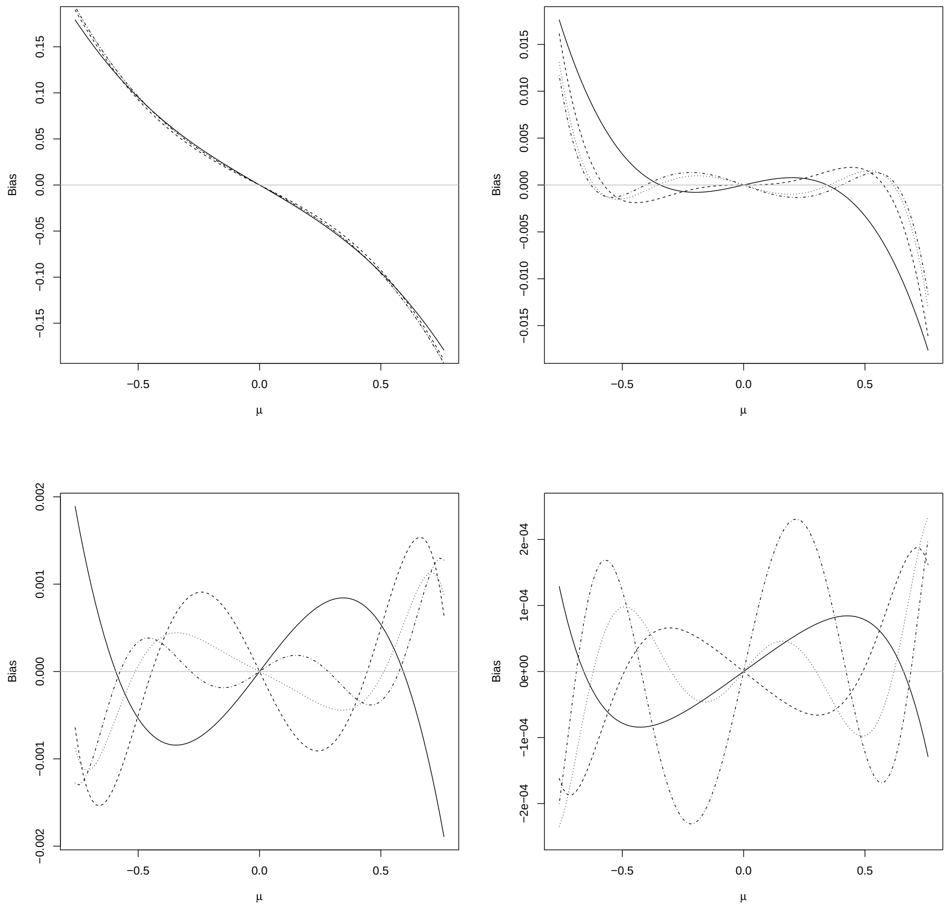

Figure 5.7.1: Bias of the estimator of μ with $R = 10$ (top left), $R = 100$ (top right), $R = 1,000$ (bottom left), $R = 10,000$ (bottom right) and $N = 980$, for different configurations of parameters: $p = 2$ (solid line), $p = 3$ (dashed line), $p = 4$ (dotted line) and $p = 5$ (dash-dot line).

and the variance an even function of μ.

The bias of the estimator of μ is negligible and tends to decrease when the number of simulations, $S = R$, increases (see Figure 5.7.1). The prominent reason for this phenomenon seems to be the fact that, while the statistics themselves are unbiased, the polynomials are nonlinear functions of the statistics and this introduces a bias that is stronger when R is small. It is worth noting that if the number of simulated observations is much greater than the real-world observations, i.e. $R \gg S$, the bias term will disappear almost surely. In our simulation experiment we consider a “stress-test” scenario in which $S = R$. In this situation, the behavior of the estimator could be tricky, and the bias may not be negligible. A solution to this drawback could be to consider a debiasing procedure (see, e.g., [91, 481]).

Figure 5.7.2 shows that the variance of the parameter behaves well, as it approaches the CRLB. Therefore, the estimation method seems to be efficient. This is due to the fact that the parameter to

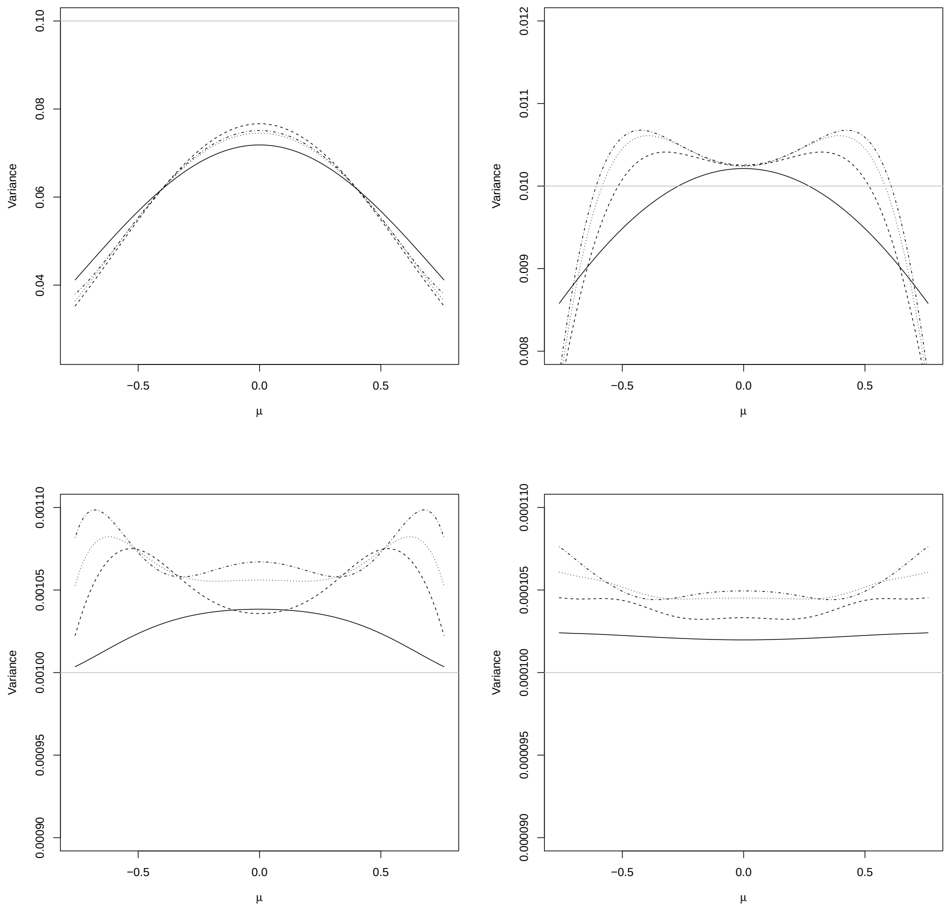

Figure 5.7.2: Variance of the estimator of μ with $R = 10$ (top left), $R = 100$ (top right), $R = 1,000$ (bottom left), $R = 10,000$ (bottom right) and $N = 980$, for different configurations of parameters: $p = 2$ (solid line), $p = 3$ (dashed line), $p = 4$ (dotted line) and $p = 5$ (dash-dot line).

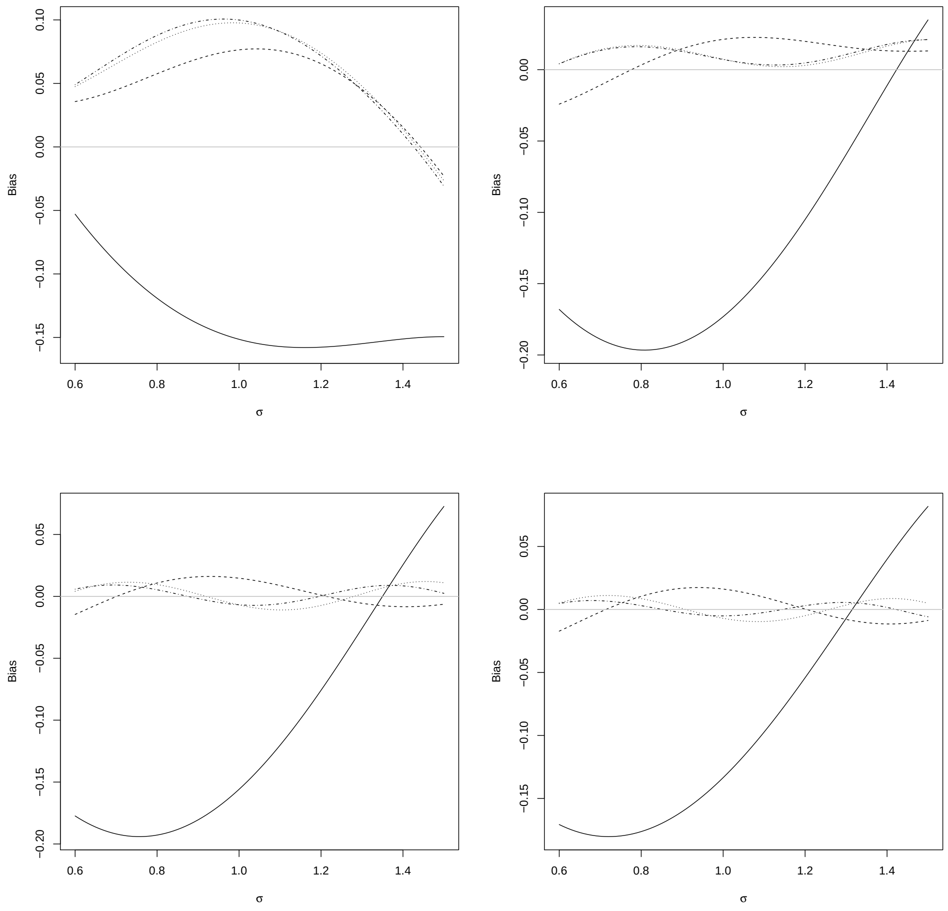

Figure 5.7.3: Bias of the estimator of σ with $R = 10$ (top left), $R = 100$ (top right), $R = 1,000$ (bottom left), $R = 10,000$ (bottom right) and $N = 980$, for different configurations of parameters: $p = 2$ (solid line), $p = 3$ (dashed line), $p = 4$ (dotted line) and $p = 5$ (dash-dot line).

be estimated is a function of the limiting value of the statistics. Hence, we can estimate $f_P^{(j)}$ so fast that the estimation of the function does not influence the asymptotic distribution of the estimator itself.

These results are not surprising, but rather they confirm our expectations as μ is a parametric function of the limiting value s_1.

5.7.2 Estimation of the standard deviation of a Gaussian random variable

The behavior of the bias and the variance of the estimator of σ are reported in Figures 5.7.3 and 5.7.4 respectively. The grey horizontal line in the bias plots represents the value 0. In the variance plots, the CRLB for each value of S is represented by the thin grey line.

When dealing with a nonparametric function of the statistics, as in the case of σ, the bias is not

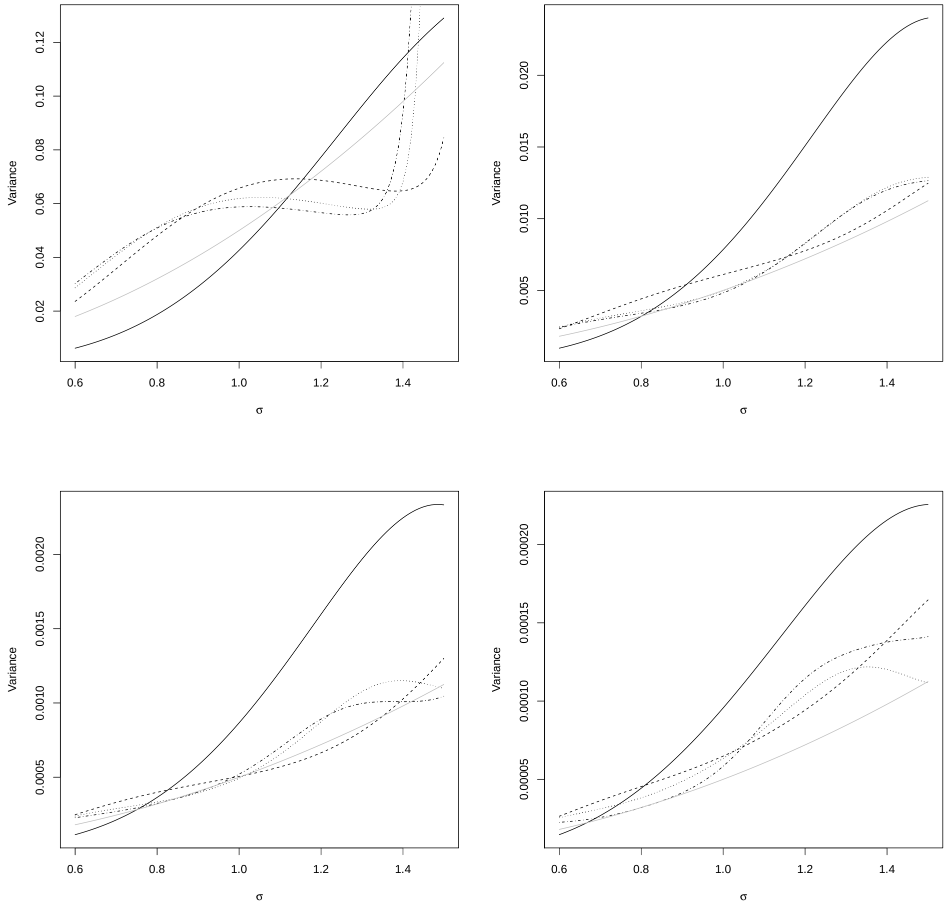

Figure 5.7.4: Variance of the estimator of σ with $R = 10$ (top left), $R = 100$ (top right), $R = 1,000$ (bottom left), $R = 10,000$ (bottom right) and $N = 980$, for different configurations of parameters: $p = 2$ (solid line), $p = 3$ (dashed line), $p = 4$ (dotted line) and $p = 5$ (dash-dot line).

negligible but decreases as the order of polynomials, p, increases. The shape of the bias as a function of the parameter values is not directly interpretable (see Figure 5.7.3), as it is due to the interaction of two separate sources, the error of approximation arising when the nonparametric function f is approximated by a polynomial and the bias induced by the nonlinear transformation of the unbiased statistics.

In this case too, the variance of the estimator of σ is still aligned with the CRLB (see Figure 5.7.4), and we can conclude that the estimator seems to be efficient.

The same observations concerning the behavior of the estimator of μ in terms of bias and efficiency hold also in this case.

5.8 Conclusions

In this work, we develop a nonparametric technique to estimate the parameters of simulation models, without involving any optimization algorithm. The parameters of the simulated models are estimated combining OLS regression and sieve estimation. The nonparametric element allows to detect the nonlinear and unknown relations linking the statistics used as explanatory variables in the regression and the parameters to be estimated. Since we do not rely on optimization, our framework overcomes some issues generally met when estimating simulation models via classical simulation-based econometric methods (i.e., roughness of the objective function and lack of stochastic equicontinuity).

The full asymptotic theory of the estimator $\widehat{\theta}^{(j)}$ is explicitly and rigorously characterized, including the order of the bias, confidence intervals and hypotheses tests. As we consider a nonparametric setup, we obtain a sub-optimal asymptotic rate of convergence of $\widehat{\theta}^{(j)}$ to $\theta^{(j)}$. Nevertheless, we compensate this loss of efficiency with the flexibility deriving from the nonparametric estimation of the function linking the statistics and the parameters.

At last, the simulation study shows that the estimator behaves differently when the parameter to be estimated is a parametric function of the limiting values of the statistics, rather than when we deal with a nonparametric relation (e.g., a parameter that can be represented as a linear combination of an infinite number of basis functions). In the first case, the bias is negligible and tends to decrease when the number of observations, S, and the number of simulations, R, increase. In the second case, although the bias is not negligible and its shape as a function of the parameters values is not directly interpretable, it seems to decrease as the order of polynomials increases. In both cases the variance of the estimator approaches the Cramér-Rao lower bound, hence the estimator seems to be efficient.

5.9 Appendix

We define the matrices:

$$\mathbf{\Pi} = \mathbf{\Pi}_{NP} := p_P\left(\mathbf{S}\right) \cdot p'_P\left(\mathbf{S}\right),$$
$$\widehat{\mathbf{\Pi}} = \widehat{\mathbf{\Pi}}_{NP} := p_P\left(\widehat{\mathbf{S}}\right) \cdot p'_P\left(\widehat{\mathbf{S}}\right).$$

We note that under Assumption A3, for fixed P:

$$\lim_{N\to\infty}\frac{1}{N}\mathbf{\Pi}=\mathbf{\Pi}_{0P}.$$

As in [354] and [40] among others, we assume without loss of generality that the matrix $\mathbf{\Pi}_{0P}$ is the identity matrix of dimension P, $\mathbf{I}_P$. This implies that both $\lambda_{\min}\left(\frac{1}{N}\mathbf{\Pi}\right)$ and $\lambda_{\max}\left(\frac{1}{N}\mathbf{\Pi}\right)$ converge to 1.

We start collecting some results that we will use repeatedly in the following.

Lemma 5.1. *We have:*

- *without any assumptions:*

$$\begin{aligned}\left\|\mathbf{I}_N-p_P'\left(\mathbf{S}\right)\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\right\|=&1,\\ \left\|p_P\left(\mathbf{s}\right)\right\|\leq&\zeta_0\left(P\right),\\ \left\|\frac{\partial p_P\left(\mathbf{s}\right)}{\partial\mathbf{s}'}\right\|\leq&M^{\frac{1}{2}}\zeta_1\left(P\right);\end{aligned}$$

- *under A1,* $\left\|\boldsymbol{\theta}^{(j)}\right\|=O\left(N^{\frac{1}{2}}\right)$*;*
- *under A3:*

$$\begin{aligned}\left\|\mathbf{\Pi}^{-1}\right\|\simeq&N^{-1},\\ \left\|p_P\left(\mathbf{S}\right)\right\|\simeq&N^{\frac{1}{2}};\end{aligned}$$

- *under A9,* $\left\|p_P\left(\widehat{\mathbf{S}}\right)-p_P\left(\mathbf{S}\right)\right\|=O_{\mathbb{P}}\left(R^{-\frac{1}{2}}N^{\frac{1}{2}}M\zeta_1\left(P\right)\right)$*;*
- *under A7:*

$$\begin{aligned}\left\|\widehat{\mathbf{s}}-\mathbf{s}\right\|=O_{\mathbb{P}}\left(S^{-\frac{1}{2}}M^{\frac{1}{2}}\right),\\ \left\|p_P\left(\widehat{\mathbf{s}}\right)-p_P\left(\mathbf{s}\right)\right\|=O_{\mathbb{P}}\left(S^{-\frac{1}{2}}M\zeta_1\left(P\right)\right);\end{aligned}$$

- *under A1 and A3:*

$$\begin{aligned}\left\|\boldsymbol{\gamma}_{NPj}\right\|=&O\left(1\right),\\ \left\|\boldsymbol{\gamma}_{Pj}\right\|=&O\left(1+E_{NP}^{(2)}\right),\\ \left\|\boldsymbol{\gamma}_{NPj}-\boldsymbol{\gamma}_{Pj}\right\|\lesssim&E_{NP}.\end{aligned}$$

Proof. Under A1, $\left\|\boldsymbol{\theta}^{(j)}\right\| = \left(\sum_{n=1}^{N} \theta_n^{(j)}\right)^{\frac{1}{2}}$ is $O\left(N^{\frac{1}{2}}\right)$. Under A3, we have:

$$\begin{aligned}\left\|\boldsymbol{\Pi}^{-1}\right\| =& \lambda_{\max}\left(\boldsymbol{\Pi}^{-1}\right) = \lambda_{\min}^{-1}\left(\boldsymbol{\Pi}\right) = N^{-1}\lambda_{\min}^{-1}\left(\frac{1}{N}\boldsymbol{\Pi}\right)\\ \simeq& N^{-1}\lambda_{\min}^{-1}\left(\boldsymbol{\Pi}_{0P}\right) = N^{-1}.\end{aligned}$$

For the same reason, $\left\|p_P\left(\mathbf{S}\right)\right\| = \lambda_{\max}^{\frac{1}{2}}\left(\boldsymbol{\Pi}\right) \simeq N^{\frac{1}{2}}$. We also have $\left\|\mathbf{I}_N - p_P'\left(\mathbf{S}\right)\cdot\boldsymbol{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\right\| = 1$, as the matrix is idempotent.

The definition of ζ_s implies that $\left\|p_P\left(\mathbf{s}\right)\right\| \le \zeta_0\left(P\right)$ and that:

$$\left\|\frac{\partial p_P\left(\mathbf{s}\right)}{\partial \mathbf{s}'}\right\|^2 \le \left\|\frac{\partial p_P\left(\mathbf{s}\right)}{\partial \mathbf{s}'}\right\|_F^2 = \sum_{j=1}^{P}\sum_{k=1}^{M}\left(\frac{\partial p_{jP}\left(\mathbf{s}\right)}{\partial s_k}\right)^2 \le M\zeta_1^2\left(P\right).$$

We define:

$$\mathrm{d}\mathbf{P} := p_P\left(\widehat{\mathbf{S}}\right) - p_P\left(\mathbf{S}\right). \tag{5.9.1}$$

Now:

$$\begin{aligned}\left\|\mathrm{d}\mathbf{P}\right\|^2 \le& \left\|\mathrm{d}\mathbf{P}\right\|_F^2\\ =& \sum_{n=1}^{N}\sum_{j=1}^{P}\left\{p_{jP}\left(\widehat{\mathbf{S}}_n\right) - p_{jP}\left(\mathbf{S}_n\right)\right\}^2\\ \simeq& \sum_{n=1}^{N}\sum_{j=1}^{P}\left\{\sum_{k=1}^{M}\frac{\partial p_{jP}\left(\mathbf{S}_n\right)}{\partial S_{kn}}\cdot\left(\widehat{S}_{kn} - S_{kn}\right)\right\}^2\\ \le& M\cdot\sum_{n=1}^{N}\sum_{k=1}^{M}\left(\widehat{S}_{kn} - S_{kn}\right)^2\cdot\sum_{j=1}^{P}\left(\frac{\partial p_{jP}\left(\mathbf{S}_n\right)}{\partial S_{kn}}\right)^2\\ \le& M\cdot\zeta_1^2\left(P\right)\cdot\sum_{n=1}^{N}\sum_{k=1}^{M}\left(\widehat{S}_{kn} - S_{kn}\right)^2\end{aligned} \tag{5.9.2}$$

where we have used $\left(\sum_{i=1}^{n} x_i\right)^2 \le n\sum_{i=1}^{n} x_i^2$ and the definition of $\zeta_1^2\left(P\right)$. Then, from A9:

$$\mathbb{E}\left\|\mathrm{d}\mathbf{P}\right\|^2 \lesssim M\cdot\zeta_1^2\left(P\right)\cdot\sum_{n=1}^{N}\sum_{k=1}^{M}\left(\widehat{S}_{kn} - S_{kn}\right)^2 \le c\cdot\frac{NM^2}{R}\cdot\zeta_1^2\left(P\right)$$

and:

$$\left\|\mathrm{d}\mathbf{P}\right\| = O_{\mathbb{P}}\left(\frac{N^{\frac{1}{2}}M}{R^{\frac{1}{2}}}\cdot\zeta_1\left(P\right)\right).$$

Under A7, we have:

$$\mathbb{E}\left\|\widehat{\mathbf{s}} - \mathbf{s}\right\|^2 = \mathbb{E}\sum_{k=1}^{M}\left(\widehat{s}_k - s_k\right)^2 \le \frac{cM}{S}$$

from which:

$$\|\widehat{\mathbf{s}}-\mathbf{s}\| = O_{\mathbb{P}}\left(S^{-\frac{1}{2}}M^{\frac{1}{2}}\right).$$

The result for $\|p_P(\widehat{\mathbf{s}})-p_P(\mathbf{s})\|$ comes from:

$$\begin{aligned}\|p_P(\widehat{\mathbf{s}})-p_P(\mathbf{s})\| \leq & \|\widehat{\mathbf{s}}-\mathbf{s}\| \cdot \left\|\frac{\partial p_P(\mathbf{s}^\star)}{\partial \mathbf{s}'}\right\| \\ \leq & O_{\mathbb{P}}\left(S^{-\frac{1}{2}}M\zeta_1(P)\right)\end{aligned}$$

where $\mathbf{s}^\star$ lies between $\widehat{\mathbf{s}}$ and $\mathbf{s}$.

Under A1 and A3:

$$\begin{aligned}\left\|\boldsymbol{\gamma}_{NPj}\right\| &\leq \left\|\boldsymbol{\Pi}^{-1}\right\| \cdot \|p_P(\mathbf{S})\| \cdot \left\|\boldsymbol{\theta}^{(j)}\right\| = O(1), \\ \left\|\boldsymbol{\gamma}_{Pj}\right\| &\leq \left\|\boldsymbol{\Pi}_{0P}^{-1}\right\| \cdot \|\boldsymbol{\pi}_{0P}\| = \|\boldsymbol{\pi}_{0P}\| \\ &\leq \left\|\boldsymbol{\pi}_{0P} - N^{-1}p_P(\mathbf{S}) \cdot \boldsymbol{\theta}^{(j)}\right\| + \left\|N^{-1}p_P(\mathbf{S}) \cdot \boldsymbol{\theta}^{(j)}\right\| \\ &\leq E_{NP}^{(2)} + N^{-1}\|p_P(\mathbf{S})\| \cdot \left\|\boldsymbol{\theta}^{(j)}\right\| \simeq E_{NP}^{(2)} + 1.\end{aligned}$$

Moreover:

$$\begin{aligned}&\boldsymbol{\gamma}_{NPj} - \boldsymbol{\gamma}_{Pj} \\ &\quad= \left(\frac{1}{N}\boldsymbol{\Pi}\right)^{-1} \cdot \frac{1}{N}p_P(\mathbf{S}) \cdot \boldsymbol{\theta}^{(j)} - \boldsymbol{\Pi}_{0P}^{-1} \cdot \boldsymbol{\pi}_{0P} \\ &\quad= \left(\frac{1}{N}\boldsymbol{\Pi}\right)^{-1} \cdot \left[\boldsymbol{\Pi}_{0P} - N^{-1}\boldsymbol{\Pi}\right] \cdot \boldsymbol{\Pi}_{0P}^{-1} \cdot \frac{1}{N}p_P(\mathbf{S}) \cdot \boldsymbol{\theta}^{(j)} \\ &\qquad+ \boldsymbol{\Pi}_{0P}^{-1} \cdot \left[\frac{1}{N}p_P(\mathbf{S}) \cdot \boldsymbol{\theta}^{(j)} - \boldsymbol{\pi}_{0P}\right]\end{aligned}$$

and:

$$\begin{aligned}&\left\|\boldsymbol{\gamma}_{NPj} - \boldsymbol{\gamma}_{Pj}\right\| \\ &\quad\leq \left\|\left(N^{-1}\boldsymbol{\Pi}\right)^{-1}\right\| \cdot \left\|\boldsymbol{\Pi}_{0P} - N^{-1}\boldsymbol{\Pi}\right\| \cdot \left\|\boldsymbol{\Pi}_{0P}^{-1}\right\| \cdot \left\|N^{-1}p_P(\mathbf{S})\right\| \cdot \left\|\boldsymbol{\theta}^{(j)}\right\| \\ &\qquad+ \left\|\boldsymbol{\Pi}_{0P}^{-1}\right\| \cdot \left\|\frac{1}{N}p_P(\mathbf{S}) \cdot \boldsymbol{\theta}^{(j)} - \boldsymbol{\pi}_{0P}\right\| \\ &\quad\lesssim E_{NP}^{(1)} + E_{NP}^{(2)} = E_{NP}.\end{aligned}$$

QED

Lemma 5.2. *Under Assumptions A1-A3, A5 and A9, we have:*

$$\begin{aligned}\widehat{\theta}^{(j)} - \theta^{(j)} =& \left(p_P\left(\widehat{\mathbf{s}}\right) - p_P\left(\mathbf{s}\right)\right)' \boldsymbol{\gamma}_{Pj} \\ & - p_P'\left(\widehat{\mathbf{s}}\right) \cdot \mathbf{\Pi}^{-1} \cdot p_P\left(\mathbf{S}\right) \cdot \left(p_P\left(\widehat{\mathbf{S}}\right) - p_P\left(\mathbf{S}\right)\right)' \cdot \mathbf{\Pi}^{-1} \cdot p_P\left(\mathbf{S}\right) \cdot \boldsymbol{\theta}^{(j)} \\ & + p_P'\left(\widehat{\mathbf{s}}\right) \cdot \mathbf{\Pi}^{-1} \cdot \left(p_P\left(\widehat{\mathbf{S}}\right) - p_P\left(\mathbf{S}\right)\right) \cdot \left[\mathbf{I}_P - p_P'\left(\mathbf{S}\right) \cdot \mathbf{\Pi}^{-1} \cdot p_P\left(\mathbf{S}\right)\right] \cdot \boldsymbol{\theta}^{(j)} \\ & + O\left(N_P + \zeta_0\left(P\right) \cdot E_{NP}\right) + O_{\mathbb{P}}\left(\frac{M^2}{R} \cdot \zeta_0\left(P\right) \zeta_1^2\left(P\right)\right).\end{aligned}$$

Proof. The matrix $\widehat{\mathbf{\Pi}}^{-1}$ can be written as:

$$\begin{aligned}\widehat{\mathbf{\Pi}}^{-1} =& \left(\left(p_P\left(\mathbf{S}\right) + \mathrm{d}\mathbf{P}\right) \cdot \left(p_P'\left(\mathbf{S}\right) + \mathrm{d}\mathbf{P}'\right)\right)^{-1} \\ =& \frac{1}{N}\left(\frac{1}{N}\mathbf{\Pi} + \frac{1}{N}\left[p_P\left(\mathbf{S}\right) \cdot \mathrm{d}\mathbf{P}' + \mathrm{d}\mathbf{P} \cdot p_P'\left(\mathbf{S}\right) + \mathrm{d}\mathbf{P} \cdot \mathrm{d}\mathbf{P}'\right]\right)^{-1}\end{aligned}$$

where d$\mathbf{P}$ is defined in (5.9.1). Now, through Lemma 5.1, under A3 $\left\|\frac{1}{N}\mathbf{\Pi}\right\| \simeq 1$ and under A3 and A9:

$$\begin{aligned}&\left\|\frac{1}{N}\left[p_P\left(\mathbf{S}\right) \cdot \mathrm{d}\mathbf{P}' + \mathrm{d}\mathbf{P} \cdot p_P'\left(\mathbf{S}\right) + \mathrm{d}\mathbf{P} \cdot \mathrm{d}\mathbf{P}'\right]\right\| \\ &\quad \leq N^{-1} \cdot \left(2\left\|\mathrm{d}\mathbf{P}\right\| \cdot \left\|p_P\left(\mathbf{S}\right)\right\| + \left\|\mathrm{d}\mathbf{P}\right\|^2\right) \\ &\quad \leq O_{\mathbb{P}}\left(\frac{M}{R^{\frac{1}{2}}} \cdot \zeta_1\left(P\right) + \frac{M^2}{R} \cdot \zeta_1^2\left(P\right)\right) = o_{\mathbb{P}}\left(1\right)\end{aligned}$$

where the last step comes from A5.

From [321, p. 168], we then have:

$$\begin{aligned}\widehat{\mathbf{\Pi}}^{-1} =& \left(\left(p_P\left(\mathbf{S}\right) + \mathrm{d}\mathbf{P}\right) \cdot \left(p_P'\left(\mathbf{S}\right) + \mathrm{d}\mathbf{P}'\right)\right)^{-1} \\ =& \frac{1}{N}\left(\frac{1}{N}\mathbf{\Pi} + \frac{1}{N}\left[p_P\left(\mathbf{S}\right) \cdot \mathrm{d}\mathbf{P}' + \mathrm{d}\mathbf{P} \cdot p_P'\left(\mathbf{S}\right) + \mathrm{d}\mathbf{P} \cdot \mathrm{d}\mathbf{P}'\right]\right)^{-1} \\ =& \mathbf{\Pi}^{-1} - \mathbf{\Pi}^{-1}\left(p_P\left(\mathbf{S}\right) \cdot \mathrm{d}\mathbf{P}' + \mathrm{d}\mathbf{P} \cdot p_P'\left(\mathbf{S}\right) + \mathrm{d}\mathbf{P} \cdot \mathrm{d}\mathbf{P}'\right)\mathbf{\Pi}^{-1} + \mathbf{R}_1. \end{aligned} \tag{5.9.3}$$

Here, from Lemma 5.1, under A3 $\left\|\mathbf{\Pi}^{-1}\right\| \simeq N^{-1}$ and:

$$\begin{aligned}\left\|\mathbf{\Pi}^{-1} \cdot p_P\left(\mathbf{S}\right) \cdot \mathrm{d}\mathbf{P}' \cdot \mathbf{\Pi}^{-1}\right\| \leq& \left\|\mathbf{\Pi}^{-1}\right\|^2 \cdot \left\|p_P\left(\mathbf{S}\right)\right\| \cdot \left\|\mathrm{d}\mathbf{P}\right\| \\ =& O_{\mathbb{P}}\left(\frac{M}{NR^{\frac{1}{2}}} \cdot \zeta_1\left(P\right)\right) = o_{\mathbb{P}}\left(\left\|\mathbf{\Pi}^{-1}\right\|\right) \\ \left\|\mathbf{\Pi}^{-1} \cdot \mathrm{d}\mathbf{P} \cdot \mathrm{d}\mathbf{P}' \cdot \mathbf{\Pi}^{-1}\right\| \leq& \left\|\mathbf{\Pi}^{-1}\right\|^2 \cdot \left\|\mathrm{d}\mathbf{P}\right\|^2 \\ =& O_{\mathbb{P}}\left(\frac{M^2}{NR} \cdot \zeta_1^2\left(P\right)\right) = o_{\mathbb{P}}\left(\left\|\mathbf{\Pi}^{-1} \cdot p_P\left(\mathbf{S}\right) \cdot \mathrm{d}\mathbf{P}' \cdot \mathbf{\Pi}^{-1}\right\|\right)\end{aligned}$$

where we have used A3, A5 and A9. Moreover (see [321, p. 169]), using A3, A5 and A9:

$$\begin{aligned}\mathbf{R}_1 &\simeq \left\{\mathbf{\Pi}^{-1}\left[p_P\left(\mathbf{S}\right)\cdot \mathrm{d}\mathbf{P'} + \mathrm{d}\mathbf{P}\cdot p_P'\left(\mathbf{S}\right) + \mathrm{d}\mathbf{P}\cdot \mathrm{d}\mathbf{P'}\right]\right\}^2 \mathbf{\Pi}^{-1},\\ \left\|\mathbf{R}_1\right\| =& O\left(\left(\left\|p_P\left(\mathbf{S}\right)\right\|^2\cdot\left\|\mathrm{d}\mathbf{P}\right\|^2 + \left\|\mathrm{d}\mathbf{P}\right\|^4\right)\cdot\left\|\mathbf{\Pi}^{-1}\right\|^3\right)\\ =& O_{\mathbb{P}}\left(\frac{M^2}{NR}\cdot\zeta_1^2\left(P\right)\cdot\left(1+\frac{M^2}{R}\cdot\zeta_1^2\left(P\right)\right)\right) = o_{\mathbb{P}}\left(\left\|\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\cdot\mathrm{d}\mathbf{P'}\cdot\mathbf{\Pi}^{-1}\right\|\right).\end{aligned}$$

Therefore, we can neglect the term $\mathbf{R}_2 := -\mathbf{\Pi}^{-1}\cdot\mathrm{d}\mathbf{P}\cdot\mathrm{d}\mathbf{P'}\cdot\mathbf{\Pi}^{-1}+\mathbf{R}_1$ where $\left\|\mathbf{R}_2\right\| = O_{\mathbb{P}}\left(\frac{M^2}{NR}\cdot\zeta_1^2\left(P\right)\right)$.

Using (5.5.3), (5.9.1) and (5.9.3), the coefficient $\widehat{\boldsymbol{\beta}}_j$ can be written as:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_j =& \widehat{\mathbf{\Pi}}^{-1}\cdot p_P\left(\widehat{\mathbf{S}}\right)\cdot\boldsymbol{\theta}^{(j)} = \widehat{\mathbf{\Pi}}^{-1}\cdot\left(p_P\left(\mathbf{S}\right)+\mathrm{d}\mathbf{P}\right)\cdot\boldsymbol{\theta}^{(j)}\\ =& \mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\cdot\boldsymbol{\theta}^{(j)} + \mathbf{\Pi}^{-1}\cdot\mathrm{d}\mathbf{P}\cdot\boldsymbol{\theta}^{(j)}\\ &-\mathbf{\Pi}^{-1}\cdot\left(p_P\left(\mathbf{S}\right)\cdot\mathrm{d}\mathbf{P'}+\mathrm{d}\mathbf{P}\cdot p_P'\left(\mathbf{S}\right)\right)\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\cdot\boldsymbol{\theta}^{(j)}\\ &-\mathbf{\Pi}^{-1}\cdot\left(p_P\left(\mathbf{S}\right)\cdot\mathrm{d}\mathbf{P'}+\mathrm{d}\mathbf{P}\cdot p_P'\left(\mathbf{S}\right)\right)\cdot\mathbf{\Pi}^{-1}\cdot\mathrm{d}\mathbf{P}\cdot\boldsymbol{\theta}^{(j)}\\ &+\mathbf{R}_2\cdot\left(p_P\left(\mathbf{S}\right)+\mathrm{d}\mathbf{P}\right)\cdot\boldsymbol{\theta}^{(j)}\\ =& \mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\cdot\boldsymbol{\theta}^{(j)}\\ &-\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\cdot\mathrm{d}\mathbf{P'}\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\cdot\boldsymbol{\theta}^{(j)}\\ &+\mathbf{\Pi}^{-1}\cdot\mathrm{d}\mathbf{P}\cdot\left[\mathbf{I}_P - p_P'\left(\mathbf{S}\right)\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\right]\cdot\boldsymbol{\theta}^{(j)}\\ &+\mathbf{R}_3\end{aligned}$$

where

$$\mathbf{R}_3 := -\mathbf{\Pi}^{-1}\cdot\left(p_P\left(\mathbf{S}\right)\cdot\mathrm{d}\mathbf{P'}+\mathrm{d}\mathbf{P}\cdot p_P'\left(\mathbf{S}\right)\right)\cdot\mathbf{\Pi}^{-1}\cdot\mathrm{d}\mathbf{P}\cdot\boldsymbol{\theta}^{(j)}+\mathbf{R}_2\cdot\left(p_P\left(\mathbf{S}\right)+\mathrm{d}\mathbf{P}\right)\cdot\boldsymbol{\theta}^{(j)}.$$

Now, under A1, A3, A5 and A9:

$$\begin{aligned}\left\|\mathbf{R}_3\right\| \leq& 2\left\|\mathbf{\Pi}^{-1}\right\|^2\cdot\left\|\mathrm{d}\mathbf{P}\right\|^2\cdot\left\|p_P\left(\mathbf{S}\right)\right\|\cdot\left\|\boldsymbol{\theta}^{(j)}\right\| + \left\|\mathbf{R}_2\right\|\cdot\left(\left\|p_P\left(\mathbf{S}\right)\right\|+\left\|\mathrm{d}\mathbf{P}\right\|\right)\cdot\left\|\boldsymbol{\theta}^{(j)}\right\|\\ =& O_{\mathbb{P}}\left(\frac{M^2}{R}\cdot\zeta_1^2\left(P\right)\right).\end{aligned}$$

We consider $\boldsymbol{\gamma}_{NPj} = \mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\cdot\boldsymbol{\theta}^{(j)}$ as in (5.5.5). As a result, the forecast value of $\boldsymbol{\theta}$ is given, for $j = 1,\ldots,K$, by (5.5.4):

$$\begin{aligned}\widehat{\theta}^{(j)} =& p_P'\left(\widehat{\mathbf{s}}\right)\cdot\widehat{\boldsymbol{\beta}}_j\\ =& p_P'\left(\widehat{\mathbf{s}}\right)\cdot\boldsymbol{\gamma}_{NPj} - p_P'\left(\widehat{\mathbf{s}}\right)\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\cdot\mathrm{d}\mathbf{P'}\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\cdot\boldsymbol{\theta}^{(j)}\\ &+p_P'\left(\widehat{\mathbf{s}}\right)\cdot\mathbf{\Pi}^{-1}\cdot\mathrm{d}\mathbf{P}\cdot\left[\mathbf{I}_P - p_P'\left(\mathbf{S}\right)\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\right]\cdot\boldsymbol{\theta}^{(j)}\\ &+p_P'\left(\widehat{\mathbf{s}}\right)\cdot\mathbf{R}_3.\end{aligned}$$

However, $\boldsymbol{\theta}$ is given, for $j = 1, \ldots, K$:

$$\begin{aligned}\theta^{(j)} =& f_P^{(j)}\left(\widehat{\mathbf{s}}\right) + \left[f_P^{(j)}\left(\mathbf{s}\right) - f_P^{(j)}\left(\widehat{\mathbf{s}}\right)\right] + \left[f^{(j)}\left(\mathbf{s}\right) - f_P^{(j)}\left(\mathbf{s}\right)\right] \\ =& p_P'\left(\widehat{\mathbf{s}}\right)\boldsymbol{\gamma}_{Pj} + \left[p_P'\left(\mathbf{s}\right)\boldsymbol{\gamma}_{Pj} - p_P'\left(\widehat{\mathbf{s}}\right)\boldsymbol{\gamma}_{Pj}\right] + \left[f^{(j)}\left(\mathbf{s}\right) - p_P'\left(\mathbf{s}\right)\boldsymbol{\gamma}_{Pj}\right]\end{aligned}$$

where $\boldsymbol{\gamma}_{Pj}$ is given by (5.5.6). We will analyze it through the following decomposition:

$$\begin{aligned}\widehat{\theta}^{(j)} - \theta^{(j)} =& \left(p_P'\left(\widehat{\mathbf{s}}\right) - p_P'\left(\mathbf{s}\right)\right)\boldsymbol{\gamma}_{Pj} - \left(f^{(j)}\left(\mathbf{s}\right) - p_P'\left(\mathbf{s}\right)\boldsymbol{\gamma}_{Pj}\right) \\ &+ p_P'\left(\widehat{\mathbf{s}}\right) \cdot \left(\boldsymbol{\gamma}_{NPj} - \boldsymbol{\gamma}_{Pj}\right) \\ &- p_P'\left(\widehat{\mathbf{s}}\right) \cdot \mathbf{\Pi}^{-1} \cdot p_P\left(\mathbf{S}\right) \cdot \mathrm{d}\mathbf{P}' \cdot \mathbf{\Pi}^{-1} \cdot p_P\left(\mathbf{S}\right) \cdot \boldsymbol{\theta}^{(j)} \\ &+ p_P'\left(\widehat{\mathbf{s}}\right) \cdot \mathbf{\Pi}^{-1} \cdot \mathrm{d}\mathbf{P} \cdot \left[\mathbf{I}_P - p_P'\left(\mathbf{S}\right) \cdot \mathbf{\Pi}^{-1} \cdot p_P\left(\mathbf{S}\right)\right] \cdot \boldsymbol{\theta}^{(j)} \\ &+ p_P'\left(\widehat{\mathbf{s}}\right) \cdot \mathbf{R}_3.\end{aligned}$$

The second term is majorized by N_P. As concerns the third term, from Lemma 5.1, under A1 and A3 we have:

$$\left\|p_P'\left(\mathbf{s}\right) \cdot \left(\boldsymbol{\gamma}_{NPj} - \boldsymbol{\gamma}_{Pj}\right)\right\| \leq \left\|p_P\left(\mathbf{s}\right)\right\| \cdot \left\|\boldsymbol{\gamma}_{NPj} - \boldsymbol{\gamma}_{Pj}\right\| = O\left(\zeta_0\left(P\right) \cdot E_{NP}\right). \qquad (5.9.4)$$

As concerns the sixth term, from Lemma 5.1 and under A1, A3, A5 and A9, we have:

$$\left\|p_P'\left(\widehat{\mathbf{s}}\right) \cdot \mathbf{R}_3\right\| \leq \left\|p_P\left(\widehat{\mathbf{s}}\right)\right\| \cdot \left\|\mathbf{R}_3\right\| = O_{\mathbb{P}}\left(\frac{M^2}{R} \cdot \zeta_0\left(P\right)\zeta_1^2\left(P\right)\right).$$

QED

Proof of Theorem 5.1. Let us start from the second and third terms in the statement of Lemma 5.2. They can be majorized as:

$$\begin{aligned}&\left\|p_P'\left(\widehat{\mathbf{s}}\right) \cdot \mathbf{\Pi}^{-1} \cdot p_P\left(\mathbf{S}\right) \cdot \mathrm{d}\mathbf{P}' \cdot \mathbf{\Pi}^{-1} \cdot p_P\left(\mathbf{S}\right) \cdot \boldsymbol{\theta}^{(j)}\right\| \\ &\quad \leq \left\|p_P\left(\widehat{\mathbf{s}}\right)\right\| \cdot \left\|\mathbf{\Pi}^{-1}\right\|^2 \cdot \left\|p_P\left(\mathbf{S}\right)\right\|^2 \cdot \left\|\mathrm{d}\mathbf{P}\right\| \cdot \left\|\boldsymbol{\theta}^{(j)}\right\| \\ &\quad = O_{\mathbb{P}}\left(\frac{M}{R^{\frac{1}{2}}} \cdot \zeta_0\left(P\right)\zeta_1\left(P\right)\right)\end{aligned}$$

and:

$$\begin{aligned}&\left\|p_P'\left(\widehat{\mathbf{s}}\right) \cdot \mathbf{\Pi}^{-1} \cdot \mathrm{d}\mathbf{P} \cdot \left\{\mathbf{I}_N - p_P'\left(\mathbf{S}\right) \cdot \mathbf{\Pi}^{-1} \cdot p_P\left(\mathbf{S}\right)\right\} \cdot \boldsymbol{\theta}^{(j)}\right\| \\ &\quad \leq \left\|p_P\left(\widehat{\mathbf{s}}\right)\right\| \cdot \left\|\mathbf{\Pi}^{-1}\right\| \cdot \left\|\mathrm{d}\mathbf{P}\right\| \cdot \left\|\mathbf{I}_N - p_P'\left(\mathbf{S}\right) \cdot \mathbf{\Pi}^{-1} \cdot p_P\left(\mathbf{S}\right)\right\| \cdot \left\|\boldsymbol{\theta}^{(j)}\right\| \\ &\quad = O_{\mathbb{P}}\left(\frac{M}{R^{\frac{1}{2}}} \cdot \zeta_0\left(P\right)\zeta_1\left(P\right)\right)\end{aligned}$$

where we have used A1, A3 and A9.

Therefore, we have:

$$\begin{aligned}\widehat{\theta}^{(j)}-\theta^{(j)} =& \left(p_P'\left(\widehat{\mathbf{s}}\right)-p_P'\left(\mathbf{s}\right)\right)\boldsymbol{\gamma}_{Pj}\\ &+O\left(N_P+\zeta_0\left(P\right)\cdot E_{NP}\right)+O_{\mathbb{P}}\left(\frac{M}{R^{\frac{1}{2}}}\cdot\zeta_0\left(P\right)\zeta_1\left(P\right)\right).\end{aligned}$$

Now:

$$p_P\left(\widehat{\mathbf{s}}\right)-p_P\left(\mathbf{s}\right)=\underset{P\times M}{\frac{\partial p_P\left(\mathbf{s}^{\star}\right)}{\partial\mathbf{s}'}}\cdot\underset{M\times 1}{\left(\widehat{\mathbf{s}}-\mathbf{s}\right)}$$

where $\mathbf{s}^{\star}$ lies between $\widehat{\mathbf{s}}$ and $\mathbf{s}$. Under A1, A3 and A7, we have:

$$\begin{aligned}&\left|\left(p_P'\left(\widehat{\mathbf{s}}\right)-p_P'\left(\mathbf{s}\right)\right)\boldsymbol{\gamma}_{Pj}\right|\\ &\quad\leq\left\|p_P\left(\widehat{\mathbf{s}}\right)-p_P\left(\mathbf{s}\right)\right\|\cdot\left\|\boldsymbol{\gamma}_{Pj}\right\|\\ &\quad\leq\left\|\frac{\partial p_P\left(\mathbf{s}^{\star}\right)}{\partial\mathbf{s}'}\right\|\cdot\left\|\widehat{\mathbf{s}}-\mathbf{s}\right\|\cdot\left\|\boldsymbol{\gamma}_{Pj}\right\|\\ &\quad\simeq O_{\mathbb{P}}\left(S^{-\frac{1}{2}}M\zeta_1\left(P\right)\cdot\left(1+E_{NP}^{(2)}\right)\right)\end{aligned}$$

where we have majorized $\left\|\boldsymbol{\gamma}_{Pj}\right\|$ and $\left\|\frac{\partial p_P\left(\mathbf{s}^{\star}\right)}{\partial\mathbf{s}}\right\|$ through Lemma 5.1. From this, the result follows. QED

Proof of Theorem 5.2. We use Lemma 5.2 as well as the first steps in the proof of Theorem 5.1 to get, under A1-A3, A5 and A9:

$$\begin{aligned}S^{\frac{1}{2}}\left(\widehat{\theta}^{(j)}-\theta^{(j)}\right)=&S^{\frac{1}{2}}\boldsymbol{\gamma}_{Pj}'\left(p_P\left(\widehat{\mathbf{s}}\right)-p_P\left(\mathbf{s}\right)\right)\\ &+O\left(S^{\frac{1}{2}}N_P+S^{\frac{1}{2}}\zeta_0\left(P\right)\cdot E_{NP}\right)+O_{\mathbb{P}}\left(\frac{S^{\frac{1}{2}}M}{R^{\frac{1}{2}}}\cdot\zeta_0\left(P\right)\zeta_1\left(P\right)\right).\end{aligned}\tag{5.9.5}$$

Now, we can write:

$$\begin{aligned}S^{\frac{1}{2}}\left(\widehat{\theta}^{(j)}-\theta^{(j)}\right)=&S^{\frac{1}{2}}\left(f\left(\widehat{\mathbf{s}}\right)-f\left(\mathbf{s}\right)\right)+S^{\frac{1}{2}}\left(\boldsymbol{\gamma}_{Pj}'p_P\left(\widehat{\mathbf{s}}\right)-f\left(\widehat{\mathbf{s}}\right)\right)-S^{\frac{1}{2}}\left(\boldsymbol{\gamma}_{Pj}'p_P\left(\mathbf{s}\right)-f\left(\mathbf{s}\right)\right)\\ &+O\left(S^{\frac{1}{2}}N_P+S^{\frac{1}{2}}\zeta_0\left(P\right)\cdot E_{NP}\right)+O_{\mathbb{P}}\left(\frac{S^{\frac{1}{2}}M}{R^{\frac{1}{2}}}\cdot\zeta_0\left(P\right)\zeta_1\left(P\right)\right).\end{aligned}$$

The second and the third terms are both majorized by $S^{\frac{1}{2}}N_P$. As to the first term, $f\left(\widehat{\mathbf{s}}\right)-f\left(\mathbf{s}\right)=\frac{\partial f\left(\mathbf{s}^{\star}\right)}{\partial\mathbf{s}'}\cdot\left(\widehat{\mathbf{s}}-\mathbf{s}\right)$, where $\mathbf{s}^{\star}$ lies between $\widehat{\mathbf{s}}$ and $\mathbf{s}$. Therefore:

$$\begin{aligned}S^{\frac{1}{2}}\left(\widehat{\theta}^{(j)}-\theta^{(j)}\right)=&\frac{\partial f\left(\mathbf{s}^{\star}\right)}{\partial\mathbf{s}'}\cdot S^{\frac{1}{2}}\left(\widehat{\mathbf{s}}-\mathbf{s}\right)\\ &+O\left(S^{\frac{1}{2}}N_P+S^{\frac{1}{2}}\zeta_0\left(P\right)\cdot E_{NP}\right)+O_{\mathbb{P}}\left(\frac{S^{\frac{1}{2}}M}{R^{\frac{1}{2}}}\cdot\zeta_0\left(P\right)\zeta_1\left(P\right)\right).\end{aligned}$$

From A2 and A8, the first part can be shown to converge to $\mathcal{N}\left(\mathbf{0},\frac{\partial f(\mathbf{s})}{\partial\mathbf{s}'}\cdot\boldsymbol{\Sigma}\cdot\frac{\partial f(\mathbf{s})}{\partial\mathbf{s}}\right)$. QED

Proof of Theorem 5.3. Consider the expression in Lemma 5.2. Reasoning as in Theorem 5.1, under A1, A3 and A7, we have:

$$\left|\left(p_P'\left(\widehat{\mathbf{s}}\right)-p_P'\left(\mathbf{s}\right)\right)\boldsymbol{\gamma}_{Pj}\right|=O_{\mathbb{P}}\left(S^{-\frac{1}{2}}M\zeta_1\left(P\right)\right).$$

The asymptotic behavior in this case is determined by the second and third terms in the statement of Lemma 5.2. We replace there $p_P\left(\widehat{\mathbf{s}}\right)$ with $p_P\left(\mathbf{s}\right)$. We have:

$$\begin{aligned}&-p_P'\left(\widehat{\mathbf{s}}\right)\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\cdot\left(p_P\left(\widehat{\mathbf{S}}\right)-p_P\left(\mathbf{S}\right)\right)'\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\cdot\boldsymbol{\theta}^{(j)}\\&\quad=-p_P'\left(\mathbf{s}\right)\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\cdot\left(p_P\left(\widehat{\mathbf{S}}\right)-p_P\left(\mathbf{S}\right)\right)'\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\cdot\boldsymbol{\theta}^{(j)}\\&\qquad-\left(p_P'\left(\widehat{\mathbf{s}}\right)-p_P'\left(\mathbf{s}\right)\right)\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\cdot\left(p_P\left(\widehat{\mathbf{S}}\right)-p_P\left(\mathbf{S}\right)\right)'\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\cdot\boldsymbol{\theta}^{(j)}\end{aligned}$$

and:

$$\begin{aligned}&p_P'\left(\widehat{\mathbf{s}}\right)\cdot\mathbf{\Pi}^{-1}\cdot\left(p_P\left(\widehat{\mathbf{S}}\right)-p_P\left(\mathbf{S}\right)\right)\cdot\left[\mathbf{I}_P-p_P'\left(\mathbf{S}\right)\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\right]\cdot\boldsymbol{\theta}^{(j)}\\&\quad=p_P'\left(\mathbf{s}\right)\cdot\mathbf{\Pi}^{-1}\cdot\left(p_P\left(\widehat{\mathbf{S}}\right)-p_P\left(\mathbf{S}\right)\right)\cdot\left[\mathbf{I}_P-p_P'\left(\mathbf{S}\right)\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\right]\cdot\boldsymbol{\theta}^{(j)}\\&\qquad+\left(p_P'\left(\widehat{\mathbf{s}}\right)-p_P'\left(\mathbf{s}\right)\right)\cdot\mathbf{\Pi}^{-1}\cdot\left(p_P\left(\widehat{\mathbf{S}}\right)-p_P\left(\mathbf{S}\right)\right)\cdot\left[\mathbf{I}_P-p_P'\left(\mathbf{S}\right)\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\right]\cdot\boldsymbol{\theta}^{(j)}\end{aligned}$$

where, under A1, A3, A7 and A9:

$$\begin{aligned}&\left|\left(p_P'\left(\widehat{\mathbf{s}}\right)-p_P'\left(\mathbf{s}\right)\right)\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\cdot\left(p_P\left(\widehat{\mathbf{S}}\right)-p_P\left(\mathbf{S}\right)\right)'\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\cdot\boldsymbol{\theta}^{(j)}\right|\\&\quad\leq\left\|p_P\left(\widehat{\mathbf{s}}\right)-p_P\left(\mathbf{s}\right)\right\|\cdot\left\|\mathbf{\Pi}^{-1}\right\|^2\cdot\left\|p_P\left(\mathbf{S}\right)\right\|^2\cdot\left\|p_P\left(\widehat{\mathbf{S}}\right)-p_P\left(\mathbf{S}\right)\right\|\cdot\left\|\boldsymbol{\theta}^{(j)}\right\|\\&\quad=O_{\mathbb{P}}\left(S^{-\frac{1}{2}}R^{-\frac{1}{2}}M^2\zeta_1^2\left(P\right)\right),\\&\left|\left(p_P'\left(\widehat{\mathbf{s}}\right)-p_P'\left(\mathbf{s}\right)\right)\cdot\mathbf{\Pi}^{-1}\cdot\left(p_P\left(\widehat{\mathbf{S}}\right)-p_P\left(\mathbf{S}\right)\right)\cdot\left[\mathbf{I}_P-p_P'\left(\mathbf{S}\right)\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\right]\cdot\boldsymbol{\theta}^{(j)}\right|\\&\quad\leq\left\|p_P\left(\widehat{\mathbf{s}}\right)-p_P\left(\mathbf{s}\right)\right\|\cdot\left\|\mathbf{\Pi}^{-1}\right\|\cdot\left\|p_P\left(\widehat{\mathbf{S}}\right)-p_P\left(\mathbf{S}\right)\right\|\cdot\left\|\mathbf{I}_P-p_P'\left(\mathbf{S}\right)\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\right\|\cdot\left\|\boldsymbol{\theta}^{(j)}\right\|\\&\quad=O_{\mathbb{P}}\left(S^{-\frac{1}{2}}R^{-\frac{1}{2}}M^2\zeta_1^2\left(P\right)\right).\end{aligned}$$

From this, using A5:

$$\begin{aligned}\widehat{\theta}^{(j)}-\theta^{(j)}=&-p_P'\left(\mathbf{s}\right)\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\cdot\left(p_P\left(\widehat{\mathbf{S}}\right)-p_P\left(\mathbf{S}\right)\right)'\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\cdot\boldsymbol{\theta}^{(j)}\\&+p_P'\left(\mathbf{s}\right)\cdot\mathbf{\Pi}^{-1}\cdot\left(p_P\left(\widehat{\mathbf{S}}\right)-p_P\left(\mathbf{S}\right)\right)\cdot\left[\mathbf{I}_P-p_P'\left(\mathbf{S}\right)\cdot\mathbf{\Pi}^{-1}\cdot p_P\left(\mathbf{S}\right)\right]\cdot\boldsymbol{\theta}^{(j)}\\&+O\left(N_P+\zeta_0\left(P\right)\cdot E_{NP}\right)+O_{\mathbb{P}}\left(\left(\frac{M}{R}\cdot\zeta_0\left(P\right)\zeta_1\left(P\right)+\frac{1}{S^{\frac{1}{2}}}\right)\cdot M\zeta_1\left(P\right)\right).\qquad(5.9.6)\end{aligned}$$

We rewrite the leading two terms as:

$$
\begin{aligned}
&p_P'(\mathbf{s})\cdot\boldsymbol{\Pi}^{-1}\cdot \mathrm{d}\mathbf{P}\cdot\left\{\mathbf{I}_N-p_P'(\mathbf{S})\cdot\boldsymbol{\Pi}^{-1}\cdot p_P(\mathbf{S})\right\}\cdot\boldsymbol{\theta}^{(j)}\\
&-\boldsymbol{\theta}^{(j),\prime}\cdot p_P'(\mathbf{S})\cdot\boldsymbol{\Pi}^{-1}\cdot \mathrm{d}\mathbf{P}\cdot p_P'(\mathbf{S})\cdot\boldsymbol{\Pi}^{-1}\cdot p_P(\mathbf{s})\\
&\quad=p_P'(\mathbf{s})\cdot\boldsymbol{\Pi}^{-1}\cdot \mathrm{d}\mathbf{P}\cdot\sum_{n=1}^{N}\mathbf{e}_n\mathbf{e}_n'\cdot\left\{\mathbf{I}_N-p_P'(\mathbf{S})\cdot\boldsymbol{\Pi}^{-1}\cdot p_P(\mathbf{S})\right\}\cdot\boldsymbol{\theta}^{(j)}\\
&\qquad-\boldsymbol{\theta}^{(j),\prime}\cdot p_P'(\mathbf{S})\cdot\boldsymbol{\Pi}^{-1}\cdot \mathrm{d}\mathbf{P}\cdot\sum_{n=1}^{N}\mathbf{e}_n\mathbf{e}_n'\cdot p_P'(\mathbf{S})\cdot\boldsymbol{\Pi}^{-1}\cdot p_P(\mathbf{s})\\
&\quad=\sum_{n=1}^{N}p_P'(\mathbf{s})\cdot\boldsymbol{\Pi}^{-1}\cdot\left(p_P\left(\widehat{\mathbf{S}}_n\right)-p_P(\mathbf{S}_n)\right)\cdot\left[\mathbf{e}_n'\cdot\left\{\mathbf{I}_N-p_P'(\mathbf{S})\cdot\boldsymbol{\Pi}^{-1}\cdot p_P(\mathbf{S})\right\}\cdot\boldsymbol{\theta}^{(j)}\right]\\
&\qquad-\sum_{n=1}^{N}\boldsymbol{\theta}^{(j),\prime}\cdot p_P'(\mathbf{S})\cdot\boldsymbol{\Pi}^{-1}\cdot\left(p_P\left(\widehat{\mathbf{S}}_n\right)-p_P(\mathbf{S}_n)\right)\cdot\left[\mathbf{e}_n'\cdot p_P'(\mathbf{S})\cdot\boldsymbol{\Pi}^{-1}\cdot p_P(\mathbf{s})\right]\\
&\quad=\mathbf{a}'\cdot\sum_{n=1}^{N}\left(p_P\left(\widehat{\mathbf{S}}_n\right)-p_P(\mathbf{S}_n)\right)\cdot a_n\\
&\qquad-\mathbf{b}'\cdot\sum_{n=1}^{N}\left(p_P\left(\widehat{\mathbf{S}}_n\right)-p_P(\mathbf{S}_n)\right)\cdot b_n\\
&\quad=\begin{bmatrix}\mathbf{a}' & -\mathbf{b}'\end{bmatrix}\cdot\begin{bmatrix}\sum_{n=1}^{N}\left(p_P\left(\widehat{\mathbf{S}}_n\right)-p_P(\mathbf{S}_n)\right)\cdot a_n\\ \sum_{n=1}^{N}\left(p_P\left(\widehat{\mathbf{S}}_n\right)-p_P(\mathbf{S}_n)\right)\cdot b_n\end{bmatrix}
\end{aligned}
\tag{5.9.7}
$$

where:

$$
\begin{aligned}
a_n=&\mathbf{e}_n'\cdot\left\{\mathbf{I}_N-p_P'(\mathbf{S})\cdot\boldsymbol{\Pi}^{-1}\cdot p_P(\mathbf{S})\right\}\cdot\boldsymbol{\theta}^{(j)}\\
b_n=&\mathbf{e}_n'\cdot p_P'(\mathbf{S})\cdot\boldsymbol{\Pi}^{-1}\cdot p_P(\mathbf{s})\\
\mathbf{a}'=&p_P'(\mathbf{s})\cdot\boldsymbol{\Pi}^{-1}\\
\mathbf{b}'=&\boldsymbol{\theta}^{(j),\prime}\cdot p_P'(\mathbf{S})\cdot\boldsymbol{\Pi}^{-1}.
\end{aligned}
$$

We note some facts that will be used in the following. From Lemma 5.1, respectively under A3 and under A1 and A3:

$$
\begin{aligned}
\|\mathbf{a}\|&\le\|p_P(\mathbf{s})\|\cdot\left\|\boldsymbol{\Pi}^{-1}\right\|=O\left(N^{-1}\zeta_0(P)\right),\\
\|\mathbf{b}\|&\le\left\|\boldsymbol{\theta}^{(j)}\right\|\cdot\|p_P(\mathbf{S})\|\cdot\left\|\boldsymbol{\Pi}^{-1}\right\|=O(1).
\end{aligned}
$$

From the same source, respectively under A1 and A3, and under A3:

$$\begin{aligned}
|a_n| =& \left|\mathbf{e}_n' \cdot \left\{\mathbf{I}_N - p_P' \left(\mathbf{S}\right) \cdot \mathbf{\Pi}^{-1} \cdot p_P \left(\mathbf{S}\right)\right\} \cdot \boldsymbol{\theta}^{(j)}\right| \\
=& \left|\theta_n^{(j)} - p_P' \left(\mathbf{S}_n\right) \cdot \boldsymbol{\gamma}_{NPj}\right| \\
=& \left|\left(f^{(j)} \left(\mathbf{S}_n\right) - p_P' \left(\mathbf{S}_n\right) \cdot \boldsymbol{\gamma}_{Pj}\right) + p_P' \left(\mathbf{S}_n\right) \cdot \left(\boldsymbol{\gamma}_{Pj} - \boldsymbol{\gamma}_{NPj}\right)\right| \\
\leq& N_P + \left\|p_P \left(\mathbf{S}_n\right)\right\| \cdot \left\|\boldsymbol{\gamma}_{Pj} - \boldsymbol{\gamma}_{NPj}\right\| \\
=& O\left(N_P + \zeta_0 \left(P\right) \cdot E_{NP}\right), \\
|b_n| =& \left|p_P' \left(\mathbf{S}_n\right) \cdot \mathbf{\Pi}^{-1} \cdot p_P \left(\mathbf{s}\right)\right| \\
\leq& \left\|p_P \left(\mathbf{S}_n\right)\right\| \cdot \left\|\mathbf{\Pi}^{-1}\right\| \cdot \left\|p_P \left(\mathbf{s}\right)\right\| \\
\leq& N^{-1} \zeta_0^2 \left(P\right).
\end{aligned}$$

We are led to consider the vectors:

$$\mathbf{X}_n := \left[\begin{array}{c} \left(p_P\left(\widehat{\mathbf{S}}_n\right) - p_P\left(\mathbf{S}_n\right)\right) \cdot a_n \\ \left(p_P\left(\widehat{\mathbf{S}}_n\right) - p_P\left(\mathbf{S}_n\right)\right) \cdot b_n \end{array}\right]$$

for $n = 1, \ldots, N$. We rewrite (5.9.7) as:

$$\left[\begin{array}{cc} \mathbf{a}' & -\mathbf{b}' \end{array}\right] \cdot \left[\begin{array}{c} \sum_{n=1}^N \left(p_P\left(\widehat{\mathbf{S}}_n\right) - p_P\left(\mathbf{S}_n\right)\right) \cdot a_n \\ \sum_{n=1}^N \left(p_P\left(\widehat{\mathbf{S}}_n\right) - p_P\left(\mathbf{S}_n\right)\right) \cdot b_n \end{array}\right] = \left[\begin{array}{cc} \mathbf{a}' & -\mathbf{b}' \end{array}\right] \cdot \sum_{n=1}^N \mathbf{X}_n.$$

Let $\overline{\mathbf{X}}_n := \mathbf{X}_n - \mathbb{E}\mathbf{X}_n$. We start writing this as:

$$\left[\begin{array}{cc} \mathbf{a}' & -\mathbf{b}' \end{array}\right] \cdot \sum_{n=1}^N \mathbf{X}_n = \left[\begin{array}{cc} \mathbf{a}' & -\mathbf{b}' \end{array}\right] \cdot \sum_{n=1}^N \overline{\mathbf{X}}_n + \left[\begin{array}{cc} \mathbf{a}' & -\mathbf{b}' \end{array}\right] \cdot \sum_{n=1}^N \mathbb{E}\mathbf{X}_n.$$

This implies that (5.9.6) becomes, under A1-A3, A5, A7 and A9:

$$\begin{aligned}
\widehat{\theta}^{(j)} - \theta^{(j)} =& \left[\begin{array}{cc} \mathbf{a}' & -\mathbf{b}' \end{array}\right] \cdot \sum_{n=1}^N \overline{\mathbf{X}}_n + \left[\begin{array}{cc} \mathbf{a}' & -\mathbf{b}' \end{array}\right] \cdot \sum_{n=1}^N \mathbb{E}\mathbf{X}_n \\
&+ O\left(N_P + \zeta_0\left(P\right) \cdot E_{NP}\right) + O_{\mathbb{P}}\left(\left(\frac{M}{R} \cdot \zeta_0\left(P\right) \zeta_1\left(P\right) + \frac{1}{S^{\frac{1}{2}}}\right) \cdot M \zeta_1\left(P\right)\right). \qquad (5.9.8)
\end{aligned}$$

Now, under A1 and A3:

$$\begin{aligned}
&\left|\left[\begin{array}{cc}\mathbf{a}' & -\mathbf{b}'\end{array}\right]\cdot\sum_{n=1}^{N}\mathbb{E}\mathbf{X}_n\right| \\
&\quad\leq\left\|\left[\begin{array}{c}\mathbf{a}\\ \mathbf{b}\end{array}\right]\right\|\cdot\sum_{n=1}^{N}\left\|\mathbb{E}\mathbf{X}_n\right\| \\
&\quad\leq\left(\left\|\mathbf{a}\right\|^2+\left\|\mathbf{b}\right\|^2\right)^{\frac{1}{2}}\cdot\sum_{n=1}^{N}\left\|\mathbb{E}p_P\left(\widehat{\mathbf{S}}_n\right)-p_P\left(\mathbf{S}_n\right)\right\|\cdot\left(a_n^2+b_n^2\right)^{\frac{1}{2}} \\
&\quad=O\left(\left(1+N^{-1}\zeta_0\left(P\right)\right)\cdot\left(N_P+\zeta_0\left(P\right)\cdot E_{NP}+N^{-1}\zeta_0^2\left(P\right)\right)\right) \\
&\qquad\cdot\sum_{n=1}^{N}\left\{\sum_{j=1}^{P}\left[\mathbb{E}p_{jP}\left(\widehat{\mathbf{S}}_n\right)-p_{jP}\left(\mathbf{S}_n\right)\right]^2\right\}^{\frac{1}{2}}.
\end{aligned}$$

We take a limited development of $p_{jP}\left(\widehat{\mathbf{S}}_n\right)$ around $\mathbf{S}_n$:

$$p_{jP}\left(\widehat{\mathbf{S}}_n\right)\simeq p_{jP}\left(\mathbf{S}_n\right)+\sum_{h=1}^{M}\frac{\partial p_{jP}\left(\mathbf{S}_n\right)}{\partial S_{nh}}\left(\widehat{S}_{nh}-S_{nh}\right)$$

and, through the definition of $\zeta_1\left(P\right)$ and A9:

$$\begin{aligned}
&\sum_{j=1}^{P}\left[\mathbb{E}p_{jP}\left(\widehat{\mathbf{S}}_n\right)-p_{jP}\left(\mathbf{S}_n\right)\right]^2 \\
&\quad\simeq\sum_{j=1}^{P}\left[\sum_{h=1}^{M}\frac{\partial p_{jP}\left(\mathbf{S}_n\right)}{\partial S_{nh}}\mathbb{E}\left(\widehat{S}_{nh}-S_{nh}\right)\right]^2 \\
&\quad\leq M\sum_{j=1}^{P}\sum_{h=1}^{M}\left[\frac{\partial p_{jP}\left(\mathbf{S}_n\right)}{\partial S_{nh}}\mathbb{E}\left(\widehat{S}_{nh}-S_{nh}\right)\right]^2 \\
&\quad\leq M\sum_{j=1}^{P}\sum_{h=1}^{M}\left|\frac{\partial p_{jP}\left(\mathbf{S}_n\right)}{\partial S_{nh}}\right|^2\mathbb{E}\left|\widehat{S}_{nh}-S_{nh}\right|^2 \\
&\quad\leq\frac{cM^2}{R}\sup_h\sum_{j=1}^{P}\left|\frac{\partial p_{jP}\left(\mathbf{S}_n\right)}{\partial S_{nh}}\right|^2\leq\frac{cM^2\zeta_1^2\left(P\right)}{R}. \qquad (5.9.9)
\end{aligned}$$

At last, using A6:

$$\begin{aligned}
&\left|\left[\begin{array}{cc}\mathbf{a}' & -\mathbf{b}'\end{array}\right]\cdot\sum_{n=1}^{N}\mathbb{E}\mathbf{X}_n\right| \\
&\quad=O\left(\frac{NM\zeta_1\left(P\right)}{R^{\frac{1}{2}}}\cdot\max\left\{1,N^{-1}\zeta_0\left(P\right)\right\}\cdot\left(N_P+\zeta_0\left(P\right)\cdot E_{NP}+N^{-1}\zeta_0^2\left(P\right)\right)\right) \\
&\quad=O\left(\frac{NM\zeta_1\left(P\right)}{R^{\frac{1}{2}}}\cdot\left(N_P+\zeta_0\left(P\right)\cdot E_{NP}+N^{-1}\zeta_0^2\left(P\right)\right)\right)
\end{aligned}$$

from which (5.9.8) becomes, under A1-A3, A5-A7 and A9:

$$\begin{aligned}\widehat{\theta}^{(j)} - \theta^{(j)} = & \left[\begin{array}{cc} \mathbf{a}' & -\mathbf{b}' \end{array}\right] \cdot \sum_{n=1}^{N} \overline{\mathbf{X}}_n + O_{\mathbb{P}}\left(\left(\frac{M\zeta_0(P)\zeta_1(P)}{R} + \frac{1}{S^{\frac{1}{2}}}\right) \cdot M\zeta_1(P)\right) \\ & + O\left(\left(1 + \frac{NM\zeta_1(P)}{R^{\frac{1}{2}}}\right) \cdot (N_P + \zeta_0(P) \cdot E_{NP}) + \frac{M\zeta_0^2(P)\zeta_1(P)}{R^{\frac{1}{2}}}\right). \end{aligned} \tag{5.9.10}$$

We approximate $\sum_{n=1}^{N} \overline{\mathbf{X}}_n$ with a Gaussian vector having the same mean and the same variance as $\sum_{n=1}^{N} \overline{\mathbf{X}}_n$, say $\mathbf{G}_N$. Then we would like to find conditions for $\left[\begin{array}{cc} \mathbf{a}' & -\mathbf{b}' \end{array}\right] \cdot \sum_{n=1}^{N} \overline{\mathbf{X}}_n$ to be near to $\left[\begin{array}{cc} \mathbf{a}' & -\mathbf{b}' \end{array}\right] \cdot \mathbf{G}_N$. We have, using A6:

$$\begin{aligned} & \left| \left[\begin{array}{cc} \mathbf{a}' & -\mathbf{b}' \end{array}\right] \cdot \sum_{n=1}^{N} \overline{\mathbf{X}}_n - \left[\begin{array}{cc} \mathbf{a}' & -\mathbf{b}' \end{array}\right] \cdot \mathbf{G}_N \right| \\ & \quad = \left| \left[\begin{array}{cc} \mathbf{a}' & -\mathbf{b}' \end{array}\right] \cdot \left(\sum_{n=1}^{N} \overline{\mathbf{X}}_n - \mathbf{G}_N \right) \right| \\ & \quad \leq \left\| \left[\begin{array}{c} \mathbf{a} \\ \mathbf{b} \end{array}\right] \right\| \cdot \left\| \sum_{n=1}^{N} \overline{\mathbf{X}}_n - \mathbf{G}_N \right\| = \left(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 \right)^{\frac{1}{2}} \cdot \left\| \sum_{n=1}^{N} \overline{\mathbf{X}}_n - \mathbf{G}_N \right\| \\ & \quad = O\left(\left(1 + N^{-1}\zeta_0(P)\right) \cdot \left\| \sum_{n=1}^{N} \overline{\mathbf{X}}_n - \mathbf{G}_N \right\| \right) = O\left(\left\| \sum_{n=1}^{N} \overline{\mathbf{X}}_n - \mathbf{G}_N \right\| \right). \end{aligned} \tag{5.9.11}$$

Now, $\left\| \sum_{n=1}^{N} \overline{\mathbf{X}}_n - \mathbf{G}_N \right\|$ can be dealt with using Yurisnkii's coupling. We recall that this is the inequality in [390, Theorem 10, p. 244].

Theorem. *Let $\mathbf{X}_n$, for $n = 1, \ldots, N$, be independent random P-vectors with $\mathbb{E}\mathbf{X}_n = \mathbf{0}$, for $n = 1, \ldots, N$, and let $\beta := \sum_{n=1}^{N} \mathbb{E} \|\mathbf{X}_n\|^3$. For each $\delta > 0$ there exists a random vector $\mathbf{G}_N$ with a $\mathcal{N}\left(\mathbf{0}, \mathbb{V}\left(\sum_{n=1}^{N} \mathbf{X}_n\right)\right)$ distribution such that:*

$$\mathbb{P}\left\{ \left\| \sum_{n=1}^{N} \mathbf{X}_n - \mathbf{G}_N \right\| > \delta \right\} \leq C_0 \frac{\beta P}{\delta^3} \cdot \left(1 + \frac{\left|\ln\left(\delta^3/\beta P\right)\right|}{P}\right)$$

for some universal constant C_0.

Note that the statement of Yurinskii's coupling can be written as:

$$\mathbb{P}\left\{ \left\| \sum_{n=1}^{N} \mathbf{X}_n - \mathbf{G}_N \right\| > \left(\frac{\beta P}{\varepsilon}\right)^{\frac{1}{3}} \right\} \leq C_0 \varepsilon \cdot \left(1 + \frac{|\ln \varepsilon|}{P}\right)$$

where $\varepsilon \downarrow 0$. This means, provided $\frac{\varepsilon \cdot |\ln \varepsilon|}{P} \downarrow 0$ too, that:

$$\left\| \sum_{n=1}^{N} \mathbf{X}_n - \mathbf{G}_N \right\| = o_{\mathbb{P}}\left(\left(\frac{\beta P}{\varepsilon}\right)^{\frac{1}{3}} \right). \tag{5.9.12}$$

This is the statement we are going to use in the following. Therefore, A1-A3, A5-A7 and A9, (5.9.10) becomes:

$$\begin{aligned}\widehat{\theta}^{(j)}-\theta^{(j)}=&\left[\begin{array}{cc}\mathbf{a}' & -\mathbf{b}'\end{array}\right]\cdot\mathbf{G}_N+o_{\mathbb{P}}\left(\left(\frac{\beta P}{\varepsilon}\right)^{\frac{1}{3}}\right)\\&+O_{\mathbb{P}}\left(\left(\frac{M\zeta_0\left(P\right)\zeta_1\left(P\right)}{R}+\frac{1}{S^{\frac{1}{2}}}\right)\cdot M\zeta_1\left(P\right)\right)\\&+O\left(\left(1+\frac{NM\zeta_1\left(P\right)}{R^{\frac{1}{2}}}\right)\cdot\left(N_P+\zeta_0\left(P\right)\cdot E_{NP}\right)+\frac{M\zeta_0^2\left(P\right)\zeta_1\left(P\right)}{R^{\frac{1}{2}}}\right)\end{aligned}\tag{5.9.13}$$

provided $\varepsilon\downarrow 0$ and $\frac{\varepsilon\cdot|\ln\varepsilon|}{P}\downarrow 0$.

Then we define:

$$\begin{aligned}\beta_n:=&\mathbb{E}\left\|\overline{\mathbf{X}}_n\right\|^3=\mathbb{E}\left|\overline{\mathbf{X}}_n'\overline{\mathbf{X}}_n\right|^{\frac{3}{2}}\\\leq&\mathbb{E}\left|\left[\begin{array}{cc}\left(p_P\left(\widehat{\mathbf{S}}_n\right)-\mathbb{E}p_P\left(\widehat{\mathbf{S}}_n\right)\right)'\cdot a_n & \left(p_P\left(\widehat{\mathbf{S}}_n\right)-\mathbb{E}p_P\left(\widehat{\mathbf{S}}_n\right)\right)'\cdot b_n\end{array}\right]\right.\\&\left.\cdot\left[\begin{array}{c}\left(p_P\left(\widehat{\mathbf{S}}_n\right)-\mathbb{E}p_P\left(\widehat{\mathbf{S}}_n\right)\right)\cdot a_n\\\left(p_P\left(\widehat{\mathbf{S}}_n\right)-\mathbb{E}p_P\left(\widehat{\mathbf{S}}_n\right)\right)\cdot b_n\end{array}\right]\right|^{\frac{3}{2}}\\\leq&\mathbb{E}\left|\left(p_P\left(\widehat{\mathbf{S}}_n\right)-\mathbb{E}p_P\left(\widehat{\mathbf{S}}_n\right)\right)'\left(p_P\left(\widehat{\mathbf{S}}_n\right)-\mathbb{E}p_P\left(\widehat{\mathbf{S}}_n\right)\right)\right|^{\frac{3}{2}}\cdot\left(a_n^2+b_n^2\right)^{\frac{3}{2}}.\end{aligned}$$

The terms of this equation can be dealt with as follows. We have:

$$\begin{aligned}&\mathbb{E}\left|\left(p_P\left(\widehat{\mathbf{S}}_n\right)-\mathbb{E}p_P\left(\widehat{\mathbf{S}}_n\right)\right)'\left(p_P\left(\widehat{\mathbf{S}}_n\right)-\mathbb{E}p_P\left(\widehat{\mathbf{S}}_n\right)\right)\right|^{\frac{3}{2}}\\&\quad=\mathbb{E}\left\|p_P\left(\widehat{\mathbf{S}}_n\right)-\mathbb{E}p_P\left(\widehat{\mathbf{S}}_n\right)\right\|^3\\&\quad=\mathbb{E}\left\|\left[p_P\left(\widehat{\mathbf{S}}_n\right)-p_P\left(\mathbf{S}_n\right)\right]+\left[p_P\left(\mathbf{S}_n\right)-\mathbb{E}p_P\left(\widehat{\mathbf{S}}_n\right)\right]\right\|^3\\&\quad\leq\mathbb{E}\left\{\left\|p_P\left(\widehat{\mathbf{S}}_n\right)-p_P\left(\mathbf{S}_n\right)\right\|+\left\|p_P\left(\mathbf{S}_n\right)-\mathbb{E}p_P\left(\widehat{\mathbf{S}}_n\right)\right\|\right\}^3\\&\quad\leq 4\left\{\mathbb{E}\left\|p_P\left(\widehat{\mathbf{S}}_n\right)-p_P\left(\mathbf{S}_n\right)\right\|^3+\left\|p_P\left(\mathbf{S}_n\right)-\mathbb{E}p_P\left(\widehat{\mathbf{S}}_n\right)\right\|^3\right\}\\&\quad\leq 4\mathbb{E}\left\|p_P\left(\widehat{\mathbf{S}}_n\right)-p_P\left(\mathbf{S}_n\right)\right\|^3+O\left(\frac{M^3\zeta_1^3\left(P\right)}{R^{\frac{3}{2}}}\right)\end{aligned}$$

where the last step uses A9 and follows the development of (5.9.9). The first term is:

$$\begin{aligned}
&\mathbb{E}\left\|p_P\left(\widehat{\mathbf{S}}_n\right)-p_P\left(\mathbf{S}_n\right)\right\|^3\\
&\quad\simeq\mathbb{E}\left\|\frac{\partial p_P\left(\mathbf{S}_n\right)}{\partial\mathbf{s}'}\left(\widehat{\mathbf{S}}_n-\mathbf{S}_n\right)\right\|^3\\
&\quad\leq\mathbb{E}\left[\sum_{k=1}^{M}\left(\widehat{S}_{kn}-S_{kn}\right)^2\right]^{\frac{3}{2}}\cdot\left\|\frac{\partial p_P\left(\mathbf{S}_n\right)}{\partial\mathbf{s}'}\right\|^3\\
&\quad\leq M^{\frac{1}{2}}\sum_{k=1}^{M}\mathbb{E}\left|\widehat{S}_{kn}-S_{kn}\right|^3\cdot\left\|\frac{\partial p_P\left(\mathbf{S}_n\right)}{\partial\mathbf{s}'}\right\|^3\\
&\quad\leq M^{\frac{3}{2}}\sup_k\mathbb{E}\left|\widehat{S}_{kn}-S_{kn}\right|^3\cdot M^{\frac{3}{2}}\zeta_1^3\left(P\right)\\
&\quad=M^3\zeta_1^3\left(P\right)\sup_k\mathbb{E}\left|\widehat{S}_{kn}-S_{kn}\right|^3\\
&\quad\leq\frac{cM^3\zeta_1^3\left(P\right)}{R^{\frac{3}{2}}}
\end{aligned}$$

where we have majorized $\left\|\frac{\partial p_P(\mathbf{S}_n)}{\partial\mathbf{s}'}\right\|$ as in Lemma 5.1 and we have used A10. Using the expression for a_n and b_n above, this implies that, under A1, A3 and A10:

$$\beta_n=O\left(\frac{M^3\zeta_1^3\left(P\right)}{R^{\frac{3}{2}}}\cdot\left(N_P+\zeta_0\left(P\right)\cdot E_{NP}+N^{-1}\zeta_0^2\left(P\right)\right)^3\right)$$

and:

$$\beta=O\left(\frac{NM^3\zeta_1^3\left(P\right)}{R^{\frac{3}{2}}}\cdot\left(N_P+\zeta_0\left(P\right)\cdot E_{NP}+N^{-1}\zeta_0^2\left(P\right)\right)^3\right).$$

From (5.9.11) and (5.9.12):

$$\begin{aligned}
&\left|\left[\begin{array}{cc}\mathbf{a}' & -\mathbf{b}'\end{array}\right]\cdot\sum_{n=1}^{N}\overline{\mathbf{X}}_n-\left[\begin{array}{cc}\mathbf{a}' & -\mathbf{b}'\end{array}\right]\cdot\mathbf{G}_N\right|\\
&\quad\leq O\left(\left\|\sum_{n=1}^{N}\overline{\mathbf{X}}_n-\mathbf{G}_N\right\|\right)=o_{\mathbb{P}}\left(\left(\frac{\beta P}{\varepsilon}\right)^{\frac{1}{3}}\right)\\
&\quad=o_{\mathbb{P}}\left(\frac{N^{\frac{1}{3}}P^{\frac{1}{3}}M\zeta_1\left(P\right)}{\varepsilon^{\frac{1}{3}}R^{\frac{1}{2}}}\cdot\left(N_P+\zeta_0\left(P\right)\cdot E_{NP}+N^{-1}\zeta_0^2\left(P\right)\right)\right).
\end{aligned}$$

Therefore, from (5.9.10), under A1-A3, A5-A7 and A9-A10, if $\varepsilon \downarrow 0$ and $\frac{\varepsilon \cdot |\ln \varepsilon|}{P} \downarrow 0$:

$$\begin{aligned}
\widehat{\theta}^{(j)} - \theta^{(j)} = & \left[\begin{array}{cc} \mathbf{a}' & -\mathbf{b}' \end{array}\right] \cdot \mathbf{G}_N \\
& + o_{\mathbb{P}}\left(\frac{N^{\frac{1}{3}} P^{\frac{1}{3}} M \zeta_1(P)}{\varepsilon^{\frac{1}{3}} R^{\frac{1}{2}}} \cdot \left(N_P + \zeta_0(P) \cdot E_{NP} + N^{-1} \zeta_0^2(P)\right)\right) \\
& + O_{\mathbb{P}}\left(\left(\frac{M \zeta_0(P) \zeta_1(P)}{R} + \frac{1}{S^{\frac{1}{2}}}\right) \cdot M \zeta_1(P)\right) \\
& + O\left(\left(1 + \frac{N M \zeta_1(P)}{R^{\frac{1}{2}}}\right) \cdot \left(N_P + \zeta_0(P) \cdot E_{NP}\right) + \frac{M \zeta_0^2(P) \zeta_1(P)}{R^{\frac{1}{2}}}\right).
\end{aligned}$$

The result follows taking $\varepsilon = B_{NP}^{-1}$.

Now we turn to the variance of $\left[\begin{array}{cc} \mathbf{a}' & -\mathbf{b}' \end{array}\right] \cdot \mathbf{G}_N$ and its order in probability. We have:

$$\begin{aligned}
& \mathbb{V}\left\{\left[\begin{array}{cc} \mathbf{a}' & -\mathbf{b}' \end{array}\right] \cdot \mathbf{G}_N\right\} \\
& = \mathbb{V}\left\{\sum_{n=1}^{N} \left(a_n \mathbf{a}' - b_n \mathbf{b}'\right) \cdot \left(p_P\left(\widehat{\mathbf{S}}_n\right) - p_P\left(\mathbf{S}_n\right)\right)\right\} \\
& = \sum_{n=1}^{N} \left(a_n \mathbf{a}' - b_n \mathbf{b}'\right) \cdot \mathbb{V}\left(p_P\left(\widehat{\mathbf{S}}_n\right)\right) \cdot \left(a_n \mathbf{a} - b_n \mathbf{b}\right) \\
& \simeq \sum_{n=1}^{N} \left(a_n \mathbf{a}' - b_n \mathbf{b}'\right) \cdot \frac{\partial p_P\left(\mathbf{S}_n\right)}{\partial \mathbf{s}'} \cdot \mathbb{V}\left(\widehat{\mathbf{S}}_n\right) \cdot \left(\frac{\partial p_P\left(\mathbf{S}_n\right)}{\partial \mathbf{s}'}\right)' \cdot \left(a_n \mathbf{a} - b_n \mathbf{b}\right).
\end{aligned}$$

As to the order in probability, we write:

$$\begin{aligned}
\mathbb{V}\left(\sum_{n=1}^{N} \mathbf{X}_n\right) = & \mathbb{V}\left(\sum_{n=1}^{N} \left[\begin{array}{cc} a_n \mathbf{I}_P & \mathbf{0}_{P \times P} \\ \mathbf{0}_{P \times P} & b_n \mathbf{I}_P \end{array}\right] \left[\begin{array}{c} p_P\left(\widehat{\mathbf{S}}_n\right) - p_P\left(\mathbf{S}_n\right) \\ p_P\left(\widehat{\mathbf{S}}_n\right) - p_P\left(\mathbf{S}_n\right) \end{array}\right]\right) \\
= & \sum_{n=1}^{N} \left[\begin{array}{cc} a_n \mathbf{I}_P & \mathbf{0}_{P \times P} \\ \mathbf{0}_{P \times P} & b_n \mathbf{I}_P \end{array}\right] \left\{\mathbf{U}_2 \otimes \mathbb{V}\left(p_P\left(\widehat{\mathbf{S}}_n\right) - p_P\left(\mathbf{S}_n\right)\right)\right\} \left[\begin{array}{cc} a_n \mathbf{I}_P & \mathbf{0}_{P \times P} \\ \mathbf{0}_{P \times P} & b_n \mathbf{I}_P \end{array}\right]
\end{aligned}$$

and:

$$\begin{aligned}
& \lambda_{\max}\left(\mathbb{V}\left(\sum_{n=1}^{N} \mathbf{X}_n\right)\right) \\
& \leq \lambda_{\max}\left(\sum_{n=1}^{N} \left[\begin{array}{cc} a_n \mathbf{I}_P & \mathbf{0}_{P \times P} \\ \mathbf{0}_{P \times P} & b_n \mathbf{I}_P \end{array}\right] \left\{\mathbf{U}_2 \otimes \mathbb{V}\left(p_P\left(\widehat{\mathbf{S}}_n\right) - p_P\left(\mathbf{S}_n\right)\right)\right\} \left[\begin{array}{cc} a_n \mathbf{I}_P & \mathbf{0}_{P \times P} \\ \mathbf{0}_{P \times P} & b_n \mathbf{I}_P \end{array}\right]\right) \\
& \leq \lambda_{\max}\left(\left[\begin{array}{cc} \sum_{n=1}^{N} a_n^2 \mathbf{I}_P & \mathbf{0}_{P \times P} \\ \mathbf{0}_{P \times P} & \sum_{n=1}^{N} b_n^2 \mathbf{I}_P \end{array}\right]\right) \cdot \sup_n \lambda_{\max}\left(\mathbb{V}\left(p_P\left(\widehat{\mathbf{S}}_n\right) - p_P\left(\mathbf{S}_n\right)\right)\right) \\
& \leq \max\left\{\sum_{n=1}^{N} a_n^2, \sum_{n=1}^{N} b_n^2\right\} \cdot \sup_n \left\|\frac{\partial p_P\left(\mathbf{S}_n\right)}{\partial \mathbf{s}'}\right\|^2 \cdot \sup_n \lambda_{\max}\left(\mathbb{V}\left(\widehat{\mathbf{S}}_n - \mathbf{S}_n\right)\right).
\end{aligned}$$

Now, respectively under A1 and under A3:

$$\begin{aligned}\sum_{n=1}^{N} a_n^2 =&\boldsymbol{\theta}^{(j),\prime}\cdot\left\{\mathbf{I}_N - p_P'(\mathbf{S})\cdot\boldsymbol{\Pi}^{-1}\cdot p_P(\mathbf{S})\right\}\cdot\sum_{n=1}^{N}\mathbf{e}_n\mathbf{e}_n'\cdot\left\{\mathbf{I}_N - p_P'(\mathbf{S})\cdot\boldsymbol{\Pi}^{-1}\cdot p_P(\mathbf{S})\right\}\cdot\boldsymbol{\theta}^{(j)}\\ =&\boldsymbol{\theta}^{(j),\prime}\cdot\left\{\mathbf{I}_N - p_P'(\mathbf{S})\cdot\boldsymbol{\Pi}^{-1}\cdot p_P(\mathbf{S})\right\}\cdot\boldsymbol{\theta}^{(j)}\\ \leq&\left\|\boldsymbol{\theta}^{(j)}\right\|^2 = O(N),\\ \sum_{n=1}^{N} b_n^2 =&p_P'(\mathbf{s})\cdot\boldsymbol{\Pi}^{-1}\cdot p_P(\mathbf{S})\cdot\sum_{n=1}^{N}\mathbf{e}_n\mathbf{e}_n'\cdot p_P'(\mathbf{S})\cdot\boldsymbol{\Pi}^{-1}\cdot p_P(\mathbf{s})\\ =&p_P'(\mathbf{s})\cdot\boldsymbol{\Pi}^{-1}\cdot p_P(\mathbf{s})\leq\|p_P(\mathbf{s})\|^2\cdot\left\|\boldsymbol{\Pi}^{-1}\right\| = O\left(N^{-1}\zeta_0^2(P)\right),\end{aligned}$$

so that, under A6, $\max\left\{\sum_{n=1}^{N}a_n^2,\sum_{n=1}^{N}b_n^2\right\} = O\left(N\cdot\left(1+N^{-2}\zeta_0^2(P)\right)\right) = O(N)$. Now, from Lemma 5.1, $\left\|\frac{\partial p_P(\mathbf{S}_n)}{\partial\mathbf{s}'}\right\|\leq M^{\frac{1}{2}}\zeta_1(P)$ and, under A9:

$$\lambda_{\max}\left(\mathbb{V}\left(\widehat{\mathbf{S}}_n-\mathbf{S}_n\right)\right)\leq\operatorname{tr}\left(\mathbb{V}\left(\widehat{\mathbf{S}}_n-\mathbf{S}_n\right)\right)=\sum_{k=1}^{M}\mathbb{V}\left(\widehat{S}_{kn}-S_{kn}\right)\leq\frac{cM}{R}.$$

Therefore, under A1, A3, A6 and A9:

$$\lambda_{\max}\left(\mathbb{V}\left(\sum_{n=1}^{N}\mathbf{X}_n\right)\right)=O\left(\frac{NM^2}{R}\cdot\zeta_1^2(P)\right).$$

At last, under A1, A3, A6 and A9:

$$\begin{aligned}&\mathbb{V}\left\{\sum_{n=1}^{N}\left(a_n\mathbf{a}'-b_n\mathbf{b}'\right)\cdot\left(p_P\left(\widehat{\mathbf{S}}_n\right)-p_P(\mathbf{S}_n)\right)\right\}\\ &=\begin{bmatrix}\mathbf{a}' & -\mathbf{b}'\end{bmatrix}\cdot\mathbb{V}\left(\sum_{n=1}^{N}\mathbf{X}_n\right)\cdot\begin{bmatrix}\mathbf{a}\\ -\mathbf{b}\end{bmatrix}\\ &\leq\left\|\begin{bmatrix}\mathbf{a}\\ -\mathbf{b}\end{bmatrix}\right\|^2\cdot\lambda_{\max}\left(\mathbb{V}\left(\sum_{n=1}^{N}\mathbf{X}_n\right)\right)\\ &=\left(\|\mathbf{a}\|^2+\|\mathbf{b}\|^2\right)\cdot\lambda_{\max}\left(\mathbb{V}\left(\sum_{n=1}^{N}\mathbf{X}_n\right)\right)\\ &=O\left(\left(1+N^{-2}\zeta_0^2(P)\right)\cdot\frac{M^2N}{R}\cdot\zeta_1^2(P)\right)=O\left(\frac{M^2N}{R}\cdot\zeta_1^2(P)\right).\end{aligned}$$

QED

Proof of Proposition 5.1. Let us start from $E_{NP}^{(1)}$. We can write $N^{-1}\boldsymbol{\Pi}=N^{-1}\sum_{n=1}^{N}p_P(\mathbf{S}_n)\,p_P'(\mathbf{S}_n)$. The generic element $\left[N^{-1}\boldsymbol{\Pi}\right]_{(i,j)}$ of this matrix is:

$$\frac{1}{N}\sum_{n=1}^{N}p_{iP}(\mathbf{S}_n)\,p_{jP}'(\mathbf{S}_n).$$

Any S_{kn} is a deterministic function of $\boldsymbol{\theta}_n$, say $S_{kn} = g^{(k)}\left(\boldsymbol{\theta}_n\right)$. We write the vector version of this equality as $\mathbf{S}_n = \mathbf{g}\left(\boldsymbol{\theta}_n\right)$. Therefore:

$$\frac{1}{N}\sum_{n=1}^{N} p_{iP}\left(\mathbf{S}_n\right) p'_{jP}\left(\mathbf{S}_n\right) = \frac{1}{N}\sum_{n=1}^{N} p_{iP}\left(\mathbf{g}\left(\boldsymbol{\theta}_n\right)\right) p'_{jP}\left(\mathbf{g}\left(\boldsymbol{\theta}_n\right)\right).$$

It is clear that:

$$\frac{1}{N}\sum_{n=1}^{N} p_{iP}\left(\mathbf{S}_n\right) p'_{jP}\left(\mathbf{S}_n\right) = \int_{\mathbb{R}^K} p_{iP}\left(\mathbf{g}\left(\mathbf{x}\right)\right) p_{jP}\left(\mathbf{g}\left(\mathbf{x}\right)\right) \mathbb{P}_N\left(\mathrm{d}\mathbf{x}\right).$$

Under A12, we have:

$$\left[N^{-1}\boldsymbol{\Pi} - \boldsymbol{\Pi}_{0P}\right]_{(i,j)} = \int_{\mathbb{R}^M} p_{iP}\left(\mathbf{g}\left(\mathbf{x}\right)\right) p_{jP}\left(\mathbf{g}\left(\mathbf{x}\right)\right) \left(\mathbb{P}_N - \mathbb{P}\right)\left(\mathrm{d}\mathbf{x}\right).$$

Using the Koksma-Hlawka inequality stated in [192, Corollary 2.4], [225, p. 165], [6]:

$$\left|\left[N^{-1}\boldsymbol{\Pi} - \boldsymbol{\Pi}_{0P}\right]_{(i,j)}\right| \leq V_{HK}\left(p_{iP}\left(\mathbf{g}\left(\cdot\right)\right) p_{jP}\left(\mathbf{g}\left(\cdot\right)\right)\right) \cdot D_{N,\mathbb{P}}$$

where $V_{HK}\left(f\right)$ is the Hardy-Krause total variation (see, e.g., [370]) and $D_{N,\mathbb{P}}$ is the non-uniform unanchored discrepancy defined in [192, p. 100] or [225, p. 165]:

$$D_{N,\mathbb{P}} := \sup_{A\subseteq[0,1]^K} \left|\mathbb{P}_N\left(A\right) - \mathbb{P}\left(A\right)\right|$$

where A is any axis-parallel rectangle. This implies that:

$$\begin{aligned} & \left\|N^{-1}\boldsymbol{\Pi} - \boldsymbol{\Pi}_{0P}\right\|^2 \\ \leq & \left\|N^{-1}\boldsymbol{\Pi} - \boldsymbol{\Pi}_{0P}\right\|_F^2 \\ = & \sum_{i=1}^{P}\sum_{j=1}^{P} \left|\left[N^{-1}\boldsymbol{\Pi} - \boldsymbol{\Pi}_{0P}\right]_{(i,j)}\right|^2 \\ \leq & D_{N,\mathbb{P}}^2 \cdot \sum_{i=1}^{P}\sum_{j=1}^{P} \left[V_{HK}\left(p_{iP}\left(\mathbf{g}\left(\cdot\right)\right) p_{jP}\left(\mathbf{g}\left(\cdot\right)\right)\right)\right]^2. \end{aligned}$$

We will make use of the following inequality:

$$V_{HK}\left(fg\right) \leq \left(3^K + 1 - 2^{K+1}\right) \cdot V_{HK}\left(f\right) V_{HK}\left(g\right) + \left\|f\right\|_\infty V_{HK}\left(g\right) + \left\|g\right\|_\infty V_{HK}\left(f\right), \qquad (5.9.14)$$

derived from the last formula on p. 251 in [62] for $\varphi(x) = x$. Then we have:

$$\begin{aligned}
&\sum_{i=1}^{P}\sum_{j=1}^{P}\left[V_{HK}\left(p_{iP}\left(\mathbf{g}\left(\cdot\right)\right)p_{jP}\left(\mathbf{g}\left(\cdot\right)\right)\right)\right]^2 \\
&\quad\leq \sum_{i=1}^{P}\sum_{j=1}^{P}\left[\left(3^M + 1 - 2^{M+1}\right)\cdot V_{HK}\left(p_{iP}\circ\mathbf{g}\right)V_{HK}\left(p_{jP}\circ\mathbf{g}\right)\right. \\
&\qquad\left. + \left\|p_{iP}\circ\mathbf{g}\right\|_\infty V_{HK}\left(p_{jP}\circ\mathbf{g}\right) + \left\|p_{jP}\circ\mathbf{g}\right\|_\infty V_{HK}\left(p_{iP}\circ\mathbf{g}\right)\right]^2 \\
&\quad\leq \left(3^K + 1 - 2^{K+1}\right)^2\cdot 3\left\{\sum_{i=1}^{P}\left[V_{HK}\left(p_{iP}\circ\mathbf{g}\right)\right]^2\right\}^2 \qquad (5.9.15)\\
&\qquad + 6\sum_{i=1}^{P}\left\|p_{iP}\right\|_\infty^2\cdot\sum_{j=1}^{P}\left[V_{HK}\left(p_{jP}\circ\mathbf{g}\right)\right]^2
\end{aligned}$$

where we have used the inequality $\left(\sum_{i=1}^{n} x_i\right)^2 \leq n\sum_{i=1}^{n} x_i^2$.

Now, from Eq. (3) in [33, p. 1948] (see also [370, p. 61]):

$$\begin{aligned}
V_{HK}\left(f\right) &\leq \sum_{u\neq\emptyset}\int_{[0,1]^u}\left|\frac{\partial^{|u|}f\left(\boldsymbol{\theta}_u;\mathbf{1}_{-u}\right)}{\partial\boldsymbol{\theta}_u}\right|\mathrm{d}\boldsymbol{\theta}_u \\
&\leq \sum_{u\neq\emptyset}\sup_{\boldsymbol{\theta}_u\in[0,1]^u}\left|\frac{\partial^{|u|}f\left(\boldsymbol{\theta}_u;\mathbf{1}_{-u}\right)}{\partial\boldsymbol{\theta}_u}\right| \\
&\leq \sum_{u\neq\emptyset}\sup_{\boldsymbol{\theta}\in[0,1]^K}\left|\frac{\partial^{|u|}f\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}_u}\right|,
\end{aligned}$$

where:

$$\frac{\partial^{|u|}f\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}_u} = \frac{\partial^{|u|}p_{iP}\left(\mathbf{g}\left(\boldsymbol{\theta}\right)\right)}{\partial\boldsymbol{\theta}_u}.$$

In order to compute this derivative of a composite function, we apply the results in [101] and [33]. In the following we prefer to follow the notation in [101] as this paper contains some results that we will find useful, but we will modify their formula according to the lines of [33]. Theorem 2.1 in [101] considers $h\left(\mathbf{x}\right) = f\left(\mathbf{g}\left(\mathbf{x}\right)\right)$ where $\mathbf{x}\in\mathbb{R}^d$ and $\mathbf{g}\left(\mathbf{x}\right)\in\mathbb{R}^m$. Suppose that a vector $\boldsymbol{\nu} = \left(\nu_1,\ldots,\nu_d\right)\in\mathbb{N}_0^d$ with $n := |\boldsymbol{\nu}| \neq 0$ is given. For a vector $\mathbf{m} = \left(m_1,\ldots,m_d\right)$, we denote $f_{\mathbf{m}} = \frac{\partial^{|\mathbf{m}|}f}{\partial x_1^{m_1}\cdots\partial x_d^{m_d}}$. We use [101, Theorem 2.1] or [33, Theorem 2].

In our case, the vector $\boldsymbol{\nu}$ (that we call u) belongs to the set $\{0,1\}^d\setminus\{\mathbf{0}\}$. As a result we can apply Lemma 3 and the reasoning leading to Eq. (9) in [33], taking into account the different notations of the two papers. As a consequence, $\boldsymbol{\ell}_j\in\{0,1\}^d\setminus\{\mathbf{0}\}$ and $\mathbf{k}_j\in\{0,1\}^m$ with $|\mathbf{k}_j| = 1$. Therefore, we identify $\mathbf{k}_j$ with the (only) integer k_j for which $\mathbf{k}_j$ takes the value 1. Applying this, we get:

$$\left[\mathbf{g}_{\boldsymbol{\ell}_j}\right]^{\mathbf{k}_j} = \frac{\partial g^{(k_j)}}{\partial x_{\ell_j}}.$$

Under A13, for $n = |\boldsymbol{\nu}| \neq 0$, we have:

$$\begin{aligned}
|h_{\boldsymbol{\nu}}| \leq & \boldsymbol{\nu}! \sum_{1 \leq |\boldsymbol{\lambda}| \leq n} |f_{\boldsymbol{\lambda}}| \sum_{s=1}^{n} \sum_{p_s(\boldsymbol{\nu},\boldsymbol{\lambda})} \prod_{j=1}^{s} \frac{\left| \left[\mathbf{g}_{\boldsymbol{\ell}_j} \right]^{\mathbf{k}_j} \right|}{(\mathbf{k}_j!) \left[\boldsymbol{\ell}_j! \right]^{|\mathbf{k}_j|}} \\
& \leq (\mu \vee \mu^n) \, \boldsymbol{\nu}! \sum_{k=1}^{n} \sum_{|\boldsymbol{\lambda}|=k} |f_{\boldsymbol{\lambda}}| \sum_{s=1}^{n} \sum_{p_s(\boldsymbol{\nu},\boldsymbol{\lambda})} \prod_{j=1}^{s} \frac{1}{(\mathbf{k}_j!) \left[\boldsymbol{\ell}_j! \right]^{|\mathbf{k}_j|}} \\
& \leq (\mu \vee \mu^n) \, \boldsymbol{\nu}! \sum_{k=1}^{n} \sqrt{ \sum_{|\boldsymbol{\lambda}|=k} |f_{\boldsymbol{\lambda}}|^2 \cdot \sum_{|\boldsymbol{\lambda}|=k} \left\{ \sum_{s=1}^{n} \sum_{p_s(\boldsymbol{\nu},\boldsymbol{\lambda})} \prod_{j=1}^{s} \frac{1}{(\mathbf{k}_j!) \left[\boldsymbol{\ell}_j! \right]^{|\mathbf{k}_j|}} \right\}^2 } \\
& \leq (\mu \vee \mu^n) \, \boldsymbol{\nu}! \sum_{k=1}^{n} \sqrt{ \left\{ \sum_{|\boldsymbol{\lambda}|=k} |f_{\boldsymbol{\lambda}}|^2 \right\} \cdot \left\{ \sum_{|\boldsymbol{\lambda}|=k} \sum_{s=1}^{n} \sum_{p_s(\boldsymbol{\nu},\boldsymbol{\lambda})} \prod_{j=1}^{s} \frac{1}{(\mathbf{k}_j!) \left[\boldsymbol{\ell}_j! \right]^{|\mathbf{k}_j|}} \right\}^2 } \\
& \leq (\mu \vee \mu^n) \sum_{k=1}^{n} \left(\sum_{|\boldsymbol{\lambda}|=k} |f_{\boldsymbol{\lambda}}|^2 \right)^{\frac{1}{2}} \cdot m^k S_n^k
\end{aligned}$$

where the second inequality uses A13, the third uses Cauchy-Schwarz inequality, the fourth uses $\sum_{i=1}^{n} x_i^2 \leq \left(\sum_{i=1}^{n} x_i \right)^2$, and the fifth uses Corollary 2.9 in [101, p. 511]. In the previous formulas, S_n^k is a Stirling number of the second kind. In our case, $h(\cdot) = f(\mathbf{g}(\cdot))$ is replaced by $p_{iP}(\mathbf{g}(\cdot))$, $\boldsymbol{\nu}$ is replaced by u, n is identified with $|u|$, m is replaced by M:

$$\begin{aligned}
\sum_{i=1}^{P} \left[V_{HK} (p_{iP} \circ \mathbf{g}) \right]^2 & \leq \sum_{i=1}^{P} \left[\sum_{u \neq \emptyset} (\mu \vee \mu^n) \sum_{k=1}^{n} \left(\sum_{|\boldsymbol{\lambda}|=k} |f_{\boldsymbol{\lambda}}|^2 \right)^{\frac{1}{2}} \cdot M^k S_n^k \right]^2 \\
& \leq \left(\mu \vee \mu^K \right)^2 \sum_{i=1}^{P} \left[\sum_{u \neq \emptyset} \sum_{k=1}^{n} \left(\sum_{|\boldsymbol{\lambda}|=k} |f_{\boldsymbol{\lambda}}|^2 \right)^{\frac{1}{2}} \cdot M^k S_n^k \right]^2 \\
& \leq \left(\mu \vee \mu^K \right)^2 \sum_{i=1}^{P} \left(\sum_{u \neq \emptyset} \sum_{k=1}^{n} \sum_{|\boldsymbol{\lambda}|=k} |f_{\boldsymbol{\lambda}}|^2 \right) \cdot \left(\sum_{u \neq \emptyset} \sum_{k=1}^{n} M^{2k} S_n^{2k} \right) \\
& \leq \left(\mu \vee \mu^K \right)^2 \left(\sum_{u \neq \emptyset} \sum_{k=1}^{n} \sum_{|\boldsymbol{\lambda}|=k} \sum_{i=1}^{P} \left| \partial^{\boldsymbol{\lambda}} p_{iP} \right|^2 \right) \cdot \left(\sum_{u \neq \emptyset} \sum_{k=1}^{n} M^{2k} S_n^{2k} \right) \\
& \leq \left(\mu \vee \mu^K \right)^2 \left(\sum_{u \neq \emptyset} \sum_{k=1}^{n} \sum_{|\boldsymbol{\lambda}|=k} \zeta_k^2 (P) \right) \cdot \left(\sum_{u \neq \emptyset} \sum_{k=1}^{n} M^{2k} S_n^{2k} \right)
\end{aligned}$$

where the third step uses Cauchy-Schwarz inequality while the others are simply rearrangements.

Now:

$$\begin{aligned}
\sum_{u\neq\emptyset}\sum_{k=1}^{n}\sum_{|\boldsymbol{\lambda}|=k}\zeta_k^2(P) =& \sum_{n=1}^{K}\binom{K}{n}\sum_{k=1}^{n}\binom{K+k-1}{k}\zeta_k^2(P)\\
=& \sum_{k=1}^{K}\binom{K+k-1}{k}\zeta_k^2(P)\sum_{n=k}^{K}\binom{K}{n}\\
\leq& \zeta_K^2(P)\cdot\sum_{k=1}^{K}\binom{K+k-1}{k}\sum_{n=k}^{K}\binom{K}{n}
\end{aligned}$$

and:

$$\begin{aligned}
\sum_{u\neq\emptyset}\sum_{k=1}^{n}M^{2k}S_n^{2k} =& \sum_{n=1}^{K}\binom{K}{n}\sum_{k=1}^{n}M^{2k}S_n^{2k}\\
=& \sum_{k=1}^{K}M^{2k}S_n^{2k}\sum_{n=k}^{K}\binom{K}{n}.
\end{aligned}$$

At last:

$$\begin{aligned}
\sum_{i=1}^{P}\left[V_{HK}\left(p_{iP}\circ\mathbf{g}\right)\right]^2 \leq& \left(\mu\vee\mu^K\right)^2\zeta_K^2(P)\cdot\left(\sum_{k=1}^{K}\binom{K+k-1}{k}\sum_{n=k}^{K}\binom{K}{n}\right)\\
&\cdot\left(\sum_{k=1}^{K}M^{2k}S_n^{2k}\sum_{n=k}^{K}\binom{K}{n}\right)\\
=:& \left(\mu\vee\mu^K\right)^2\zeta_K^2(P)\cdot C(M,K)
\end{aligned}$$

for a constant $C(M,K)$ depending only on M and K.

Replacing this into (5.9.15), we get:

$$\begin{aligned}
&\sum_{i=1}^{P}\sum_{j=1}^{P}\left[V_{HK}\left(p_{iP}\left(\mathbf{g}(\cdot)\right)p_{jP}\left(\mathbf{g}(\cdot)\right)\right)\right]^2\\
&\quad\leq\left(3^K+1-2^{K+1}\right)^2\cdot 3\left\{\left(\mu\vee\mu^K\right)^4\zeta_K^4(P)\cdot C^2(M,K)\right\}\\
&\qquad+6\sum_{i=1}^{P}\|p_{iP}\|_\infty^2\cdot\left(\mu\vee\mu^K\right)^2\zeta_K^2(P)\cdot C(M,K)\\
&\quad\leq C(M,K)\cdot\left(\mu\vee\mu^K\right)^2\zeta_K^2(P)\cdot\left\{\left(\mu\vee\mu^K\right)^2\zeta_K^2(P)+\sum_{i=1}^{P}\|p_{iP}\|_\infty^2\right\}.
\end{aligned}$$

From this the final result for $E_{NP}^{(1)}$ follows.

The result for $E_{NP}^{(2)}$ is obtained remarking that, underr A12:

$$\left[N^{-1}p_P(\mathbf{S})\cdot\boldsymbol{\theta}^{(j)}-\boldsymbol{\pi}_{0P}\right]_{(j)}=\int_{\mathbb{R}^M}p_{iP}\left(\mathbf{g}(\mathbf{x})\right)x_j\left(\mathbb{P}_N-\mathbb{P}\right)(\mathrm{d}\mathbf{x})$$

Then:

$$\left\| N^{-1} p_P\left(\mathbf{S}\right) \cdot \boldsymbol{\theta}^{(j)} - \boldsymbol{\pi}_{0P} \right\|^2 \leq D_{N,\mathbb{P}}^2 \cdot \sum_{i=1}^{P} \left[V_{HK}\left(p_{iP}\left(\mathbf{g}\left(\cdot\right)\right) x_j\right)\right]^2$$

where:

$$\begin{aligned} V_{HK}\left(p_{iP}\left(\mathbf{g}\left(\cdot\right)\right) x_j\right) &\leq \left(3^K + 1 - 2^{K+1}\right) \cdot V_{HK}\left(p_{iP}\left(\mathbf{g}\left(\cdot\right)\right)\right) V_{HK}\left(x_j\right) + \left\|p_{iP}\right\|_\infty V_{HK}\left(x_j\right) \\ &\quad + \left\|x_j\right\|_\infty V_{HK}\left(p_{iP}\left(\mathbf{g}\left(\cdot\right)\right)\right) \\ &\leq \left(3^K + 1 - 2^{K+1}\right) \cdot V_{HK}\left(p_{iP}\left(\mathbf{g}\left(\cdot\right)\right)\right) + \left\|p_{iP}\right\|_\infty + V_{HK}\left(p_{iP}\left(\mathbf{g}\left(\cdot\right)\right)\right) \\ &= \left(3^K + 2 - 2^{K+1}\right) \cdot V_{HK}\left(p_{iP}\left(\mathbf{g}\left(\cdot\right)\right)\right) + \left\|p_{iP}\right\|_\infty . \end{aligned}$$

This leads to:

$$\begin{aligned} \left\| N^{-1} p_P\left(\mathbf{S}\right) \cdot \boldsymbol{\theta}^{(j)} - \boldsymbol{\pi}_{0P} \right\|^2 \leq & C\left(M,K\right) \cdot D_{N,\mathbb{P}}^2 \cdot \left\{ \sum_{i=1}^{P} \left[V_{HK}\left(p_{iP}\left(\mathbf{g}\left(\cdot\right)\right)\right)\right]^2 + \sum_{i=1}^{P} \left\|p_{iP}\right\|_\infty^2 \right\} \\ \leq & C\left(M,K\right) \cdot D_{N,\mathbb{P}}^2 \cdot \left\{ \left(\mu \vee \mu^K\right)^2 \cdot \zeta_K^2\left(P\right) + \sum_{i=1}^{P} \left\|p_{iP}\right\|_\infty^2 \right\} \end{aligned}$$

from which the final result follows.

The results on the discrepancy can be found in [5] and [6]. Note that some of these results are stated for the anchored discrepancy while others are stated for the unanchored one, but the rate of decrease in N is the same. QED

Chapter 6

Nonparametric Moment-based Estimation of Simulated Models via Regularized Regression[1]

This chapter is similar in spirit to the work by [86]. The statistical framework is close to the one exposed in Chapter 5 but, instead of OLS, we estimate the parameters of a simulated models via a nonparametric least absolute shrinkage and selection operator (Lasso) regression. The nonparametric element is introduced to capture the nonlinear relations between the statistics and the parameters. This implies some advantages, when comparing the method to the previous chapter, that will be clarified in the following. First of all, the Lasso allows the joint estimation and automatic selection of a subset of the basis functions used to model the nonparametric function linking the statistics and the parameters to be estimated. Second, in Lasso regression the number of basis function in the dictionary is not upper bounded by the number of points chosen out of the parameter space, while in OLS it is. Third, the oracle property of the Lasso suggests that the researcher may run this algorithm, identify the nonnull coefficients of the regression, and run a nonparametric sieve regression estimated by OLS containing only the retained coefficients, thus making the inferential tools of Chapter 5 available in the present situation. Furthermore, we explicitly and rigorously characterize the asymptotic behavior of the estimator. We end the chapter with a small simulation study showing the correct behavior of the method.

6.1 Introduction

In the present work, we develop a new estimation technique for simulated models inspired by the one of [86]. We aim at characterizing the parameters of a simulation model generating moments that are similar to real-world/benchmark statistics.

[1] This chapter is jointly written with Raffaello Seri.

Most econometric methods used to estimate simulation-based models exploit indirect inference ([196, 460, 195]), method of simulated moments ([335, 372, 133]), simulated minimum-distance ([217, 463, 191, 48]), approximate ([292, 293, 157, 2]) and nonparametric simulated maximum likelihood ([156, 272]), and approximate Bayesian computations ([148, 25, 131, 175, 168, 458]).

The parameters of an agent-based model are generally estimated applying the above-mentioned techniques (see [444, p. 3]). Many instances have been published in the last decades. The majority of them use indirect-based inference (see [189, 190, 54, 150]), MSM (see [499, 172, 206, 90]), SMD (see [401, 282]), simulated maximum likelihood (see [277, 314]), ABC (see [207]) and surrogate meta-models (see [422, 283]).

A drawback of simulation-based econometric methods is that estimation is performed by optimizing complex objective functions that can often be seen as distances between simulated and real-world/benchmark data. This creates some issues in terms of computational time (see [472, 307]), choice of the statistical distance (see [184, 301, 402, 32, 329, 426]) and/or determination of the statistics to match (see [179, 499, 111, 75]).

Furthermore, as detailed in Chapter 5, when estimating ABM via classic simulation-based econometric techniques we meet three important problems. First of all, in order to establish the asymptotic properties of the estimators, the objective function must respect a stochastic equicontinuity hypothesis (see [335], [372] and [355, pp. 2136-2137]). To comply with this assumption, simulation-based estimators must use the same draw of (pseudo-)random numbers involved in the approximation of the objective function for different values of the parameter $\boldsymbol{\theta}$ (see [335, p. 999], [195, p. 16], [272, p. 78] and [143, p. 346]). As a consequence, they work well with models characterized by recursive equations (see [196] and [195]), but not with ABM, in which new samples of (pseudo-)random numbers are drawn for any $\boldsymbol{\theta}$ (i.e. random numbers are not "recycled"). Second, since the (pseudo-)random numbers are not recycled, the objective function in the case of ABM is rugged and nowhere differentiable. This implies that the researcher tackles some difficulties deriving from the presence of multiple local minima, identification issues leading to large standard deviations, etc. (see [277]). Therefore, in order to obtain the numerical convergence of the objective function, the derivative-free algorithms used for standard optimization routines must be integrated with other approaches (see, e.g., [189] and [190]). Finally, the asymptotic properties and the corresponding inferential tools developed for classic simulation-based methods do not hold.

To deal with these problems, we draw on [86] and we develop a method that does not involve any optimization of an objective function. Instead, the moments are selected using a least absolute shrinkage and selection operator nonparametric regression. Other regression-based frameworks have been previously developed by [403, 449, 110, 266, 400, 396]. Nevertheless, these methods are more suitable for prediction rather than estimation.

Differently from the above-mentioned contributions, [86] use a regularized linear regression for parameter estimation. Despite the advantages deriving from its simplicity and versatility, the work of [86] presents some minor weaknesses. First of all, their work assumes and, consequently, estimates a linear relation linking the statistics to the parameters, but it is reasonable to suppose that the relation is nonlinear. Second, they provide no asymptotic properties for their method.

In this paper, we propose solutions to both problems. First, we improve their contribution adding sieve estimation (see [208, 354, 88]) to the Lasso regression. Second, we explicitly and rigorously characterize the asymptotic behavior of the estimator $\widehat{\theta}^{(j)}$.

Let us introduce how our method works. We have a simulation model indexed by a parameter vector $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^K$, where $\theta^{(j)}$, for $j = 1, \ldots, K$, is the generic element of $\boldsymbol{\theta}$. We select N parameter values in the parameter space Θ and, for each value in the parameter space, we simulate some runs and we calculate $M \geq K$ statistics on these runs. We aim at obtaining an estimation of $\theta^{(j)}$. We estimate the following Lasso nonparametric regression:

$$\theta_n^{(j)} = f_P^{(j)} \left(\widehat{S}_{1n}, \ldots, \widehat{S}_{Mn} \right) + \eta_n^{(j)},$$

where $n = 1, \ldots, N$, $f_P^{(j)}$ is a linear combination of basis functions (also called *dictionary*) and $\widehat{S}_{1n}, \ldots, \widehat{S}_{Mn}$ are the moments obtained from the data simulated using $\boldsymbol{\theta}_n$ as parameter. When we consider the statistics $\widehat{s}_1, \ldots, \widehat{s}_M$ computed on real-world/benchmark observations, we estimate $\theta^{(j)}$ as:

$$\widehat{\theta}^{(j)} = f_P^{(j)} \left(\widehat{s}_1, \ldots, \widehat{s}_M \right).$$

We conceive that our framework can bring some improvements to the literature. Most of them already arise when nonparametric ordinary least squares estimation is used instead of nonparametric Lasso and have been examined in Chapter 5. Therefore, in the next lines, we elucidate only the specific advantages of using Lasso.

First of all, our method, as the one in [86], allows joint estimation and automatic selection of a subset of the basis functions used to model the nonparametric dependence. This is one of the focal points of the contribution: as our aim is to estimate highly parametrized simulation models and, in particular, ABM, the possibility of considering a number of basis functions larger than the number of observations and of reducing them contextually with estimation is of great interest.

Second, differently from OLS, in Lasso regression the number of basis functions in the dictionary is not bounded from above by the number N of points chosen out of the parameter space. This implies that also the number of statistics that are used, namely M, can be an asymptotic parameter.

Third, the Lasso has generally an oracle property, i.e. it is asymptotically able to identify the elements of the dictionary whose coefficients in $f_P^{(j)}$ are not zero. This opens up the possibility that the researcher may estimate a model through our nonparametric Lasso, identify the regression coefficients that are nonnull, and estimate a OLS regression inserting only the elements of the dictionary whose coefficient is not zero. It is sensible to suppose that the parameters estimated by this procedure maintains the asymptotic properties of OLS estimates. This makes the large-sample theory of Chapter 5 available for the construction of inferential tools (tests, confidence intervals, etc.).

To be fair, also our technique has some weaknesses. First, we only provide an asymptotic rate of convergence of $\widehat{\theta}^{(j)}$ to $\theta^{(j)}$, we consider its implications in different cases, and we discuss at some lengths the assumptions that are needed to obtain it, but we do not give a proof of the above-mentioned oracle property. Second, as we study the properties of a nonparametric estimator, the convergence speed of $\widehat{\theta}^{(j)}$ to $\theta^{(j)}$ may be slower than the parametric one. On the other hand, the nonparametric component allows us to gain some flexibility from the estimation of the function linking the statistics and the parameters. Finally, the method seems to work very well when we deal with many observations. Indeed, when we consider a few noisy observations, the oracle property is not respected and the Lasso may select parameters that are equal to 0.

The structure of the paper is the following. Section 6.2 clarifies the notations used in the contribution. Section 6.3 outlines the statistical framework and provides some auxiliary results on the statistics and the dictionary. In Section 6.4, the main results of the work are derived, i.e. the error in the estimation of the coefficients and the prediction. Section 6.5 presents a simulation study to test the correct behavior of the estimation method. Section 6.6 wraps up the conclusions. In Section 6.7, the proofs of the theorems and the corollaries are provided.

6.2 Notation

We will write $\mathbb{N}$ for the positive integers, $\mathbb{N}_0$ for the non-negative integers, $\mathbb{R}$ for the real numbers and $\mathbb{C}$ for the complex numbers. Capital bold letters, such as $\mathbf{A}$, denote matrices while lowercase bold letters, such as $\mathbf{a}$, usually denote vectors. The i-th element of vector $\mathbf{a}$ is generally denoted a_i. $\mathbf{u}_n$ is a n-vector composed of ones. $\mathbf{I}_n$ is the $(n \times n)$-identity matrix. $\mathbf{U}_n$ is a $(n \times n)$-matrix composed of ones. $\mathbf{e}_{i,n}$ is a n-vector of zeros with a one in the i-th position; when the length is clear from the context we simply use $\mathbf{e}_i$. $\mathbf{0}_{m\times n}$ is a $(m \times n)$-matrix composed of zeros. We do not indicate the dimensions when they are clear from the context. $\operatorname{diag}(\mathbf{a})$ is a diagonal matrix with $\mathbf{a}$ on its diagonal. $\mathbf{A}'$ and $\mathbf{A}^{-1}$ are respectively the transpose and the classical inverse of the matrix $\mathbf{A}$, provided they exist.

6.3 Framework

We consider a parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^K$. We suppose that each component of the K-vector can be expressed as an unknown function of M statistics:

$$\theta^{(j)} = f^{(j)}\left(S_1, S_2, \ldots, S_M\right), \qquad j = 1, \ldots, K, \tag{6.3.1}$$

where $f^{(j)} : \mathbb{R}^M \to \mathbb{R}$ and $M \geq K$.

Our aim is to get an estimate of the function $f^{(j)}$ and use it to predict the value of $\boldsymbol{\theta}$. In order to obtain an estimate, we extract N configurations of parameters $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\} \subset \Theta \subset \mathbb{R}^K$ indexed with n. For each point $\boldsymbol{\theta}_n$, we simulate one or more runs and, on the basis of these observations, we estimate the M statistics. The k-th statistic based on the run generated by $\boldsymbol{\theta}_n$ is written as $\widehat{S}_{kn}$ and is an estimate of the fixed value S_{kn}. We define:

$$\theta_n^{(j)} = f^{(j)}\left(S_{1n}, \ldots, S_{Mn}\right) = f^{(j)}\left(\widehat{S}_{1n}, \ldots, \widehat{S}_{Mn}\right) + \varepsilon_n^{(j)}, \qquad n = 1, \ldots, N, j = 1, \ldots, K,$$

where $\varepsilon_n^{(j)}$ captures the error due to the estimation of the statistics. We approximate $f^{(j)}$ through a function $f_P^{(j)}$ given by an expansion in a series of P basis functions:

$$f_P^{(j)}(\mathbf{x}) = p_P'(\mathbf{x}) \cdot \boldsymbol{\beta}^{(j)}.$$

We then estimate a Lasso nonparametric regression of the form:

$$\theta_n^{(j)} = f_P^{(j)}\left(\widehat{S}_{1n}, \dots, \widehat{S}_{Mn}\right) + \eta_n^{(j)}, \qquad n = 1, \dots, N, j = 1, \dots, K,$$

where

$$\eta_n^{(j)} := f^{(j)}\left(\widehat{S}_{1n}, \dots, \widehat{S}_{Mn}\right) - f_P^{(j)}\left(\widehat{S}_{1n}, \dots, \widehat{S}_{Mn}\right) + \varepsilon_n^{(j)}.$$

Note that each parameter can be estimated separately, therefore in the following we will remove the index j when clear from the context.

6.3.1 Link between statistics and parameters

Let us start from the case of a single statistic. The simulation model produces a value of the statistic for any value of the parameter $\boldsymbol{\theta} \in \Theta$. The link is a function $g : \mathbb{R}^K \to \mathbb{R}$:

$$S = g\left(\theta^{(1)}, \dots, \theta^{(K)}\right) = g(\boldsymbol{\theta}).$$

The statistic is a fixed number, not a random variable, a fact that is often associated with model ergodicity (see [233]).

We suppose to observe $M \geq K$ statistics and we will mark this fact by introducing a progressive subscript both on S and on g:

$$S^{(j)} = g^{(j)}\left(\theta^{(1)}, \dots, \theta^{(K)}\right) = g^{(j)}(\boldsymbol{\theta}), \qquad j = 1, \dots, M. \tag{6.3.2}$$

In vector coordinates, $\mathbf{S} = \mathbf{g}(\boldsymbol{\theta})$, where $\mathbf{g} = (g_1, \dots, g_M)' : \mathbb{R}^K \to \mathbb{R}^M$. We need the following assumption.

Der For any k and j, $\left|\frac{\partial g^{(k)}}{\partial \theta^{(j)}}\right| \leq \mu$ in a neighborhood of the true parameter value.

We suppose that the relation linking $\boldsymbol{\theta}$ to $\mathbf{S}$ can be inverted into a function linking each $\theta^{(k)}$ with $\mathbf{S}$:

$$\theta^{(k)} = f^{(k)}(\mathbf{S}), \qquad k = 1, \dots, K.$$

The simplest way to do this is to select K statistics out of M belonging to $\mathbf{S}$, thus creating the system $\mathbf{S}^\star = \mathbf{g}^\star(\boldsymbol{\theta})$, where $\mathbf{S}^\star \in \mathbb{R}^K$ and $\mathbf{g}^\star = (g_1, \dots, g_K)' : \mathbb{R}^K \to \mathbb{R}^K$. We suppose that the conditions for the inversion of the system are respected and we get $\boldsymbol{\theta} = \mathbf{f}^\star(\mathbf{S}^\star)$. As we cannot be sure that the statistics selected in $\mathbf{S}^\star$ are the right ones, we extend the system as $\boldsymbol{\theta} = \mathbf{f}(\mathbf{S})$.[2]

We suppose to choose some values $\{\boldsymbol{\theta}_n, n = 1, \dots, N\}$ in the parameter space Θ. We define the *design measure* (see [109, p. 714]) as the discrete uniform distribution supported by the values

[2] We will let the Lasso set to 0 the coefficients of the irrelevant variables.

$\{\boldsymbol{\theta}_n, n = 1, \ldots, N\}$:

$$\mathbb{P}_N(A) := N^{-1} \sum_{n=1}^{N} 1\{\boldsymbol{\theta}_n \in A\}$$

for a Borel set $A \subset \mathbb{R}^K$. We suppose that $\mathbb{P}_N$ converges to an asymptotic design measure $\mathbb{P}$.

The following assumption requires that the parameter space can be reduced to a hypercube. In econometrics this condition is often stated as the requirement that the parameters are variation free (see [147], or Assumptions 8 and 11 in [354]).

Par The parameter space is given by the product of the parameter space of each single component of $\boldsymbol{\theta}$. Moreover, each component of the parameter vector can be rescaled to the interval $[0, 1]$. The design measure $\mathbb{P}_N$ converges to an asymptotic design measure $\mathbb{P}$.

We define the *(non-uniform) unanchored discrepancy* as a measure of distance between $\mathbb{P}_N$ and $\mathbb{P}$:

$$D_{N,\mathbb{P}} := \sup_{A \subseteq [0,1]^K} |\mathbb{P}_N(A) - \mathbb{P}(A)| \tag{6.3.3}$$

where A is any axis-parallel rectangle in $[0, 1]^K$. For any N, it is possible to find a point-set such that $D_{N,\mathbb{P}} = O\left(\frac{(\ln N)^{d-1}}{N}\right)$ if $\mathbb{P}$ is the uniform measure on Θ and such that $D_{N,\mathbb{P}} = O\left(\frac{(\ln N)^{\frac{3d+1}{2}}}{N}\right)$ if $\mathbb{P}$ is not the uniform measure on Θ; there is a sequence such that $D_{N,\mathbb{P}} = O\left(\frac{(\ln N)^{d}}{N}\right)$ if $\mathbb{P}$ is the uniform measure on Θ and such that $D_{N,\mathbb{P}} = O\left(\frac{(\ln N)^{\frac{3d+4}{2}}}{N}\right)$ if $\mathbb{P}$ is not the uniform measure on Θ. These results can be found in [5] and [6] (some of these results are stated for the anchored discrepancy, but they have the same rate of decrease in N).

To each value $\boldsymbol{\theta}_n$ for $n = 1, \ldots, N$, a statistic vector $\mathbf{S}_n$ is associated as $\mathbf{S}_n = \mathbf{g}(\boldsymbol{\theta}_n)$. These vectors are stacked in a $(P \times N)$-matrix $\mathbf{S}$.

6.3.2 Statistics estimated through simulation

We suppose to estimate each $\mathbf{S}_n$ through an estimator $\widehat{\mathbf{S}}_n$ based on R observations simulated from a model. These observations can be independent, if they are obtained as the result of R runs of the model, or dependent, if they are obtained aggregating one or more segments from a smaller number of runs. This is largely irrelevant for the computation of the rate of convergence, but may affect the computing time and the precision of the estimators. Indeed, a larger number of runs generally requires more time but should provide more precise estimates.

We define the error induced by estimation as:

$$\varepsilon_n := \theta_n - f\left(\widehat{\mathbf{S}}_n\right) = f(\mathbf{S}_n) - f\left(\widehat{\mathbf{S}}_n\right).$$

In the following we will introduce some hypotheses concerning the distribution of ε_n. They can be easily generalized to the case of *α-sub-exponential* random variables (see, e.g., [193]), but we will not pursue this topic here.

A centered random variable X is *sub-Gaussian* with variance proxy $\sigma^2 \in [0, +\infty)$ if:

$$\mathbb{E}\exp\{uX\} \leq \exp\left\{\frac{u^2\sigma^2}{2}\right\}$$

for all $u \in \mathbb{R}$. A random vector is sub-Gaussian iff all of its coordinates are.

subG We suppose that the following inequality holds true for any n:

$$\mathbb{E}\exp\{u\varepsilon_n\} = \mathbb{E}\exp\left\{u\left(\theta_n - f\left(\widehat{\mathbf{S}}_n\right)\right)\right\} \leq \exp\left\{u\mu_n + \frac{u^2\sigma_R^2}{2}\right\}$$

where σ_R^2 is a variance proxy dependent on R but not on n. Moreover, $\sum_{n=1}^N \mu_n^2 \leq N\mu_R^2$.

By taking a limited development on both sides around $u = 0$ it is easy to see that $\mu_n = \mathbb{E}\varepsilon_n = \theta_n - \mathbb{E}f\left(\widehat{\mathbf{S}}_n\right)$ and that:

$$\mathbb{V}(\varepsilon_n) = \mathbb{V}\left(f\left(\widehat{\mathbf{S}}_n\right)\right) \leq \sigma_R^2.$$

Through an informal reasoning, it is possible to get some more information about μ_R and σ_R^2. Assuming that $\widehat{\mathbf{S}}_n - \mathbf{S}_n = O_{\mathbb{P}}\left(R^{-\frac{1}{2}}\right)$ and using a mean value theorem, we have:

$$\mu_n = \mathbb{E}\varepsilon_n = \mathbb{E}f\left(\widehat{\mathbf{S}}_n\right) - f(\mathbf{S}_n) = O\left(\left\|\text{bias}\left(\widehat{\mathbf{S}}_n\right)\right\|_2 + R^{-1}\right)$$

and:

$$\mathbb{V}(\varepsilon_n) = \mathbb{V}\left(f\left(\widehat{\mathbf{S}}_n\right)\right) \leq \sigma_R^2$$

where $\mathbb{V}(\varepsilon_n) = O\left(R^{-1}\right)$. Supposing that $\left\|\text{bias}\left(\widehat{\mathbf{S}}_n\right)\right\|_2 = O\left(R^{-1}\right)$, we have:

$$\begin{aligned}\mu_R =& O\left(R^{-1}\right),\\ \sigma_R \geq& O\left(R^{-\frac{1}{2}}\right).\end{aligned}$$

We note that the coordinates of $\boldsymbol{\varepsilon}$ are independent of each other. Therefore, we have:

$$\mathbb{E}\exp\{\mathbf{u}'(\boldsymbol{\varepsilon} - \boldsymbol{\mu})\} \leq \exp\left\{\frac{\|\mathbf{u}\|_2^2 \cdot \sigma_R^2}{2}\right\}.$$

Note that the requirement that σ_R^2 depends on R but not on n is not strictly necessary, but simplifies the interpretations of the bounds.

Another hypothesis concerns sub-exponentiality for ε_n. A centered random variable X is *sub-exponential* if:

$$\mathbb{E}\exp\{uX\} \leq \exp\left\{\frac{u^2\sigma^2}{2}\right\} \quad \text{for } |u| \leq \frac{1}{c}.$$

A random vector is sub-exponential iff all of its coordinates are.

Note that a sub-Gaussian random variable is formally equivalent to a sub-exponential random variable with $c \equiv 0$. For this reason, in the main statements we will only consider the following assumption.

subE We suppose that the following inequality holds true for any n:

$$\mathbb{E}\exp\left\{u\left(\varepsilon_n-\mu_n\right)\right\}=\mathbb{E}\exp\left\{u\left(\theta_n-f\left(\widehat{\mathbf{S}}_n\right)-\mu_n\right)\right\}\leq\exp\left\{\frac{u^2\sigma_R^2}{2}\right\}\quad\text{for }\ |u|\leq\frac{1}{c_R}$$

where σ_R^2 is a variance proxy and c_R a threshold parameter dependent on R but not on n. Moreover, $\sum_{n=1}^N\mu_n^2\leq N\mu_R^2$.

It is important to recall that a variable respecting **subE** has a bound of the following form:

$$\mathbb{P}\left\{\varepsilon_n\geq\mu_n+t\right\}\leq\exp\left\{-\frac{1}{2}\min\left(\frac{t}{c_R},\frac{t^2}{\sigma_R^2}\right)\right\}.$$

We will see below that bounds of this kind can arise in some cases of real interest.

As the coordinates of $\boldsymbol{\varepsilon}$ are independent of each other, we have:

$$\mathbb{E}\exp\left\{\mathbf{u}'\left(\boldsymbol{\varepsilon}-\boldsymbol{\mu}\right)\right\}\leq\exp\left\{\frac{\|\mathbf{u}\|_2^2\cdot\sigma_R^2}{2}\right\}$$

under $\|\mathbf{u}\|_\infty\leq c_R^{-1}$.

As Assumption **subE** may look daunting at first, the following proposition covers a case in which a variant of the hypothesis holds true. The difference with **subE** is the presence of a leading multiplicative (non-absolute) constant. In the remarks following the main results of the paper, we will allow for the case in which **subE** is replaced by a version with a multiplicative constant.

Proposition 6.1. *Suppose that the estimator $\widehat{\mathbf{S}}$ of $\mathbf{S}\in\mathbb{R}^M$ can be written as $\widehat{\mathbf{S}}:=\frac{1}{R}\sum_{r=1}^R\mathbf{X}_r$ where $\mathbf{X}_r$ are iid with mean $\mathbb{E}\mathbf{X}_r=\mathbf{S}$ and $\boldsymbol{\Sigma}:=\mathbb{E}\mathbf{X}_r\mathbf{X}_r'$, and such that $\|\mathbf{X}_r-\mathbf{S}\|_2\leq B$. Let $f:\mathbb{R}^M\to\mathbb{R}$ be a Lipschitz function, i.e. $|f(\mathbf{x})-f(\mathbf{y})|\leq L\cdot\|\mathbf{x}-\mathbf{y}\|_2$. Then:*

1. *for $|u|\leq c_R^{-1}<\frac{C_2R}{M^2BL}$:*

$$\mathbb{E}e^{u\left\{f\left(\widehat{\mathbf{S}}\right)-f(\mathbf{S})\right\}}\leq C\left(R,B,L,M,\operatorname{tr}\left(\boldsymbol{\Sigma}\right)\right)\cdot\exp\left\{\frac{L^2\lambda_{\max}\left(\boldsymbol{\Sigma}\right)}{R}u^2\right\},$$

where the precise value of $C\left(R,B,L,M,\operatorname{tr}\left(\boldsymbol{\Sigma}\right)\right)$ is derived in the proof of the result;

2. *with probability $1-\delta$:*

$$\begin{aligned}\left|f\left(\widehat{\mathbf{S}}\right)-f\left(\mathbf{S}\right)\right|\leq&\frac{M^2BL}{C_2R}\cdot\ln\frac{2C_1M^2}{\delta}+\sqrt{\frac{2L^2\lambda_{\max}\left(\boldsymbol{\Sigma}\right)}{R}\ln\frac{4}{\delta}}+LR^{-\frac{1}{2}}\cdot\left(\operatorname{tr}\left(\boldsymbol{\Sigma}\right)\right)^{\frac{1}{2}}\\\leq&\max\left\{\frac{2M^2BL}{C_2R}\cdot\ln\frac{\max\left\{2C_1M^2,4\right\}}{\delta},\right.\\&\left.\sqrt{\frac{8L^2\lambda_{\max}\left(\boldsymbol{\Sigma}\right)}{R}\cdot\ln\frac{\max\left\{2C_1M^2,4\right\}}{\delta}}\right\}+LR^{-\frac{1}{2}}\cdot\left(\operatorname{tr}\left(\boldsymbol{\Sigma}\right)\right)^{\frac{1}{2}}\end{aligned}$$

for two absolute constants $C_1>0$ and $C_2>0$;

3. for $t \geq 0$:

$$\begin{aligned}
&\mathbb{P}\left\{\left|f\left(\widehat{\mathbf{S}}\right)-f\left(\mathbf{S}\right)\right|>t+LR^{-\frac{1}{2}}\cdot\left(\operatorname{tr}\left(\mathbf{\Sigma}\right)\right)^{\frac{1}{2}}\right\}\\
\leq&\max\left\{2C_1M^2,4\right\}\cdot\exp\left\{-\min\left\{\frac{C_2R}{2M^2BL}\cdot t,\frac{R}{8L^2\lambda_{\max}\left(\mathbf{\Sigma}\right)}\cdot t^2\right\}\right\}.
\end{aligned}$$

Remark 6.1. (i) The assumption of boundedness of $\mathbf{X}_r$ can be weakened (see Theorem 1.1 and Remark 1.1 in [505]).
(ii) If all the other arguments are fixed, the constant $C\left(R,B,L,M,\operatorname{tr}\left(\mathbf{\Sigma}\right)\right)$ converges to 1 provided $R^{\frac{1}{2}}c_R\to\infty$. Moreover, it asymptotically behaves as

$$C\left(R,B,L,M,\operatorname{tr}\left(\mathbf{\Sigma}\right)\right)\simeq 1+\frac{L\left(\operatorname{tr}\left(\mathbf{\Sigma}\right)\right)^{\frac{1}{2}}}{R^{\frac{1}{2}}c_R}+O\left(\frac{1}{R\left(c_R\wedge c_R^2\right)}\right).$$

(iii) The origin of the constants C_1 and C_2 is explained in the proof of the result.

6.3.3 Real-world statistics

For the real-world data that are used to calibrate the model, we formulate an assumption that is similar to **subE** above.

subE' We suppose that the following inequality holds true:

$$\mathbb{E}\exp\left\{u\left(f\left(\widehat{\mathbf{s}}\right)-\mathbb{E}f\left(\widehat{\mathbf{s}}\right)\right)\right\}\leq\exp\left\{\frac{u^2\sigma_S^2}{2}\right\}\quad\text{for }\ |u|\leq\frac{1}{c_S}$$

where σ_S^2 is a variance proxy and c_S a threshold parameter dependent on S but not on n. Moreover, $\left|\mathbb{E}f\left(\widehat{\mathbf{s}}\right)-f\left(\mathbf{s}\right)\right|\leq\mu_S$.

If $c_S=0$, then we will say that **subG'** holds true.

6.3.4 Approximating dictionaries

A major difference of our approach with respect to the one in [86] is that, in the regression, we replace the statistics $(S_1,S_2,\dots,S_M)$ with a set of functions of these statistics. In order to approximate the functions $f^{(k)}$ for $k=1,\dots,K$, we consider a dictionary $\mathcal{P}_P=\left\{p_{1P}\left(\cdot\right),\dots,p_{PP}\left(\cdot\right)\right\}$, indexed by P, composed of functions $p_{jP}:\mathbb{R}^M\to\mathbb{R}$. The index P refers to the number of items in the dictionary. On the basis of the set $\mathcal{P}_P$, it is possible to build the vector $p_P\left(\cdot\right)=\left(p_{1P}\left(\cdot\right),\dots,p_{PP}\left(\cdot\right)\right)'$. In most cases, $\mathcal{P}_P$ will be a subset of size P of an infinite sequence $\mathcal{P}$.

Example 6.1. [Linear form] The method of [86] is obtained when $p_{jP}\left(\mathbf{x}\right)=x_j$ for any j. Therefore, $\mathcal{P}_P=\left\{1,x_1,x_2,\dots,x_M\right\}$.

In the most common case, the functions are basis functions such that a linear combination of these functions can approximate f. The number of these functions, P, is allowed to increase with the number of observations.

Example 6.2. [Monomials] One of the simplest examples concerns low-degree monomials. Suppose to have just one variable defined over $[0,1]$. Then we have:

$$\mathcal{P} = \left\{1, x_1, x_1^2, x_1^3, \dots\right\}.$$

Example 6.3. [Orthogonal polynomials] A solution with better properties is to use orthogonal polynomials. As an example, Legendre polynomials on $[0,1]$ are:

$$\mathcal{P} = \left\{1, 2x_1 - 1, 6x_1^2 - 6x_1 + 1, \dots\right\}$$

but several other options are possible.

Example 6.4. [Tensor product space of monomials] When more than one variable is involved, a common solution is to consider a set like the one in Example 6.2 for each x_m in $(x_1, x_2, \dots, x_M)$, say $\mathcal{P}_{m,p} = \left\{1, x_m, x_m^2, \dots, x_m^p\right\}$. Then one takes the tensor product of these sets, i.e. the set of all interactions obtained multiplying the elements of each $\mathcal{P}_{m,p}$. This can be represented using multi-indexes. Let $\boldsymbol{\lambda} \in \mathbb{N}_0^M$ be a multi-index of size M, and let Λ_P be a set of multi-indices of cardinality P. For a multi-index $\boldsymbol{\lambda} \in \mathbb{N}_0^M$ and a vector $\mathbf{x} = (x_1, \dots, x_M)$, we define the monomial $\mathbf{x}^{\boldsymbol{\lambda}} = \prod_{j=1}^M x_j^{\lambda_j}$. For a set of multi-indexes Λ_P, we define:

$$\mathcal{P}_P = \left\{\mathbf{x}^{\boldsymbol{\lambda}}, \boldsymbol{\lambda} \in \Lambda_P\right\}. \tag{6.3.4}$$

The choice $\Lambda_P = \left\{\boldsymbol{\lambda} \in \mathbb{N}_0^M, \lambda_j \leq p, j = 1, \dots, M\right\}$ provides the desired set. Therefore, $P = (p+1)^M$. If we define the row p-vectors $\mathbf{x}_i := \begin{bmatrix} 1 & x_i & x_i^2 & \cdots & x_i^{p-1} \end{bmatrix}$ for $i = 1, \dots, M$, we can write the tensor product as:

$$p_P(\mathbf{x}) = \bigotimes_{i=1}^{M} \mathbf{x}_i \tag{6.3.5}$$

where $\bigotimes$ denotes the Kronecker product.

Example 6.5. [Total degree space of monomials] The previous solution has the defect that the highest degree is equal to pM, but it does not contain all the monomials of this degree (e.g., it contains $\prod_{m=1}^M x_m^p$ but not x_m^{pM}). Another solution is to arrange the previous set of monomials according to the degree and cut off all those that are larger than p. This is called a total degree space in [340]. In this case we have $\Lambda_P = \left\{\boldsymbol{\lambda} \in \mathbb{N}_0^M, \sum_{j=1}^M \lambda_j \leq p\right\}$ where the number of terms is:

$$P = \frac{(M+p)!}{p!M!}.$$

For large p, we have $P \sim \frac{p^M}{M!}$.

Example 6.6. [Tensor product space of orthogonal polynomials] Very much in the same way as we defined the tensor product of monomials, we can define a tensor product of orthogonal polynomials.

Let $\{\pi_0, \pi_1, \dots\}$ be a sequence of orthogonal polynomials with scalar argument. Then we define:

$$\mathcal{P}_P = \left\{ \prod_{j=1}^{M} \pi_{\lambda_j}(x_j), \boldsymbol{\lambda} \in \Lambda_P \right\}. \tag{6.3.6}$$

With the choice $\Lambda_P = \left\{ \boldsymbol{\lambda} \in \mathbb{N}_0^M, \lambda_j \leq p, j = 1, \dots, M \right\}$, these tensor products can be written as $p_P(\mathbf{x}) = \bigotimes_{i=1}^{M} p_p(x_i)$.

Other examples are in [109], [13], [354], [241], [88] and [40].

We define a measure of approximability of the function by the dictionary as:

$$N_P := \sup_{\mathbf{x} \in \mathbb{R}^M} \inf_{\boldsymbol{\beta} \in \mathbb{R}^P} \left| f(\mathbf{x}) - p_P'(\mathbf{x}) \cdot \boldsymbol{\beta} \right|.$$

Several results concerning the rate of decreases of N_P are available in the literature.

Example 6.7. [Tensor product space of monomials] From [305, Theorem 8, p. 90], if d is the number of continuous derivatives of $f : \mathbb{R}^M \to \mathbb{R}$, we can see that:

$$N_P \lesssim M \cdot p^{-d} \sim M \cdot P^{-\frac{d}{M}}.$$

The values taken by the dictionary $\mathcal{P}_P$ in a point $\boldsymbol{\theta}_n$ is a P-vector $p_P(\mathbf{S}_n)$. These values are stacked in a $(P \times N)$-matrix $p_P(\mathbf{S})$.

We need two further definitions. First, we consider ξ_P defined by:

$$\xi_P := \sup_{1 \leq j \leq P} \sqrt{\frac{1}{N} \sum_{n=1}^{N} p_{jP}^2(\mathbf{S}_n)}.$$

The quantity ξ_P^2 is the maximum value on the diagonal of the Gram matrix $\boldsymbol{\Xi}_N := \frac{1}{N} p_P(\mathbf{S}) p_P'(\mathbf{S})$, or the entrywise maximum norm of $\boldsymbol{\Xi}_N$. Note that a suitable normalization of the design matrix can make $\xi_P = 1$, but this introduces dependence among the rows of the matrix. The next Proposition 6.2 gives an upper bound on ξ_P^2.

Proposition 6.2. *Under **Par** and **Der**, we have:*

$$\sup_{1 \leq j \leq P} \left| \frac{1}{N} \sum_{n=1}^{N} p_{jP}^2(\mathbf{S}_n) - \int_{\mathbb{R}^K} p_{jP}^2(\mathbf{g}(\mathbf{x})) \mathbb{P}(\mathrm{d}\mathbf{x}) \right| \leq C(M, K) \cdot \zeta_K^2(P) \cdot D_{N,\mathbb{P}}$$

or:

$$\xi_P^2 \leq \sup_{1 \leq j \leq P} \int_{\mathbb{R}^K} p_{jP}^2(\mathbf{g}(\mathbf{x})) \mathbb{P}(\mathrm{d}\mathbf{x}) + C(M, K) \cdot \zeta_K^2(P) \cdot D_{N,\mathbb{P}}$$

and:

$$\xi_P^2 \geq \sup_{1 \leq j \leq P} \int_{\mathbb{R}^K} p_{jP}^2(\mathbf{g}(\mathbf{x})) \mathbb{P}(\mathrm{d}\mathbf{x}) - C(M, K) \cdot \zeta_K^2(P) \cdot D_{N,\mathbb{P}}.$$

We also need a function measuring the growth of the derivatives of the approximating polynomials. Let us define, for $f : \mathbb{R}^n \to \mathbb{R}$ and a multi-index $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)$ with $|\boldsymbol{\lambda}| := \lambda_1 + \cdots + \lambda_n$, the partial derivative:

$$\partial^{\boldsymbol{\lambda}} f\left(\mathbf{x}\right) = \frac{\partial^{|\boldsymbol{\lambda}|} f\left(\mathbf{x}\right)}{\partial x_1^{\lambda_1} \ldots \partial x_n^{\lambda_n}}.$$

Then we define:

$$\zeta_k\left(P\right) := \sup_{1 \leq j \leq P} \sup_{\boldsymbol{\lambda} : |\boldsymbol{\lambda}| \leq k} \sup_{\mathbf{x}} \left|\partial^{\boldsymbol{\lambda}} p_{jP}\right|.$$

It is clear that $\xi_P \leq \zeta_0\left(P\right)$ and that $\zeta_k\left(P\right)$ is increasing in k.

As to $\zeta_k\left(P\right)$, we provide some hints about its computation in some special cases.

Example 6.8. [Monomials] For a monomial, the j-th element of $\mathcal{P}_P$ is x_1^{j-1} and its k-th derivative is $\frac{(j-1)!}{(j-1-k)!} x_1^{j-1-k}$. Therefore:

$$\zeta_k\left(P\right) = \frac{(P-1)!}{(P-1-k)!} \leq \left(P-1\right)^k.$$

Example 6.9. [Orthogonal polynomials] For orthogonal polynomials, we can use Markov brothers' inequality (see, e.g., [446]), valid for any polynomial of degree n:

$$\sup_{x \in [0,1]} \left|\pi^{(k)}\left(x\right)\right| \leq 2^k \frac{n^2\left(n^2-1\right)\cdots\left(n^2-\left(k-1\right)^2\right)}{1 \cdot 3 \cdots \left(2k-1\right)} \sup_{x \in [0,1]} \left|\pi\left(x\right)\right| \leq n^{2k} \sup_{x \in [0,1]} \left|\pi\left(x\right)\right|.$$

If we impose the normalization $\sup_{1 \leq j \leq P} \sup_{x \in [0,1]} \left|p_{jP}\left(x\right)\right| \leq 1$, we have:

$$\zeta_k\left(P\right) = \sup_{1 \leq j \leq P} \sup_{\boldsymbol{\lambda} : |\boldsymbol{\lambda}| \leq k} \sup_{x} \left|\partial^{\boldsymbol{\lambda}} p_{jP}\right| \leq c_k\left(P-1\right)^{2k}.$$

In this case the exponent of P is much worse than in Example 6.8, but better exponents can be obtained in special cases.

Example 6.10. [Tensor product space of monomials] In the case of a tensor product, let $\partial^{\boldsymbol{\lambda}} p_{jP}\left(\mathbf{x}\right) = \prod_{i=1}^M \frac{\partial^{\lambda_i} x_i^{k_i}}{\partial x_1^{\lambda_i}}$ for a multi-index $\boldsymbol{\lambda}$. We have:

$$\begin{aligned}\partial^{\boldsymbol{\lambda}} p_{jP} = & \prod_{i=1}^M \frac{\partial^{\lambda_i} x_i^{k_i}}{\partial x_1^{\lambda_i}} = \prod_{i=1}^M \frac{k_i!}{(k_i-\lambda_i)!} x_1^{k_i-\lambda_i} \\ \leq & \prod_{i=1}^M \frac{k_i!}{(k_i-\lambda_i)!} \leq \prod_{i=1}^M k_i^{\lambda_i} \leq p^{\sum_{i=1}^M \lambda_i} = p^{|\boldsymbol{\lambda}|}\end{aligned}$$

from which $\zeta_k\left(P\right) = p^k = \left(P^{\frac{1}{M}} - 1\right)^k \leq P^{\frac{k}{M}}$.

Example 6.11. [Total degree space of monomials] In this case, it is easy to see that $\zeta_k\left(P\right) = p^k$. Using the expression of P for large p, we see that $\zeta_k\left(P\right) \sim \left(M!P\right)^{\frac{k}{M}}$.

6.3.5 The restricted eigenvalue condition

In order to clarify this condition, we repeat the reasoning in [57, p. 1710].

Let $\mathbf{X}$ be, in this paragraph, the $N \times P$ design matrix of the regression problem that we denoted elsewhere as $p_P(\mathbf{S})$. OLS is based on the assumption that the matrix $\mathbf{X}'\mathbf{X}$ is invertible or that its smallest eigenvalue is bounded away from 0. Let us define the *Gram matrix* $\Xi_N := \frac{1}{N}\mathbf{X}'\mathbf{X}$. The *Rayleigh-Ritz quotient* of Ξ_N is:

$$\frac{\boldsymbol{\delta}'\Xi_N\boldsymbol{\delta}}{\boldsymbol{\delta}'\boldsymbol{\delta}} = \frac{\boldsymbol{\delta}'\mathbf{X}'\mathbf{X}\boldsymbol{\delta}}{N\boldsymbol{\delta}'\boldsymbol{\delta}} = \left(\frac{\|\mathbf{X}\boldsymbol{\delta}\|_2}{\sqrt{N}\,\|\boldsymbol{\delta}\|_2}\right)^2.$$

In terms of this quotient, the smallest eigenvalue of the Gram matrix is:

$$\lambda_{\min}(\Xi_N) = \min_{\boldsymbol{\delta}\neq\mathbf{0}} \sqrt{\frac{\boldsymbol{\delta}'\mathbf{X}'\mathbf{X}\boldsymbol{\delta}}{N\boldsymbol{\delta}'\boldsymbol{\delta}}} = \min_{\boldsymbol{\delta}\neq\mathbf{0}} \frac{\|\mathbf{X}\boldsymbol{\delta}\|_2}{\sqrt{N}\,\|\boldsymbol{\delta}\|_2}.$$

A consequence of this assumption is that the matrix $\mathbf{X}$ must have more rows than columns or, otherwise stated, that the number of observations must be larger than the number of variables.

For the Lasso this condition can be weakened as the minimum can be replaced over a smaller set of possible vectors (or directions) $\boldsymbol{\delta}$. In particular, it has been shown (and we will show it in the proofs of the results below) that the following set of directions can be considered. Let J_0 be the set of non-zero coefficients of the vector $\boldsymbol{\beta}$. The Lasso respects with a probability close to 1 the following condition:

$$\left\|\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)_{J_0^c}\right\|_1 \leq c_0 \left\|\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)_{J_0}\right\|_1$$

for a strictly positive constant c_0. This means that the error in the estimates of the zero coefficients (i.e. $\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)_{J_0^c}$) is somewhat smaller than the error in the estimates of the non-zero coefficients (i.e. $\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)_{J_0}$). Now, in the proof below we will need conditions on $\left\|\mathbf{X}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right\|_2$, thus suggesting that we can identify $\boldsymbol{\delta}$ above with $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$. Therefore, we will introduce a condition on the Rayleigh-Ritz quotient of Ξ_N for those directions $\boldsymbol{\delta}$ such that $\left\|\boldsymbol{\delta}_{J_0^c}\right\|_1 \leq c_0 \left\|\boldsymbol{\delta}_{J_0}\right\|_1$. This explains why we are interested in the (modified) Rayleigh-Ritz quotient of Ξ_N:

$$\min_{\boldsymbol{\delta}\neq\mathbf{0},\left\|\boldsymbol{\delta}_{J_0^c}\right\|_1 \leq c_0\left\|\boldsymbol{\delta}_{J_0}\right\|_1} \frac{\|\mathbf{X}\boldsymbol{\delta}\|_2^2}{N\left\|\boldsymbol{\delta}_{J_0}\right\|_2^2} = \left(\min_{\boldsymbol{\delta}\neq\mathbf{0},\left\|\boldsymbol{\delta}_{J_0^c}\right\|_1 \leq c_0\left\|\boldsymbol{\delta}_{J_0}\right\|_1} \frac{\|\mathbf{X}\boldsymbol{\delta}\|_2}{\sqrt{N}\left\|\boldsymbol{\delta}_{J_0}\right\|_2}\right)^2.$$

As we do not know the true J_0, we will express the condition as a minimum over all possible J_0. At last, we will limit the relation to those J_0 with $|J_0| \leq s$. Here s is a measure of "sparsity", i.e. of the maximum number of non-zero coefficients we are allowed to consider.

We have the following assumption:

RE(s, c_0) For some integer s such that $1 \leq s \leq P$ and a positive number c_0, the following condition holds:

$$\kappa(s, c_0) := \min_{J_0 \subseteq \{1,\ldots,P\}, |J_0| \leq s} \; \min_{\boldsymbol{\delta}\neq\mathbf{0},\left\|\boldsymbol{\delta}_{J_0^c}\right\|_1 \leq c_0\left\|\boldsymbol{\delta}_{J_0}\right\|_1} \frac{\|\mathbf{X}\boldsymbol{\delta}\|_2}{\sqrt{N}\left\|\boldsymbol{\delta}_{J_0}\right\|_2} > 0.$$

We pass to the second assumption. Consider integers s, m such that $1 \leq s \leq P/2$, $m \geq s$ and

$s+m \leq P$. Take a vector $J_0 \subseteq \{1, \ldots, P\}$ with $|J_0| \leq s$. Let J_1 be the subset of $J \backslash J_0$ containing the m coordinates largest in absolute value. Let $J_{01} := J_0 \cup J_1$. We have the following assumption:

RE(s, m, c_0) For two integers s and m such that $1 \leq s \leq P/2$, $m \geq s$ and $s+m \leq P$, and a positive number c_0, the following condition holds:

$$\kappa\left(s, m, c_0\right) := \min_{J_0 \subseteq \{1,\ldots,P\}, |J_0| \leq s} \; \min_{\boldsymbol{\delta} \neq \mathbf{0}, \left\|\boldsymbol{\delta}_{J_0^c}\right\|_1 \leq c_0 \left\|\boldsymbol{\delta}_{J_0}\right\|_1} \frac{\left\|\mathbf{X}\boldsymbol{\delta}\right\|_2}{\sqrt{N}\left\|\boldsymbol{\delta}_{J_{01}}\right\|_2} > 0$$

where J_{01} is defined above.

Provided s respects the conditions necessary for **RE**(s, m, c_0) to be well defined, the only difference is in the denominators, as $\left\|\boldsymbol{\delta}_{J_{01}}\right\|_2 \geq \left\|\boldsymbol{\delta}_{J_0}\right\|_2$. As a result, $\kappa\left(s, m, c_0\right) \leq \kappa\left(s, c_0\right)$ and **RE**(s, c_0) is weaker than **RE**(s, m, c_0). Some further considerations on these and other assumptions are in [57, pp. 1711-1714] or in [482].

In the following proposition we provide a lower bound for the restricted eigenvalue. Let $\boldsymbol{\Xi}$ be the matrix whose (j,k)-element is:

$$[\boldsymbol{\Xi}]_{(j,k)} = \int_{\mathbb{R}^M} p_{jP}\left(\mathbf{g}\left(\mathbf{x}\right)\right) p_{kP}\left(\mathbf{g}\left(\mathbf{x}\right)\right) \mathbb{P}\left(\mathrm{d}\mathbf{x}\right)$$

where $\mathbb{P}$ is the probability defined in **Par**. Let $\lambda_{\min}\left(\boldsymbol{\Xi}\right)$ be its smallest eigenvalue.

Proposition 6.3. *Under **Par**, **Der**, **RE**$(s, m, 3)$, we have:*

$$\kappa\left(s, m, 3\right) \geq \lambda_{\min}\left(\boldsymbol{\Xi}\right) - \sqrt{C\left(M, K\right) \cdot s \cdot D_{N,\mathbb{P}}} \cdot \zeta_K\left(P\right).$$

Characterizing the behavior of $\lambda_{\min}\left(\boldsymbol{\Xi}\right)$ is more difficult. However, we can provide some hints through some special cases. We will frequently use the following two tricks.

First, the $(P \times N)$-matrix $p_P\left(\mathbf{S}\right)$ in our applications is often characterized by the fact of having P comparable with or even larger than N. However, in order to study $\lambda_{\min}\left(\boldsymbol{\Xi}\right)$ we can use results for $p_P\left(\mathbf{S}\right)$ with N larger than P and consider what happens to $\boldsymbol{\Xi}_N := \frac{1}{N} p_P\left(\mathbf{S}\right) p_P'\left(\mathbf{S}\right)$ when N diverges. This is eased by the fact that most results quoted are really independent of N.

Second, we recall that, for a generic matrix $\mathbf{A}$, the (*Euclidean*) *condition number* of $\mathbf{A}$ is given by:

$$\kappa\left(\mathbf{A}\right) := \sqrt{\frac{\lambda_{\max}\left(\mathbf{A}'\mathbf{A}\right)}{\lambda_{\min}\left(\mathbf{A}'\mathbf{A}\right)}}.$$

If the matrix $\mathbf{A}$ is symmetric, then:

$$\kappa\left(\mathbf{A}\right) = \frac{\lambda_{\max}\left(\mathbf{A}\right)}{\lambda_{\min}\left(\mathbf{A}\right)}.$$

Writing:

$$\lambda_{\min}\left(\boldsymbol{\Xi}\right) = \frac{\lambda_{\min}\left(\boldsymbol{\Xi}\right)}{\lambda_{\max}\left(\boldsymbol{\Xi}\right)} \cdot \lambda_{\max}\left(\boldsymbol{\Xi}\right) = \frac{\lambda_{\max}\left(\boldsymbol{\Xi}\right)}{\kappa\left(\boldsymbol{\Xi}\right)},$$

we reduce the problem to bounding $\lambda_{\max}\left(\boldsymbol{\Xi}\right)$ from below and $\kappa\left(\boldsymbol{\Xi}\right)$ from above. Now, the maximum eigenvalue $\lambda_{\max}(\boldsymbol{\Xi})$ can be bounded from above by the trace:

$$\lambda_{\max}(\boldsymbol{\Xi}) \leq \operatorname{tr}(\boldsymbol{\Xi}) \leq P \cdot \sup_{1\leq j\leq P} \int_{\mathbb{R}^K} p_{jP}^2(\mathbf{g}(\mathbf{x}))\,\mathbb{P}(\mathrm{d}\mathbf{x})$$

and from below, as in [468, Eq. (2.2)], by the largest diagonal element:

$$\lambda_{\max}(\boldsymbol{\Xi}) \geq \sup_{1\leq j\leq P} \int_{\mathbb{R}^K} p_{jP}^2(\mathbf{g}(\mathbf{x}))\,\mathbb{P}(\mathrm{d}\mathbf{x}).$$

We recall that a *Vandermonde matrix* is a $(N \times P)$-matrix of the form:

$$\mathbf{V}_P(\mathbf{x}) := \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{P-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{P-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^{P-1} \end{bmatrix}$$

where $\mathbf{x}$ is a generic row P-vector $\begin{bmatrix} 1 & x & x^2 & \cdots & x^{P-1} \end{bmatrix}$, while a *polynomial Vandermonde* or *Vandermonde-like matrix* (see [183]) is a $(N \times P)$-matrix of the form:

$$\mathbf{V}_P(p_P(\mathbf{x})) := \begin{bmatrix} p_0(x_1) & p_1(x_1) & p_2(x_1) & \cdots & p_{P-1}(x_1) \\ p_0(x_2) & p_1(x_2) & p_2(x_2) & \cdots & p_{P-1}(x_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_0(x_N) & p_1(x_N) & p_2(x_N) & \cdots & p_{P-1}(x_N) \end{bmatrix}.$$

These matrices may be relevant because:

$$\boldsymbol{\Xi} = \lim_{N\to\infty} \frac{1}{N} \mathbf{V}'_P(\mathbf{x}) \cdot \mathbf{V}_P(\mathbf{x}).$$

Now, we consider some special cases. The analysis is mostly restricted to the case $M = 1$, but the results suggest that the case $M > 1$ should not be too different. Moreover, most examples cover the case $M = K$.

Example 6.12. [Monomials] In the case of monomials, $\boldsymbol{\Xi}$ is a Hankel matrix defined by the equality:

$$[\boldsymbol{\Xi}]_{(j,k)} = \int_{\mathbb{R}} x^{j+k-2}\,\mathbb{P}(\mathrm{d}x) = \mu_{j+k-2}$$

where μ_n is the n-th non-central moment of $\mathbb{P}(\cdot)$. The matrix $\boldsymbol{\Xi}$ is well studied in the literature and is related to the moment problem. A rather general result is proved in [497] where it is shown that, under certain conditions on $\mathbb{P}$:

$$\lambda_{\min}(\boldsymbol{\Xi}) \sim A P^{\frac{1}{2}} B^P$$

for constants $A > 0$ and $0 < B < 1$ that can be characterized. Several other cases have been considered in the literature, see for example [497, 89, 43, 44]; in particular, [43, 44] consider the case in which $\lambda_{\min}(\Xi)$ is bounded away from 0 (the so-called indeterminate case of the moment problem). One can have a confirmation through the literature on condition numbers of Hankel and Vandermonde matrices. Theorems 3.6 and 4.1 in [37] (see also Corollaries 5.3 and 5.5 in [38]) state that $\kappa(\Xi)$ grows at least exponentially in P when the support of the points is any bounded or unbounded interval. The literature does not seem to contain upper bounds applicable to our case.

Example 6.13. [Orthogonal polynomials] The picture changes when we turn away from monomials to orthogonal polynomials. In this case the matrix can be well conditioned or even optimally conditioned (i.e. all eigenvalues are equal to 1). Theorems 2 and 4 in [276] apply to polynomial Vandermonde matrices and show that a careful choice of the integration measure–rather difficult to achieve in pratice–can lead to $\lambda_{\min}(\Xi) = 1$. A more general result (see [40, Proposition 2.1]) shows that $\lambda_{\min}(\Xi)$ may be bounded away from 0. Suppose that $M = K$ and $\mathcal{P}_P$ is a sequence of orthonormal polynomials with respect to a measure μ. Let $\mathbb{Q}$ be the image measure of $\mathbb{P}$ under $\mathbf{g}$ and suppose that the Radon-Nikodým derivative of $\mathbb{Q}$ with respect to μ is bounded from below on its support, i.e. ${}^{\mathrm{d}\mathbb{Q}}/_{\mathrm{d}\mu} \geq c > 0$. Then:

$$\begin{aligned}\boldsymbol{\lambda}'\Xi\boldsymbol{\lambda} =& \int_{\mathbb{R}} \left[\boldsymbol{\lambda}' p_P(y)\right]^2 \mathbb{Q}(\mathrm{d}y) = \int_{\mathbb{R}} \left[\boldsymbol{\lambda}' p_P(y)\right]^2 \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mu}\mu(\mathrm{d}y)\\ \geq& c\int_{\mathbb{R}} \left[\boldsymbol{\lambda}' p_P(y)\right]^2 \mu(\mathrm{d}y) = c\boldsymbol{\lambda}'\boldsymbol{\lambda}\end{aligned}$$

from which $\lambda_{\min}(\Xi) \geq c > 0$.

Example 6.14. [Tensor products] As in (6.3.5), suppose that $p_P(\mathbf{x}) = \bigotimes_{i=1}^M p_p(x_i)$. Let $\mathbb{Q}$ be the image measure of $\mathbb{P}$ under $\mathbf{g}$ and suppose that the Radon-Nikodým derivative of $\mathbb{Q}$ with respect to a product measure $\mu = \bigotimes_{i=1}^M \mu_i$ is bounded from below on its support, i.e. ${}^{\mathrm{d}\mathbb{Q}}/_{\mathrm{d}\mu} \geq c > 0$. Then:

$$\begin{aligned}\boldsymbol{\lambda}'\Xi\boldsymbol{\lambda} =& \int_{\mathbb{R}} \left[\boldsymbol{\lambda}' p_P(\mathbf{y})\right]^2 \mathbb{Q}(\mathrm{d}\mathbf{y}) = \int_{\mathbb{R}} \left[\boldsymbol{\lambda}' \bigotimes_{i=1}^M p_p(y_i)\right]^2 \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mu}\mu(\mathrm{d}\mathbf{y})\\ \geq& c\boldsymbol{\lambda}'\left\{\int_{\mathbb{R}} \left[\bigotimes_{i=1}^M p_p(y_i)\cdot \bigotimes_{i=1}^M p_p'(y_i)\right] \bigotimes_{i=1}^M \mu_i(\mathrm{d}y_i)\right\}\boldsymbol{\lambda}\\ =& c\boldsymbol{\lambda}'\left\{\int_{\mathbb{R}} \left[\bigotimes_{i=1}^M p_p(y_i)\, p_p'(y_i)\right] \bigotimes_{i=1}^M \mu_i(\mathrm{d}y_i)\right\}\boldsymbol{\lambda}\\ =& c\boldsymbol{\lambda}'\left(\bigotimes_{i=1}^M \Xi_i\right)\boldsymbol{\lambda} \geq c\boldsymbol{\lambda}'\boldsymbol{\lambda}\prod_{i=1}^M \lambda_{\min}(\Xi_i)\end{aligned}$$

where $\Xi_i := \int_{\mathbb{R}} p_p(y_i)\, p_p'(y_i)\, \mu_i(\mathrm{d}y_i)$.

The previous examples show that there is an advantage in passing from a linear combination of monomials to a combination of orthogonal polynomials. However, we can do much better using a trick that is quite common in econometrics (see [353, 354]). The trick consists in exploiting the invariance of the estimator to the choice of the basis by modifying the dictionary from $p_P(\cdot)$ to $p_P^\star(\cdot)$. The objective is to use this new dictionary $p_P^\star(\cdot)$ to compute an improved rate of convergence. We are not proposing to use $p_P^\star(\cdot)$ in estimation, but simply to replace $p_P(\cdot)$ with $p_P^\star(\cdot)$ in the proofs of the properties of $\left|\widehat{\theta}-\theta\right|$. There may be computational advantages in using $p_P^\star(\cdot)$ but, apart from these advantages, it is largely immaterial whether one uses $p_P(\cdot)$ or $p_P^\star(\cdot)$ because of the invariance of the estimator to the choice of the basis. Here is the main assumption.

Ort For every P there is a constant and nonsingular matrix $\mathbf{B}$ such that, for $p_P^\star(\cdot)=\mathbf{B}p_P(\cdot)$, the smallest eigenvalue of the matrix $\boldsymbol{\Xi}^\star$ defined by:

$$\boldsymbol{\Xi}^\star := \int_{\mathbb{R}^M} p_P^\star(\mathbf{g}(\mathbf{x}))\, p_P^{\star,\prime}(\mathbf{g}(\mathbf{x}))\, \mathbb{P}(\mathrm{d}\mathbf{x}) = \int_{\mathbb{R}^M} p_P^\star(\mathbf{y})\, p_P^{\star,\prime}(\mathbf{y})\, \mathbb{Q}(\mathrm{d}\mathbf{y})$$

is bounded away from zero uniformly in P.

The following proposition is a reformulation of Lemma A.15 in [353]. We make the hypothesis that the multi-index set Λ_P is *downward closed* with respect to the partial order in $\mathbb{N}_0^M$, i.e. if $\boldsymbol{\lambda}\in\Lambda_P$ and $\mu\le\boldsymbol{\lambda}$ then $\mu\in\Lambda_P$ (see [340, p. 140]).

Proposition 6.4. *Suppose that the multi-index set Λ_P is downward closed with respect to the partial order in $\mathbb{N}_0^M$. Suppose that the space can be reduced to the unit hypercube $[0,1]^M$. Let the density of $\mathbb{Q}$ be bounded from below by $C\prod_{j=1}^M y_j^\nu(1-y_j)^\nu$. Then, there is a choice of $\mathbf{B}$ such that **Ort** holds true, N_P is unchanged and one can take $\zeta_P(k)\le C\cdot P^{\frac{1}{2}+\nu+2k}$.*

6.4 Main results

The following is our main result.

Theorem 6.1. *Let $1>\delta>0$ be a constant. Under **subE**, let:*

$$\tau := 2\left\{\max\left\{\frac{\sigma_R\xi_P}{N^{\frac{1}{2}}}\sqrt{2\ln\frac{2P}{\delta}}, \frac{2c_R\zeta_0(P)}{N}\ln\frac{2P}{\delta}\right\}+\xi_P(\mu_R+N_P)\right\}.$$

*Under **RE**$(s,3)$, with probability at least $1-\delta$:*

$$\left\|\widehat{\boldsymbol{\beta}}_{J_0}-\boldsymbol{\beta}_{J_0}\right\|_2\le\frac{3\tau s^{\frac{1}{2}}}{\kappa^2(s,3)}$$

and:

$$\left\|p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\|_2\le\frac{3\tau N^{\frac{1}{2}}s^{\frac{1}{2}}}{\kappa(s,3)}.$$

*Under **RE**$(s,m,3)$, with probability at least $1-\delta$:*

$$\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_2\le\left(1+9\frac{s}{m}\right)^{\frac{1}{2}}\frac{3\tau s^{\frac{1}{2}}}{\kappa^2(s,m,3)}.$$

Under **subE**, **subE'** *and* **RE**$(s,m,3)$*, with probability at least* $1-\delta-\Delta$*:*

$$\left|\widehat{\theta}-\theta\right| \leq \left(1+9\frac{s}{m}\right)^{\frac{1}{2}} \frac{3\zeta_0\left(P\right)\tau s^{\frac{1}{2}} P^{\frac{1}{2}}}{\kappa^2\left(s,m,3\right)} + N_P + \max\left\{\sigma_S\sqrt{2\ln\frac{2}{\Delta}}, 2c_S\ln\frac{2}{\Delta}\right\} + \mu_S.$$

Remark 6.2. (i) The results under Assumptions **subG** and **subG'** are obtained setting, respectively, $c_R \equiv 0$ and $c_S \equiv 0$.
(ii) If Assumptions **subG** and **subE** are supposed to hold with a leading constant, i.e. $\mathbb{E}\exp\left\{u\left(\varepsilon_n-\mu_n\right)\right\} \leq C_R\exp\left\{\frac{u^2\sigma_R^2}{2}\right\}$ either for all u of for $|u| \leq \frac{1}{c_R}$, the result of Theorem 6.1 still holds by replacing $\ln\frac{2P}{\delta}$ in the formula for τ with $\ln\frac{2C_RP}{\delta}$. The same holds true for Assumptions **subG'** and **subE'**.

Let us suppose that $\mu_R = O\left(R^{-1}\right)$, $\sigma_R = O\left(R^{-\frac{1}{2}}\right)$, $\mu_S = O\left(S^{-1}\right)$, $\sigma_S = O\left(S^{-\frac{1}{2}}\right)$, $\xi_P \simeq 1$ and $N_P = O\left(P^{-\alpha}\right)$. Then, under **subG** and **subG'**, we have:

$$\tau \asymp \sqrt{\frac{\ln P - \ln\delta}{NR}} + R^{-1} + P^{-\alpha}$$

and, with probability at least $1-\delta-\Delta$:

$$\begin{aligned}\left|\widehat{\theta}-\theta\right| &\leq \left(1+9\frac{s}{m}\right)^{\frac{1}{2}} \frac{3\zeta_0\left(P\right)\tau s^{\frac{1}{2}} P^{\frac{1}{2}}}{\kappa^2\left(s,m,3\right)} + N_P + \max\left\{\sigma_S\sqrt{2\ln\frac{2}{\Delta}}, 2c_S\ln\frac{2}{\Delta}\right\} + \mu_S \\ &\lesssim \left(1+\frac{s}{m}\right)^{\frac{1}{2}} \frac{\zeta_0\left(P\right)s^{\frac{1}{2}}}{\kappa^2\left(s,m,3\right)}\left(\sqrt{\frac{P\left(\ln P-\ln\delta\right)}{NR}} + P^{\frac{1}{2}}R^{-1} + P^{\frac{1}{2}-\alpha}\right) + P^{-\alpha} + S^{-\frac{1}{2}}\sqrt{\left|\ln\Delta\right|} + S^{-1}.\end{aligned}$$

Moreover, suppose that **RE**$(s,m,3)$ holds true. Now we consider two scenarios.

- Suppose that θ can be expressed in such a way that J_0 has finite cardinality. In this case, s is finite and:

  $$\left|\widehat{\theta}-\theta\right| \lesssim \frac{\zeta_0\left(P\right)}{\kappa^2\left(s,m,3\right)}\left(\sqrt{\frac{P\left(\ln P-\ln\delta\right)}{NR}} + P^{\frac{1}{2}}R^{-1} + P^{\frac{1}{2}-\alpha}\right) + P^{-\alpha} + S^{-\frac{1}{2}}\sqrt{\left|\ln\Delta\right|} + S^{-1}.$$

  Let us fix δ and Δ:

  $$\left|\widehat{\theta}-\theta\right| \lesssim \frac{\zeta_0\left(P\right)}{\kappa^2\left(s,m,3\right)}\left(\sqrt{\frac{P\ln P}{NR}} + R^{-1} + P^{-\alpha}\right) + P^{-\alpha} + S^{-\frac{1}{2}}.$$

- Suppose that θ can be expressed as a function of a subset of dimension $M_0 =: \chi M$, for $\chi \in (0,1)$, of the M statistics involved in the estimation process. We take the tensor products of monomials as dictionary. If $p-1$ is the maximal degree of each variable, we have $P \sim p^M$. Moreover, $s \sim p^{M_0}$ and we take $s = m$. Then:

  $$\left|\widehat{\theta}-\theta\right| \lesssim \frac{\zeta_0\left(P\right)}{\kappa^2\left(s,s,3\right)}p^{M_0}\left(\sqrt{\frac{p^M\left(M\ln p-\ln\delta\right)}{NR}} + p^{\frac{M}{2}}R^{-1} + p^{\left(\frac{1}{2}-\alpha\right)M}\right) + p^{-\alpha M} + S^{-\frac{1}{2}}\sqrt{\left|\ln\Delta\right|} + S^{-1}.$$

Let us fix δ and Δ:

$$\left|\widehat{\theta}-\theta\right| \lesssim \frac{\zeta_0\left(P\right)}{\kappa^2\left(s,s,3\right)}\left(\sqrt{\frac{p^{(2\chi+1)M}\ln p}{NR}}+p^{\left(\chi+\frac{1}{2}\right)M}R^{-1}+p^{\left(\chi+\frac{1}{2}-\alpha\right)M}\right)+p^{-\alpha M}+S^{-\frac{1}{2}}.$$

6.5 Simulation experiments

The simulation study described in this section, follows the procedure exposed in Section 5.7 of Chapter 5 on nonparametric moment-based estimation via OLS. Below we detail two examples, the estimation of the mean μ and of the standard deviation σ of a Gaussian random variable. As explained in Section 5.7 of Chapter 5, we choose to estimate μ, in order to cover the parametric case, and σ to deal with the nonparametric situation. For the sake of clarity, σ can be expressed as a linear combination of an infinite number of basis functions, but this number is of a smaller order than the order of elements of the dictionary. Both the estimation of μ and σ are therefore characterized by sparsity. Nevertheless, in both cases the variance seems to reach the Cramér-Rao Lower Bound (CRLB), whose come.

The algorithm follows the steps elucidated in Section 5.7 of Chapter 5. The main difference relies on the estimation of function $f_P^{(j)}$, that is carried out exploiting Lasso regression with a tuning parameter (λ) selected through cross-validation.

We replicate the procedure $10,000$ times, with 16 different configurations. Table 6.5.1 describes the parametrization of the simulation study. The values of μ and σ for the training sample are selected according to equispaced grids (see Table 6.5.1 for the range and the cardinality of the grid). A two-dimensional set of points in Θ is then built through a full factorial design of the values of μ and σ. The parameter values can be chosen according to other methods; we pursue this technique for simplicity. The value of μ for the test sample is determined fixing $\sigma\equiv 1$ and taking μ on a grid covering $[-0.76,+0.76]$. The values of σ for the test sample are determined fixing $\mu\equiv 0$ and taking σ on a grid covering $[0.6,1.5]$. The cardinalities of μ and σ for the test samples are specified in Table 6.5.1.

6.5.1 Estimation of the mean of a Gaussian random variable

Figure 6.5.1 and 6.5.2 describe, respectively, the behavior of the bias and the variance of the estimator of μ. The thin line with the same style as the corresponding variance of the estimator, in the variance plot, represent the CRLB.

Figure 6.5.1 shows that the bias of the estimator of μ is small and decreases when the number of simulations, $S=R$, and the number of training points, N, increase. Two phenomena are at work here: first, for small S and R the regressors are biased estimators of the same quantities for S and R equal to infinity, as the polynomials are nonlinear functions of unbiased statistics; second, the Lasso is in itself a biased estimator of the function $f_P^{(j)}$.

The variance of the estimator of μ, exhibited in Figure 6.5.2, seems to reach the CRLB. Hence, the estimator seems to be efficient.

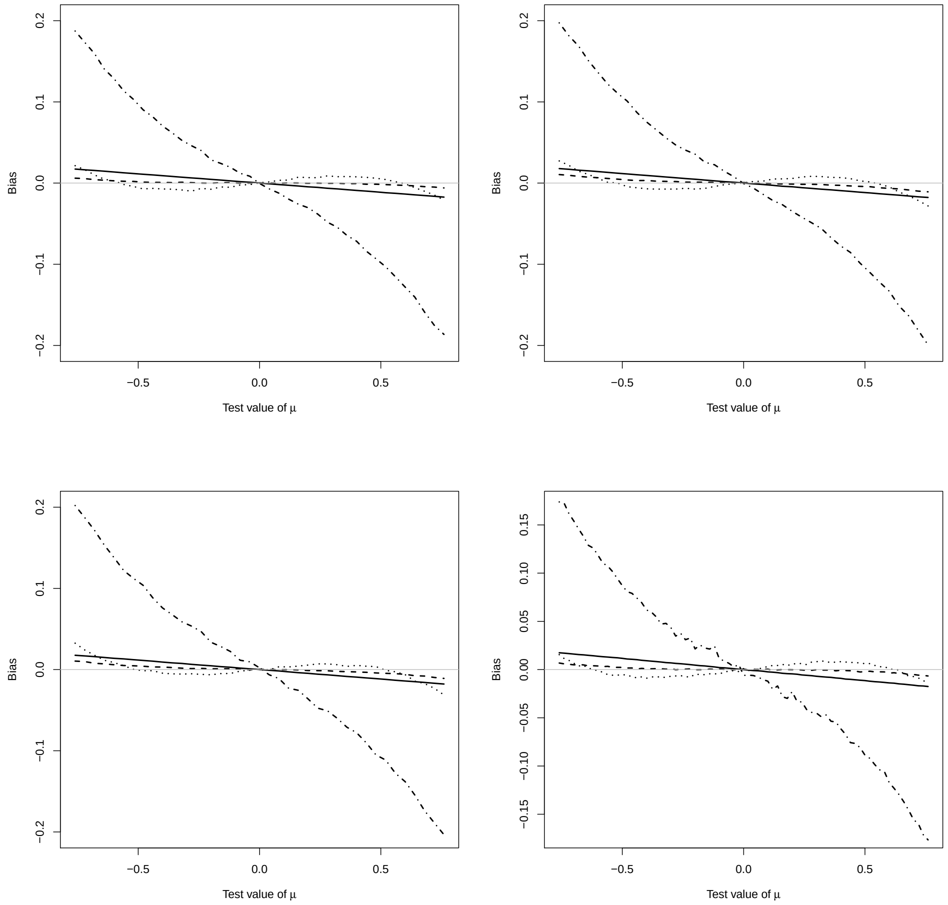

Figure 6.5.1: Bias of the estimator of μ with $R = 10$ (dash-dot line), $R = 100$ (dotted line), $R = 1,000$ (dashed line) and $R = 10,000$ (solid lime), for different configurations of parameters: $p = 4$ and $N = 980$ (top left), $p = 6$ and $N = 980$ (top right), $p = 6$ and $N = 3960$ (bottom left) and $p = 9$ and $N = 980$ (bottom right).

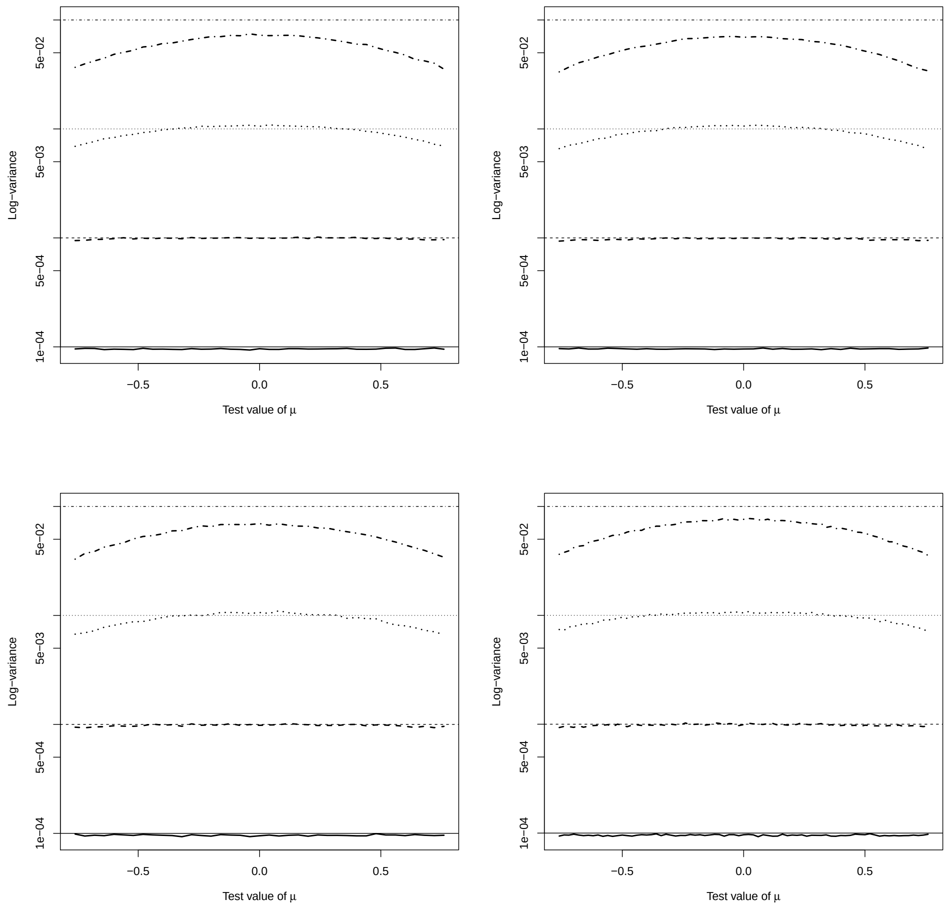

Figure 6.5.2: Variance of the estimator of μ with $R = 10$ (dash-dot line), $R = 100$ (dotted line), $R = 1,000$ (dashed line) and $R = 10,000$ (solid line), for different configurations of parameters: $p = 4$ and $N = 980$ (top left), $p = 6$ and $N = 980$ (top right), $p = 9$ and $N = 980$ (bottom left) and $p = 6$ and $N = 3,960$ (bottom right).

Table 6.5.1: Parametrization of the simulation experiment

			Training sample				Test sample	Test sample
p	$R = S$	N	μ		σ		for μ	for σ
			card.	range	card.	range	card.	card.
4	10	980	49	$[-0.96, +0.96]$	20	$[1, 2]$	39	10
4	100	980	49	$[-0.96, +0.96]$	20	$[1, 2]$	39	10
4	1,000	980	49	$[-0.96, +0.96]$	20	$[1, 2]$	39	10
4	10,000	980	49	$[-0.96, +0.96]$	20	$[1, 2]$	39	10
6	10	980	49	$[-0.96, +0.96]$	20	$[1, 2]$	39	10
6	100	980	49	$[-0.96, +0.96]$	20	$[1, 2]$	39	10
6	1,000	980	49	$[-0.96, +0.96]$	20	$[1, 2]$	39	10
6	10,000	980	49	$[-0.96, +0.96]$	20	$[1, 2]$	39	10
9	10	980	49	$[-0.96, +0.96]$	20	$[1, 2]$	39	10
9	100	980	49	$[-0.96, +0.96]$	20	$[1, 2]$	39	10
9	1,000	980	49	$[-0.96, +0.96]$	20	$[1, 2]$	39	10
9	10,000	980	49	$[-0.96, +0.96]$	20	$[1, 2]$	39	10
6	10	3960	99	$[-0.98, +0.98]$	40	$[1, 2]$	77	19
6	100	3960	99	$[-0.98, +0.98]$	40	$[1, 2]$	77	19
6	1,000	3960	99	$[-0.98, +0.98]$	40	$[1, 2]$	77	19
6	10,000	3960	99	$[-0.98, +0.98]$	40	$[1, 2]$	77	19

Note: “card.” stands for “cardinality”.

6.5.2 Estimation of the standard deviation of a Gaussian random variable

Figure 6.5.3 and 6.5.4 describe, respectively, the behavior of the bias and the variance of the estimator of σ. The thin line with the same style as the corresponding variance of the estimator, in the variance plot, represent the CRLB.

When we move to the estimation of σ, which is a nonparametric function of the statistics, the bias of the estimator is not negligible. Furthermore, its shape as a function of the parameter values, is not immediately intelligible (see 6.5.3). Nonetheless, it decreases as the order of polynomials, P, and the number of simulated observations, $S = R$, increase. In this case three phenomena are at work: as above, for small S and R the regressors are biased estimators of their asymptotic values and the Lasso is a biased estimator of $f_P^{(j)}$; however, the function $f_P^{(j)}$ is a biased estimator of the function $f^{(j)}$, as the approximation of the function by a linear combination of basis functions neglect a bias term given by $f^{(j)} - f_P^{(j)}$.

Also in this case, the variance of the estimator of σ reaches the CRLB (see 6.5.4). Therefore, the estimator seems to be efficient.

6.6 Conclusions

In this paper, we propose a new method for the nonparametric estimation of simulated models. Starting from the work of [86], we draw on their contribution adding sieve estimation to the Lasso regression, in order to capture the nonlinear and unknown relations between the statistics used as regressors and the parameters to be estimated, and rigorously characterizing the asymptotic behavior

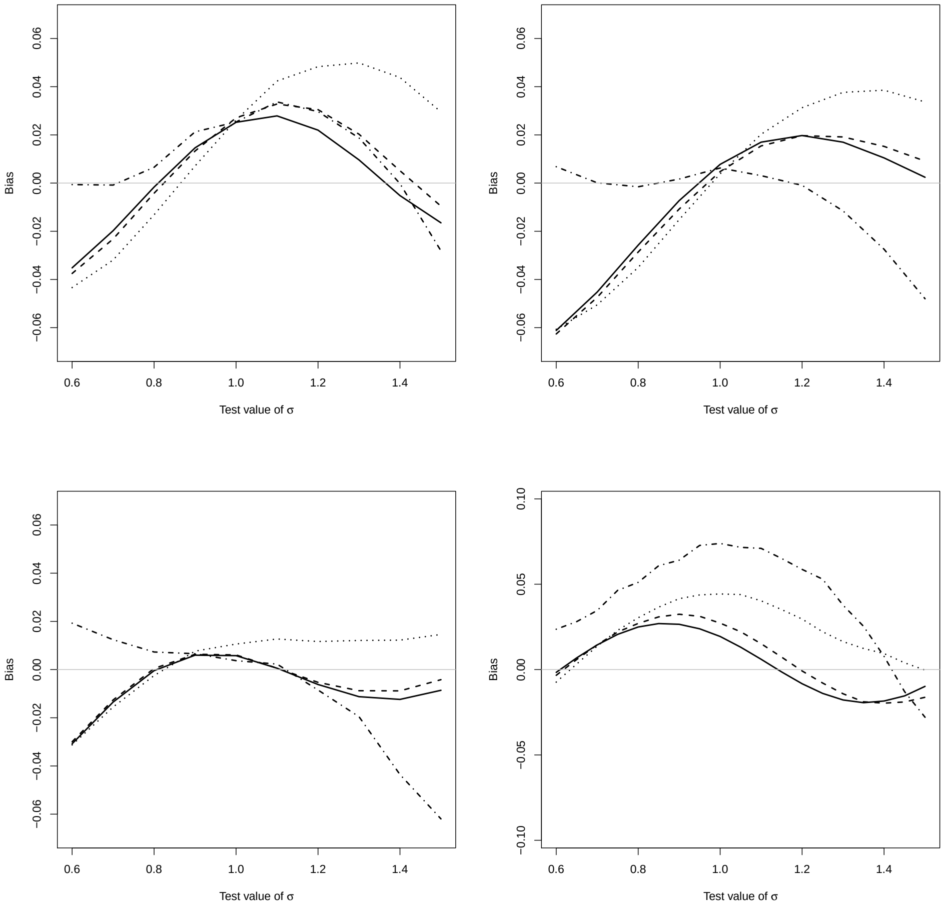

Figure 6.5.3: Bias of the estimator of σ with $R = 10$ (dash-dot line), $R = 100$ (dotted line), $R = 1,000$ (dashed line) and $R = 10,000$ (solid line), for different configurations of parameters: $p = 4$ and $N = 980$ (top left), $p = 6$ and $N = 980$ (top right), $p = 9$ and $N = 980$ (bottom left) and $p = 6$ and $N = 3960$ (bottom right).

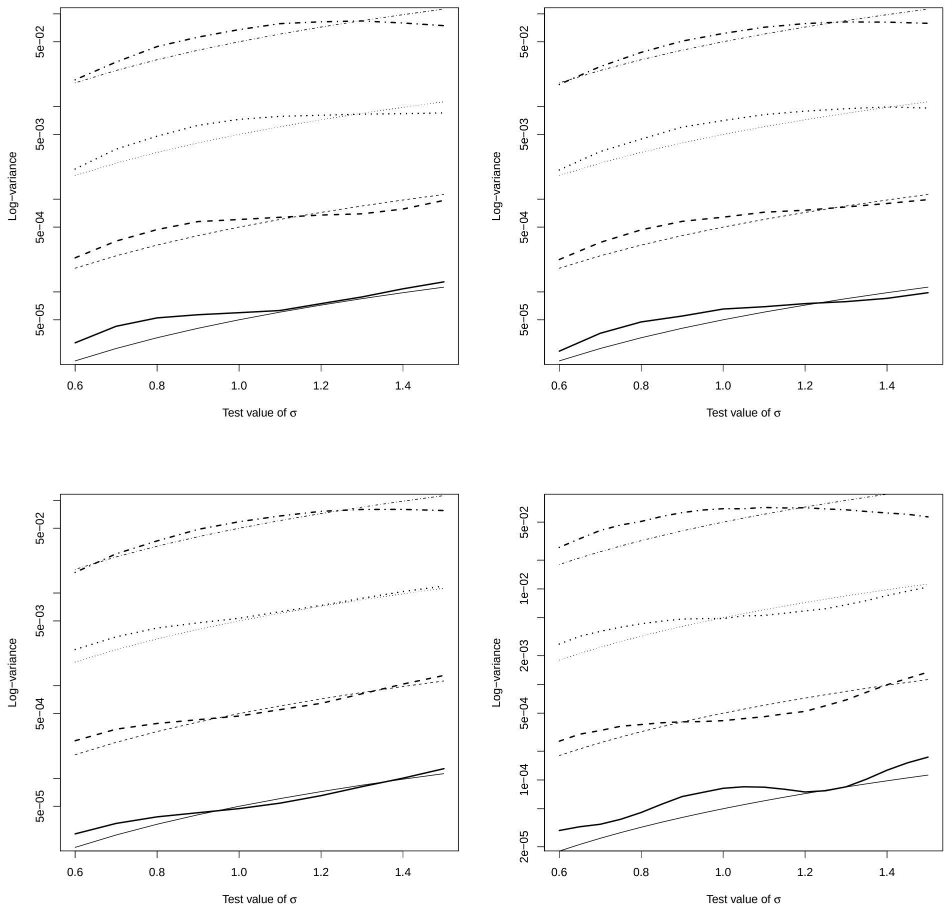

Figure 6.5.4: Variance of the estimator of σ with $R = 10$ (dash-dot line), $R = 100$ (dotted line), $R = 1,000$ (dashed line) and $R = 10,000$ (solid line), for different configurations of parameters: $p = 4$ and $N = 980$ (top left), $p = 6$ and $N = 980$ (top right), $p = 9$ and $N = 980$ (bottom left) and $p = 6$ and $N = 3,960$ (bottom right).

of the estimator $\widehat{\theta}^{(j)}$.

The asymptotic rate of convergence of $\widehat{\theta}^{(j)}$ to $\theta^{(j)}$ is generally sub-optimal as we consider a nonparametric estimation technique, but the loss of efficiency in terms of convergence speed is compensated by the flexibility deriving from the nonparametric estimation of the function linking the statistics and the parameters. The asymptotic results on the error in the estimation of the coefficients and the prediction are good and quite general.

Finally, the simulation experiment concerning the finite-sample properties of the estimator shows that, when the parameter to be estimated is a parametric function of the limiting values of the statistics (e.g., can be expressed as a finite linear combination of the elements of the dictionary used in the regression), the bias is negligible and tends to decrease when S, the number of real-world observations, R, the number of simulations, and N, the number of training points, increase. Instead, when we deal with a parameter that can be represented as a linear combination of an infinite number of basis functions (e.g., a nonparametric relation between the parameter to be estimated and the statistics is involved), although the bias seems to decrease as the order of polynomials and the number of simulated observations increase, it is not negligible and its shape as a function of the parameter values is not directly interpretable. In both cases the variance of the estimator reaches the Cramér-Rao lower bound, hence the estimator seems to be efficient.

6.7 Proofs

6.7.1 Auxiliary results

Lemma 6.1. *Let X be a non-negative random variable with a tail inequality $1 - F_X(x) \leq ce^{-Cx}$ for $x > 0$. Then, for $0 \leq u < C$:*

$$\int_0^{\infty} e^{ux} \mathrm{d}F_X(x) \leq c^{\frac{u}{C}} \frac{C}{C-u}.$$

Proposition 6.5. *Under **Par**, **Der**, we have:*

$$\left| \left[\widehat{\mathbf{\Pi}} \right]_{(j,k)} - [\mathbf{\Pi}]_{(j,k)} \right| \leq C(M,K) \cdot \zeta_K(P) \cdot (\zeta_K(P) + \zeta_0(P)) \cdot D_{N,\mathbb{P}}.$$

6.7.2 Proofs of auxiliary results

Proof of Lemma 6.1. We use the alternative expectation formula applied to the mgf of a positive random variable. By integration by parts:

$$
\begin{aligned}
-\int_0^\infty e^{ux}\mathrm{d}F_X(x) &= \int_0^\infty e^{ux}\mathrm{d}\left(1-F_X(x)\right)\\
&= e^{ux}\left(1-F_X(x)\right)\Big|_0^\infty - u\int_0^\infty \left(1-F_X(x)\right)e^{ux}\mathrm{d}x,\\
\int_0^\infty e^{ux}\mathrm{d}F_X(x) &= -\,e^{ux}\left(1-F_X(x)\right)\Big|_0^\infty + u\int_0^\infty \left(1-F_X(x)\right)e^{ux}\mathrm{d}x.
\end{aligned}
$$

To guarantee that the first term is bounded, we use a bound of the form $1-F_X(x)\le ce^{-Cx}$. The term is bounded, provided $u<C$, and yields:

$$
\int_0^\infty e^{ux}\mathrm{d}F_X(x) = 1 + u\int_0^\infty \left(1-F_X(x)\right)e^{ux}\mathrm{d}x.
$$

Using the bound $1-F_X(x)\le ce^{-Cx}$ again, we get:

$$
1-F_X(x)\le \min\left\{1, ce^{-Cx}\right\} = \begin{cases} 1 & x\le \frac{\ln c}{C}\\ ce^{-Cx} & x>\frac{\ln c}{C}\end{cases}
$$

and:

$$
\int_0^\infty e^{ux}\mathrm{d}F_X(x)\le 1+u\int_0^{\frac{\ln c}{C}} e^{ux}\mathrm{d}x + cu\int_{\frac{\ln c}{C}}^\infty e^{(u-C)x}\mathrm{d}x = 1+u\frac{c^{\frac{u}{C}}-1}{u} + cu\frac{c^{\frac{u}{C}-1}}{C-u} = c^{\frac{u}{C}}\frac{C}{C-u}.
$$

QED

Proof of Proposition 6.1. Let $\mathbf{Y}_r := \mathbf{X}_r - \mathbf{S}$, let $\mathbf{G}$ be a zero-mean Gaussian vector with variance $\mathbf{\Sigma}$, and let $\mathbf{G}^\star := \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{G}$. We can write $\widehat{\mathbf{S}} = \frac{1}{R}\sum_{r=1}^R \mathbf{Y}_r + \mathbf{S}$. Therefore, we have the following decomposition that will be used in the following:

$$
\begin{aligned}
f\left(\frac{1}{R}\sum_{r=1}^R \mathbf{Y}_r + \mathbf{S}\right) - f\left(\mathbf{S}\right) =& f\left(\frac{1}{R}\sum_{r=1}^R \mathbf{Y}_r + \mathbf{S}\right) - f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)\\
&+ f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right) - \mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)\\
&+ \mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right) - f\left(\mathbf{S}\right).
\end{aligned}
$$

Note that the first and the third terms can be majorized as:

$$\left|f\left(\frac{1}{R}\sum_{r=1}^{R}\mathbf{Y}_r+\mathbf{S}\right)-f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)\right| \leq \frac{L}{R}\cdot\left\|\sum_{r=1}^{R}\mathbf{Y}_r-R^{\frac{1}{2}}\mathbf{G}\right\|_2, \tag{6.7.1}$$

$$\begin{aligned}\left|\mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)-f\left(\mathbf{S}\right)\right| \leq& L\cdot\mathbb{E}\left\|R^{-\frac{1}{2}}\mathbf{G}\right\|_2 \leq L\cdot\left(\mathbb{E}\left\|R^{-\frac{1}{2}}\mathbf{G}\right\|_2^2\right)^{\frac{1}{2}}\\ =&LR^{-\frac{1}{2}}\cdot\left(\mathbb{E}\mathbf{G}'\mathbf{G}\right)^{\frac{1}{2}} = LR^{-\frac{1}{2}}\cdot\left(\mathrm{tr}\left(\mathbf{\Sigma}\right)\right)^{\frac{1}{2}}.\end{aligned} \tag{6.7.2}$$

Let us start from the mgf. Using the Cauchy-Schwarz inequality, we have:

$$\begin{aligned}\mathbb{E}e^{u\left\{f\left(\frac{1}{R}\sum_{r=1}^{R}\mathbf{Y}_r+\mathbf{S}\right)-f(\mathbf{S})\right\}} \leq& \sqrt{\mathbb{E}e^{2u\left\{f\left(\frac{1}{R}\sum_{r=1}^{R}\mathbf{Y}_r+\mathbf{S}\right)-f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)\right\}}}\\ &\cdot\sqrt{\mathbb{E}e^{2u\left\{f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)-\mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)\right\}}}\cdot e^{u\left\{\mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)-f(\mathbf{S})\right\}}.\end{aligned} \tag{6.7.3}$$

From [386, pp. 180-181], the second term in (6.7.3) can be majorized as:

$$\begin{aligned}\mathbb{E}e^{2u\left\{f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)-\mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)\right\}} &\leq \exp\left\{2\frac{L^2\lambda_{\max}\left(\mathbf{\Sigma}\right)}{R}u^2\right\},\\ \sqrt{\mathbb{E}e^{2u\left\{f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)-\mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)\right\}}} &\leq \exp\left\{\frac{L^2\lambda_{\max}\left(\mathbf{\Sigma}\right)}{R}u^2\right\}.\end{aligned}$$

The first term in (6.7.3) can be majorized through (6.7.1) as:

$$\mathbb{E}e^{2u\left\{f\left(\frac{1}{R}\sum_{r=1}^{R}\mathbf{Y}_r+\mathbf{S}\right)-f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)\right\}} \leq \mathbb{E}e^{\frac{2|u|L}{R}\cdot\left\|\sum_{r=1}^{R}\mathbf{Y}_r-R^{\frac{1}{2}}\mathbf{G}\right\|_2}.$$

We apply Lemma 6.1 with $c=C_1M^2$ and $C=\frac{C_2R}{M^2BL}$. Therefore:

$$\begin{aligned}\sqrt{\mathbb{E}e^{2u\left\{f\left(\frac{1}{R}\sum_{r=1}^{R}\mathbf{Y}_r+\mathbf{S}\right)-f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)\right\}}} \leq& \sqrt{\mathbb{E}e^{\frac{2|u|L}{R}\cdot\left\|\sum_{r=1}^{R}\mathbf{Y}_r-R^{\frac{1}{2}}\mathbf{G}\right\|_2}}\\ \leq& \sqrt{\frac{C_2R}{C_2R-2M^2BL\left|u\right|}}\cdot\exp\left\{\frac{M^2BL\left|u\right|}{C_2R}\ln\left(C_1M^2\right)\right\}.\end{aligned}$$

The third term in (6.7.3) can be majorized through (6.7.2). At last, we have:

$$\mathbb{E}e^{u\left\{f\left(\widehat{\mathbf{S}}\right)-f(\mathbf{S})\right\}} \leq \sqrt{\frac{C_2R}{C_2R-2M^2BL\left|u\right|}}\cdot\exp\left\{\left[\frac{M^2BL}{C_2R}\ln\left(C_1M^2\right)+\frac{L\left(\mathrm{tr}\left(\mathbf{\Sigma}\right)\right)^{\frac{1}{2}}}{R^{\frac{1}{2}}}\right]\left|u\right|+\frac{L^2\lambda_{\max}\left(\mathbf{\Sigma}\right)}{R}u^2\right\},$$

for $|u|<\frac{C_2R}{M^2BL}$. Over $|u|\leq c_R^{-1}$, with $c_R^{-1}<\frac{C_2R}{M^2BL}$:

$$\mathbb{E}e^{u\left\{f\left(\widehat{\mathbf{S}}\right)-f(\mathbf{S})\right\}} \leq \sqrt{\frac{C_2Rc_R}{C_2Rc_R-2M^2BL}}\cdot\exp\left\{\left[\frac{M^2BL}{C_2R}\ln\left(C_1M^2\right)+\frac{L\left(\mathrm{tr}\left(\mathbf{\Sigma}\right)\right)^{\frac{1}{2}}}{R^{\frac{1}{2}}}\right]c_R^{-1}+\frac{L^2\lambda_{\max}\left(\mathbf{\Sigma}\right)}{R}u^2\right\}.$$

Now we turn to the majorization of the tail probability. We have:

$$\begin{aligned}
&\left| f\left(\frac{1}{R}\sum_{r=1}^{R}\mathbf{Y}_r + \mathbf{S}\right) - f\left(\mathbf{S}\right)\right| \\
\leq &\left| f\left(\frac{1}{R}\sum_{r=1}^{R}\mathbf{Y}_r + \mathbf{S}\right) - \mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G} + \mathbf{S}\right)\right| + \left|\mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G} + \mathbf{S}\right) - f\left(\mathbf{S}\right)\right| \\
\leq &\left| f\left(\frac{1}{R}\sum_{r=1}^{R}\mathbf{Y}_r + \mathbf{S}\right) - f\left(R^{-\frac{1}{2}}\mathbf{G} + \mathbf{S}\right)\right| \\
&\quad + \left| f\left(R^{-\frac{1}{2}}\mathbf{G} + \mathbf{S}\right) - \mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G} + \mathbf{S}\right)\right| + \left|\mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G} + \mathbf{S}\right) - f\left(\mathbf{S}\right)\right| \\
\leq &\frac{L}{R}\cdot\left\|\sum_{r=1}^{R}\mathbf{Y}_r - R^{\frac{1}{2}}\mathbf{G}\right\|_2 + \left| f\left(R^{-\frac{1}{2}}\mathbf{G} + \mathbf{S}\right) - \mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G} + \mathbf{S}\right)\right| \\
&\quad + \left|\mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G} + \mathbf{S}\right) - f\left(\mathbf{S}\right)\right|.
\end{aligned}$$

We start with the first two terms. We write:

$$\begin{aligned}
&\mathbb{P}\left\{\left| f\left(\frac{1}{R}\sum_{r=1}^{R}\mathbf{Y}_r + \mathbf{S}\right) - \mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G} + \mathbf{S}\right)\right| > t\right\} \\
\leq &\mathbb{P}\left\{\frac{L}{R}\cdot\left\|\sum_{r=1}^{R}\mathbf{Y}_r - R^{\frac{1}{2}}\mathbf{G}\right\|_2 + \left| f\left(R^{-\frac{1}{2}}\mathbf{G} + \mathbf{S}\right) - \mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G} + \mathbf{S}\right)\right| > t\right\} \\
\leq &\mathbb{P}\left\{\left\|\sum_{r=1}^{R}\mathbf{Y}_r - R^{\frac{1}{2}}\mathbf{G}\right\|_2 > \frac{R}{L}\cdot t_1\right\} + \mathbb{P}\left\{\left| f\left(R^{-\frac{1}{2}}\mathbf{G} + \mathbf{S}\right) - \mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G} + \mathbf{S}\right)\right| > t_2\right\} \qquad (6.7.4)
\end{aligned}$$

where $t_1 + t_2 = t$. Now, from [51, p. 155], the first term in (6.7.4) can be bounded as:

$$\mathbb{P}\left\{\left\|\sum_{r=1}^{R}\mathbf{Y}_r - R^{\frac{1}{2}}\mathbf{G}\right\|_2 > \frac{R}{L}\cdot t_1\right\} \leq C_1 M^2 \exp\left\{-\frac{C_2 R}{M^2 BL}\cdot t_1\right\}$$

or, with probability at least $1 - \delta_1$:

$$\left| f\left(\frac{1}{R}\sum_{r=1}^{R}\mathbf{Y}_r + \mathbf{S}\right) - f\left(R^{-\frac{1}{2}}\mathbf{G} + \mathbf{S}\right)\right| \leq \frac{M^2 BL}{C_2 R}\cdot\ln\frac{C_1 M^2}{\delta_1}.$$

Now, we consider the second term in (6.7.4). We plan to apply [291, Ch. 2.3, Eq. (2.35)]. The result applies to a function of a standard normal vector, i.e. to $h\left(\mathbf{x}\right) := f\left(R^{-\frac{1}{2}}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{x} + \mathbf{S}\right)$. We show that h is Lipschitz with constant $LR^{-\frac{1}{2}}\lambda_{\max}^{\frac{1}{2}}\left(\mathbf{\Sigma}\right)$:

$$\left|h\left(\mathbf{x}\right) - h\left(\mathbf{y}\right)\right| \leq \left| f\left(R^{-\frac{1}{2}}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{x} + \mathbf{S}\right) - f\left(R^{-\frac{1}{2}}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{y} + \mathbf{S}\right)\right| \leq LR^{-\frac{1}{2}}\lambda_{\max}^{\frac{1}{2}}\left(\mathbf{\Sigma}\right)\cdot\left\|\mathbf{x} - \mathbf{y}\right\|_2.$$

Therefore, we have:

$$\mathbb{P}\left\{\left|f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)-\mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)\right|>t_2\right\}\leq 2\exp\left\{-\frac{Rt_2^2}{2L^2\lambda_{\max}\left(\mathbf{\Sigma}\right)}\right\}$$

or, with probability at least $1-\delta_2$:

$$\left|f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)-\mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)\right|\leq\sqrt{\frac{2L^2\lambda_{\max}\left(\mathbf{\Sigma}\right)}{R}\ln\frac{2}{\delta_2}}.$$

As a result, taking $\delta_1=\delta_2={}^{\delta}\!/_{2}$:

$$\left|f\left(\frac{1}{R}\sum_{r=1}^{R}\mathbf{Y}_r+\mathbf{S}\right)-\mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)\right|\leq\frac{M^2BL}{C_2R}\cdot\ln\frac{2C_1M^2}{\delta}+\sqrt{\frac{2L^2\lambda_{\max}\left(\mathbf{\Sigma}\right)}{R}\ln\frac{4}{\delta}}.$$

To get the final result, we use (6.7.2).

The tail probability result can be obtained from (6.7.4), letting $t_1=t_2=\frac{t}{2}$:

$$\begin{aligned}
&\mathbb{P}\left\{\left|f\left(\frac{1}{R}\sum_{r=1}^{R}\mathbf{Y}_r+\mathbf{S}\right)-\mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)\right|>t\right\}\\
\leq&\mathbb{P}\left\{\left\|\sum_{r=1}^{R}\mathbf{Y}_r-R^{\frac{1}{2}}\mathbf{G}\right\|_2>\frac{R}{L}\cdot t_1\right\}+\mathbb{P}\left\{\left|f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)-\mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)\right|>t_2\right\}\\
\leq&C_1M^2\exp\left\{-\frac{C_2R}{M^2BL}\cdot t_1\right\}+2\exp\left\{-\frac{R}{2L^2\lambda_{\max}\left(\mathbf{\Sigma}\right)}\cdot t_2^2\right\}\\
\leq&\max\left\{C_1M^2,2\right\}\cdot\left\{\exp\left\{-\frac{C_2R}{M^2BL}\cdot t_1\right\}+\exp\left\{-\frac{R}{2L^2\lambda_{\max}\left(\mathbf{\Sigma}\right)}\cdot t_2^2\right\}\right\}\\
\leq&2\max\left\{C_1M^2,2\right\}\cdot\exp\left\{-\min\left\{\frac{C_2R}{M^2BL}\cdot t_1,\frac{R}{2L^2\lambda_{\max}\left(\mathbf{\Sigma}\right)}\cdot t_2^2\right\}\right\}\\
\leq&\max\left\{2C_1M^2,4\right\}\cdot\exp\left\{-\min\left\{\frac{C_2R}{2M^2BL}\cdot t,\frac{R}{8L^2\lambda_{\max}\left(\mathbf{\Sigma}\right)}\cdot t^2\right\}\right\}.
\end{aligned}$$

At last, through (6.7.2):

$$\begin{aligned}
&\mathbb{P}\left\{\left|f\left(\frac{1}{R}\sum_{r=1}^{R}\mathbf{Y}_r+\mathbf{S}\right)-\mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)\right|>t\right\}\\
=&\mathbb{P}\left\{\left|f\left(\frac{1}{R}\sum_{r=1}^{R}\mathbf{Y}_r+\mathbf{S}\right)-f\left(\mathbf{S}\right)+f\left(\mathbf{S}\right)-\mathbb{E}f\left(R^{-\frac{1}{2}}\mathbf{G}+\mathbf{S}\right)\right|>t\right\}\\
\geq&\mathbb{P}\left\{\left|f\left(\frac{1}{R}\sum_{r=1}^{R}\mathbf{Y}_r+\mathbf{S}\right)-f\left(\mathbf{S}\right)\right|-LR^{-\frac{1}{2}}\cdot\left(\mathrm{tr}\left(\mathbf{\Sigma}\right)\right)^{\frac{1}{2}}>t\right\}.
\end{aligned}$$

QED

Proof of Proposition 6.5. Under **Par**, we have:

$$\left[\widehat{\mathbf{\Pi}}-\mathbf{\Pi}\right]_{(i,j)}=\int_{\mathbb{R}^M}p_{iP}\left(\mathbf{g}\left(\mathbf{x}\right)\right)p_{jP}\left(\mathbf{g}\left(\mathbf{x}\right)\right)\left(\mathbb{P}_N-\mathbb{P}\right)\left(\mathrm{d}\mathbf{x}\right).$$

The Koksma-Hlawka inequality of [192, Corollary 2.4], [225, p. 165], [6] yields:

$$\left|\left[N^{-1}\mathbf{\Pi}-\mathbf{\Pi}_{0P}\right]_{(i,j)}\right|\leq V_{HK}\left(p_{iP}\left(\mathbf{g}\left(\cdot\right)\right)p_{jP}\left(\mathbf{g}\left(\cdot\right)\right)\right)\cdot D_{N,\mathbb{P}}$$

where $V_{HK}\left(f\right)$ is the Hardy-Krause total variation (see, e.g., [370]) and $D_{N,\mathbb{P}}$ is the non-uniform unanchored discrepancy defined in (6.3.3) (see also [192, p. 100] or [225, p. 165]).

The inequality derived from the last formula on p. 251 in [62] for $\varphi\left(x\right)=x$:

$$V_{HK}\left(fg\right)\leq\left(3^K+1-2^{K+1}\right)\cdot V_{HK}\left(f\right)V_{HK}\left(g\right)+\left\|f\right\|_\infty V_{HK}\left(g\right)+\left\|g\right\|_\infty V_{HK}\left(f\right),\tag{6.7.5}$$

yields:

$$\begin{aligned}&V_{HK}\left(p_{iP}\left(\mathbf{g}\left(\cdot\right)\right)p_{jP}\left(\mathbf{g}\left(\cdot\right)\right)\right)\\&\quad\leq\left(3^M+1-2^{M+1}\right)\cdot V_{HK}\left(p_{iP}\circ\mathbf{g}\right)V_{HK}\left(p_{jP}\circ\mathbf{g}\right)\\&\qquad+\left\|p_{iP}\circ\mathbf{g}\right\|_\infty V_{HK}\left(p_{jP}\circ\mathbf{g}\right)+\left\|p_{jP}\circ\mathbf{g}\right\|_\infty V_{HK}\left(p_{iP}\circ\mathbf{g}\right)\\&\quad\leq\left(3^M+1-2^{M+1}\right)\cdot V_{HK}\left(p_{iP}\circ\mathbf{g}\right)V_{HK}\left(p_{jP}\circ\mathbf{g}\right)\\&\qquad+\zeta_0\left(P\right)\cdot V_{HK}\left(p_{jP}\circ\mathbf{g}\right)+\zeta_0\left(P\right)\cdot V_{HK}\left(p_{iP}\circ\mathbf{g}\right).\end{aligned}\tag{6.7.6}$$

Eq. (3) in [33, p. 1948] (see also [370, p. 61]) gives:

$$V_{HK}\left(f\right)\leq\sum_{u\neq\emptyset}\int_{[0,1]^u}\left|\frac{\partial^{|u|}f\left(\boldsymbol{\theta}_u;\mathbf{1}_{-u}\right)}{\partial\boldsymbol{\theta}_u}\right|\mathrm{d}\boldsymbol{\theta}_u\leq\sum_{u\neq\emptyset}\sup_{\boldsymbol{\theta}_u\in[0,1]^u}\left|\frac{\partial^{|u|}f\left(\boldsymbol{\theta}_u;\mathbf{1}_{-u}\right)}{\partial\boldsymbol{\theta}_u}\right|\leq\sum_{u\neq\emptyset}\sup_{\boldsymbol{\theta}\in[0,1]^K}\left|\frac{\partial^{|u|}f\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}_u}\right|,$$

where:

$$\frac{\partial^{|u|}f\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}_u}=\frac{\partial^{|u|}p_{iP}\left(\mathbf{g}\left(\boldsymbol{\theta}\right)\right)}{\partial\boldsymbol{\theta}_u}.$$

As this is a composite function, we apply the results in [101] and [33]. We follow the notation in [101] but we will use some results from [33]. Theorem 2.1 in [101] applies to $h\left(\mathbf{x}\right)=f\left(\mathbf{g}\left(\mathbf{x}\right)\right)$ where $\mathbf{x}\in\mathbb{R}^d$ and $\mathbf{g}\left(\mathbf{x}\right)\in\mathbb{R}^m$. Let $\boldsymbol{\nu}=\left(\nu_1,\dots,\nu_d\right)\in\mathbb{N}_0^d$ be a vector with $n:=\left|\boldsymbol{\nu}\right|\neq0$. For $\mathbf{m}=\left(m_1,\dots,m_d\right)$, let $f_{\mathbf{m}}=\frac{\partial^{|\mathbf{m}|}f}{\partial x_1^{m_1}\cdots\partial x_d^{m_d}}$.

We use [101, Theorem 2.1] or [33, Theorem 2]. As our vector $\boldsymbol{\nu}$ (that we call u in the above passages) belongs to the set $\left\{0,1\right\}^d\setminus\left\{\mathbf{0}\right\}$, we can apply Lemma 3 and the reasoning leading to Eq. (9) in [33]. Therefore, $\boldsymbol{\ell}_j\in\left\{0,1\right\}^d\setminus\left\{\mathbf{0}\right\}$ and $\mathbf{k}_j\in\left\{0,1\right\}^m$ with $\left|\mathbf{k}_j\right|=1$. It is easy to identify $\mathbf{k}_j$ with the (only) integer k_j for which $\mathbf{k}_j$ takes the value 1:

$$\left[\mathbf{g}_{\boldsymbol{\ell}_j}\right]^{\mathbf{k}_j}=\frac{\partial g^{(k_j)}}{\partial x_{\ell_j}}.$$

Under **Der**, for $n = |\boldsymbol{\nu}| \neq 0$, we have:

$$\begin{aligned}
|h_{\boldsymbol{\nu}}| \leq & \boldsymbol{\nu}! \sum_{1 \leq |\boldsymbol{\lambda}| \leq n} |f_{\boldsymbol{\lambda}}| \sum_{s=1}^{n} \sum_{p_s(\boldsymbol{\nu},\boldsymbol{\lambda})} \prod_{j=1}^{s} \frac{\left| \left[\mathbf{g}_{\boldsymbol{\ell}_j} \right]^{\mathbf{k}_j} \right|}{(\mathbf{k}_j!) \left[\boldsymbol{\ell}_j! \right]^{|\mathbf{k}_j|}} \\
\leq & \left(\mu \vee \mu^n \right) \boldsymbol{\nu}! \sum_{k=1}^{n} \sum_{|\boldsymbol{\lambda}|=k} |f_{\boldsymbol{\lambda}}| \sum_{s=1}^{n} \sum_{p_s(\boldsymbol{\nu},\boldsymbol{\lambda})} \prod_{j=1}^{s} \frac{1}{(\mathbf{k}_j!) \left[\boldsymbol{\ell}_j! \right]^{|\mathbf{k}_j|}} \\
\leq & \left(\mu \vee \mu^n \right) \boldsymbol{\nu}! \sum_{k=1}^{n} \sqrt{ \sum_{|\boldsymbol{\lambda}|=k} |f_{\boldsymbol{\lambda}}|^2 \cdot \sum_{|\boldsymbol{\lambda}|=k} \left\{ \sum_{s=1}^{n} \sum_{p_s(\boldsymbol{\nu},\boldsymbol{\lambda})} \prod_{j=1}^{s} \frac{1}{(\mathbf{k}_j!) \left[\boldsymbol{\ell}_j! \right]^{|\mathbf{k}_j|}} \right\}^2 } \\
\leq & \left(\mu \vee \mu^n \right) \boldsymbol{\nu}! \sum_{k=1}^{n} \sqrt{ \left\{ \sum_{|\boldsymbol{\lambda}|=k} |f_{\boldsymbol{\lambda}}|^2 \right\} \cdot \left\{ \sum_{|\boldsymbol{\lambda}|=k} \sum_{s=1}^{n} \sum_{p_s(\boldsymbol{\nu},\boldsymbol{\lambda})} \prod_{j=1}^{s} \frac{1}{(\mathbf{k}_j!) \left[\boldsymbol{\ell}_j! \right]^{|\mathbf{k}_j|}} \right\}^2 } \\
\leq & \left(\mu \vee \mu^n \right) \sum_{k=1}^{n} \left(\sum_{|\boldsymbol{\lambda}|=k} |f_{\boldsymbol{\lambda}}|^2 \right)^{\frac{1}{2}} \cdot m^k S_n^k
\end{aligned}$$

where the second inequality comes from **Der**, the third from Cauchy-Schwarz inequality, the fourth from $\sum_{i=1}^{n} x_i^2 \leq \left(\sum_{i=1}^{n} x_i \right)^2$, and the fifth from Corollary 2.9 in [101, p. 511]. Note that S_n^k is a Stirling number of the second kind. We replace $h(\cdot) = f(\mathbf{g}(\cdot))$ with $p_{iP}(\mathbf{g}(\cdot))$, $\boldsymbol{\nu}$ with u, m with M, and n equals $|u|$, and we use the definition of $\zeta_k(P)$ to get:

$$\begin{aligned}
V_{HK}\left(p_{iP} \circ \mathbf{g} \right) \leq & \sum_{u \neq \emptyset} \left(\mu \vee \mu^n \right) \sum_{k=1}^{n} \left(\sum_{|\boldsymbol{\lambda}|=k} |f_{\boldsymbol{\lambda}}|^2 \right)^{\frac{1}{2}} \cdot M^k S_n^k \\
\leq & \left(\mu \vee \mu^K \right) \left[\sum_{u \neq \emptyset} \sum_{k=1}^{n} \left(\sum_{|\boldsymbol{\lambda}|=k} \zeta_k^2(P) \right)^{\frac{1}{2}} \cdot M^k S_n^k \right] \\
\leq & \left(\mu \vee \mu^K \right) \cdot C(M,K) \cdot \zeta_K(P)
\end{aligned}$$

for a constant $C(M,K)$ depending only on M and K.

Replacing this into (6.7.6), we get:

$$\begin{aligned}
& V_{HK}\left(p_{iP}(\mathbf{g}(\cdot)) \, p_{jP}(\mathbf{g}(\cdot)) \right) \\
& \quad \leq C(M,K) \cdot \zeta_K(P) \cdot \left(\mu \vee \mu^K \right) \cdot \left(\zeta_K(P) \cdot \left(\mu \vee \mu^K \right) + \zeta_0(P) \right) \\
& \quad \leq C(M,K) \cdot \zeta_K(P) \cdot \left(\zeta_K(P) + \zeta_0(P) \right).
\end{aligned}$$

QED

Proof of Proposition 6.2. The proposition is simply a corollary of Proposition 6.5. Indeed, the result derives from:

$$\sup_{1\leq j\leq P}\left(\frac{1}{N}\sum_{n=1}^{N}p_{jP}^{2}\left(\mathbf{S}_{n}\right)-\int_{\mathbb{R}^{K}}p_{jP}^{2}\left(\mathbf{g}\left(\mathbf{x}\right)\right)\mathbb{P}\left(\mathrm{d}\mathbf{x}\right)\right)\leq C\left(M,K\right)\cdot\zeta_{K}^{2}\left(P\right)\cdot D_{N,\mathbb{P}}.$$

QED

Proof of Proposition 6.3. Let us define $\boldsymbol{\Xi}_N=\frac{1}{N}p_P'\left(\mathbf{S}\right)p_P\left(\mathbf{S}\right)$. We define a distance as:

$$d_{\infty}\left(\boldsymbol{\Xi}_N,\boldsymbol{\Xi}\right):=\max_{j,k}\left|\left[\boldsymbol{\Xi}_N\right]_{(j,k)}-\left[\boldsymbol{\Xi}\right]_{(j,k)}\right|.$$

Then, from Corollary 10.1 in [482]:

$$\kappa\left(s,m,c_0\right)\geq\lambda_{\min}\left(\boldsymbol{\Xi}\right)-4\sqrt{d_{\infty}\left(\boldsymbol{\Xi}_N,\boldsymbol{\Xi}\right)\cdot s}.$$

Applying Proposition 6.5, we get:

$$\kappa\left(s,m,c_0\right)\geq\lambda_{\min}\left(\boldsymbol{\Xi}\right)-\sqrt{C\left(M,K\right)\cdot s\cdot\zeta_K\left(P\right)\cdot\left(\zeta_K\left(P\right)+\zeta_0\left(P\right)\right)\cdot D_{N,\mathbb{P}}}.$$

QED

Proof of Proposition 6.4. The space of polynomials spanned by $\mathcal{P}_P$ in (6.3.4) and $\mathcal{P}_P$ in (6.3.6) is the same, provided the set Λ_P is downward closed with respect to the partial order in $\mathbb{N}_0^M$. Moreover, the space spanned does not depend on the choice of the polynomials. Therefore, as in [353], we take the explicit choice of normalized Gegenbauer or ultraspherical polynomials $C_n^{(\alpha)}, n\in\mathbb{N}$ (see [366, Chapter 18]). These are the orthogonal polynomials with respect to the weighting function $\left(1-x^2\right)^{\alpha-\frac{1}{2}}$ on $[-1,+1]$ and can be made orthonormal by a simple normalization. In the following, we will use the notation $C_n^{(\alpha)}$ to denote the orthonormal ones. Moreover, as the support of the polynomials is $[-1,+1]$, a scaling of the argument is necessary. As a final result, we consider:

$$\mathcal{P}_P=\left\{\prod_{j=1}^{M}C_{\lambda_j}^{(\alpha)}\left(2x_j-1\right),\lambda\in\Lambda_P\right\}.$$

Let ℓ_P be the largest element of any multi-index in Λ_P. Then we define:

$$\widetilde{\Lambda}_P:=\left\{\boldsymbol{\lambda}\in\mathbb{N}_0^M:\lambda_j\leq\ell_P,1\leq j\leq M\right\}.$$

Note that $\widetilde{\Lambda}_P$ is the set of all multi-indexes with elements ranging from 1 to ℓ_P. Let $\widetilde{\mathcal{P}}_P$ and $\widetilde{p}_P\left(\cdot\right)$ be the set $\mathcal{P}_P$ and the vector $p_P\left(\cdot\right)$ when Λ_P is replaced by $\widetilde{\Lambda}_P$. It is clear that $\widetilde{p}_P\left(\mathbf{y}\right)=\bigotimes_{j=1}^{M}\widetilde{p}_{\ell_P}\left(y_j\right)$. Then, by the Cauchy interlacing theorem, we have:

$$\lambda_{\min}\left\{\int_{\mathbb{R}}p_P\left(\mathbf{y}\right)p_P'\left(\mathbf{y}\right)\mathbb{Q}\left(\mathrm{d}\mathbf{y}\right)\right\}\geq\lambda_{\min}\left\{\int_{\mathbb{R}}\tilde{p}_P\left(\mathbf{y}\right)\tilde{p}_P'\left(\mathbf{y}\right)\mathbb{Q}\left(\mathrm{d}\mathbf{y}\right)\right\}$$

and, by the variational property of the eigenvalues and the positive semi-definiteness of the matrices:

$$\begin{aligned}
&\lambda_{\min}\left\{\int_{\mathbb{R}^M}\tilde{p}_P\left(\mathbf{y}\right)\tilde{p}_P'\left(\mathbf{y}\right)\mathbb{Q}\left(\mathrm{d}\mathbf{y}\right)\right\}\\
&\quad=\lambda_{\min}\left\{\int_{\mathbb{R}^M}\tilde{p}_P\left(\mathbf{y}\right)\tilde{p}_P'\left(\mathbf{y}\right)\frac{\mathbb{Q}\left(\mathrm{d}\mathbf{y}\right)}{\mathrm{d}\mathbf{y}}\mathrm{d}\mathbf{y}\right\}\\
&\quad\geq\lambda_{\min}\left\{\int_{\mathbb{R}^M}\tilde{p}_P\left(\mathbf{y}\right)\tilde{p}_P'\left(\mathbf{y}\right)C\prod_{j=1}^{M}y_j^{\nu}\left(1-y_j\right)^{\nu}\mathrm{d}\mathbf{y}\right\}\\
&\qquad+\lambda_{\min}\left\{\int_{\mathbb{R}^M}\tilde{p}_P\left(\mathbf{y}\right)\tilde{p}_P'\left(\mathbf{y}\right)\left[\frac{\mathbb{Q}\left(\mathrm{d}\mathbf{y}\right)}{\mathrm{d}\mathbf{y}}-C\prod_{j=1}^{M}y_j^{\nu}\left(1-y_j\right)^{\nu}\right]\mathrm{d}\mathbf{y}\right\}\\
&\quad\geq C\lambda_{\min}\left\{\int_{\mathbb{R}^M}\bigotimes_{j=1}^{M}\widetilde{p}_{\ell_P}\left(y_j\right)\widetilde{p}_{\ell_P}'\left(y_j\right)\prod_{j=1}^{M}y_j^{\nu}\left(1-y_j\right)^{\nu}\mathrm{d}\mathbf{y}\right\}\\
&\quad=C\lambda_{\min}\left\{\bigotimes_{j=1}^{M}\int_{\mathbb{R}}\widetilde{p}_{\ell_P}\left(y_j\right)\widetilde{p}_{\ell_P}'\left(y_j\right)y_j^{\nu}\left(1-y_j\right)^{\nu}\mathrm{d}y_j\right\}=C.
\end{aligned}$$

Therefore, **Ort** holds true. The value of N_P is clearly unchanged, as it is invariant with respect to a different choice of polynomials, provided they span the same space. As to $\zeta_P\left(k\right)$, this comes from [353, p. 271]. QED

6.7.3 Proofs of main results

Proof of Theorem 6.1. From the definition of the Lasso, we have, for any $\boldsymbol{\theta}\in\mathbb{R}^N$:

$$\frac{1}{N}\left\|\boldsymbol{\theta}-p_P'\left(\widehat{\mathbf{S}}\right)\cdot\widehat{\boldsymbol{\beta}}\right\|_2^2\leq\frac{1}{N}\left\|\boldsymbol{\theta}-p_P'\left(\widehat{\mathbf{S}}\right)\cdot\boldsymbol{\beta}\right\|_2^2+2\tau\left\|\boldsymbol{\beta}\right\|_1-2\tau\left\|\widehat{\boldsymbol{\beta}}\right\|_1.$$

We multiply by N and we use $\boldsymbol{\theta}=f\left(\mathbf{S}\right)=f\left(\widehat{\mathbf{S}}\right)+\boldsymbol{\varepsilon}$:

$$\begin{aligned}
\left\|f\left(\widehat{\mathbf{S}}\right)+\boldsymbol{\varepsilon}-p_P'\left(\widehat{\mathbf{S}}\right)\cdot\widehat{\boldsymbol{\beta}}\right\|_2^2&\leq\left\|f\left(\widehat{\mathbf{S}}\right)+\boldsymbol{\varepsilon}-p_P'\left(\widehat{\mathbf{S}}\right)\cdot\boldsymbol{\beta}\right\|_2^2+2N\tau\left\|\boldsymbol{\beta}\right\|_1-2N\tau\left\|\widehat{\boldsymbol{\beta}}\right\|_1\\
\left\|f\left(\widehat{\mathbf{S}}\right)-p_P'\left(\widehat{\mathbf{S}}\right)\cdot\widehat{\boldsymbol{\beta}}\right\|_2^2-\left\|f\left(\widehat{\mathbf{S}}\right)-p_P'\left(\widehat{\mathbf{S}}\right)\cdot\boldsymbol{\beta}\right\|_2^2&\leq2\boldsymbol{\varepsilon}'p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)+2N\tau\left\|\boldsymbol{\beta}\right\|_1-2N\tau\left\|\widehat{\boldsymbol{\beta}}\right\|_1.
\end{aligned}$$

We write $\boldsymbol{\eta}:=f\left(\widehat{\mathbf{S}}\right)-p_P'\left(\widehat{\mathbf{S}}\right)\cdot\boldsymbol{\beta}$ for the approximation error. We have:

$$\begin{aligned}
\left\|\boldsymbol{\eta}+p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right\|_2^2-\left\|\boldsymbol{\eta}\right\|_2^2&\leq2\boldsymbol{\varepsilon}'p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)+2N\tau\left\|\boldsymbol{\beta}\right\|_1-2N\tau\left\|\widehat{\boldsymbol{\beta}}\right\|_1\\
\left\|p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right\|_2^2+2\boldsymbol{\eta}'\cdot p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)&\leq2\boldsymbol{\varepsilon}'p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)+2N\tau\left\|\boldsymbol{\beta}\right\|_1-2N\tau\left\|\widehat{\boldsymbol{\beta}}\right\|_1\\
\left\|p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right\|_2^2&\leq2\left(\boldsymbol{\varepsilon}'+\boldsymbol{\eta}'\right)\cdot p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)+2N\tau\left\|\boldsymbol{\beta}\right\|_1-2N\tau\left\|\widehat{\boldsymbol{\beta}}\right\|_1.
\end{aligned}$$

Now we maximize $2\left(\boldsymbol{\varepsilon}'+\boldsymbol{\eta}'\right)\cdot p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$. We start from the term containing $\boldsymbol{\eta}$, i.e. $\boldsymbol{\eta}' p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$. Through Hölder's inequality:

$$\begin{aligned}\boldsymbol{\eta}' p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) &\le \left\|p_P\left(\widehat{\mathbf{S}}\right)\boldsymbol{\eta}\right\|_\infty\cdot\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_1 \le \max_{1\le j\le P}\left|\left[p_P\left(\widehat{\mathbf{S}}\right)\right]_{(j,\cdot)}\boldsymbol{\eta}\right|\cdot\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_1\\
&\le \max_{1\le j\le P}\left\|\left[p_P\left(\widehat{\mathbf{S}}\right)\right]_{(j,\cdot)}\right\|_2\|\boldsymbol{\eta}\|_2\cdot\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_1 \le \xi_P N^{\frac{1}{2}}\|\boldsymbol{\eta}\|_2\cdot\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_1\\
&\le N\xi_P N_P\cdot\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_1.\end{aligned}$$

For $\boldsymbol{\varepsilon}' p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$ we use Hölder's and triangle inequalities to get:

$$\begin{aligned}\boldsymbol{\varepsilon}' p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) &\le \left\|p_P\left(\widehat{\mathbf{S}}\right)\boldsymbol{\varepsilon}\right\|_\infty\cdot\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_1\\
&\le \left\|p_P\left(\widehat{\mathbf{S}}\right)\left(\boldsymbol{\varepsilon}-\boldsymbol{\mu}\right)\right\|_\infty\cdot\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_1+\left\|p_P\left(\widehat{\mathbf{S}}\right)\boldsymbol{\mu}\right\|_\infty\cdot\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_1.\end{aligned}$$

The second term can be majorized through the Cauchy-Schwarz inequality and through $\left\|\left[p_P\left(\widehat{\mathbf{S}}\right)\right]_{(j,\cdot)}\right\|_2\le N^{\frac{1}{2}}\xi_P$, for any j, and $\sum_{n=1}^N\mu_n^2\le N\mu_R^2$ (**subG** and **subE**):

$$\left\|p_P\left(\widehat{\mathbf{S}}\right)\boldsymbol{\mu}\right\|_\infty=\max_{1\le j\le P}\left|\left[p_P\left(\widehat{\mathbf{S}}\right)\right]_{(j,\cdot)}\boldsymbol{\mu}\right|\le\max_{1\le j\le P}\left\|\left[p_P\left(\widehat{\mathbf{S}}\right)\right]_{(j,\cdot)}\right\|_2\|\boldsymbol{\mu}\|_2\le N\xi_P\mu_R.$$

Now we maximize $\left\|p_P\left(\widehat{\mathbf{S}}\right)\left(\boldsymbol{\varepsilon}-\boldsymbol{\mu}\right)\right\|_\infty$ using **subE**. We first note that we can write:

$$\left\|p_P\left(\widehat{\mathbf{S}}\right)\left(\boldsymbol{\varepsilon}-\boldsymbol{\mu}\right)\right\|_\infty=\max_{1\le j\le P}\left|\left[p_P\left(\widehat{\mathbf{S}}\right)\right]_{(j,\cdot)}\left(\boldsymbol{\varepsilon}-\boldsymbol{\mu}\right)\right|.$$

Let us suppose that **subE** holds true. Using $\left\|\left[p_P\left(\widehat{\mathbf{S}}\right)\right]_{(j,\cdot)}\right\|_2\le N^{\frac{1}{2}}\xi_P$, for any j:

$$\begin{aligned}&\mathbb{P}\left\{\left\|p_P\left(\widehat{\mathbf{S}}\right)\left(\boldsymbol{\varepsilon}-\boldsymbol{\mu}\right)\right\|_\infty\ge t\right\}\\
=&\mathbb{P}\left\{\max_{1\le j\le P}\left|\left[p_P\left(\widehat{\mathbf{S}}\right)\right]_{(j,\cdot)}\left(\boldsymbol{\varepsilon}-\boldsymbol{\mu}\right)\right|\ge t\right\}\le\sum_{j=1}^P\mathbb{P}\left\{\left|\left[p_P\left(\widehat{\mathbf{S}}\right)\right]_{(j,\cdot)}\left(\boldsymbol{\varepsilon}-\boldsymbol{\mu}\right)\right|\ge t\right\}\\
\le&\sum_{j=1}^P 2e^{-ts}\mathbb{E}\exp\left\{s\left[p_P\left(\widehat{\mathbf{S}}\right)\right]_{(j,\cdot)}\left(\boldsymbol{\varepsilon}-\boldsymbol{\mu}\right)\right\}\le\sum_{j=1}^P 2e^{-ts}\exp\left(\frac{1}{2}\sigma_R^2 s^2\left\|\left[p_P\left(\widehat{\mathbf{S}}\right)\right]_{(j,\cdot)}\right\|_2^2\right)\\
\le&2P\exp\left(\frac{1}{2}\sigma_R^2 s^2\max_{1\le j\le P}\left\|\left[p_P\left(\widehat{\mathbf{S}}\right)\right]_{(j,\cdot)}\right\|_2^2-ts\right)\le 2P\exp\left(\frac{1}{2}\sigma_R^2 s^2\xi_P^2 N-ts\right)\end{aligned}$$

under the constraint that $\left\|s\left[p_P\left(\widehat{\mathbf{S}}\right)\right]_{(j,\cdot)}\right\|_\infty\le c_R^{-1}$ or $|s|\le\frac{1}{c_R\xi_P'}$. The s minimizing the bound without any constraint is:

$$s^\star=\frac{t}{\sigma_R^2\xi_P^2 N}.$$

If $\frac{t}{\sigma_R^2 \xi_P^2 N} \leq \frac{1}{c_R \xi'_P}$, we have:

$$\mathbb{P}\left\{\left\|p_P\left(\widehat{\mathbf{S}}\right)(\boldsymbol{\varepsilon}-\boldsymbol{\mu})\right\|_\infty \geq t\right\} \leq 2P\exp\left(-\frac{t^2}{2\sigma_R^2\xi_P^2 N}\right).$$

When $\frac{t}{\sigma_R^2 \xi_P^2 N} > \frac{1}{c_R \xi'_P}$ or $\frac{\sigma_R^2 \xi_P^2 N}{c_R \xi'_P} < t$, the maximum occurs at $s^\star = \frac{1}{c_R \xi'_P}$ and we have:

$$\begin{aligned}\mathbb{P}\left\{\left\|p_P\left(\widehat{\mathbf{S}}\right)(\boldsymbol{\varepsilon}-\boldsymbol{\mu})\right\|_\infty \geq t\right\} &\leq 2P\exp\left(\frac{1}{2}\sigma_R^2\frac{\xi_P^2 N}{c_R^2 \xi_P'^{\,2}} - t\frac{1}{c_R\xi'_P}\right)\\ &\leq 2P\exp\left(-\frac{t}{2c_R\xi'_P}\right).\end{aligned}$$

At the end we have:

$$\mathbb{P}\left\{\left\|p_P\left(\widehat{\mathbf{S}}\right)(\boldsymbol{\varepsilon}-\boldsymbol{\mu})\right\|_\infty \geq t\right\} \leq 2P\exp\left(-\min\left\{\frac{t}{2c_R\xi'_P}, \frac{t^2}{2\sigma_R^2\xi_P^2 N}\right\}\right).$$

Therefore, with probability $1-\eta$:

$$\left\|p_P\left(\widehat{\mathbf{S}}\right)\boldsymbol{\varepsilon}\right\|_\infty \leq \max\left\{\sqrt{2\sigma_R^2\xi_P^2 N \ln\frac{2P}{\eta}}, 2c_R\xi'_P\ln\frac{2P}{\eta}\right\} + \xi_P N \mu_R$$

and:

$$\boldsymbol{\varepsilon}' p'_P\left(\widehat{\mathbf{S}}\right)\cdot\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \leq \left\{\max\left\{\sqrt{2\sigma_R^2\xi_P^2 N \ln\frac{2P}{\eta}}, 2c_R\xi'_P\ln\frac{2P}{\eta}\right\} + \xi_P N \mu_R\right\}\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_1.$$

At last, on $\mathcal{A}$:

$$2\left(\boldsymbol{\varepsilon}'+\boldsymbol{\eta}'\right)\cdot p'_P\left(\widehat{\mathbf{S}}\right)\cdot\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \leq 2\left\{\max\left\{\sqrt{2\sigma_R^2\xi_P^2 N \ln\frac{2P}{\eta}}, 2c_R\xi'_P\ln\frac{2P}{\eta}\right\} + N\xi_P\mu_R + N\xi_P N_P\right\}\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_1.$$

Let:

$$\tau := 2\left\{\max\left\{\sqrt{\frac{2\sigma_R^2\xi_P^2}{N}\ln\frac{2P}{\eta}}, \frac{2c_R\xi'_P}{N}\ln\frac{2P}{\eta}\right\} + \xi_P\left(\mu_R + N_P\right)\right\}.$$

In both cases, on $\mathcal{A}$:

$$\left\|p'_P\left(\widehat{\mathbf{S}}\right)\cdot\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right\|_2^2 \leq \tau N\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_1 + 2N\tau\left\|\boldsymbol{\beta}\right\|_1 - 2N\tau\left\|\widehat{\boldsymbol{\beta}}\right\|_1.$$

We can write:

$$\left\|p'_P\left(\widehat{\mathbf{S}}\right)\cdot\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right\|_2^2 \leq 2\tau N\left(\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_1 + \left\|\boldsymbol{\beta}\right\|_1 - \left\|\widehat{\boldsymbol{\beta}}\right\|_1\right) - \tau N\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_1.$$

Let the support J_0 of $\boldsymbol{\theta}$ be the list of indexes of $\boldsymbol{\theta}$ such that the corresponding parameter is not 0.

Then:

$$\begin{aligned}\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_1+\|\boldsymbol{\beta}\|_1-\left\|\widehat{\boldsymbol{\beta}}\right\|_1&=\left\|\widehat{\boldsymbol{\beta}}_{J_0}-\boldsymbol{\beta}_{J_0}\right\|_1+\left\|\widehat{\boldsymbol{\beta}}_{J_0^c}-\boldsymbol{\beta}_{J_0^c}\right\|_1+\|\boldsymbol{\beta}_{J_0}\|_1+\left\|\boldsymbol{\beta}_{J_0^c}\right\|_1-\left\|\widehat{\boldsymbol{\beta}}_{J_0}\right\|_1-\left\|\widehat{\boldsymbol{\beta}}_{J_0^c}\right\|_1\\&=\left\|\widehat{\boldsymbol{\beta}}_{J_0}-\boldsymbol{\beta}_{J_0}\right\|_1+\left\|\widehat{\boldsymbol{\beta}}_{J_0^c}\right\|_1+\|\boldsymbol{\beta}_{J_0}\|_1-\left\|\widehat{\boldsymbol{\beta}}_{J_0}\right\|_1-\left\|\widehat{\boldsymbol{\beta}}_{J_0^c}\right\|_1\\&=\left\|\widehat{\boldsymbol{\beta}}_{J_0}-\boldsymbol{\beta}_{J_0}\right\|_1+\|\boldsymbol{\beta}_{J_0}\|_1-\left\|\widehat{\boldsymbol{\beta}}_{J_0}\right\|_1\le 2\left\|\widehat{\boldsymbol{\beta}}_{J_0}-\boldsymbol{\beta}_{J_0}\right\|_1.\end{aligned}$$

Therefore, on $\mathcal{A}$:

$$\left\|p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right\|_2^2+\tau N\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_1\le 4\tau N\left\|\widehat{\boldsymbol{\beta}}_{J_0}-\boldsymbol{\beta}_{J_0}\right\|_1. \tag{6.7.7}$$

On the one hand, this implies that, on $\mathcal{A}$:

$$\begin{aligned}4\left\|\widehat{\boldsymbol{\beta}}_{J_0}-\boldsymbol{\beta}_{J_0}\right\|_1&\ge\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_1\\&=\left\|\widehat{\boldsymbol{\beta}}_{J_0}-\boldsymbol{\beta}_{J_0}\right\|_1+\left\|\widehat{\boldsymbol{\beta}}_{J_0^c}-\boldsymbol{\beta}_{J_0^c}\right\|_1,\\3\left\|\widehat{\boldsymbol{\beta}}_{J_0}-\boldsymbol{\beta}_{J_0}\right\|_1&\ge\left\|\widehat{\boldsymbol{\beta}}_{J_0^c}-\boldsymbol{\beta}_{J_0^c}\right\|_1.\end{aligned}$$

Therefore we can apply Assumption $\mathbf{RE}(s,3)$ to get:

$$\left\|\widehat{\boldsymbol{\beta}}_{J_0}-\boldsymbol{\beta}_{J_0}\right\|_2\le\frac{\left\|p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\|_2}{\sqrt{N}\kappa\left(s,3\right)}.$$

On the other hand, (6.7.7) becomes:

$$\left\|p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right\|_2^2\le 3\tau N\left\|\widehat{\boldsymbol{\beta}}_{J_0}-\boldsymbol{\beta}_{J_0}\right\|_1\le 3\tau N\left\|\boldsymbol{\beta}\right\|_0^{\frac{1}{2}}\left\|\widehat{\boldsymbol{\beta}}_{J_0}-\boldsymbol{\beta}_{J_0}\right\|_2. \tag{6.7.8}$$

Combining the two, we have:

$$N\kappa^2\left(s,3\right)\left\|\widehat{\boldsymbol{\beta}}_{J_0}-\boldsymbol{\beta}_{J_0}\right\|_2^2\le\left\|p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\|_2^2\le 3\tau N\left\|\boldsymbol{\beta}\right\|_0^{\frac{1}{2}}\left\|\widehat{\boldsymbol{\beta}}_{J_0}-\boldsymbol{\beta}_{J_0}\right\|_2$$

or:

$$\left\|\widehat{\boldsymbol{\beta}}_{J_0}-\boldsymbol{\beta}_{J_0}\right\|_2\le\frac{3\tau\left\|\boldsymbol{\beta}\right\|_0^{\frac{1}{2}}}{\kappa^2\left(s,3\right)}=\frac{3\tau s^{\frac{1}{2}}}{\kappa^2\left(s,3\right)}$$

and:

$$\left\|p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\|_2\le\frac{3\tau N^{\frac{1}{2}}\left\|\boldsymbol{\beta}\right\|_0^{\frac{1}{2}}}{\kappa\left(s,3\right)}=\frac{3\tau N^{\frac{1}{2}}s^{\frac{1}{2}}}{\kappa\left(s,3\right)}.$$

Now we turn to $\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_2$. The proof follows the one of Theorem 7.1 in [57]. The k-th largest in absolute value element of $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}$ is such that:

$$\left|\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)_{J_0^c}\right]_{(k)}\right|\le\frac{\left\|\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)_{J_0^c}\right\|_1}{k}.$$

As a result:

$$\left\|\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)_{J_{01}^c}\right\|_2^2 \leq \sum_{k=1}^{|J\setminus J_{01}|}\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)_{J_0^c}\right]_{(k)}^2 \leq \sum_{k\geq m+1}\frac{\left\|\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)_{J_0^c}\right\|_1^2}{k^2}=\frac{\left\|\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)_{J_0^c}\right\|_1^2}{m}.$$

Now:

$$\left\|\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)_{J_{01}^c}\right\|_2 \leq \frac{1}{\sqrt{m}}\left\|\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)_{J_0^c}\right\|_1 \leq \frac{3}{\sqrt{m}}\left\|\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)_{J_0}\right\|_1 \leq 3\sqrt{\frac{s}{m}}\left\|\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)_{J_0}\right\|_2 \leq 3\sqrt{\frac{s}{m}}\left\|\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)_{J_{01}}\right\|_2$$

and:

$$\begin{aligned}\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_2^2 &= \left\|\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)_{J_{01}}\right\|_2^2+\left\|\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)_{J_{01}^c}\right\|_2^2 \leq \left\|\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)_{J_{01}}\right\|_2^2+9\frac{s}{m}\left\|\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)_{J_{01}}\right\|_2^2\\ &\leq \left(1+9\frac{s}{m}\right)\left\|\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)_{J_{01}}\right\|_2^2. \end{aligned} \tag{6.7.9}$$

From (6.7.8):

$$\left\|p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right\|_2^2 \leq 3\tau N\left\|\boldsymbol{\beta}\right\|_0^{\frac{1}{2}}\left\|\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)_{J_0}\right\|_2 \leq 3\tau N\sqrt{s}\left\|\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)_{J_{01}}\right\|_2.$$

Through $\mathbf{RE}(s,m,3)$:

$$\left\|\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)_{J_{01}}\right\|_2^2 \leq \frac{\left\|p_P'\left(\widehat{\mathbf{S}}\right)\cdot\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right\|_2^2}{N\kappa^2\left(s,m,3\right)} \leq \frac{3\tau\sqrt{s}\left\|\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)_{J_{01}}\right\|_2}{\kappa^2\left(s,m,3\right)}$$

and:

$$\left\|\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)_{J_{01}}\right\|_2 \leq \frac{3\tau\sqrt{s}}{\kappa^2\left(s,m,3\right)}.$$

With this inequality, (6.7.9) becomes:

$$\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_2^2 \leq \left(1+9\frac{s}{m}\right)\frac{9\tau^2 s}{\kappa^4\left(s,m,3\right)}.$$

At last we see what happens to the forecast. We have:

$$\begin{aligned}\widehat{\theta}-\theta =& p_P'\left(\widehat{\mathbf{s}}\right)\cdot\widehat{\boldsymbol{\beta}}-f\left(\mathbf{s}\right)\\ =&\left\{p_P'\left(\widehat{\mathbf{s}}\right)\cdot\widehat{\boldsymbol{\beta}}-p_P'\left(\widehat{\mathbf{s}}\right)\cdot\boldsymbol{\beta}\right\}+\left\{p_P'\left(\widehat{\mathbf{s}}\right)\cdot\boldsymbol{\beta}-f\left(\widehat{\mathbf{s}}\right)\right\}+\left\{f\left(\widehat{\mathbf{s}}\right)-f\left(\mathbf{s}\right)\right\}\\ =&\left\{p_P'\left(\widehat{\mathbf{s}}\right)\cdot\widehat{\boldsymbol{\beta}}-p_P'\left(\widehat{\mathbf{s}}\right)\cdot\boldsymbol{\beta}\right\}+\left\{p_P'\left(\widehat{\mathbf{s}}\right)\cdot\boldsymbol{\beta}-f\left(\widehat{\mathbf{s}}\right)\right\}+\left\{f\left(\widehat{\mathbf{s}}\right)-f\left(\mathbf{s}\right)\right\}.\end{aligned}$$

The first term is majorized as follows:

$$\left|p_P'\left(\widehat{\mathbf{s}}\right)\cdot\widehat{\boldsymbol{\beta}}-p_P'\left(\widehat{\mathbf{s}}\right)\cdot\boldsymbol{\beta}\right| \leq \left\|p_P\left(\widehat{\mathbf{s}}\right)\right\|_2\cdot\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_2$$

where:

$$\|p_P(\widehat{\mathbf{s}})\|_2^2 = \sum_{j=1}^{P} p_{jP}^2(\mathbf{s}) \leq P \cdot \xi_P^{\prime,2}.$$

The second term is majorized by N_P. The third term can be majorized as:

$$|f(\widehat{\mathbf{s}}) - f(\mathbf{s})| = |\boldsymbol{\varepsilon}| \leq |\boldsymbol{\varepsilon} - \mu| + |\mu|.$$

Under Assumption **subE'**, we have $|\mu| \leq \mu_S$ and, reasoning as above:

$$\mathbb{P}\{|\boldsymbol{\varepsilon} - \mu| \geq t\} \leq 2\exp\left(-\min\left\{\frac{t^2}{2\sigma_S^2}, \frac{t}{2c_S}\right\}\right).$$

Taking $t = \max\left\{\sigma_S\sqrt{2\ln\frac{2}{\Delta}}, 2c_S\ln\frac{2}{\Delta}\right\}$, with probability $1-\Delta$:

$$|\boldsymbol{\varepsilon} - \mu| \leq \max\left\{\sigma_S\sqrt{2\ln\frac{2}{\Delta}}, 2c_S\ln\frac{2}{\Delta}\right\}.$$

QED

Bibliography

[1] Robert P. Abelson and Alex Bernstein. A Computer Simulation Model of Community Referendum Controversies. *Public Opinion Quarterly*, 27(1):93–122, 1963.

[2] Daniel Ackerberger, John Geweke, and Jinyong Hahn. Comments on ”Convergence Properties of the Likelihood of Computed Dynamic Models”. *Econometrica*, 77(6):2009–2017, 2009.

[3] Vikas A. Aggarwal, Nicolaj Siggelkow, and Harbir Singh. Governing collaborative activity: interdependence and the impact of coordination and exploration. *Strategic Management Journal*, 32(7):705–730, 2011.

[4] Ibrahim A. Ahmad and Pi-Erh Lin. A nonparametric estimation of the entropy for absolutely continuous distributions (Corresp.). *IEEE Transactions on Information Theory*, 22(3):372–375, 1976.

[5] Christoph Aistleitner and Josef Dick. Low-discrepancy point sets for non-uniform measures. *Acta Arithmetica*, 163(4):345–369, 2014.

[6] Christoph Aistleitner and Josef Dick. Functions of bounded variation, signed measures, and a general Koksma–Hlawka inequality. *Acta Arithmetica*, 167(2):143–171, 2015.

[7] David J. Albers and George Hripcsak. Estimation of time-delayed mutual information and bias for irregularly and sparsely sampled time-series. *Chaos, Solitons & Fractals*, 45(6):853–860, 2012.

[8] Paul H. Algoet and Thomas M. Cover. Sandwich Proof of the Shannon-McMillan-Breiman Theorem. *The Annals of Probability*, 16(2):899–909, 1988.

[9] Syed M. Ali and Samuel D. Silvey. A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.

[10] Ingo Althöfer and Klaus-Uwe Koschnick. On the convergence of “Threshold Accepting”. *Applied Mathematics and Optimization*, 24(1):183–195, 1991.

[11] Philip Warren Anderson. More Is Different. *Science*, 177(4047):393–396, 1972.

[12] Donald W. K. Andrews. Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. Technical Report 877R, Cowles Foundation for Research in Economics, Yale University, 1989.

[13] Donald W. K. Andrews. Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models. *Econometrica*, 59(2):307–345, 1991.

[14] Donald W. K. Andrews. Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. *Econometrica*, 59(3):817–858, 1991.

[15] András Antos and Ioannis Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures and Algorithms*, 19(3-4):163–193, 2001.

[16] András Antos and Ioannis Kontoyiannis. Estimating the entropy of discrete distributions. In *Proceedings. 2001 IEEE International Symposium on Information Theory (IEEE Cat. No.01CH37252)*, page 45, Washington, DC, USA, 2001. IEEE.

[17] Masanao Aoki. New macroeconomic modeling approaches: Hierarchical dynamics and mean-field approximation. *Journal of economic dynamics and control*, 18(3-4):865–877, 1994.

[18] Tom M. Apostol. An Elementary View of Euler's Summation Formula. *The American Mathematical Monthly*, 106(5):409–418, 1999.

[19] Sami Assaf, Persi Diaconis, and Kannan Soundararajan. A rule of thumb for riffle shuffling. *The Annals of Applied Probability*, 21(3):843–875, 2011.

[20] Robert Axelrod. The Dissemination of Culture: A Model with Local Convergence and Global Polarization. *Journal of Conflict Resolution*, 41(2):203–226, 1997.

[21] Robert Axelrod and Leigh Tesfatsion. Appendix A - A Guide for Newcomers to Agent-Based Modeling in the Social Sciences. In Leigh Tesfatsion and Kenneth L. Judd, editors, *Handbook of Computational Economics*, volume 2, pages 1647–1659. Elsevier, 2006.

[22] Robert Azencott. Grandes deviations et applications. In Paul L. Hennequin, editor, *Ecole d'Eté de Probabilités de Saint-Flour VIII-1978*, volume 774, pages 1–176. Springer-Verlag, Berlin, Heidelberg, 1980.

[23] Andrew D. Back, Daniel Angus, and Janet Wiles. Determining the Number of Samples Required to Estimate Entropy in Natural Sequences. *IEEE Transactions on Information Theory*, 65(7):4345–4352, 2019.

[24] Jerry Banks, John S. II Carson, Barry L. Nelson, and David M. Nicol. *Discrete-event system simulation.* Prentice-Hall international series in industrial and systems engineering. Pearson Prentice Hall, Upper Saddle River, NJ, fourth edition, 2005.

[25] Stuart Barber, Jochen Voss, and Mark Webster. The rate of convergence for approximate Bayesian computation. *Electronic Journal of Statistics*, 9(1):80–105, 2015.

[26] Sylvain Barde. Direct comparison of agent-based models of herding in financial markets. *Journal of Economic Dynamics and Control*, 73:329–353, 2016.

[27] Sylvain Barde. A Practical, Accurate, Information Criterion for Nth Order Markov Processes. *Computational Economics*, 50(2):281–324, 2017.

[28] Sylvain Barde. Macroeconomic simulation comparison with a multivariate extension of the Markov information criterion. *Journal of Economic Dynamics and Control*, 111:103795, 2020.

[29] Sylvain Barde and Sander van der Hoog. An Empirical Validation Protocol for Large-Scale Agent-Based Models. *SSRN Electronic Journal*, 2017.

[30] Emanuele Bardone and Davide Secchi. Inquisitiveness: distributing rational thinking. *Team Performance Management: An International Journal*, 23(1/2):66–81, 2017.

[31] Georgij P. Basharin. On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables. *Theory of Probability & Its Applications*, 4(3):333–336, 1959.

[32] Michèle Basseville. Divergence measures for statistical data processing—An annotated bibliography. *Signal Processing*, 93(4):621–633, 2013.

[33] Kinjal Basu and Art B. Owen. Transformations and Hardy–Krause Variation. *SIAM Journal on Numerical Analysis*, 54(3):1946–1966, 2016.

[34] Manel Baucells and Emanuele Borgonovo. Invariant Probabilistic Sensitivity Analysis. *Management Science*, 59(11):2536–2549, 2013.

[35] Oliver Baumann, Jens Schmidt, and Nils Stieglitz. Effective search in rugged performance landscapes: A review and outlook. *Journal of Management*, 45(1):285 – 318, 2019.

[36] Dave Bayer and Persi Diaconis. Trailing the Dovetail Shuffle to its Lair. *The Annals of Applied Probability*, 2(2):294–313, 1992.

[37] Bernhard Beckermann. The condition number of real Vandermonde, Krylov and positive definite Hankel matrices:. *Numerische Mathematik*, 85(4):553–577, 2000.

[38] Bernhard Beckermann and Alex Townsend. On the Singular Values of Matrices with Displacement Structure. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1227–1248, 2017.

[39] Jan Beirlant, Edward J. Dudewicz, László Györfi, and Edward C. van der Meulen. Nonparametric entropy estimation: an overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, 1997.

[40] Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366, 2015.

[41] Gérard Ben Arous, Leonid V. Bogachev, and Stanislav A. Molchanov. Limit theorems for sums of random exponentials. *Probability Theory and Related Fields*, 132(4):579–612, 2005.

[42] Rudolf Beran. 30 Minimum distance procedures. In Paruchuri R. Krishnaiah and Pranab K. Sen, editors, *Handbook of Statistics*, volume 4, pages 741–754. Elsevier, 1984.

[43] Christian Berg, Yang Chen, and Mourad E. H. Ismail. Small eigenvalues of large Hankel matrices: The indeterminate case. *Mathematica Scandinavica*, 91(1):67–81, 2002.

[44] Christian Berg and Ryszard Szwarc. The Smallest Eigenvalue of Hankel Matrices. *Constructive Approximation*, 34(1):107–133, 2011.

[45] Elwyn R. Berlekamp, John Horton Conway, and Richard K. Guy. *Winning ways for your mathematical plays.* A.K. Peters, Natick, MA, second edition, 2001.

[46] Michele Bernasconi, Christine Choirat, and Raffaello Seri. A re-examination of the algebraic properties of the AHP as a ratio-scaling technique. *Journal of Mathematical Psychology*, 55(2):152–165, 2011.

[47] Donald J. Berndt and James Clifford. Using Dynamic Time Warping to Find Patterns in Time Series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, AAAIWS'94, pages 359–370. AAAI Press, 1994.

[48] Espen Bernton, Pierre E. Jacob, Mathieu Gerber, and Christian P. Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 2019.

[49] Thomas B. Berrett, Richard J. Samworth, and Ming Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *The Annals of Statistics*, 47(1):288–318, 2019.

[50] Ludwig van Bertalanffy. *General system theory: foundations, development, applications.* Braziller, New York, N.Y., 1968.

[51] Philippe Berthet and David M. Mason. Revisiting two strong approximation results of Dudley and Philipp. In *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, pages 155–172. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2006.

[52] Francesco Berto and Jacopo Tagliabue. Cellular automata. In Edward N. Zalta, editor, *The stanford encyclopedia of philosophy.* Metaphysics Research Lab, Stanford University, fall 2017 edition, 2017.

[53] Rabi N. Bhattacharya and Ngai H. Chan. Comparisons of Chisquare, Edgeworth Expansions and Bootstrap Approximations to the Distribution of the Frequency Chisquare. *Sankhyā: The Indian Journal of Statistics, Series A*, 58(1):57–68, 1996.

[54] Carlo Bianchi, Pasquale Cirillo, Mauro Gallegati, and Pietro A. Vagliasindi. Validating and Calibrating Agent-Based Models: A Case Study. *Computational Economics*, 30(3):245–264, 2007.

[55] Carlo Bianchi, Pasquale Cirillo, Mauro Gallegati, and Pietro A. Vagliasindi. Validation in agent-based models: An investigation on the CATS model. *Journal of Economic Behavior & Organization*, 67(3-4):947–964, 2008.

[56] Peter J. Bickel and Kjell A. Doksum. *Mathematical statistics: basic ideas and selected topics*. Texts in statistical science. CRC Press, Boca Raton, FL, second edition, 2015.

[57] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[58] Torsten Biemann and Eric Kearney. Size Does Matter: How Varying Group Sizes in a Sample Affect the Most Common Measures of Group Diversity. *Organizational Research Methods*, 13(3):582–599, 2010.

[59] Andrew J. Black and Alan J. McKane. Stochastic formulation of ecological models and their applications. *Trends in Ecology & Evolution*, 27(6):337–345, 2012.

[60] David Blitz and Mario Bunge. *Emergent Evolution: Qualitative Novelty and the Levels of Reality*. Springer Netherlands, Dordrecht, 2010.

[61] Michael G. B. Blum, Matthew A. Nunes, Dennis Prangle, and Scott A. Sisson. A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation. *Statistical Science*, 28(2):189–208, 2013.

[62] Martin Blümlinger and Robert F. Tichy. Topological algebras of functions of bounded variation I. *Manuscripta Mathematica*, 65(2):245–255, 1989.

[63] Riccardo Boero and Flaminio Squazzoni. Does Empirical Embeddedness Matter? Methodological Issues on Agent-Based Models for Analytical Social Science. *Journal of Artificial Societies and Social Simulation*, 8(4):6, 2005.

[64] Juan A. Bonachela, Haye Hinrichsen, and Miguel A. Munoz. Entropy estimates of small data sets. *Journal of Physics A: Mathematical and Theoretical*, 41(20):202001, 2008.

[65] Emanuele Borgonovo and Elmar Plischke. Sensitivity analysis: A review of recent advances. *European Journal of Operational Research*, 248(3):869–887, 2016.

[66] Emanuele Borgonovo, Stefano Tarantola, Elmar Plischke, and Max D. Morris. Transformations and invariance in the sensitivity analysis of computer experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(5):925–947, 2014.

[67] Michael J. Box. A Comparison of Several Current Optimization Methods, and the use of Transformations in Constrained Problems. *The Computer Journal*, 9(1):67–77, 1966.

[68] Ray Bradbury. A Sound of Thunder. *Collier's*, (June 28):20–21, 60–61, 1952.

[69] Luca Brandolini, Leonardo Colzani, Giacomo Gigante, and Giancarlo Travaglini. On the Koksma–Hlawka inequality. *Journal of Complexity*, 29(2):158–172, 2013.

[70] Thomas Brenner. Chapter 18 Agent Learning Representation: Advice on Modelling Economic Learning. In *Handbook of Computational Economics*, volume 2, pages 895–947. Elsevier, 2006.

[71] Dermot Breslin, Daniela Romano, and James Percival. Conceptualizing and modeling multilevel organizational co-evolution. In Davide Secchi and Martin Neumann, editors, *Agent-based simulation of organizational behavior*, pages 137–157. Springer, 2016.

[72] William A. Brock and Cars H. Hommes. Heterogeneous beliefs and routes to chaos in a simple asset pricing model. *Journal of Economic Dynamics and Control*, 22(8-9):1235–1274, 1998.

[73] Donald J. Brown and Marten H. Wegkamp. Weighted Minimum Mean-Square Distance from Independence Estimation. *Econometrica*, 70(5):2035–2051, 2002.

[74] Robert G. Brown. Dieharder: A Random Number Test Suite, 2019.

[75] Marianne Bruins, James A. Duffy, Michael P. Keane, and Anthony A. Smith. Generalized indirect inference for discrete choice models. *Journal of Econometrics*, 205(1):177–203, 2018.

[76] Johannes Brumm and Simon Scheidegger. Using Adaptive Sparse Grids to Solve High-Dimensional Dynamic Models. *Econometrica*, 85(5):1575–1612, 2017.

[77] Georges-Louis Leclerc Buffon. Geometrie [Résolution des problémes qui regardent le jeu du franc-carreau]. *Histoire de l'Académie Royale des Sciences*, Année 1733:43–45, 1735.

[78] Georges-Louis Leclerc Buffon. Essais d'Arithmétique morale. In *Histoire naturelle, générale et particulière, Supplément, Tome Quatrième*, pages 46–123. Imprimerie Royale, Paris, 1777.

[79] Mario Bunge. Mechanism and Explanation. *Philosophy of the Social Sciences*, 27(4):410–465, 1997.

[80] Eugene Burdick. *The 480*. McGraw Hill, New York, N.Y., first edition, 1964.

[81] Daniel A. Butts and Daniel S. Rokhsar. The Information Content of Spontaneous Retinal Waves. *The Journal of Neuroscience*, 21(3):961–973, 2001.

[82] Alessandro Caiani, Antoine Godin, Eugenio Caverzasi, Mauro Gallegati, Stephen Kinsella, and Joseph E. Stiglitz. Agent based-stock flow consistent macroeconomics: Towards a benchmark model. *Journal of Economic Dynamics and Control*, 69:375–408, 2016.

[83] Colin F. Camerer and Ernst Fehr. When Does "Economic Man" Dominate Social Behavior? *Science*, 311(5757):47–52, 2006.

[84] Fabio Canova and Luca Sala. Back to square one: Identification issues in DSGE models. *Journal of Monetary Economics*, 56(4):431–449, 2009.

[85] A. G. Carlton. On the bias of information estimates. *Psychological Bulletin*, 71(2):108–109, 1969.

[86] Ernesto Carrella, Richard Bailey, and Jens Madsen. Calibrating Agent-Based Models with Linear Regressions. *Journal of Artificial Societies and Social Simulation*, 23(1):7, 2020.

[87] Shu-Heng Chen, Chia-Ling Chang, and Ye-Rong Du. Agent-based economic models and econometrics. *The Knowledge Engineering Review*, 27(2):187–219, 2012.

[88] Xiaohong Chen. Chapter 76 - Large Sample Sieve Estimation of Semi-Nonparametric Models. In *Handbook of Econometrics*, volume 6, pages 5549–5632. Elsevier, 2007.

[89] Yang Chen and Nigel Lawrence. Small eigenvalues of large Hankel matrices. *Journal of Physics A: Mathematical and General*, 32(42):7305–7315, 1999.

[90] Zhenxi Chen and Thomas Lux. Estimation of Sentiment Effects in Financial Markets: A Simulated Method of Moments Approach. *Computational Economics*, 52(3):711–744, 2018.

[91] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[92] Christine Choirat, Christian Hess, and Raffaello Seri. A functional version of the Birkhoff ergodic theorem for a normal integrand: A variational approach. *The Annals of Probability*, 31(1):63–92, 2003.

[93] Christine Choirat and Raffaello Seri. Estimation in Discrete Parameter Models. *Statistical Science*, 27(2):278–293, 2012.

[94] Christine Choirat and Raffaello Seri. Computational aspects of Cui-Freeden statistics for equidistribution on the sphere. *Mathematics of Computation*, 82(284):2137–2156, 2013.

[95] Mihai Ciucu. No-feedback card guessing for dovetail shuffles. *The Annals of Applied Probability*, 8(4):1251–1269, 1998.

[96] Michael D. Cohen, James G. March, and Johan P. Olsen. A Garbage Can Model of Organizational Choice. *Administrative Science Quarterly*, 17(1):1–25, 1972.

[97] David Colander, Peter Howitt, Alan Kirman, Axel Leijonhufvud, and Perry Mehrling. Beyond DSGE Models: Toward an Empirically Based Macroeconomics. *American Economic Review*, 98(2):236–240, 2008.

[98] James S. Coleman. Social Theory, Social Research, and a Theory of Action. *American Journal of Sociology*, 91(6):1309–1335, 1986.

[99] James S. Coleman. *Foundations of social theory.* Belknap Press of Harvard University Press, Cambridge, MA, 1990.

[100] Henri Comman. Differentiability-free conditions on the free-energy function implying large deviations. *Confluentes Mathematici*, 01(02):181–196, 2009.

[101] Gregory M. Constantine and Thomas H. Savits. A Multivariate Faa Di Bruno Formula With Applications. *Transactions of the American Mathematical Society*, 348(02):503–521, 1996.

[102] Rosaria Conte. Agent-based modeling for understanding social intelligence. *Proceedings of the National Academy of Sciences*, 99(suppl 3):7189–7190, 2002.

[103] Rosaria Conte and Mario Paolucci. On agent-based modeling and computational social science. *Frontiers in Psychology*, 5, 2014.

[104] Thomas F. Cooley. Calibrated models. *Oxford Review of Economic Policy*, 13(3):55–69, 1997.

[105] I. P. Cornfeld and Ya. G. Sinai. Entropy Theory of Dynamical Systems. In R. V. Gamkrelidze and Ya. G. Sinai, editors, *Dynamical Systems II*, volume 2, pages 36–58. Springer, Berlin, 1989.

[106] Rand Corporation, editor. *A million random digits with 100,000 normal deviates.* Free Press, Glencoe, IL, 1955.

[107] Franck Courchamp, Ivan Jaric, Céline Albert, Yves Meinard, William J. Ripple, and Guillaume Chapron. The paradoxical extinction of the most charismatic animals. *PLOS Biology*, 16(4):e2003997, 2018.

[108] Sthephen J. Cowley and Frédéric Vallée-Tourangeau, editors. *Cognition Beyond the Brain: Computation, Interactivity and Human Artifice.* Springer-Verlag, London, second edition, 2017.

[109] Dennis D. Cox. Approximation of Least Squares Regression on Nested Subspaces. *The Annals of Statistics*, 16(2):713–732, 1988.

[110] Michael Creel. Neural nets for indirect inference. *Econometrics and Statistics*, 2:36–49, 2017.

[111] Michael Creel and Dennis Kristensen. On selection of statistics for approximate Bayesian computing (or the method of simulated moments). *Computational Statistics & Data Analysis*, 100:99–114, 2016.

[112] Andrew Crooks, Christian Castle, and Michael Batty. Key challenges in agent-based modelling for geo-spatial simulation. *Computers, Environment and Urban Systems*, 32(6):417–430, 2008.

[113] Katalin Csilléry, Michael G.B. Blum, Oscar E. Gaggiotti, and Olivier François. Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, 25(7):410–418, 2010.

[114] Imre Csiszár. Eine Informationstheoretische Ungleichung und ihre Anwendung auf Beweis der Ergodizität von Markoffschen Ketten. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, 8:84–108, 1963.

[115] Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

[116] Bryon Cunningham. The reemergence of 'emergence'. *Philosophy of Science*, 68:S62–S75, 2001.

[117] Gianni Dal Maso. *An Introduction to* Γ*-Convergence*. Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser, Boston, MA, first edition, 1993.

[118] James Davidson. *Stochastic limit theory: an introduction for econometricians*. Advanced texts in econometrics. Oxford University Press, Oxford, 1994.

[119] Scott de Marchi and Scott E. Page. Agent-Based Models. *Annual Review of Political Science*, 17(1):1–20, 2014.

[120] Stefano De Marco, Antoine Jacquier, and Patrick Roome. Two examples of non strictly convex large deviations. *Electronic Communications in Probability*, 21, 2016.

[121] Augustus De Morgan. Supplement to the Budget of Paradoxes (No. IV). *The Athenæum*, (2017):835–836, 1866.

[122] Augustus De Morgan. *A Budget of Paradoxes*. Longmans, Green, and Co., 1872.

[123] Domenico Delli Gatti, Corrado Di Guilmi, Mauro Gallegati, and Simone Landini. Reconstructing Aggregate Dynamics in Heterogeneous Agents Models: A Markovian Approach. *Revue de l'OFCE*, 124(5):117, 2012.

[124] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*, volume 38 of *Stochastic Modelling and Applied Probability*. Springer, Berlin, second edition, 2010.

[125] Giuseppe Di Molfetta. On the parameter identifiability problem in Agent Based economical models. *arXiv:1602.01271 [q-fin]*, 2016. arXiv: 1602.01271.

[126] Roberto Dieci and Xue-Zhong He. Chapter 5 - Heterogeneous Agent Models in Finance. In *Handbook of Computational Economics*, volume 4, pages 257–328. Elsevier, 2018.

[127] Karen Doore and Paul Fishwick. Prototyping an analog computing representation of predator prey dynamics. In *Proceedings of the Winter Simulation Conference 2014*, pages 3561–3571a, Savanah, GA, 2014. IEEE.

[128] Ali Dorri, Salil S. Kanhere, and Raja Jurdak. Multi-Agent Systems: A Survey. *IEEE Access*, 6:28573–28593, 2018.

[129] Paul Doukhan, Pascal Massart, and Emmanuel Rio. The functional central limit theorem for strongly mixing processes. *Annales de l'I.H.P. Probabilités et statistiques*, 30(1):63–82, 1994.

[130] Ramdan Dridi, Alain Guay, and Eric Renault. Indirect inference and calibration of dynamic stochastic general equilibrium models. *Journal of Econometrics*, 136(2):397–430, 2007.

[131] Christopher C. Drovandi, Anthony N. Pettitt, and Anthony Lee. Bayesian Indirect Inference Using a Parametric Auxiliary Model. *Statistical Science*, 30(1):72–95, 2015.

[132] Gunter Dueck and Tobias Scheuer. Threshold accepting: A general purpose optimization algorithm appearing superior to simulated annealing. *Journal of Computational Physics*, 90(1):161–175, 1990.

[133] Darrell Duffie and Kenneth J. Singleton. Simulated Moments Estimation of Markov Models of Asset Prices. *Econometrica*, 61(4):929–952, 1993.

[134] Nick G. Duffield, John T. Lewis, Neil O'Connell, Raymond Russell, and Fergal Toomey. Entropy of ATM traffic streams: a tool for estimating QoS parameters. *IEEE Journal on Selected Areas in Communications*, 13(6):981–990, 1995.

[135] Ken Duffy and Anthony P. Metcalfe. How to Estimate the Rate Function of a Cumulative Process. *Journal of Applied Probability*, 42(4):1044–1052, 2005.

[136] Ken Duffy and Anthony P. Metcalfe. The Large Deviations of Estimating Rate Functions. *Journal of Applied Probability*, 42(1):267–274, 2005.

[137] Ken Duffy and Brendan D. Williamson. Estimating large deviation rate functions. *arXiv:1511.02295 [math]*, 2015. arXiv: 1511.02295.

[138] Émile Durkheim. De la Méthode Objective en Sociologie. *Revue de Synthèse Historique*, II(1):3–17, 1901.

[139] Émile Durkheim. *Les Règles de la méthode sociologique, revue et augmentée d'une préface nouvelle*. Bibliothèque de philosophie contemporaine. Alcan, Paris, second edition, 1901.

[140] Émile Durkheim. *The Rules of Sociological Method*. The Free Press, New York, N.Y., 1982.

[141] Roger Eckhardt. Stan Ulam, John von Neumann, and the Monte Carlo method. In *Los Alamos Sci.*, number 15, Special Issue, pages 131–137. 1987.

[142] Bruce Edmonds and Scott Moss. From KISS to KIDS – An 'Anti-simplistic' Modelling Approach. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Paul Davidsson, Brian Logan, and Keiki Takadama, editors, *Multi-Agent and Multi-Agent-Based Simulation*, volume 3415, pages 130–144. Springer-Verlag, Berlin, Heidelberg, 2005.

[143] Philipp Eisenhauer, James J. Heckman, and Stefano Mosso. Estimation of Dynamic Discrete Choice Models by Maximum Likelihood and the Simulated Method of Moments. *International Economic Review*, 56(2):331–357, 2015.

[144] Ivar Ekeland. *Au hasard: la chance, la science et le monde*. Seuil, Paris, 1991.

[145] Ivar Ekeland. *The broken dice, and other mathematical tales of chance*. University of Chicago Press, Chicago, IL, 1993.

[146] James Elwick. Containing Multitudes: Herbert Spencer, organisms social and orders of individuality. In Mark Francis and Michael W. Taylor, editors, *Herbert Spencer: Legacies*, pages 89–110. Routledge, London, 2014.

[147] Robert F. Engle, David F. Hendry, and Jean-Francois Richard. Exogeneity. *Econometrica*, 51(2):277–304, 1983.

[148] Bjørn Eraker. MCMC Analysis of Diffusion Models With Application to Finance. *Journal of Business & Economic Statistics*, 19(2):177–191, 2001.

[149] Carl-Gustav Esséen. Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law. *Acta Mathematica*, 77(0):1–125, 1945.

[150] Annalisa Fabretti. On the problem of calibrating an agent based model for financial markets. *Journal of Economic Interaction and Coordination*, 8(2):277–293, 2013.

[151] Giorgio Fagiolo, Mattia Guerini, Francesco Lamperti, Alessio Moneta, and Andrea Roventini. Validation of Agent-Based Models in Economics and Finance. In Claus Beisbart and Nicole J. Saam, editors, *Computer Simulation Validation*, pages 763–787. Springer, Cham, 2019.

[152] Giorgio Fagiolo, Alessio Moneta, and Paul Windrum. A Critical Guide to Empirical Validation of Agent-Based Models in Economics: Methodologies, Procedures, and Open Problems. *Computational Economics*, 30(3):195–226, 2007.

[153] Giorgio Fagiolo and Andrea Roventini. Macroeconomic Policy in DSGE and Agent-Based Models Redux: New Developments and Challenges Ahead. *Journal of Artificial Societies and Social Simulation*, 20(1):1, 2017.

[154] J. Doyne Farmer and Duncan Foley. The economy needs agent-based modelling. *Nature*, 460(7256):685–686, 2009.

[155] J.Doyne Farmer and Shareen Joshi. The price dynamics of common trading strategies. *Journal of Economic Behavior & Organization*, 49(2):149–171, 2002.

[156] Jean-David Fermanian and Bernard Salanié. A Nonparametric Simulated Maximum Likelihood Estimation Method. *Econometric Theory*, 20(4), 2004.

[157] Jesus Fernandez-Villaverde, Juan F. Rubio-Ramirez, and Manuel S. Santos. Convergence Properties of the Likelihood of Computed Dynamic Models. *Econometrica*, 74(1):93–119, 2006.

[158] Andrey Feuerverger. On The Empirical Saddlepoint Approximation. *Biometrika*, 76(3):457–464, 1989.

[159] Francesco Figari, Alari Paulus, and Holly Sutherland. Chapter 24 - Microsimulation and Policy Analysis. In Anthony B. Atkinson and François Bourguignon, editors, *Handbook of Income Distribution*, volume 2 of *Handbook of Income Distribution*, pages 2141–2221. Elsevier, 2015.

[160] Guido Fioretti. Agent-Based Simulation Models in Organization Science. *Organizational Research Methods*, 16(2):227–242, 2013.

[161] Guido Fioretti. Emergent organizations. In Davide Secchi and Martin Neumann, editors, *Agent-based simulation of organizational behavior. New frontiers of social science research*, 19-41. Springer, New York, N.Y., 2016.

[162] Guido Fioretti and Alessandro Lomi. An Agent-Based Representation of the Garbage Can Model of Organizational Choice. *Journal of Artificial Societies and Social Simulation*, 11(1):1, 2008.

[163] Guido Fioretti and Alessandro Lomi. Passing the buck in the garbage can model of organizational choice. *Computational and Mathematical Organization Theory*, 16(2):113–143, 2010.

[164] Ronald A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 1922.

[165] Ronald A. Fisher. Theory of Statistical Estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5):700–725, 1925.

[166] Jean-Pierre Florens, Michel Mouchart, and Jean-Marie Rolin. Noncausality and Marginalization of Markov Processes. *Econometric Theory*, 9(2):241–262, 1993.

[167] Duncan K. Foley. A statistical equilibrium theory of markets. *Journal of Economic Theory*, 62(2):321–345, 1994.

[168] Jean-Jacques Forneron and Serena Ng. The ABC of simulation estimation with auxiliary statistics. *Journal of Econometrics*, 205(1):112–139, 2018.

[169] Jay Wright Forrester. *Principles of Systems*. MIT Press, Cambridge, MA, 1968.

[170] Jay Wright Forrester. Counterintuitive behavior of social systems. *Technological Forecasting and Social Change*, 3:1–22, 1971.

[171] Jay Wright Forrester. The beginning of system dynamics. *The McKinsey Quarterly*, 1995(4):4–16, 1995.

[172] Reiner Franke and Frank Westerhoff. Structural stochastic volatility in asset pricing dynamics: Estimation and model contest. *Journal of Economic Dynamics and Control*, 36(8):1193–1211, 2012.

[173] Andrew M. Fraser. Information and entropy in strange attractors. *IEEE Transactions on Information Theory*, 35(2):245–262, 1989.

[174] Andrew M. Fraser and Harry L. Swinney. Independent coordinates for strange attractors from mutual information. *Physical Review A*, 33(2):1134–1140, 1986.

[175] David T. Frazier, Gale M. Martin, Christian P. Robert, and Judith Rousseau. Asymptotic properties of approximate Bayesian computation. *Biometrika*, 105(3):593–607, 2018.

[176] Tak-chung Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.

[177] A. Ronald Gallant. *Nonlinear Statistical Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, N.J., 1987.

[178] A. Ronald Gallant. Effective calibration. 2001.

[179] A. Ronald Gallant and George Tauchen. Which Moments to Match? *Econometric Theory*, 12(04):657, 1996.

[180] Yun Gao, Ioannis Kontoyiannis, and Elie Bienenstock. Estimating the Entropy of Binary Time Series: Methodology, Some Theory and a Simulation Study. *Entropy*, 10(2):71–99, 2008.

[181] Martin Gardner. The fantastic combinations of John Conway's new solitaire game "life". *Scientific American*, 223(4):120–123, 1970.

[182] William A. Gardner, Antonio Napolitano, and Luigi Paura. Cyclostationarity: Half a century of research. *Signal Processing*, 86(4):639–697, 2006.

[183] Walter Gautschi. The condition of Vandermonde-like matrices involving orthogonal polynomials. *Linear Algebra and its Applications*, 52-53:293–300, 1983.

[184] Alison L. Gibbs and Francis E. Su. On Choosing and Bounding Probability Metrics. *International Statistical Review*, 70(3):419–435, 2002.

[185] Nigel Gilbert. Computational Social Science: Agent-based social simulation. In Denis Phan and Frédéric Amblard, editors, *Agent-based Modelling and Simulation*, pages 115–134. Bardwell, Oxford, 2007.

[186] Nigel Gilbert, Andreas Pyka, and Petra Ahrweiler. Innovation Networks - A Simulation Approach. *Journal of Artificial Societies and Social Simulation*, 4(3):8, 2001.

[187] Nigel Gilbert and Pietro Terna. How to build and use agent-based models in social science. *Mind and Society*, 1:57–72, 2000.

[188] Nigel Gilbert and Klaus G. Troitzsch. *Simulation for the social scientist*. Open University Press, Maidenhead, New York, N.Y., second edition, 2005.

[189] Manfred Gilli and Peter Winker. Indirect Estimation of the Parameters of Agent Based Models of Financial Markets. *SSRN Electronic Journal*, 2002.

[190] Manfred Gilli and Peter Winker. A global optimization heuristic for estimating agent based models. *Computational Statistics & Data Analysis*, 42(3):299–312, 2003.

[191] Nikolay Gospodinov, Ivana Komunjer, and Serena Ng. Simulated minimum distance estimation of dynamic models with errors-in-variables. *Journal of Econometrics*, 200(2):181–193, 2017.

[192] Mario Götz. Discrepancy and the Error in Integration. *Monatshefte für Mathematik*, 136(2):99–121, 2002.

[193] Friedrich Götze, Holger Sambale, and Arthur Sinulis. Concentration inequalities for polynomials in α-sub-exponential random variables. *arXiv:1903.05964 [math]*, 2019.

[194] Friedrich Götze and Vladimir V. Ulyanov. Asymptotic Distribution of χ^2-type Statistics. Preprintreihe der Forschergruppe spektrale Analysis und stochastische Dynamik 03–033, Universität Bielefeld, 2003. arXiv: 1708.08663.

[195] Christian Gouriéroux and Alain Monfort. *Simulation-based econometric methods*. CORE lectures. Oxford University Press, New York, N.Y., 1996.

[196] Christian Gouriéroux, Alain Monfort, and Eric Renault. Indirect inference. *Journal of Applied Econometrics*, 8(S1):S85–S118, 1993.

[197] Christian Gouriéroux, Alain Monfort, and Alain Trognon. A General Approach to Serial Correlation. *Econometric Theory*, 1(3):315–340, 1985.

[198] Clive Granger and Jin-Lung Lin. Using the Mutual Information Coefficient to Identify Lags in Nonlinear Models. *Journal of Time Series Analysis*, 15(4):371–384, 1994.

[199] Peter Grassberger. Finite sample corrections to entropy and dimension estimates. *Physics Letters A*, 128(6-7):369–373, 1988.

[200] Peter Grassberger. Estimating the information content of symbol sequences and efficient codes. *IEEE Transactions on Information Theory*, 35(3):669–675, 1989.

[201] Peter Grassberger. Entropy Estimates from Insufficient Samplings. *arXiv:physics/0307138*, 2003. arXiv: physics/0307138.

[202] Peter Grassberger and Itamar Procaccia. Estimation of the Kolmogorov entropy from a chaotic signal. *Physical Review A*, 28(4):2591–2593, 1983.

[203] Robert M. Gray. *Probability, Random Processes, and Ergodic Properties*. Springer US, Boston, MA, 2009.

[204] Robert M. Gray and John C. Kieffer. Asymptotically Mean Stationary Measures. *The Annals of Probability*, 8(5):962–973, 1980.

[205] Jakob Grazzini. Analysis of the Emergent Properties: Stationarity and Ergodicity. *Journal of Artificial Societies and Social Simulation*, 15(2), 2012.

[206] Jakob Grazzini and Matteo Richiardi. Estimation of ergodic agent-based models by simulated minimum distance. *Journal of Economic Dynamics and Control*, 51:148–165, 2015.

[207] Jakob Grazzini, Matteo G. Richiardi, and Mike Tsionas. Bayesian estimation of agent-based models. *Journal of Economic Dynamics and Control*, 77:26–47, 2017.

[208] Ulf Grenander. *Abstract inference*. Wiley series in probability and mathematical statistics. Wiley, New York, N.Y., 1981.

[209] Volker Grimm, Uta Berger, Finn Bastiansen, Sigrunn Eliassen, Vincent Ginot, Jarl Giske, John Goss-Custard, Tamara Grand, Simone K. Heinz, Geir Huse, Andreas Huth, Jane U. Jepsen, Christian Jørgensen, Wolf M. Mooij, Birgit Müller, Guy Pe'er, Cyril Piou, Steven F. Railsback, Andrew M. Robbins, Martha M. Robbins, Eva Rossmanith, Nadja Rüger, Espen Strand, Sami Souissi, Richard A. Stillman, Rune Vabø, Ute Visser, and Donald L. DeAngelis. A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, 198(1-2):115–126, 2006.

[210] Volker Grimm, Eloy Revilla, Uta Berger, Florian Jeltsch, Wolf M. Mooij, Steven F. Railsback, Hans-Hermann Thulke, Jacob Weiner, Thorsten Wiegand, and Donald L. DeAngelis. Pattern-Oriented Modeling of Agent-Based Complex Systems: Lessons from Ecology. *Science*, 310(5750):987–991, 2005.

[211] Ivo Grosse, Pedro Bernaola-Galván, Pedro Carpena, Ramón Román-Roldán, Jose Oliver, and H. Eugene Stanley. Analysis of symbolic sequences using the Jensen-Shannon divergence. *Physical Review E*, 65(4):041905, 2002.

[212] André Grow and Jan Van Bavel, editors. *Agent-Based Modelling in Population Studies*, volume 41 of *The Springer Series on Demographic Methods and Population Analysis*. Springer International Publishing, Cham, 2017.

[213] Stanislao Gualdi, Marco Tarzia, Francesco Zamponi, and Jean-Philippe Bouchaud. Tipping points in macroeconomic agent-based models. *Journal of Economic Dynamics and Control*, 50:29–61, 2015.

[214] Steinn Gudmundsson, Thomas P. Runarsson, and Sven Sigurdsson. Support vector machines and dynamic time warping for time series. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 2772–2776, Hong Kong, 2008. IEEE.

[215] Mattia Guerini and Alessio Moneta. A method for agent-based models validation. *Journal of Economic Dynamics and Control*, 82:125–141, 2017.

[216] Mattia Guerini, Mauro Napoletano, and Andrea Roventini. No man is an Island: The impact of heterogeneity and local interactions on macroeconomic dynamics. *Economic Modelling*, 68:82–95, 2018.

[217] George Hall and John Rust. Simulated Minimum Distance Estimation of a Model of Optimal Commodity Price Speculation with Endogenously Sampled Prices. 2003.

[218] Peter Hall and Sally C. Morton. On the estimation of entropy. *Annals of the Institute of Statistical Mathematics*, 45(1):69–88, 1993.

[219] D. Wade Hands. Conundrums of the representative agent. *Cambridge Journal of Economics*, 41(6):1685–1704, 2017.

[220] Edward J. Hannan, editor. *Multiple Time Series*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, 1970.

[221] Lars P. Hansen. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, 50(4):1029, 1982.

[222] Lars P. Hansen and James J. Heckman. The Empirical Foundations of Calibration. *Journal of Economic Perspectives*, 10(1):87–104, 1996.

[223] Peter R. Hansen, Asger Lunde, and James M. Nason. The Model Confidence Set. *Econometrica*, 79(2):453–497, 2011.

[224] Bernard Harris. The statistical estimation of entropy in the non-parametric case. In Imre Csiszár, editor, *Topics in information theory*, pages 323–355. Nort-Holland, Amsterdam, first edition, 1975.

[225] Jürgen Hartinger and Reinhold Kainhofer. Non-Uniform Low-Discrepancy Sequence Generation and Integration of Singular Integrands. In Harald Niederreiter and Denis Talay, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2004*, pages 163–179, Berlin, Heidelberg, 2006. Springer.

[226] Ralph V. L. Hartley. Transmission of Information. *Bell System Technical Journal*, 7(3):535–563, 1928.

[227] Peter Hedström and Petri Ylikoski. Causal Mechanisms in the Social Sciences. *Annual Review of Sociology*, 36(1):49–67, 2010.

[228] Peter Hedström and Petri Ylikoski. Analytical Sociology and Social Mechanisms, 2013.

[229] Rainer Hegselmann. Thomas C. Schelling and James M. Sakoda: The intellectual, technical, and social history of a model. *Journal of Artificial Societies and Social Simulation*, 20(3), 2017.

[230] Hanspeter Herzel, Armin O. Schmitt, and Werner Ebeling. Finite sample effects in sequence analysis. *Chaos, Solitons & Fractals*, 4(1):97–113, 1994.

[231] Christian Hess. Epi-convergence of sequences of normal integrands and strong consistency of the maximum likelihood estimator. *The Annals of Statistics*, 24(3):1298–1315, 1996.

[232] Christian Hess and Raffaello Seri. Generic Consistency for Approximate Stochastic Programming and Statistical Problems. *SIAM Journal on Optimization*, 29(1):290–317, 2019.

[233] Christian Hess, Raffaello Seri, and Christine Choirat. Ergodic theorems for extended real-valued random variables. *Stochastic Processes and their Applications*, 120(10):1908–1919, 2010.

[234] Christian Hess, Raffaello Seri, and Christine Choirat. Essential intersection and approximation results for robust optimization. *Journal of Nonlinear and Convex Analysis*, 15(5):979–1002, 2014.

[235] Are Hjørungnes and David Gesbert. Complex-Valued Matrix Differentiation: Techniques and Key Results. *IEEE Transactions on Signal Processing*, 55(6):2740–2746, 2007.

[236] John H. Holland and John H. Miller. Artificial Adaptive Agents in Economic Theory. *The American Economic Review*, 81(2):365–370, 1991.

[237] Abbey B. Holt, Eszter Kormann, Alessandro Gulberti, Monika Pötter-Nerger, Colin G. McNamara, Hayriye Cagnan, Magdalena K. Baaske, Simon Little, Johannes A. Köppen, Carsten Buhmann, Manfred Westphal, Christian Gerloff, Andreas K. Engel, Peter Brown, Wolfgang Hamel, Christian K.E. Moll, and Andrew Sharott. Phase-Dependent Suppression of Beta Oscillations in Parkinson's Disease Patients. *The Journal of Neuroscience*, 39(6):1119–1134, 2019.

[238] Cars H. Hommes. Chapter 23 - Heterogeneous Agent Models in Economics and Finance. In *Handbook of Computational Economics*, volume 2, pages 1109–1186. Elsevier, 2006.

[239] Cars H. Hommes and Florian Wagener. Chapter 4 - Complex Evolutionary Systems in Behavioral Finance. In *Handbook of Financial Markets: Dynamics and Evolution*, pages 217–276. Elsevier, 2009.

[240] Yongmiao Hong and Halbert White. Asymptotic Distribution Theory for Nonparametric Entropy Measures of Serial Dependence. *Econometrica*, 73(3):837–901, 2005.

[241] Jianhua Z. Huang. Projection estimation in multiple regression with application to functional ANOVA models. *The Annals of Statistics*, 26(1):242–272, 1998.

[242] Peter J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1(1):221–233, 1967.

[243] Thomas Henry Huxley. *On the physical basis of life.* New Haven, CO, The College Courant, 1869.

[244] Ildar A. Ibragimov and Ju. V. Linnik. *Independent and stationary sequences of random variables.* Wolters-Noordhoff Publishing, Groningen, 1971.

[245] Michael Iltis. Sharp asymptotics of large deviations in R^d. *Journal of Theoretical Probability*, 8(3):501–522, 1995.

[246] IUCN. Giraffa camelopardalis (amended version of 2016 assessment). In *The IUCN Red List of Threatened Species 2018: e.T9194A136266699.* 2018.

[247] Robert I. Jennrich. Asymptotic Properties of Non-Linear Least Squares Estimators. *The Annals of Mathematical Statistics*, 40(2):633–643, 1969.

[248] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax Estimation of Functionals of Discrete Distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

[249] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Maximum Likelihood Estimation of Functionals of Discrete Distributions. *IEEE Transactions on Information Theory*, 63(10):6774–6798, 2017.

[250] YuGwon Jo and Nam Z. Cho. Acceleration and Real Variance Reduction in Continuous-Energy Monte Carlo Whole-Core Calculation via p-CMFD Feedback. *Nuclear Science and Engineering*, 189(1):26–40, 2018.

[251] Harry Joe. Relative Entropy Measures of Multivariate Dependence. *Journal of the American Statistical Association*, 84(405):157–164, 1989.

[252] Daniel Kahneman. Maps of Bounded Rationality: Psychology for Behavioral Economics. *American Economic Review*, 93(5):1449–1475, 2003.

[253] Alexei Kaltchenko, Nina Timofeeva, and Eugeniy A. Timofeev. Bias Reduction of the Nearest Neighbor Entropy Estimator. *International Journal of Bifurcation and Chaos*, 18(12):3781–3787, 2008.

[254] Alexei Kaltchenko, En-hui Yang, and Nina Timofeeva. Entropy Estimators with Almost Sure Convergence and an $O(n^{-1})$ Variance. In *2007 IEEE Information Theory Workshop*, pages 644–649, Tahoe City, CA, 2007. IEEE.

[255] Tosio Katō. *Perturbation theory for linear operators.* Number 132 in Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen. Springer, Berlin, second edition, 1984.

[256] Stuart A. Kauffman. Cambrian explosion and Permian quiescence: implications of rugged fitness landscapes. *Evolutionary Ecology*, 3(3):274–281, 1989.

[257] Stuart A. Kauffman and Edward D. Weinberger. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology*, 141(2):211–245, 1989.

[258] John G. Kemeny. Generalization of a fundamental matrix. *Linear Algebra and its Applications*, 38:193–206, 1981.

[259] John G. Kemeny and J. Laurie Snell. *Finite markov chains: with a new appendix "generalization of a fundamental matrix".* Undergraduate texts in mathematics. Springer, New York, N.Y., reprint edition, 1983.

[260] Matthew B. Kennel, Jonathon Shlens, Henry D. I. Abarbanel, and E. J. Chichilnisky. Estimating Entropy Rates with Bayesian Confidence Intervals. *Neural Computation*, 17(7):1531–1576, 2005.

[261] Eamonn Keogh and Abdullah Mueen. Curse of Dimensionality. In Claude Sammut and Geoffrey I. Webb, editors, *Encyclopedia of Machine Learning and Data Mining*, pages 314–315. Springer US, Boston, MA, 2017.

[262] Alan P. Kirman. The Intrinsic Limits of Modern Economic Theory: The Emperor has No Clothes. *The Economic Journal*, 99(395):126–139, 1989.

[263] Alan P. Kirman. Whom or What Does the Representative Individual Represent? *Journal of Economic Perspectives*, 6(2):117–136, 1992.

[264] Alan P. Kirman. Ants And Nonoptimal Self-Organization: Lessons For Macroeconomics. *Macroeconomic Dynamics*, 20(2):601–621, 2016.

[265] Jack P.C. Kleijnen. Kriging metamodeling in simulation: A review. *European Journal of Operational Research*, 192(3):707–716, 2009.

[266] Jack P.C. Kleijnen. Regression and Kriging metamodels with their experimental designs in simulation: A review. *European Journal of Operational Research*, 256(1):1–16, 2017.

[267] Thorbjørn Knudsen, Daniel A. Levinthal, and Phanish Puranam. Editorial: A model is a model. *Strategy Science*, 4(1), 2019.

[268] Ioannis Kontoyiannis. Asymptotic Recurrence and Waiting Times for Stationary Processes. *Journal of Theoretical Probability*, 11(3):795–811, 1998.

[269] Ioannis Kontoyiannis, Paul H. Algoet, Yuri M. Suhov, and Abraham J. Wyner. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Transactions on Information Theory*, 44(3):1319–1327, 1998.

[270] Tjalling C. Koopmans and Olav Reiersøl. The Identification of Structural Characteristics. *The Annals of Mathematical Statistics*, 21(2):165–181, 1950.

[271] Daniel G. Krige. A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139, 1951.

[272] Dennis Kristensen and Yongseok Shin. Estimation of dynamic models with nonparametric simulated maximum likelihood. *Journal of Econometrics*, 167(1):76–94, 2012.

[273] Joseph B. Kruskal. An Overview of Sequence Comparison: Time Warps, String Edits, and Macromolecules. *SIAM Review*, 25(2):201–237, 1983.

[274] Sergei Kucherenko, Daniel Albrecht, and Andrea Saltelli. Exploring multi-dimensional spaces: a Comparison of Latin Hypercube and Quasi Monte Carlo Sampling Techniques. *arXiv:1505.02350 [stat]*, 2015. arXiv: 1505.02350.

[275] Dimitris Kugiumtzis. Partial transfer entropy on rank vectors. *The European Physical Journal Special Topics*, 222(2):401–420, 2013.

[276] Mykhailo Kuian, Lothar Reichel, and Sergij Shiyanovskii. Optimally Conditioned Vandermonde-Like Matrices. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1399–1424, 2019.

[277] Jiri Kukacka and Jozef Barunik. Estimation of financial agent-based models with simulated maximum likelihood. *Journal of Economic Dynamics and Control*, 85:21–45, 2017.

[278] Solomon Kullback and Richard A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[279] Mark Kuperberg. The Two Faces Of Emergence In Economics. *Soundings: An Interdisciplinary Journal*, 90(1/2):49–63, 2007.

[280] Finn E. Kydland and Edward C. Prescott. Time to Build and Aggregate Fluctuations. *Econometrica*, 50(6):1345, 1982.

[281] Francesco Lamperti. Empirical validation of simulated models through the GSL-div: an illustrative application. *Journal of Economic Interaction and Coordination*, 13(1):143–171, 2018.

[282] Francesco Lamperti. An information theoretic criterion for empirical validation of simulation models. *Econometrics and Statistics*, 5:83–106, 2018.

[283] Francesco Lamperti, Andrea Roventini, and Amir Sani. Agent-based model calibration using machine learning surrogates. *Journal of Economic Dynamics and Control*, 90:366–389, 2018.

[284] David C. Lane. The power of the bond between cause and effect: Jay Wright Forrester and the field of system dynamics. *System Dynamics Review*, 23(2-3):95–118, 2007.

[285] Pierre-Simon de Laplace. *Théorie analytique des probabilités*. Veuve Courcier, Paris, 1812.

[286] Pierre-Simon de Laplace. *Essai philosophique sur les probabilités*. Veuve Courcier, Paris, 1814.

[287] Paul R. Lawrence and Jay W. Lorsch. Differentiation and integration in complex organizations. *Administrative Science Quarterly*, 12(1):1–47, 1967.

[288] Mario Lazzarini. Un'applicazione del calcolo della probabilità alla ricerca sperimentale di un valore approssimato di π. *Periodico di Matematica per l'insegnamento secondario*, IV(II):140–143, 1901.

[289] Blake LeBaron. Chapter 24 - Agent-based Computational Finance. In *Handbook of Computational Economics*, volume 2, pages 1187–1233. Elsevier, 2006.

[290] Blake LeBaron and Leigh Tesfatsion. Modeling Macroeconomies as Open-Ended Dynamic Systems of Interacting Agents. *American Economic Review*, 98(2):246–250, 2008.

[291] Michel Ledoux. *The concentration of measure phenomenon.* Number v. 89 in Mathematical surveys and monographs. American Mathematical Society, Providence, R.I., 2001.

[292] Lung-Fei Lee. On Efficiency of Methods of Simulated Moments and Maximum Simulated Likelihood Estimation of Discrete Response Models. *Econometric Theory*, 8(4):518–552, 1992.

[293] Lung-Fei Lee. Asymptotic Bias in Simulated Maximum Likelihood Estimation of Discrete Choice Models. *Econometric Theory*, 11(3):437–483, 1995.

[294] Erich L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses.* Springer texts in statistics. Springer, New York, N.Y., third edition, 2005.

[295] Matthias Lengnick and Hans-Werner Wohltmann. Agent-based financial markets and New Keynesian macroeconomics: a synthesis. *Journal of Economic Interaction and Coordination*, 8(1):1–32, 2013.

[296] Steven R. Lerman and Charles F. Manski. On the Use of Simulated Frequencies to Approximate Choice Probabilities. In Charles F. Manski and Daniel McFadden, editors, *Structural Analysis of Discrete Data with Econometric Applications.* The MIT Press, Cambridge, MA, 1981.

[297] Daniel A. Levinthal. Adaptation on rugged landscapes. *Management Science*, 43(7):934–950, 1997.

[298] Arthur Lewbel. The Identification Zoo: Meanings of Identification in Econometrics. *Journal of Economic Literature*, 57(4):835–903, 2019.

[299] Lin Li, Il Memming Park, Sohan Seth, Justin C. Sanchez, and José C. Principe. Functional Connectivity Dynamics Among Cortical Neurons: A Dependence Analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(1):18–30, 2012.

[300] T. Warren Liao. Clustering of time series data—a survey. *Pattern Recognition*, 38(11):1857–1874, 2005.

[301] Friedrich Liese and Igor Vajda. On Divergences and Informations in Statistics and Information Theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

[302] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[303] Damiano Lombardi and Sanjay Pant. Nonparametric k-nearest-neighbor entropy estimator. *Physical Review E*, 93(1):013310, 2016.

[304] John B. Long and Charles I. Plosser. Real Business Cycles. *Journal of Political Economy*, 91(1):39–69, 1983.

[305] George G. Lorentz. *Approximation of functions.* Holt, Rinehart and Winston, New York, N.Y., 1966.

[306] Edward N. Lorenz. Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences*, 20(2):130–141, 1963.

[307] Iris Lorscheid and Matthias Meyer. Divide and conquer: Configuring submodels for valid and efficient analyses of complex simulation models. *Ecological Modelling*, 326:152–161, 2016.

[308] Alfred J. Lotka. Analytical Note on Certain Rhythmic Relations in Organic Systems. *Proceedings of the National Academy of Sciences*, 6(7):410–415, 1920.

[309] Alfred J. Lotka. *Elements of Physical Biology.* Williams & Wilkins Company, Baltimore, MD, 1925.

[310] R. Duncan Luce. A survey of the theory of selective information and some of its behavioral applications. In R. Duncan Luce, editor, *Developments in Mathematical Psychology*, pages 1–119. Free Press, Glencoe, 1960.

[311] R. Duncan Luce. Whatever Happened to Information Theory in Psychology? *Review of General Psychology*, 7(2):183–188, 2003.

[312] Yves Lucet. Faster than the Fast Legendre Transform, the Linear-time Legendre Transform. *Numerical Algorithms*, 16(2):171–185, 1997.

[313] Yves Lucet. What Shape Is Your Conjugate? A Survey of Computational Convex Analysis and Its Applications. *SIAM Journal on Optimization*, 20(1):216–250, 2009.

[314] Thomas Lux. Estimation of agent-based models using sequential Monte Carlo methods. *Journal of Economic Dynamics and Control*, 91:391–408, 2018.

[315] Thomas Lux and Remco C.J. Zwinkels. Empirical Validation of Agent-Based Models. In *Handbook of Computational Economics*, volume 4, pages 437–488. Elsevier, 2018.

[316] George W. Mackey. Ergodic theory and its significance for statistical mechanics and probability theory. *Advances in Mathematics*, 12(2):178–268, 1974.

[317] Michael W. Macy and Robert Willer. From Factors to Actors: Computational Sociology and Agent-Based Modeling. *Annual Review of Sociology*, 28(1):143–166, 2002.

[318] Jens Koed Madsen, Richard Bailey, Ernesto Carrella, and Philipp Koralus. Analytic Versus Computational Cognitive Models: Agent-Based Modeling as a Tool in Cognitive Sciences. *Current Directions in Psychological Science*, 28(3):299–305, 2019.

[319] Elena Maggi and Elena Vallino. Understanding urban mobility and the impact of public policies: The role of the agent-based models. *Research in Transportation Economics*, 55:50–59, 2016.

[320] Jan R. Magnus and Heinz Neudecker. *Matrix differential calculus with applications in statistics and econometrics.* John Wiley, New York, N.Y., rev. edition, 1999.

[321] Jan R. Magnus and Heinz Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, New York, N.Y., third edition, 2019.

[322] Leonid E. Maĭstrov. *Probability theory: a historical sketch*. Academic Press, New York, N.Y., 1974.

[323] Franco Malerba, Richard R. Nelson, Luigi Orsenigo, and Sidney G. Winter. 'History-friendly' models of industry evolution: the computer industry. *Industrial and Corporate Change*, 8(1):3–40, 1999.

[324] Franco Malerba, Richard R. Nelson, Luigi Orsenigo, and Sidney G. Winter. History-Friendly models: An overview of the case of the Computer Industry. *Journal of Artificial Societies and Social Simulation*, 4(3):6, 2001.

[325] Edmond Malinvaud. The Consistency of Nonlinear Regressions. *The Annals of Mathematical Statistics*, 41(3):956–969, 1970.

[326] Charles F. Manski. Closest Empirical Distribution Estimation. *Econometrica*, 51(2):305, 1983.

[327] Gianluca Manzo. Variables, Mechanisms, and Simulations: Can the Three Methods Be Synthesized? A Critical Analysis of the Literature. *Revue française de sociologie*, 48(5):35–71, 2007.

[328] James G. March. Exploration and exploitation in organizational learning. *Organization Science*, 2(1):71–87, 1991.

[329] Robert E. Marks. Validation and model selection: Three similarity measures compared. *Complexity Economics*, 2(1):41–61, 2013.

[330] George Marsaglia. The Marsaglia random number CDROM including the DieHard battery of tests of randomness, 1995.

[331] Michael Mäs, Andreas Flache, Károly Takács, and Karen A Jehn. In the short term we divide, in the long term we unite: Demographic crisscrossing and the effects of faultlines on subgroup polarization. *Organization science*, 24(3):716–736, 2013.

[332] Maria Mata and Jose Machado. Entropy Analysis of Monetary Unions. *Entropy*, 19(6):245, 2017.

[333] Ernst Mayr. *The growth of biological thought: diversity, evolution, and inheritance*. Harvard University Press, Cambridge, MA, 1982.

[334] Deirdre N. McCloskey. *The rhetoric of economics*. Rhetoric of the human sciences. University of Wisconsin Press, Madison, WI, second edition, 1998.

[335] Daniel McFadden. A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration. *Econometrica*, 57(5):995–1026, 1989.

[336] Donella H. Meadows, Dennis L. Meadows, Jørgen Randers, and William W. III Behrens, editors. *The Limits to growth: a report for the Club of Rome's project on the predicament of mankind*. Universe Books, New York, N.Y., 1972.

[337] Nicholas Metropolis. The beginning of the Monte Carlo method. In *Los Alamos Sci.*, number 15, Special Issue, pages 125–130. 1987.

[338] Nicholas Metropolis and Stanislaw Ulam. The Monte Carlo method. *Journal of the American Statistical Association*, 44(247):335–341, 1949.

[339] Alexander Michaelides and Serena Ng. Estimating the rational expectations model of speculative storage: A Monte Carlo comparison of three simulation estimators. *Journal of Econometrics*, 96(2):231–266, 2000.

[340] Giovanni Migliorati. Multivariate Markov-type and Nikolskii-type inequalities for polynomials associated with downward closed multi-index sets. *Journal of Approximation Theory*, 189:137–159, 2015.

[341] John S. Mill. *A System of Logic, Ratiocinative and Inductive*, volume 1. John W. Parker, West Strand, London, 1843.

[342] George A. Miller. Note on the bias of information estimates. In H. Quastler, editor, *Information Theory in Psychology; Problems and Methods*, pages 95–100. Free Press, Glencoe, IL, 1955.

[343] John H. Miller and Scott E. Page. *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*. Princeton University Press, Princeton, NJ, 2007.

[344] Kent D. Miller. Agent-Based Modeling and Organization Studies: A critical realist perspective. *Organization Studies*, 36(2):175–196, 2015.

[345] Kent D. Miller and Shu-Jou Lin. Different Truths in Different Worlds. *Organization Science*, 21(1):97–114, 2010.

[346] Kent D. Miller, Brian T. Pentland, and Seungho Choi. Dynamics of Performing and Remembering Organizational Routines: Performing and Remembering Organizational Routines. *Journal of Management Studies*, 49(8):1536–1558, 2012.

[347] Conwy L. Morgan. *Emergent evolution: the Gifford lectures, delivered in the University of St. Andrews in the year 1922*. Henry Holt and Company, Williams and Norgate, New York, N.Y., 1923.

[348] Conwy L. Morgan. *The emergence of novelty*. Williams & Norgate, London, 1933.

[349] John F. Muth. Rational Expectations and the Theory of Price Movements. *Econometrica*, 29(3):315–335, 1961.

[350] Richard E. Nance. A History of Discrete Event Simulation Programming Languages. In *The Second ACM SIGPLAN Conference on History of Programming Languages*, HOPL-II, pages 149–175, New York, N.Y., 1993. ACM.

[351] Richard E. Nance. A history of discrete event simulation programming languages. In *History of programming languages—II*, pages 369–427. ACM, New York, N.Y., 1996.

[352] Alexey A. Naumov, Vladimir G. Spokoiny, Yuri E. Tavyrikov, and Vladimir V. Ulyanov. Nonasymptotic Estimates for the Closeness of Gaussian Measures on Balls. *Doklady Mathematics*, 98(2):490–493, 2018.

[353] Whitney K. Newey. Convergence rates for series estimators. In Gangadharrao S. Maddala, Peter C. B. Phillips, and Thirukodikaval N. Srinavasan, editors, *Advances in Econometrics and Quantitative Economics: Essays in Honor of Professor C.R. Rao*, pages 254–275. Blackwell, Oxford, Cambridge, 1995.

[354] Whitney K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168, 1997.

[355] Whitney K. Newey and Daniel McFadden. Chapter 36 - Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier, 1994.

[356] Whitney K. Newey and Kenneth D. West. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3):703–708, 1987.

[357] Whitney K. Newey and Kenneth D. West. Automatic Lag Selection in Covariance Matrix Estimation. *The Review of Economic Studies*, 61(4):631–653, 1994.

[358] Peter E. Ney. Dominating Points and the Asymptotics of Large Deviations for Random Walk on R^d. *The Annals of Probability*, 11(1):158–167, 1983.

[359] Peter E. Ney. Convexity and Large Deviations. *The Annals of Probability*, 12(3):903–906, 1984.

[360] Peter E. Ney and Stephen M. Robinson. Polyhedral Approximation of Convex Sets With an Application to Large Deviation Probability Theory. *Journal of Convex Analysis*, 2(1/2):229–240, 1995.

[361] Tim Ng and Matthew Wright. Introducing the MONIAC: an early and innovative economic model. *Reserve Bank of New Zealand: Bulletin*, 70(4):46–52, 2007.

[362] Muaz Niazi and Amir Hussain. Agent-based computing from multi-agent systems to agent-based models: a visual survey. *Scientometrics*, 89(2):479–499, 2011.

[363] Harald Niederreiter. *Random number generation and quasi-Monte Carlo methods.* Number 63 in CBMS-NSF regional conference series in applied mathematics. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.

[364] Fredrik Nilsson and Vince Darley. On complex adaptive systems and agent-based modelling for improving decision-making in manufacturing and logistics settings: Experiences from a packaging company. *International Journal of Operations & Production Management*, 26(12):1351–1373, 2006.

[365] Michael J. North and Charles M. Macal. *Managing business complexity: discovering strategic solutions with agent-based modeling and simulation.* Oxford University Press, New York, N.Y., 2007.

[366] Frank W. J. Olver, Daniel W. Lozier, Ronald F. Boisvert, and Charles W. Clark, editors. *NIST handbook of mathematical functions.* Cambridge University Press, New York, N.Y., first edition, 2010.

[367] Guy H. Orcutt. A New Type of Socio-Economic System. *The Review of Economics and Statistics*, 39(2):116–123, 1957.

[368] Naomi Oreskes, Kristin Shrader-Frechette, and Kenneth Belitz. Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science*, 263(5147):641–646, 1994.

[369] Elinor Ostrom. The Ten Most Important Books. *Tidsskriftet Politik*, 4(7):36–48, 2004.

[370] Art B. Owen. Multidimensional Variation for Quasi-Monte Carlo. In *Contemporary Multivariate Analysis and Design of Experiments*, pages 49–74. World Scientific, 2005.

[371] Ariel Pakes. Patents as Options: Some Estimates of the Value of Holding European Patent Stocks. *Econometrica*, 54(4):755–784, 1986.

[372] Ariel Pakes and David Pollard. Simulation and the Asymptotics of Optimization Estimators. *Econometrica*, 57(5):1027–1057, 1989.

[373] Milan Paluš. Testing for nonlinearity using redundancies: quantitative and qualitative aspects. *Physica D: Nonlinear Phenomena*, 80(1-2):186–205, 1995.

[374] Milan Paluš. Coarse-grained entropy rates for characterization of complex time series. *Physica D: Nonlinear Phenomena*, 93(1-2):64–77, 1996.

[375] Milan Paluš. Detecting nonlinearity in multivariate time series. *Physics Letters A*, 213(3-4):138–147, 1996.

[376] Milan Paluš, Vladimír Albrecht, and Ivan Dvořák. Information theoretic test for nonlinearity in time series. *Physics Letters A*, 175(3-4):203–209, 1993.

[377] Xue Pan, Lei Hou, Mutua Stephen, Huijie Yang, and Chenping Zhu. Evaluation of Scaling Invariance Embedded in Short Time Series. *PLoS ONE*, 9(12):e116128, 2014.

[378] Liam Paninski. Estimation of Entropy and Mutual Information. *Neural Computation*, 15(6):1191–1253, 2003.

[379] Liam Paninski and Masanao Yajima. Undersmoothed Kernel Entropy Estimators. *IEEE Transactions on Information Theory*, 54(9):4384–4388, 2008.

[380] Carl F. A. Pantin. *Relations Between Sciences.* Cambridge University Press, Cambridge, 1968.

[381] Stefano Panzeri and Alessandro Treves. Analytical estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems*, 7(1):87–107, 1996.

[382] Stergios Papadimitriou, Seferina Mavroudi, and Spiridon D. Likothanassis. Mutual Information Clustering for Efficient Mining of Fuzzy Association Rules with Application to Gene Expression Data Analysis. *International Journal on Artificial Intelligence Tools*, 15(02):227–250, 2006.

[383] Maria Papapetrou and Dimitris Kugiumtzis. Markov chain order estimation with conditional mutual information. *Physica A: Statistical Mechanics and its Applications*, 392(7):1593–1601, 2013.

[384] Maria Papapetrou and Dimitris Kugiumtzis. Markov chain order estimation with parametric significance tests of conditional mutual information. *Simulation Modelling Practice and Theory*, 61:1–13, 2016.

[385] Teemu Pennanen and Matti Koivu. Epi-convergent discretizations of stochastic programs via integration quadratures. *Numerische Mathematik*, 100(1):141–163, 2005.

[386] Gilles Pisier. Probabilistic methods in the geometry of Banach spaces. In Giorgio Letta and Maurizio Pratelli, editors, *Probability and Analysis*, Lecture Notes in Mathematics, pages 167–241, Berlin, Heidelberg, 1986. Springer.

[387] Donovan Platt. A comparison of economic agent-based model calibration methods. *Journal of Economic Dynamics and Control*, 113:103859, 2020.

[388] Dimitris N. Politis. Higher-order accurate, positive semidefinite estimation of large-sample covariance and spectral density matrices. *Econometric Theory*, 27(4):703–744, 2011.

[389] David Pollard. The minimum distance method of testing. *Metrika*, 27(1):43–70, 1980.

[390] David Pollard. *A user's guide to measure theoretic probability.* Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, New York, N.Y., 2002.

[391] Bernd Pompe. Measuring statistical dependences in a time series. *Journal of Statistical Physics*, 73(3-4):587–610, 1993.

[392] Ithiel De Sola Pool and Robert P. Abelson. The Simulmatics Project. *Public Opinion Quarterly*, 25(2):167–183, 1961.

[393] Ithiel de Sola Pool, Robert P. Abelson, and Samuel L. Popkin. *Candidates, issues and strategies: a computer simulation of the 1960 and 1964 Presidential elections.* MIT Press, Cambridge, MA, 1965.

[394] Andreas Pyka and Giorgio Fagiolo. Agent-based Modelling: A Methodology for Neo-Schumpeterian Economics. In *Elgar Companion to Neo-Schumpeterian Economics*, pages 467–488. Edward Elgar Publishing, Cheltenham, Northampton, MA, 2007.

[395] Herschel Rabitz. Systems Analysis at the Molecular Scale. *Science*, 246(4927):221–226, 1989.

[396] Stefan T. Radev, Ulf K. Mertens, Andreas Voss, and Ullrich Köthe. Towards end-to-end likelihood-free inference with convolutional neural networks. *British Journal of Mathematical and Statistical Psychology*, 73(1):23–43, 2020.

[397] Steven F. Railsback and Volker Grimm. *Agent-based and individual-based modeling: a practical introduction.* Princeton University Press, Princeton, N.J., 2012.

[398] Alexander G. Ramm and Alexander I. Zaslavsky. Reconstructing singularities of a function from its Radon transform. *Mathematical and Computer Modelling*, 18(1):109–138, 1993.

[399] Werner Raub and Thomas Voss. Micro-Macro Models in Sociology: Antecedents of Coleman's Diagram. In Ben Jann and Wojtek Przepiorka, editors, *Social dilemmas, institutions, and the evolution of cooperation.* De Gruyter, Berlin, Boston, MA, 2017.

[400] Louis Raynal, Jean-Michel Marin, Pierre Pudlo, Mathieu Ribatet, Christian P Robert, and Arnaud Estoup. ABC random forests for Bayesian parameter inference. *Bioinformatics*, 35(10):1720–1728, 2019.

[401] Maria Cristina Recchioni, Gabriele Tedeschi, and Mauro Gallegati. A calibration procedure for analyzing stock price dynamics in an agent-based framework. *Journal of Economic Dynamics and Control*, 60:1–25, 2015.

[402] Mark D. Reid and Robert C. Williamson. Information, Divergence and Risk for Binary Experiments. *Journal of Machine Learning Research*, 12:731–817, 2011.

[403] M. Isabel Reis dos Santos and Pedro M. Reis dos Santos. Switching regression metamodels in stochastic simulation. *European Journal of Operational Research*, 251(1):142–147, 2016.

[404] Henggang Ren, Yue Yang, Changgui Gu, Tongfeng Weng, and Huijie Yang. A Patient Suffering From Neurodegenerative Disease May Have a Strengthened Fractal Gait Rhythm. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(9):1765–1772, 2018.

[405] Sidney I. Resnick. *A probability path.* Birkhäuser, Boston, MA, 1999.

[406] Matteo Richiardi, Roberto Leombruni, Nicole J. Saam, and Michele Sonnessa. A Common Protocol for Agent-Based Social Simulation. *Journal of Artificial Societies and Social Simulation*, 9(1), 2006.

[407] E. Rio. About the Lindeberg method for strongly mixing sequences. *ESAIM: Probability and Statistics*, 1:35–61, 1997.

[408] Emmanuel Rio. Sur le théorème de Berry-Esseen pour les suites faiblement dépendantes. *Probability Theory and Related Fields*, 104(2):255–282, 1996.

[409] Emmanuel Rio. *Asymptotic Theory of Weakly Dependent Random Processes*, volume 80 of *Probability Theory and Stochastic Modelling.* Springer, Berlin, Heidelberg, 2017.

[410] Peter M. Robinson. On the Asymptotic Properties of Estimators of Models Containing Limited Dependent Variables. *Econometrica*, 50(1):27–41, 1982.

[411] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, Heidelberg, 1998.

[412] M. S. Rogers and B. F. Green. The moments of sample information when the alternatives are equally likely. In H. Quastler, editor, *Information Theory in Psychology; Problems and Methods*, pages 101–108. Free Press, Glencoe, IL, 1955.

[413] Christian M. Rohwer, Florian Angeletti, and Hugo Touchette. Convergence of large-deviation estimators. *Physical Review E*, 92(5):052104, 2015.

[414] David Romer. *Advanced macroeconomics*. McGraw-Hill/Irwin, New York, N.Y., fourth edition, 2012.

[415] Anthony Rosato, Friedrich Prinz, Katherine J. Standburg, and Robert H. Swendsen. Monte Carlo simulation of particulate matter segregation. *Powder Technology*, 49(1):59–69, 1986.

[416] Anthony Rosato, Katherine J. Strandburg, Friedrich Prinz, and Robert H. Swendsen. Why the Brazil nuts are on top: Size segregation of particulate matter by shaking. *Physical Review Letters*, 58(10):1038–1040, 1987.

[417] Thomas J. Rothenberg. Identification in Parametric Models. *Econometrica*, 39(3):577–591, 1971.

[418] Mark S. Roulston. Estimating the errors on measured entropy and mutual information. *Physica D: Nonlinear Phenomena*, 125(3-4):285–294, 1999.

[419] David Ruelle. *Chance and chaos*. Princeton University Press, Princeton, N.J., 1993.

[420] Francisco Ruge-Murcia. Estimating nonlinear DSGE models by the simulated method of moments: With an application to business cycles. *Journal of Economic Dynamics and Control*, 36(6):914–938, 2012.

[421] Daniil Ryabko and Boris Ryabko. Nonparametric Statistical Inference for Ergodic Processes. *IEEE Transactions on Information Theory*, 56(3):1430–1435, 2010.

[422] Isabelle Salle and Murat Yıldızoğlu. Efficient Sampling and Meta-Modeling for Computational Economic Models. *Computational Economics*, 44(4):507–536, 2014.

[423] Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global Sensitivity Analysis. The Primer*. John Wiley & Sons, Ltd, Chichester, 2007.

[424] Andrea Saltelli, Stefano Tarantola, Francesca Campolongo, and Marco Ratto. *Sensitivity analysis in practice: a guide to assessing scientific models*. John Wiley & Sons, Ltd, Chichester, first edition, 2004.

[425] Manuel S. Santos and Adrian Peralta-Alva. Accuracy of Simulations for Stochastic Dynamic Models. *Econometrica*, 73(6):1939–1976, 2005.

[426] Igal Sason and Sergio Verdú. f-Divergence Inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

[427] Thomas C. Schelling. Models of Segregation. *The American Economic Review*, 59(2):488–493, 1969.

[428] Thomas C. Schelling. Dynamic models of segregation. *The Journal of Mathematical Sociology*, 1(2):143–186, 1971.

[429] Thomas C. Schelling. *Micromotives and macrobehavior.* Fels lectures on public policy analysis. Norton, New York, N.Y., first edition, 1978.

[430] Hanspeter Schmidli. Estimation of the abscissa of convergence of the moment generating function. 1994.

[431] Thomas Schürmann and Peter Grassberger. Entropy estimation of symbol sequences. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 6(3):414–427, 1996.

[432] Davide Secchi. A case for agent-based models in organizational behavior and team research. *Team Performance Management: An International Journal*, 21(1/2):37–50, 2015.

[433] Davide Secchi. Boundary Conditions for the Emergence of "Docility" in Organizations: Agent-Based Model and Simulation. In Davide Secchi and Martin Neumann, editors, *Agent-Based Simulation of Organizational Behavior*, pages 175–200. Springer International Publishing, Cham, 2016.

[434] Davide Secchi and Emanuele Bardone. Super-docility in organizations: an evolutionary model. *International Journal of Organization Theory & Behavior*, 12(3):339–379, 2009.

[435] Davide Secchi and Nicole Gullekson. Individual and organizational conditions for the emergence and evolution of bandwagons. *Computational and Mathematical Organization Theory*, 22(1):88–133, 2016.

[436] Davide Secchi and Martin Neumann, editors. *Agent-based simulation of organizational behavior. New frontiers of social science research.* Springer, New York, N.Y., 2016.

[437] Davide Secchi and Raffaello Seri. Controlling for false negatives in agent-based models: a review of power analysis in organizational research. *Computational and Mathematical Organization Theory*, 23(1):94–121, 2017.

[438] Emilio Segrè. *From X-rays to quarks: modern physicists and their discoveries.* W. H. Freeman, San Francisco, CA, 1980.

[439] Amartya Sen. Maximization and the Act of Choice. *Econometrica*, 65(4):745–779, 1997.

[440] Eugene Seneta. Sensitivity of finite Markov chains under perturbation. *Statistics & Probability Letters*, 17(2):163–168, 1993.

[441] Robert J. Serfling, editor. *Approximation Theorems of Mathematical Statistics.* Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, 1980.

[442] Raffaello Seri. Statistical properties of b-adic diaphonies. *Mathematics of Computation*, 86(304):799–828, 2016.

[443] Raffaello Seri and Christine Choirat. Scenario Approximation of Robust and Chance-Constrained Programs. *Journal of Optimization Theory and Applications*, 158(2):590–614, 2013.

[444] Raffaello Seri, Mario Martinoli, Davide Secchi, and Samuele Centorrino. Model Calibration and Validation via Confidence Sets. *Econometrics and Statistics*, 2020.

[445] Raffaello Seri and Davide Secchi. How Many Times Should One Run a Computational Simulation? In B. Edmonds and R. Meyer, editors, *Simulating Social Complexity: A Handbook*, Understanding Complex Systems, pages 229–251. Springer International Publishing, Cham, 2017.

[446] Alexei Shadrin. Twelve proofs of the Markov inequality. In Dimitar K. Dimitrov, Geno Nikolov, and Rumen Uluchev, editors, *Approximation Theory: A volume dedicated to Borislav Bojanov'*, pages 233–299. Marin Drinov Academic Publishing House, Sofia, 2004.

[447] Claude E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[448] Bruce Shawyer and Bruce Watson. *Borel's methods of summability: theory and applications.* Oxford mathematical monographs. Clarendon Press, Oxford, first edition, 1994.

[449] Sara Sheehan and Yun S. Song. Deep Learning for Population Genetic Inference. *PLOS Computational Biology*, 12(3):e1004845, 2016.

[450] Galen R. Shorack. *Probability for statisticians.* Springer texts in statistics. Springer, New York, N.Y., 2000.

[451] Kevin M. Short. Direct Calculation of Metric Entropy from Time Series. *Journal of Computational Physics*, 104(1):162–172, 1993.

[452] Herbert A. Simon. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69(1):99–118, 1955.

[453] Herbert A. Simon. Rational decision making in business organizations. *American Economic Review*, 69(4):493–513, 1979.

[454] Herbert A. Simon. A mechanism for social selection and successful altruism. *Science*, 250(4988):1665–1668, 1990.

[455] Herbert A. Simon. Altruism and Economics. *American Economic Review*, 83(2):156–61, 1993.

[456] Herbert A. Simon. *Administrative behavior: a study of decision-making processes in administrative organizations*. Free Press, New York, N.Y., fourth edition, 1997.

[457] Christopher A Sims. Macroeconomics and Methodology. *Journal of Economic Perspectives*, 10(1):105–120, 1996.

[458] Scott A. Sisson, Yanan Fan, and Mark A. Beaumont, editors. *Handbook of Approximate Bayesian Computation*. CRC Press, Boca Raton, FL, first edition, 2018.

[459] Adam Smith. *An inquiry into the nature and causes of the wealth of nations*, volume 2. Printed for W. Strahan and T. Cadell, in the Strand, London, 1776.

[460] Anthony A. Smith Jr. Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics*, 8(S1):S63–S84, 1993.

[461] Jerome Spanier. Monte Carlo Methods. In *Nuclear Computational Science*, pages 117–165. Springer Netherlands, Dordrecht, 2010.

[462] Flaminio Squazzoni. *Agent-Based Computational Sociology*. John Wiley & Sons, Ltd, Chichester, 2012.

[463] Sorawoot Srisuma. Minimum distance estimators for dynamic games: Minimum distance estimators for dynamic games. *Quantitative Economics*, 4(3):549–583, 2013.

[464] John D. Sterman. System Dynamics Modeling: Tools for Learning in a Complex World. *California Management Review*, 43(4):8–25, 2001.

[465] Steven Stern. Simulation-Based Estimation. *Journal of Economic Literature*, 35(4):2006–2039, 1997.

[466] Lei Sun and Alexander G. Nikolaev. Mutual Information Based Matching for Causal Inference with Observational Data. *Journal of Machine Learning Research*, 17(1):6990–7020, 2016.

[467] Wai-Yuan Tan. On the distribution of quadratic forms in normal random variables. *Canadian Journal of Statistics*, 5(2):241–250, 1977.

[468] James M. Taylor. The condition of Gram matrices and related problems. *Proceedings of the Royal Society of Edinburgh: Section A Mathematics*, 80(1-2):45–56, 1978.

[469] Guus ten Broeke, George van Voorn, and Arend Ligtenberg. Which Sensitivity Analysis Method Should I Use for My Agent-Based Model? *Journal of Artificial Societies and Social Simulation*, 19(1):5, 2016.

[470] Leigh Tesfatsion. Chapter 16 - Agent-Based Computational Economics: A Constructive Approach to Economic Theory. In *Handbook of Computational Economics*, volume 2, pages 831–880. Elsevier, 2006.

[471] Richard H. Thaler and Cass R. Sunstein. *Nudge: improving decisions about health, wealth, and happiness.* Penguin Books, New York, N.Y., rev. and expanded edition, 2009.

[472] Jan C. Thiele, Winfried Kurth, and Volker Grimm. Facilitating Parameter Estimation and Sensitivity Analysis of Agent-Based Models: A Cookbook Using NetLogo and 'R'. *Journal of Artificial Societies and Social Simulation*, 17(3), 2014.

[473] Yu-Chu Tian and Furong Gao. Extraction of delay information from chaotic time series based on information entropy. *Physica D: Nonlinear Phenomena*, 108(1-2):113–118, 1997.

[474] Yung L. Tong. *The Multivariate Normal Distribution.* Springer Series in Statistics. Springer-Verlag, New York, N.Y., first edition, 1990.

[475] Kenneth Train. *Discrete choice methods with simulation.* Cambridge University Press, New York, N.Y., second edition, 2009.

[476] Klaus G. Troitzsch. Perspectives and Challenges of Agent-based Simulation As a Tool for Economics and Other Social Sciences. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '09, pages 35–42, Richland, SC, 2009. International Foundation for Autonomous Agents and Multiagent Systems.

[477] Paola Tubaro and Antonio A. Casilli. "An Ethnographic Seduction": How Qualitative Research and Agent-based Models can Benefit Each Other. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 106(1):59–74, 2010.

[478] Stanislaw Ulam. On some mathematical problems connected with patterns of growth in figures. In Richard E. Bellman, editor, *Mathematical Problems in the Biological Sciences*, number 14 in Proceedings of Symposia in Applied Mathematics, pages 215–224. American Mathematical Society, Providence, RI, 1962.

[479] Michel Valadier. Stationary stochastic processes are mixing of ergodic ones: Contingency. *Journal of Convex Analysis*, 18(4):1127–1140, 2011.

[480] Gregory Valiant and Paul Valiant. Estimating the unseen: improved estimators for entropy and other properties. *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2:2157–2165, 2013.

[481] Sara van de Geer. On the asymptotic variance of the debiased Lasso. *Electronic Journal of Statistics*, 13(2):2970–3008, 2019.

[482] Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3(0):1360–1392, 2009.

[483] Aad W. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, Cambridge, first edition, 1998.

[484] Dejan Vinković and Alan P. Kirman. A physical analogue of the Schelling model. *Proceedings of the National Academy of Sciences*, 103(51):19261–19265, 2006.

[485] Vito Volterra. Fluctuations in the Abundance of a Species considered Mathematically. *Nature*, 118(2972):558–560, 1926.

[486] Vito Volterra. Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Atti della R. Accademia nazionale dei Lincei. Memorie della Classe di scienze fisiche, matematiche e naturali*, 2(III):31–113, 1926.

[487] John von Neumann. The general and logical theory of automata (with discussion). In Lloyd A. Jeffress, editor, *Cerebral mechanisms in behaviour*, pages 1–41. Wiley, Chapman & Hall, New York, N.Y., 1951.

[488] John von Neumann. Various techniques used in connection with random digits. In A. S. Householder, G. E. Forsythe, and H. H. Germond, editors, *Monte carlo method*, volume 12 of *National bureau of standards applied mathematics series*, pages 36–38. US Government Printing Office, Washington, D.C., 1951.

[489] John von Neumann and Oskar Morgenstern. *Theory of games and economic behavior.* Princeton University Press, Princeton, N.J., 1944.

[490] Jochen Voss. *An introduction to statistical computing: a simulation-based approach.* Wiley series in computational statistics. Wiley, Chichester, first edition, 2013.

[491] Abraham Wald. Note on the Consistency of the Maximum Likelihood Estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 1949.

[492] Friederike Wall. Agent-based modeling in managerial science: an illustrative survey and study. *Review of Managerial Science*, 10(1):135–193, 2016.

[493] L. Wang, Kwangwon Ahn, C. Kim, and C. Ha. Agent-based models in financial market studies. *Journal of Physics: Conference Series*, 1039:012022, 2018.

[494] Qiang Wang, Yi Shen, and Jian Qiu Zhang. A nonlinear correlation measure for multivariable data set. *Physica D: Nonlinear Phenomena*, 200(3-4):287–295, 2005.

[495] Xiaoyue Wang, Abdullah Mueen, Hui Ding, Goce Trajcevski, Peter Scheuermann, and Eamonn Keogh. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, 26(2):275–309, 2013.

[496] Gerhard Weiss, editor. *Multiagent systems: a modern approach to distributed artificial intelligence.* MIT Press, Cambridge, MA, 1999.

[497] Harold Widom and Herbert Wilf. Small eigenvalues of large Hankel matrices. *Proceedings of the American Mathematical Society*, 17(2):338–338, 1966.

[498] Paul Windrum, Giorgio Fagiolo, and Alessio Moneta. Empirical Validation of Agent-Based Models: Alternatives and Prospects. *Journal of Artificial Societies and Social Simulation*, 10(2):19, 2007.

[499] Peter Winker, Manfred Gilli, and Vahidin Jeleskovic. An objective function for simulation based inference on exchange rate data. *Journal of Economic Interaction and Coordination*, 2(2):125–145, 2007.

[500] Yihong Wu and Pengkun Yang. Minimax Rates of Entropy Estimation on Large Alphabets via Best Polynomial Approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

[501] Li Xia and Peter W. Glynn. A Generalized Fundamental Matrix for Computing Fundamental Quantities of Markov Systems. *arXiv:1604.04343 [cs, math]*, 2016.

[502] Wanting Xiong, Luca Faes, and Plamen C. Ivanov. Entropy measures, entropy estimators, and their performance in quantifying complex dynamics: Effects of artifacts, nonstationarity, and long-range correlations. *Physical Review E*, 95(6):062114, 2017.

[503] Yue Yang, Changgui Gu, Qin Xiao, and Huijie Yang. Evolution of scaling behaviors embedded in sentence series from A Story of the Stone. *PLOS ONE*, 12(2):e0171776, 2017.

[504] Minho Yoon and Keun Lee. Agent-based and "History-Friendly" Models for Explaining Industrial Evolution. *Evolutionary and Institutional Economics Review*, 6(1):45–70, 2009.

[505] Andrei Yu. Zaitsev. Estimates of the Lévy–Prokhorov Distance in the Multivariate Central Limit Theorem for Random Variables with Finite Exponential Moments. *Theory of Probability & Its Applications*, 31(2):203–220, 1987.

[506] Wenqing Zhang, Lu Qiu, Qin Xiao, Huijie Yang, Qingjun Zhang, and Jianyong Wang. Evaluation of scale invariance in physiological signals by means of balanced estimation of diffusion entropy. *Physical Review E*, 86(5):056107, 2012.

[507] Zhiyi Zhang. Asymptotic Normality of an Entropy Estimator With Exponentially Decaying Bias. *IEEE Transactions on Information Theory*, 59(1):504–508, 2013.

[508] Zhiyi Zhang and Xing Zhang. A Normal Law for the Plug-in Estimator of Entropy. *IEEE Transactions on Information Theory*, 58(5):2745–2747, 2012.

[509] Rongxi Zhou, Ru Cai, and Guanqun Tong. Applications of Entropy in Finance: A Review. *Entropy*, 15(12):4909–4931, 2013.

[510] Antoni Zygmund. *Trigonometric series*. Cambridge mathematical library. Cambridge University Press, Cambridge, third edition, 2002.