Environmental and Ecological Statistics
https://doi.org/10.1007/s10651-025-00697-z

A discrete Kumaraswamy distribution for modeling rating and ranking data

Alessandro Barbiero[1] · **Asmerilda Hitaj**[2]

Received: 31 January 2025 / Revised: 16 November 2025 / Accepted: 20 November 2025

Abstract
We propose a discrete probability distribution supported on the first k positive integers ($k \geq 3$), defined in terms of the cumulative distribution function of the continuous two-parameter Kumaraswamy distribution supported on the interval (0,1). This distribution offers an alternative to existing models, such as the discrete Beta, the Beta-Binomial, and the CUB distributions, for modeling ordinal data, which are ubiquitous in applied disciplines. The key properties of the distribution are explored, with a particular focus on the possible shapes of its probability mass function, moments, pseudo-random simulation, and inferential procedures. Regression models where the response variable follows the proposed discrete Kumaraswamy distribution are also discussed. Analyses of different real data sets, regarding individuals' perspectives on environmental matters, are provided to practically illustrate the model introduced in this work and assess its ability to fit rating or ranking data.

Keywords Discrete distribution · Latent variable · Ordinal data · Ratings

Alessandro Barbiero and Asmerilda Hitaj have contributed equally to this work and are members of GNAMPA-INdAM.

Handling Editor: Luiz Duczmal.

✉ Alessandro Barbiero
alessandro.barbiero@unimi.it

Asmerilda Hitaj
asmerilda.hitaj@uninsubria.it

[1] Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, via Conservatorio 7, 20122 Milan, MI, Italy

[2] Department of Economics, Università degli Studi dell'Insubria, Via Monte Generoso 71, 21100 Varese, VA, Italy

Published online: 20 February 2026

1 Introduction

Ordinal data are widely utilized in the social sciences and various scientific disciplines to evaluate surveys focused on ratings, preferences, and judgments. Examples include assessing product or service quality (on a scale comprising, e.g., "poor", "fair", "good", "very good", "excellent"), defect severity (e.g., "minor", "major", "critical"), food taste preferences (e.g., "too mild", "just right", "too spicy"), and levels of agreement (e.g., "strongly disagree", "disagree", "neutral", "agree", "strongly agree") (Iannario 2014). We refer to this type of data as rating data. Similarly, ordinal data arise in contexts where items, individuals, or entities are ordered or prioritized based on certain criteria or attributes: in this context, we more appropriately refer to ranking data. An example comes from preference surveys, where respondents rank their preferences for a set of options, such as favorite products, services, or features, or from sports events, where athletes or teams are ranked based on performance and results-based metrics such as scores, times, or placements. Standard techniques model these data assuming that the observed ordinal outcomes are driven by an underlying continuous variable (Agresti 2010). This latent variable represents an unobserved factor that influences the observed ordinal categories. The ordinal data are modeled by defining a set of thresholds along the scale of the latent variable, where each threshold corresponds to a cutoff point between two adjacent ordinal categories. When the latent variable exceeds a particular threshold, the observation is assigned to a higher ordinal category, and when it is below a threshold, it is assigned to a lower category. This framework allows for a flexible representation of ordinal data, as it captures the continuous nature of the underlying process while still accounting for the discrete observed categories. Statistical methods like probit or logistic regression can then be used to relate the latent variable to other explanatory variables, providing insights into the factors driving the observed ordinal outcomes. Statistical inference is usually based on a continuous latent response distribution from a known parametric family (e.g, normal or logistic) with maximum likelihood used to estimate the parameters of the latent variable model (i.e., the thresholds, the parameters of the latent variable distribution, if any, and the coefficients for the explanatory variables). Parametric assumptions about the latent variable distribution may not be appropriate in some cases. In these circumstances, a more flexible class of statistical models is needed, which can include mixture distributions, as well as kernel and spline-based density estimation methods (Ghosh et al. 2018).

Alternative models are possible that directly specify the probability distribution or probability mass function (pmf) for the ordinal variable under study: the modeling becomes much more parsimonious, still retaining a sufficient level of flexibility, if an appropriate parametric family is selected. The multinomial distribution is a natural choice for modeling ordinal data, where the probabilities of different categories are directly parametrized. However, it does not inherently incorporate the ordering of categories. The Binomial distribution can model ordinal data directly by interpreting ordinal categories as cumulative success probabilities in a sequence of trials. The Beta-Binomial distribution is another discrete probability distribution that can be used for modeling ordinal categorical data in specific contexts. It is an extension of the Binomial distribution in which the success probability p is treated as a random

variable (rv) that follows a Beta distribution. This flexibility makes it particularly suitable for situations where overdispersion is present – a common feature in ordinal data sets. The CUB/MUB (Combination/Mixture of Uniform and Binomial) distribution has been proposed in D'Elia and Piccolo (2005) to deal with ordinal data and later discussed in Corduas et al. (2009). It is particularly suited for data where responses are influenced by two key components: *feeling* (a latent preference or inclination) and *uncertainty* (a lack of confidence or randomness in the response). The former is modeled using a shifted Binomial distribution, which reflects the tendency of respondents toward specific categories due to their genuine preference or inclination. The latter is modeled using a discrete uniform distribution over the possible ordinal categories. It accounts for the lack of certainty or random guessing in the respondent's choice. The original model parameters are π (corresponding to the mixing weight of the shifted Binomial) and ξ (its failure probability); hence, $1-\xi$ and $1-\pi$ represent the feeling and the uncertainty parameters, respectively.

The modeling flexibility of the continuous two-parameter Beta distribution has led several authors to construct different discrete analogues thereof, supported on the first k positive integers (or $k+1$ non-negative integers), with the aim of modeling ordinal data. The probability density function (pdf) of a Beta rv with shape parameters $\alpha > 0$ and $\beta > 0$ is given by

$$f_B(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \quad x \in (0,1), \alpha > 0, \beta > 0, \tag{1}$$

where $B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x$ is the Beta function acting as a normalizing constant in (1). The Beta distribution is considered highly flexible because it can model a wide range of shapes depending on its two parameters. Punzo and Zini (2008); Punzo (2010), based on a re-parameterization of the pdf of the generalized Beta distribution $f(y;k,\epsilon,m,h)$, constructed a discrete version supported on $\mathcal{S} = \{0,1,\ldots,k\}$ with pmf defined as $p(x;k,\epsilon,m,h) = f(x;k,\epsilon,m,h)/\sum_{j=0}^{k} f(j;k,\epsilon,m,h)$. A similar proposal (Turner 2021b) defined the probabilities of $1,\ldots,k$ as the values of the pdf of the Beta distribution, evaluated at equally spaced points in the unit interval (0,1), normalized to sum up to 1: $p(x;\alpha,\beta,k) = f\left(\frac{x}{k+1};\alpha,\beta\right) / \sum_{i=1}^{k} f\left(\frac{i}{k+1};\alpha,\beta\right)$. This latter discrete Beta distribution is implemented in the `R` package `dbd` (Turner 2021a). While the pmf is expressed in closed-form, the cumulative distribution function (cdf) does not admit a compact closed-form expression: it is defined as the ratio of two finite sums of Beta pdf values. In practice, this computation is straightforward in statistical software such as `R`, but it requires an explicit summation at each evaluation, which may matter in repeated likelihood calculations. Another proposal (Fasola and Sciandra 2015; Sciandra et al. 2024) suggested assigning the first k positive integers the probabilities

$$p_i = F_B(i/k;\alpha,\beta) - F_B((i-1)/k;\alpha,\beta), \quad i = 1,\ldots,k,$$

where

$$F_B(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} \int_0^x t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t$$

is the cdf of the continuous Beta distribution with parameters α and β. This discrete distribution is used as a more flexible alternative to the CUB or MUB distribution for modeling rating and ranking data. It has proven to be appropriate in terms of parsimony, flexibility and ease of interpretation; by interchanging the two shape parameters, it produces distributions symmetric to each other about the midpoint of the support, as in the continuous Beta case. In the following, we will refer to it simply as the 'discrete Beta distribution'.

In this paper, we propose a new discrete probability distribution, which arises as a discrete counterpart to the continuous two-parameter Kumaraswamy distribution on the unit interval. The construction follows that of the discrete Beta model in Sciandra et al. (2024), and the resulting distribution is similar, primarily in that it is indexed by two shape parameters and can describe a variety of shapes for the pmf. However, compared to the existing model, it has the advantage of providing closed-form expressions for both the probability distribution and the quantile function, making pseudo-random simulation easier.

The remainder of the paper is organized as follows. In the next section, we recall the continuous Kumaraswamy distribution and its main properties; then, in Section 3, we introduce a discrete counterpart supported on the first k positive integers. We provide expressions for the pmf, the cdf, and the quantile function. Next, we discuss parameter estimation. The issue of incorporating covariates into the model is also addressed, by using an appropriate reparametrization of the distribution. Section 4 illustrates the application of the proposed model to data sets from surveys on opinions and perceptions regarding environmental issues. Final remarks and research perspectives are presented in the concluding section.

2 The continuous Kumaraswamy distribution

The Kumaraswamy double-bounded distribution is a family of continuous probability distributions defined on the interval (0,1) that was first proposed in Kumaraswamy (1980). It is similar to the Beta distribution, which shares many of its properties, but is much simpler to use especially in simulation studies, since its pdf, cdf and quantile function can all be expressed in closed-form (Jones 2009). The expression for its cdf is in fact the following:

$$F_K(x;a,b) = \begin{cases} 0 & x \le 0 \\ 1-(1-x^a)^b & x \in (0,1) \\ 1 & x \ge 1 \end{cases} \tag{2}$$

with $a > 0$ and $b > 0$ being the first and second shape parameters, respectively. The expression for the quantile function is easily obtained by inverting (2):

$$x_u = F^{-1}(u; a, b) = \left[1 - (1-u)^{1/b}\right]^{1/a}, u \in (0,1).$$

The expression for the pdf is

$$f_K(x; a, b) = \begin{cases} abx^{a-1}(1-x^a)^{b-1} & x \in (0,1) \\ 0 & x \notin (0,1). \end{cases} \quad (3)$$

We note from (2) and (3) that setting $b = 1$ leads to the power distribution with parameter a. In particular, setting $a = 1$ and $b = 1$ results in the standard uniform distribution. More specifically, it can be shown (Jones 2009) that the Kumaraswamy pdf has the same basic shape properties as the Beta distribution, namely:

- $a > 1; b > 1$: unimodal (a unique absolute maximum within (0,1));
- $a < 1; b < 1$: uniantimodal (a unique global minimum within (0,1));
- $a > 1; b \leq 1$: increasing;
- $a \leq 1; b > 1$: decreasing;
- $a = b = 1$: uniform.

Both Beta and Kumaraswamy pdfs are log-concave (An 1997) if and only if both their shape parameters are greater than or equal to 1. The behaviour of the Kumaraswamy pdf also matches that of the Beta pdf at the boundaries of their support: $f(x) \sim x^{a-1}$ as $x \to 0^+$; $f(x) \sim (1-x)^{b-1}$ as $x \to 1^-$. Figure 1 displays the pdf of the Kumaraswamy distribution for all the possible combinations of *a* and *b*, each taking on the values 0.5; 0.75; 1; 1.5; 2.5.

The expression for the moment of order *r* of the Kumaraswamy distribution is

$$\mathbb{E}(X^r) = bB\left(1 + \frac{r}{a}, b\right)$$

for $r > -a$. In particular, the expected value and the variance are given by the following expressions

$$\mathbb{E}(X) = bB\left(1 + \frac{1}{a}, b\right) \quad (4)$$

$$\text{var}(X) = bB\left(1 + \frac{2}{a}, b\right) - \left(bB\left(1 + \frac{1}{a}, b\right)\right)^2, \quad (5)$$

which, among other things, are similar to the analogous expressions for the Weibull distribution, where the Gamma function replaces the Beta function in (4) and (5) and the scale and shape parameters replace the *a* and *b* parameters, respectively (Johnson et al. 1994, formula (21.13)). We note that for the Kumaraswamy distribution the expressions for the expected value and variance involve the special Gamma function, whereas for the Beta distributions they are rational polynomial functions.

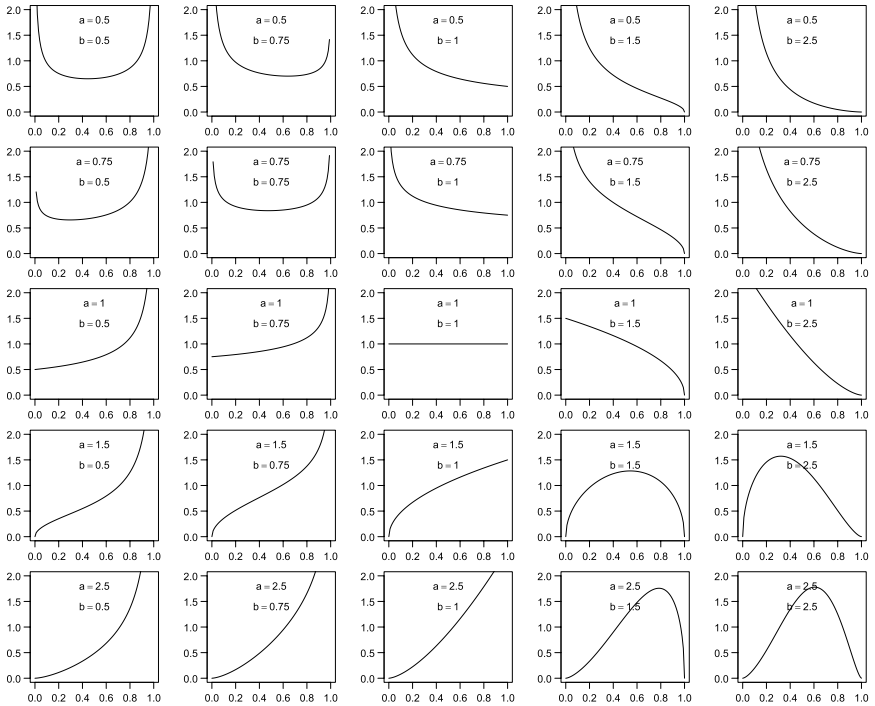

Fig. 1 Pdf of the Kumaraswamy distribution for different combinations of its parameters a and b; a and b take on the values $\{0.5, 0.75, 1, 1.5, 2.5\}$

Perhaps the least attractive feature of the Kumaraswamy distribution, at least at first thought, is that, unlike the Beta distribution, it has no symmetric special cases other than the uniform distribution ($a = b = 1$). When $a = b \neq 1$, the Kumaraswamy distribution always exhibits a non-zero degree of asymmetry, even though Fisher's coefficient of skewness may take the value zero for some pairs (a, b) with $a \neq b$. The limiting behavior of the Kumaraswamy distribution as either or both parameters tend to zero or infinity, along with other properties of the distribution in its most general form – i.e., with the addition of two boundary parameters that extend its support to a generic interval (c, d) – such as closure under linear transformations and exponentiation, are discussed in Mitnik (2013).

Several methods of estimation are available for the Kumaraswamy distribution, based on a sample $(x_1, \ldots, x_n)$, with an exhaustive account discussed in Dey et al. (2018). However, none of these methods provide a closed-form expression for the point estimators of a and b. Since the first two moments in (4) and (5) involve the Beta function, the method of moments can only return estimates of a and b numerically. As for maximum likelihood estimation, differentiating the log-likelihood with respect to b leads immediately to the relation

$$\hat{b} = -n/\sum_{i=1}^{n} \log(1 - x_i^{\hat{a}}),$$

whereas the equation to be satisfied by $\hat{a}$ arising from substituting for $\hat{b}$ in the other score equation is

$$S(a) = \frac{n}{a}\left\{1 + T_1(a) + \frac{T_2(a)}{T_3(a)}\right\} = 0,$$

where $T_1(a) = \frac{1}{n}\sum_{i=1}^{n} \frac{\log y_i}{1-y_i}$, $T_2(a) = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i \log y_i}{1-y_i}$, $T_3(a) = \frac{1}{n}\sum_{i=1}^{n} \log(1 - y_i)$, and $y_i = x_i^a$, $i = 1, \ldots, n$ (Jones 2009).

3 A discrete version of the Kumaraswamy distribution

In this section, we introduce and discuss a discrete version of the Kumaraswamy distribution that can be used for modeling ordinal data.

3.1 Definition and main properties

A discrete version of the Kumaraswamy distribution can be constructed by using methods like moment-matching via Gaussian quadrature (Golub and Welsch 1969), optimal quantization (commonly referred to as 'principal points' (Flury 1990)), or minimization of Cramér-von Mises (or Kolmogorov-Smirnov) statistical distance (Barbiero and Hitaj 2023). Beyond the specific challenges encountered by the different methods, in any case, the k support values of the discrete approximation would lie within (0,1), the support interval of the continuous Kumaraswamy distribution. Alternatively, if one is interested in defining a discrete rv supported on the first k positive integers, a discrete version of the Kumaraswamy distribution can be constructed, in a manner similar to what was proposed by Sciandra et al. (2024), by setting the probabilities equal to

$$\begin{aligned} p_i &= F_K(i/k; a, b) - F_K((i-1)/k, a, b) = 1 - (1 - (i/k)^a)^b - [1 - (1 - ((i-1)/k)^a)^b] \\ &= (1 - ((i-1)/k)^a)^b - (1 - (i/k)^a)^b, \end{aligned} \tag{6}$$

for $i = 1, \ldots, k$, where $k \geq 3$ to ensure model identifiability. In fact, letting $k = 2$, the two probabilities are $p_1 = 1 - (1 - (1/2)^a)^b$ and $p_2 = (1 - (1/2)^a)^b$ and it is easy to check that the probability distribution is the same if we consider the parameter combinations (a,b) and $\left(a^*, b^* = b\frac{\ln[1-(1/2)^a]}{\ln[1-(1/2)^{a^*}]}\right)$, with $a > 0, b > 0, a^* > 0$. Therefore, for $k = 2$ the model is not identifiable; just as, easily, it can be verified that for $k > 2$ the model is identifiable. One can immediately note that if $a = b = 1$,

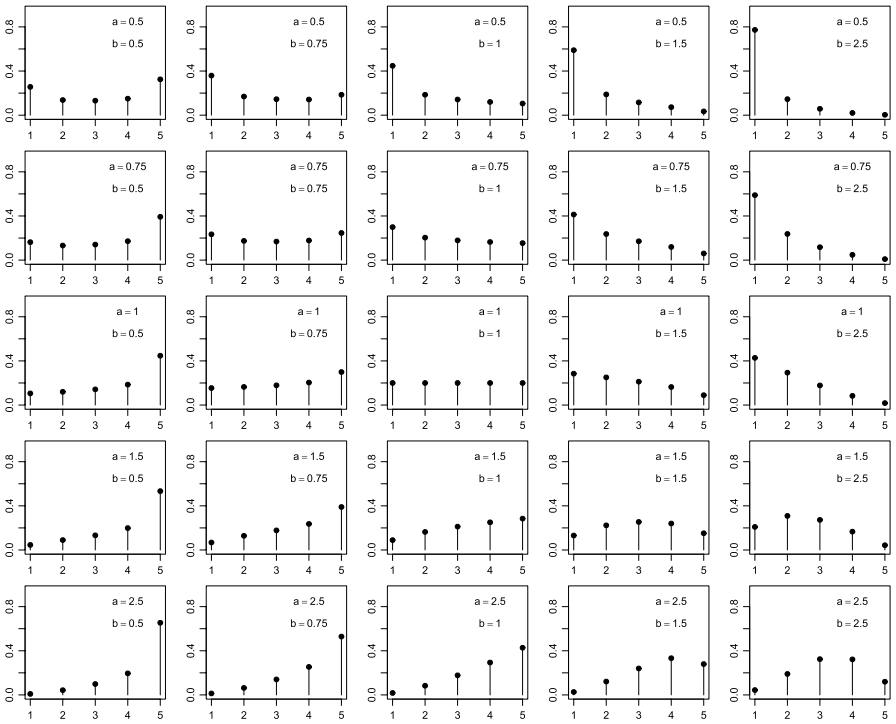

Fig. 2 Pmf of the discrete Kumaraswamy distribution for different combinations of its parameters a and b and $k = 5$; a and b take on the values $\{0.5, 0.75, 1, 1.5, 2.5\}$

then the corresponding distribution is discrete uniform ($p_i = 1/k$). Other combinations of a and b values lead to different shapes of the pmf (6): increasing, decreasing, "J"-shaped, inverted "J"-shaped, see Figure 2, for $k = 5$, and Figure 3, for $k = 7$. These shapes somewhat resemble those of the continuous counterpart distribution for the same pairs of parameter values. As with its continuous counterpart, the discrete Kumaraswamy distribution is not symmetric for any combination of the two parameters other than (1,1). This appears to be the only disadvantage of this distribution compared to the discrete Beta distribution proposed by Sciandra et al. (2024).

From (6), the cumulative probabilities of the discrete Kumaraswamy distribution are given by

$$F_i = \sum_{j=1}^{i} p_j = 1 - (1 - (i/k)^a)^b, \quad i = 1, \ldots, k. \tag{7}$$

The cdf defined by the probabilities above is readily invertible, yielding the quantile function:

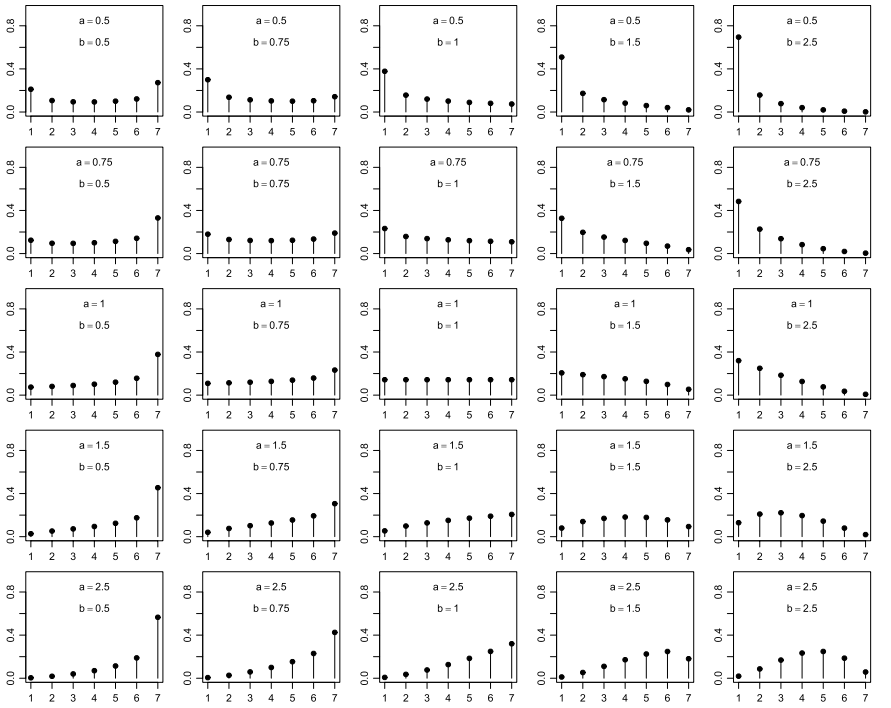

Fig. 3 Pmf of the discrete Kumaraswamy distribution for different combinations of its parameters a and b and $k = 7$; a and b take on the values $\{0.5, 0.75, 1, 1.5, 2.5\}$

$$x_u = \left\lceil k \left[1 - (1-u)^{1/b}\right]^{1/a} \right\rceil, \tag{8}$$

for $0 < u < 1$, where $\lceil \cdot \rceil$ is the ceiling function. Pseudo-random simulation is therefore straightforward: in order to draw a pseudo-random number from a discrete Kumaraswamy distribution with parameters a, b, and k, it is sufficient to generate a pseudo-random number u from the standard uniform distribution, and then calculate the corresponding u-quantile according to (8).

3.2 Moments

The expected value of the proposed discrete Kumaraswamy distribution is computed as

$$\mathbb{E}(X) = \sum_{i=1}^{k} i \cdot p_i = \sum_{i=1}^{k} i[(1 - ((i-1)/k)^a)^b - (1 - (i/k)^a)^b] = 1 + \sum_{i=1}^{k-1} [1 - (i/k)^a]^b. \tag{9}$$

The expression for the second non-central moment is

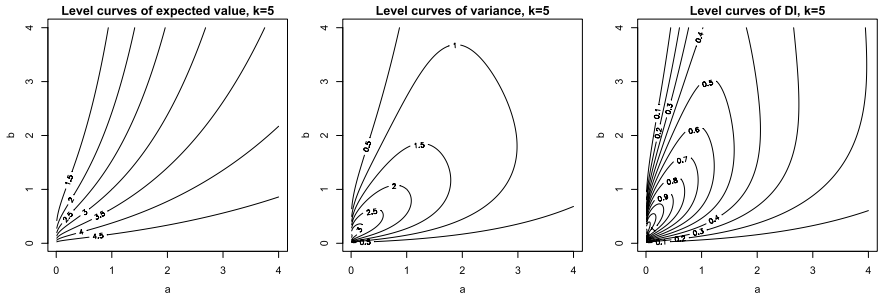

Fig. 4 Level curves of expected value, variance and dispersion index as a function of the parameters a and b of the discrete Kumaraswamy distribution with $k = 5$

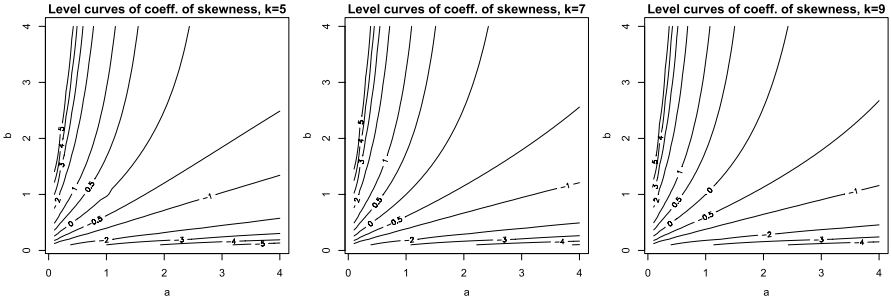

Fig. 5 Level curves of Fisher's coefficient of skewness as a function of the parameters a and b of the discrete Kumaraswamy distribution with $k = 5$, $k = 7$, and $k = 9$

$$\begin{aligned}\mathbb{E}(X^2) &= \sum_{i=1}^{k} i^2 \cdot p_i = \sum_{i=1}^{k} i^2\{[1-((i-1)/k)^a]^b - [1-(i/k)^a)]^b\} \\ &= 1 + \sum_{i=1}^{k-1}(2i+1)[1-(i/k)^a]^b,\end{aligned} \tag{10}$$

and then the variance can be computed as

$$\text{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = 1 + \sum_{i=1}^{k-1}(2i+1)[1-(i/k)^a]^b - \left(1 + \sum_{i=1}^{k-1}[1-(i/k)^a]^b\right)^2. \tag{11}$$

Figure 4 displays the level curves of the expected value, variance and dispersion index, defined as the ratio of the latter to the former, as functions of the parameters a and b, when the number of categories is $k = 5$. It is worth noting that the dispersion index is less than 1 for most parameter combinations. However, when both a and b are close to zero — that is, lying in the lower-left corner of the first quadrant near the origin — the dispersion index may exceed 1.

Figure 5 displays the level curves of Fisher's coefficient of skewness for the discrete Kumaraswamy distribution as a function of its parameters a and b, when the number of categories is $k = 5$, 7, and 9. We note that the dependence of the coefficient of skewness on the two shape parameters, given k, is quite complex, but it can be argued that when a exceeds b, it takes on negative values. On the contrary, setting $a < b$ is not a sufficient condition for obtaining a positive coefficient of skewness. It can also be observed that, when moving from $k = 5$ to $k = 7$ and $k = 9$, the coefficient of skewness does not vary significantly, provided a and b are fixed. It is worth noting that although the coefficient of skewness can take the value zero for all combinations of (a, b) lying on the level-zero curve in each panel of Fig. 5 and for the corresponding value of k, in these cases, the resulting distribution is not symmetric. For example, when $k = 5$, setting $a = 2$ and $b \approx 2.494$ yields a coefficient of skewness equal to zero, but the distribution remains asymmetric: considering the extreme categories, $p_1 \approx 0.0968$ and $p_5 \approx 0.0783 \neq p_1$.

The space of admissible expected values and variances for the discrete Kumaraswamy distribution is graphically displayed in Figure 6, when $k = 5$. This space is constructed by computing the corresponding expectation and variance for each combination of a and b, where a and b are taken from a dense grid of positive values between 0 and 100, using formulas (9) and (11). The feasible region resembles a sort of parabolic structure with a jagged base and is very similar to the analogous region for the discrete Beta distribution (see Sciandra et al. 2024, Figure 3). More importantly, the admissible region is very close to coinciding with the corresponding region that we would obtain by considering a discrete rv supported on $\{1, \ldots, k\}$, without assuming any specific parametric form for its distribution, i.e., allowing for any feasible probability distribution over this support.

3.3 Parameter estimation

We discuss possible methods for estimating the (unknown) parameters a and b of a discrete Kumaraswamy distribution. It is assumed that the k parameter of this distri-

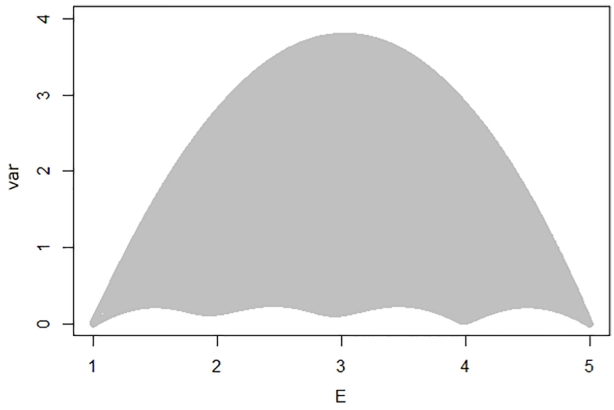

Fig. 6 Space of admissibleexpected values (E) and variances (V) for the discreteKumaraswamy distributionwith $k = 5$

bution is known and is ≥ 3, and that we have given a random sample $(x_1, x_2, \ldots, x_n)$ from this distribution.

For the model at hand, the corresponding log-likelihood function is

$$\ell(a,b) = \sum_{i=1}^{k} n_i \log\left[(1-((i-1)/k)^a)^b - (1-(i/k)^a)^b\right],$$

where n_i, $i = 1, \ldots, k$, is the sample frequency of the value i. In order to find the maximum likelihood estimates (MLEs) of a and b, one has to maximize the function above with respect to a and b (both defined on $\mathbb{R}^+$), or solve the system consisting of the two corresponding score equations, $\partial\ell(a,b)/\partial a = 0$ and $\partial\ell(a,b)/\partial b = 0$. The partial derivative with respect to a is $\frac{\partial\ell(a,b)}{\partial a} = \sum_{i=1}^{k} \frac{n_i}{p_i} \cdot \frac{\partial p_i}{\partial a}$, with

$$\frac{\partial p_i}{\partial a} = -b\left(\frac{i-1}{k}\right)^a \ln\left(\frac{i-1}{k}\right)\left[1-\left(\frac{i-1}{k}\right)^a\right]^{b-1} + b\left(\frac{i}{k}\right)^a \ln\left(\frac{i}{k}\right)\left[1-\left(\frac{i}{k}\right)^a\right]^{b-1};$$

while the partial derivative with respect to b is $\frac{\partial\ell(a,b)}{\partial b} = \sum_{i=1}^{k} \frac{n_i}{p_i} \cdot \frac{\partial p_i}{\partial b}$, with

$$\frac{\partial p_i}{\partial b} = \ln\left(\left[1-\left(\frac{i-1}{k}\right)^a\right]\right)\left[1-\left(\frac{i-1}{k}\right)^a\right]^b - \ln\left(\left[1-\left(\frac{i}{k}\right)^a\right]\right)\left[1-\left(\frac{i}{k}\right)^a\right]^b.$$

The system cannot be solved analytically; therefore, the values of the MLEs must be obtained numerically, potentially by directly maximizing the log-likelihood function.

Other estimation methods are available; although none of them generally leads to closed-form expressions for the estimates, they can be used mainly as possible sources of starting values for the MLE optimization routine. They are described in the Appendix.

3.4 Regression models with discrete Kumaraswamy response

In regression models where the response variable is assumed to follow the proposed discrete Kumaraswamy distribution, of the response variable shape parameters a and b can be modeled as functions of covariates to capture the influence of external factors. Letting $\boldsymbol{x}_i$ be the vector of covariates for the i-th observation, and $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$ the vectors of regression coefficients for a and b, respectively, a simple choice is to define

$$a_i = e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\lambda}}; b_i = e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\omega}},$$

for $i = 1, \ldots, n$, which ensures the positivity of the parameters a_i and b_i regardless of the values of the covariates. Proceeding as in Sciandra et al. (2024) and recalling the similarity between the (discrete) Beta and Kumaraswamy distributions, we apply an alternative parametrization of the discrete Kumaraswamy distribution which makes use of the transformations

$$\eta = \mathrm{logit}(a/(a+b)) = \ln a - \ln b, \quad \gamma = -\mathrm{logit}[1/(a+b+1)] = \ln(a+b), \tag{12}$$

to which correspond

$$a = e^{\gamma+\eta}/(1+e^{\eta}), \quad b = e^{\gamma}/(1+e^{\eta}). \tag{13}$$

This reparameterization enables unconstrained estimation, as both η and γ can take any value in $(-\infty, +\infty)$. As with the discrete Beta distribution, the parameter η is positively associated with the expected value of the underlying Kumaraswamy distribution (see Figure 7, left panel). When applied to rating data, η can be interpreted as a to like *indicator* (Sciandra et al. 2024), as higher values of η correspond to higher expected ratings. Similarly, the parameter γ is, at least when it takes on positive values, (approximately) inversely linked to the variance of the underlying Kumaraswamy distribution (see Figure 7, right panel), and can be considered as an *agreement indicator* (Sciandra et al. 2024), i.e., able to synthesize the degree of agreement between individuals in rating the item. Therefore, a covariate vector $\boldsymbol{x}_i$ can be introduced in the model as:

$$\eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\lambda}; \gamma_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\omega}. \tag{14}$$

This approach allows the model to incorporate covariate effects while preserving the flexible shape of the discrete Kumaraswamy distribution. The log-likelihood function, based on the sample $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, is constructed using the pmf of the discrete Kumaraswamy distribution (6), with the reparametrization (13) and including covariates through (14). The log-likelihood function is then maximized to estimate the model parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$.

Although for the continuous Kumaraswamy distribution an alternative parametrization based on the median is discussed, which turns out to be useful if one needs to include covariates in the model (Mitnik and Baek 2013), for the discrete model proposed in this work, that parametrization is not applicable. Examples of imple-

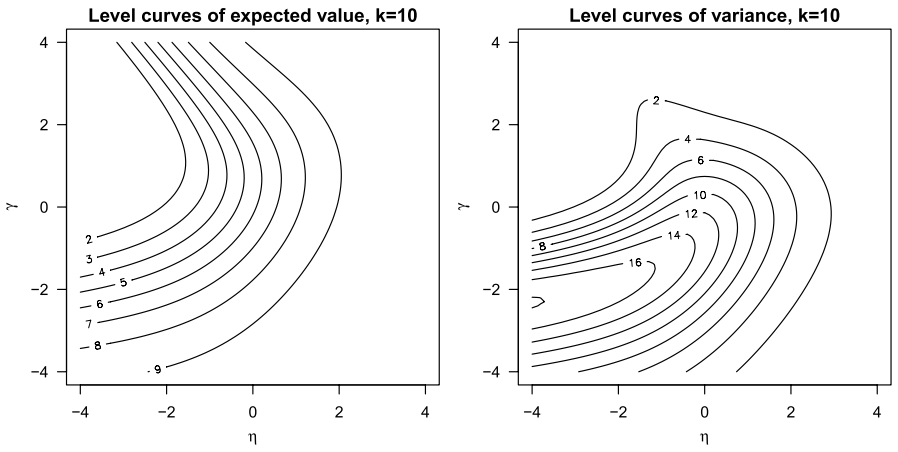

Fig. 7 Level curves of the expected value and variance of a discrete Kumaraswamy distribution with $k = 10$, shown as functions of the transformed parameters η and γ

mentation of the regression model with discrete Kumaraswamy response variable on artificial and real data will be reported in the next sections.

3.5 A Monte Carlo simulation study with covariates

We considered a regression model where the dependent variable Y follows a discrete Kumaraswamy distribution with parameter $k = 5$, and whose parameters a and b depend on two covariates, X_1 and X_2, through the following equations, which involve the parameter transforms η and γ:

$$\begin{cases} \eta_i &= \lambda_0 + \lambda_1 \cdot x_{1i} + \lambda_2 \cdot x_{2i} \\ \gamma_i &= \omega_0 + \omega_1 \cdot x_{1i} + \omega_2 \cdot x_{2i}, \end{cases} , \tag{15}$$

where $X_1 \sim \mathcal{N}(0, 1)$ and X_2, which is independent of X_1, follows a uniform distribution on $(-1, 1)$, $X_2 \sim \mathcal{U}(-1, 1)$; the regression coefficients are $\boldsymbol{\lambda} = (0, 0.5, 0.75)^\top$ and $\boldsymbol{\omega} = (\ln 2, -0.25, 0.5)^\top$. Please note that both X_1 and X_2 are symmetrically distributed around zero, and that $\lambda_0 = 0$, ensuring that η itself will be symmetrically distributed around zero. Since $\omega_0 = \ln 2$, on average, the γ parameter will be equal to $\ln 2$, whereas the sum $a + b$, equal to e^γ, will be only approximately concentrated around 2. For $i = 1, \ldots, n$, we simulate a bivariate observation for (X_1, X_2) and compute the values of η_i and γ_i, on it, according to (15); then, after computing the corresponding values (a_i, b_i), using (13), we draw a pseudo-random value y_i from a discrete Kumaraswamy distribution with that choice of parameters (and $k = 5$). On the resulting sample $(y_1, y_2, \ldots, y_n)$, we compute the MLEs of the regression coefficients. We repeat this procedure $S = 1,000$ times. For the S resulting estimates, we provide their empirical distributions through the boxplots in Figure 8, where a horizontal line is plotted at the height corresponding to the true value of the corresponding parameter. We can easily observe that all the empirical distributions of the estimators are concentrated around the true value of the corresponding parameter. We can also observe that the regression model, with the assigned parameter values,

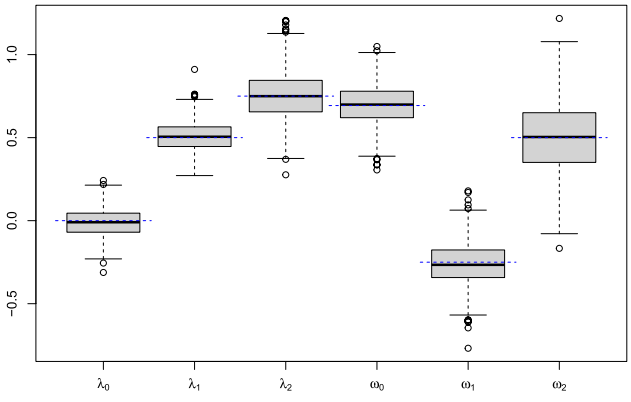

Fig. 8 Empirical distribution of the MLEs of the regression coefficients for the regression model (15) of Section 3.5 with sample size $n = 250$ and $S = 1,000$ Monte Carlo runs

Table 1 Results on the regression model of Section 3.5

Regr.par.	λ_0	λ_1	λ_2	ω_0	ω_1	ω_2
true	0	.5	.75	$\ln 2$	$-.25$	.5
mean	$-.012$	.506	.751	.700	$-.262$	.502
sd	.083	.084	.144	.122	.129	.210

Par.est.	Min.	$q_{0.25}$	Median	Mean	$q_{0.75}$	Max	sd
a	.194	.618	.969	1.075	1.470	2.398	.524
b	.026	.685	.970	1.076	1.346	6.770	.551
	a>1,b>1	a<1,b<1	a>1,b<1	a<1, b>1	Pearson	Spearman	
joint freq.	.257	.300	.225	.218	.214	.162	

The table reports the true values, along with the Monte Carlo mean and standard deviation (sd), of the regression coefficients related to the transformed parameters η and γ, see (15). Descriptive statistics of the corresponding parameters a and b of the discrete Kumaraswamy distribution are also reported, along with some bivariate measures

is capable of producing a wide range of combinations (a_i, b_i). To briefly illustrate this aspect, Table 1 shows, across all S simulated samples, some univariate summary indexes (among which, mean and standard deviation) for the values of a and b, along with the proportion of cases where $a \lessgtr 1$ and $b \lessgtr 1$ and the values of Pearson and Spearman correlation coefficients.

4 Data analysis

In this section, we analyze data sets containing rating and ranking data related to individual attitudes and perceptions about environmental issues, and we fit these data using the proposed discrete Kumaraswamy distribution.

4.1 Eurobarometer

The Eurobarometer data set contains the results of a survey on the attitudes of Europeans towards climate change that was carried out in August and September 2009. This survey concerned the attitudes of citizens in the European Union (EU) on matters regarding the environment and specifically the opinion of Europeans on several climate change-related topics. European citizens participating in the survey were asked to respond to various questions about their perception of climate change using an ordinal scale. Among these, we focused on the following question: 'How serious a problem do you think climate change is at this moment? Please use a scale from 1 to 10, with "1" meaning it is "not at all a serious problem" and "10" meaning it is "an extremely serious problem".' Figure 9 displays in the left panel the sample distribution of the response for all the $n = 25,862$ respondents. The discrete Kumaraswamy model with $k = 10$ is here fitted to these data. The estimates of the two parameters a and b are presented, along with their standard errors, in Table 2. We have that $\hat{a} = 2.282 > \hat{b} = 1.192$, which is consistent with the fact that the respondents tend to assign high importance, i.e., high ratings, to the climate change issue. The discrete Beta distribution is also fitted to the same data and shows a slightly worse fit as testified by the larger value of the AIC, see the second-to-last line of Table 2. Notably,

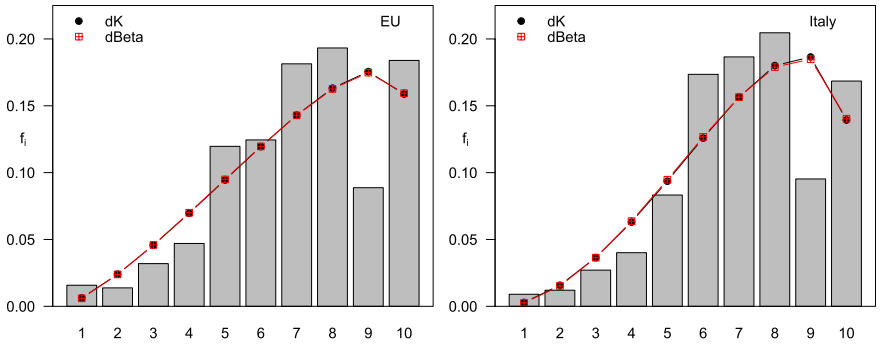

Fig. 9 Empirical distribution of the Likert variable from the Eurobarometer survey data. The heights of the bars represent the sample proportions for the $k = 10$ ordered categories. Circles and crosshatched squares are overlaid to indicate the fitted probabilities of the discrete Kumaraswamy (dK) and discrete Beta (dBeta) distributions, respectively. The left panel shows the distribution for all responses, while the right panel displays the distribution for Italian respondents only

Table 2 Parameter estimates, standard errors (SE), maximum log-likelihood, and Akaike Information Criterion (AIC) of the discrete Kumaraswamy (dK) and discrete Beta (dBeta) distributions fitted to the Eurobarometer data set

Model	$\hat{a}$	$SE(\hat{a})$	$\hat{b}$	$SE(\hat{b})$	$\max \ell(a, b)$	AIC
dK (UE)	2.282	0.020	1.192	0.012	−54265.74	108535.5
dK (Italy)	2.683	0.112	1.406	0.075	−2042.87	4089.74
Model	$\hat{\alpha}$	$SE(\hat{\alpha})$	$\hat{\beta}$	$SE(\hat{\beta})$	$\max \ell(\alpha, \beta)$	AIC
dBeta (UE)	2.335	0.023	1.180	0.011	−54266.78	108537.6
dBeta (Italy)	2.836	0.142	1.377	0.067	−2042.72	4089.44

the estimates of the parameters α and β for the discrete Beta distribution are closely matched to those of a and b for the discrete Kumaraswamy distribution. However, as we can see from the bar plot of Figure 9, the distribution of the ordinal variable is somewhat peculiar: the sample frequencies are strictly increasing from the 1-st to the 8-th category, which is the most frequent; then there is a sharp drop when moving from the 8-th to the 9-th category, followed by a positive increase when moving from the 9-th to the last. Obviously, neither the discrete Kumaraswamy nor the discrete Beta are able to perfectly fit this behavior, that we can call a "ceiling effect" (Chyung et al. 2020) or, better, a "category avoidance effect", specifically at the 9-th category, which might be psychologically or culturally perceived as ambiguous or "less meaningful". If we restrict our attention to Italian respondents, we would recognize that the sample distribution of the ratings is not very different from the aggregate distribution over all the EU countries (see Figure 9, right panel). If we fit the discrete Kumaraswamy distribution to the Italian data set, we obtain the MLEs reported in the second row of Table 2. The values of $\hat{a}$ and $\hat{b}$ are both greater than for the complete data set, thus inducing a greater "level of agreement" $\hat{\gamma} = \ln(\hat{a} + \hat{b})$ among Italian respondents than among all the European respondents.

4.2 ISSP200

The `issp2000` data set, available in the *prefmod* R package (Hatzinger and Maier 2023), is part of the International Social Survey Programme (ISSP) data, focusing on social and political attitudes. The 2000 edition specifically centers on environmental issues and people's perspectives on them. This data set allows researchers to analyze patterns and preferences in environmental attitudes across different countries. It includes variables related to personal demographics, opinions on environmental policy, and concerns about sustainability, making it valuable for studies in sociology, political science, and environmental policy analysis.

The `issp2000` data set consists of 1595 observations on 11 variables. Five of them are social and demographic variables; the other six variables are items to be answered on a 5-point rating scale (Likert type). The social and demographical variables are `SEX` (gender, 1="Male", 2="Female"), `URB` (location of residence: 1="urban area", 2="suburbs of large cities, small town, county seat", 3="rural area"), `AGE` (1="less than 40 years", 2="41 to 59 years", 3="60 or more years"), `CNTRY` (country of residence, 1="United Kingdom", 2="Austria"), and `EDU` (education, 1="below A level", 2="A level or higher"). Respondents were asked about their perception of environmental dangers. The available response categories were: (1) extremely dangerous, (2) very dangerous, (3) somewhat dangerous, (4) not very dangerous, and (5) not dangerous at all for the environment. The Likert-scale variables are named `CAR`, concerning air pollution caused by cars; `IND`, air pollution caused by industry; `FARM`, pesticides and chemicals used in farming; `WATER`, pollution of country's rivers, lakes and streams; `TEMP`, a rise in the world's temperature; `GENE`, modifying the genes of certain crops. The data set has been used for fitting paired comparison models for preferences (Hatzinger and Dittrich 2012).

Here we considered the empirical distribution of the variable `CAR` for all the respondents and fitted the discrete Kumaraswamy and the discrete Beta distribution to it by using the maximum likelihood method. The results are reported in Table 3. We observe that the discrete Kumaraswamy distribution has a better fit than the discrete Beta (for the former, the value of the maximized log-likelihood function is larger and therefore that of the AIC is smaller). Table 3 also presents the values of the dissimilarity index, which is a normalized fitting measure that compares the observed relative frequencies f_i with the expected probabilities $p_i(\boldsymbol{\theta})$, and is defined as (Leti 1983)

$$\text{diss} = \frac{1}{2}\sum_{i=1}^{k} |f_i - p_i(\boldsymbol{\theta})|,$$

Table 3 Parameter estimates, standard errors (SE), maximum log-likelihood, Akaike Information Criterion (AIC), and dissimilarity index of the discrete Kumaraswamy (dK) and discrete Beta (dBeta) distributions fitted to the variable `CAR` of the *prefmod* data set

Model	$\hat{a}$ (or $\hat{\alpha}$)	SE	$\hat{b}$ (or $\hat{\beta}$)	SE	$\max \ell$	AIC	Diss
dK	2.319	0.066	7.304	0.452	−1955.693	3915.385	0.128
dBeta	3.007	0.129	5.332	0.230	−1972.447	3948.893	0.144

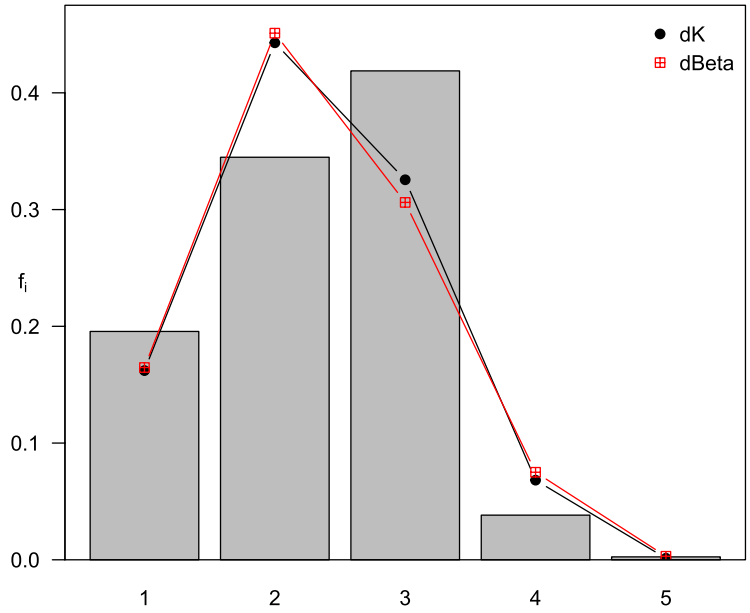

Fig. 10 Empirical distribution of the variable `CAR` of the data set `issp2000` within the package *prefmod*. The heights of the bars represent the sample proportions of the $k = 5$ ordered categories, with circles and crosshatched squares overlaid to indicate the fitted probabilities of the discrete Kumaraswamy (dK) and discrete Beta (dBeta) distributions, respectively

where $\boldsymbol{\theta} = (a, b)$ for the discrete Kumaraswamy distribution and $\boldsymbol{\theta} = (\alpha, \beta)$ for the discrete Beta distribution. For the former, the value of this measure is smaller than for the latter, confirming that the discrete Kumaraswamy distribution provides a better overall fit. The better fit of the proposed model is also more evident if we look at Figure 10, which displays the sample frequencies of the five ordinal categories of the variable `CAR` along with the fitted probabilities of the two competing models. From Table 3, we note that for the discrete Kumaraswamy model we have $1 < a < b$ and for the discrete Beta model $1 < \alpha < \beta$, which is consistent with the shape of the empirical distribution of `CAR`, which is unimodal and right-skewed, with most of the probability mass concentrated on the lower categories. We note that neither the discrete Kumaraswamy nor the discrete Beta is able to capture the mode of the empirical distribution, corresponding to the 3-rd category. The CUB distribution, fitted to the same data, yields a worse fit than both analyzed models (AIC=3995.776).

If we include in the model the covariates `SEX`, `URB`, `AGE`, `CNTRY`, and `EDU`, following the lines of Section 3.4, we obtain a model that can be fitted and whose summary is displayed in Table 4. It can be observed that the intercept estimates for both η and γ are significantly different from zero. Moreover, we observe that, for the η parameter, the coefficients corresponding to "female" and "Austria" are both negative and significant at the 5% level. This means that moving from male (baseline category for `SEX`) to female, or from UK (baseline category for `CNTRY`) to Austria, the η coefficient decreases, which translates into a shift towards lower categories for

Table 4 Estimates, standard errors (SE), and p-values of the model coefficients for the liking indicator η and the accuracy indicator γ using the `issp2000` data set, the variable `CAR`, and 5 covariates: gender (male, female), location of residence (urban area, suburbs, rural area), age (40 years or less, 41-59 years, 60 or more years), country (United Kingdom, Austria), and education (below A-level, A-level or higher)

Parameter	Estimate	SE	p-value
Eta.intercept	−0.966	0.140	0.000***
Female	−0.168	0.079	0.033*
Suburbs	−0.058	0.124	0.639
Rural area	−0.066	0.110	0.546
41-59 yrs	0.124	0.096	0.195
60+ yrs	0.103	0.098	0.294
Austria	−0.231	0.108	0.033*
A level	−0.134	0.087	0.125
Gamma.intercept	2.035	0.185	0.000***
Female	0.177	0.108	0.101
Suburbs	−0.020	0.166	0.906
Rural area	0.163	0.147	0.266
41-59 yrs	−0.158	0.131	0.227
60+ yrs	−0.194	0.136	0.152
Austria	0.434	0.142	0.002**
A level	0.097	0.120	0.421

*p = 0.01–0.05; **p = 0.001–0.01; ***$p < 0.001$

Table 5 Estimates, standard errors (SE), and p-values of the model coefficients for the liking indicator η and the accuracy indicator γ, using the `issp2000` data set, the variable `CAR`, and 2 covariates: gender and country

Parameter	Estimate	SE	p-value
Eta.intercept	−1.015	0.064	0.000***
Female	−0.133	0.077	0.085
Austriaara>	−0.160	0.078	0.041*
Gamma.intercept	1.999	0.093	0.000***
Female	0.134	0.106	0.207
Austria	0.432	0.108	0.000***

*p = 0.01–0.05; ** = 0.001–0.01; ***$p < 0.001$

the rating variable `CAR`, corresponding to a greater perception of danger towards pollution from cars. As for the γ parameter, the only regression coefficient significantly different from zero is that related to "Austria", which is positive. This means that moving from UK to Austria, the γ coefficient increases, which corresponds to an increase in the agreement among respondents.

Based on these results, instead of including all the available covariates, we can try fitted a model with only two covariates, `SEX` and `CNTRY`. The results are displayed in Table 5. Now, five out of six regression coefficients are significant at the 0.1 level. The only non-significant coefficient is related to γ and corresponds to the category "female" of the variable `SEX` (p-value ≈ 0.207). The signs of the regression coefficients are the same as those of the previous model that included all the covariates. Notably, the AIC for this latter model (3883.613) is veeeeery slightly smaller than the previous one (3883.66), indicating a better relative fit.

The fitting of these regression models was performed in the `R` environment (R Core Team 2024) using the *bbmle* package (Bolker 2023) and the function *mle2*. We observed that the MLE routine implemented via the *mle2* function can be computationally demanding, likely due to conservative default settings of the optimization parameters.

4.3 Utah air quality risk and behavioral action survey

A survey on air quality risk was conducted among Utah residents between November 2018 and January 2020. The state of Utah, USA, at times over the last 20 years has suffered from some of the worst air quality in the nation (Flowerday et al. 2023). The survey includes more than 60 questions covering demographics, daily habits, opinions about air pollution, and attitudes toward government interventions to mitigate it (Benney et al. 2020). The sample is made up of 1160 subjects. We analysed the answers to the question: Please rank the following by which causes the most air pollution in Utah (questions labelled from Q18_1 to Q18_8 in the original data set):

(i) wood burning,
(ii) automobile exhaust,
(iii) buildings (e.g., businesses and homes),
(iv) major industries (e.g., mining, airport, energy),
(v) home chemicals (e.g., aerosols, paints),
(vi) environment (e.g., wind blowing dust),
(vii) agriculture (e.g., farm equipment, animal byproducts, etc.),
(viii) government (e.g., offices, agencies, industries).

We focused on the rankings assigned to the fourth item, "Major industries (e.g., mining, airport, energy)". Table 6 reports the estimates (along with the standard errors) of the parameters a and b of the discrete Kumaraswamy and α and β of the discrete Beta distributions fitted to the `IND` variable of the `Utah` data set. We note that for the two distributions we have $\hat{b} > \hat{a} \approx 1$ and $\hat{\beta} > \hat{\alpha} \approx 1$, respectively, which was expected since the empirical distribution shows a unique mode at the lowest category, 1, and has monotonically decreasing frequencies. The fits of the two models in terms of AIC are very similar, although the proposed model is slightly worse. When comparing the dissimilarity indexes, our model proves to be slightly better. Figure 11 displays the observed ranks and fitted values derived from both the discrete Kumaraswamy model and the discrete Beta model without covariates. One can easily observe how the fitted frequencies are pairwise very close to each other (and to the corresponding empirical frequencies).

Next, we investigated if the rankings depend on some explanatory variables, specifically income and political ideology. Both are categorical variables, the former with eight categories ($< 25,000$; 25,000–34,999; 35,000–49,999; 50,000–74,999; 75,000–99,999; 100,000–149,999; $\geq 150,000$, and Not Declared) and the latter with six categories (Democrat, Independent, Libertarian, Republican, Other, and Not Declared). Table 7 summarizes the output for the parameters η and γ, introduced in

Table 6 Parameter estimates, standard errors (SE), maximum log-likelihood, Akaike Information Criterion (AIC), and dissimilarity index of the discrete Kumaraswamy (dK) and discrete Beta (dBeta) distributions fitted to the `Utah` data set, variable `IND`

Model	$\hat{a}$ (or $\hat{\alpha}$)	SE	$\hat{b}$ (or $\hat{\beta}$)	SE	$\max \ell$	AIC	Dissim
dK	1.053	0.041	3.251	0.189	−1895.335	3794.67	0.0319
dBeta	1.071	0.055	3.228	0.168	−1895.299	3794.60	0.0326

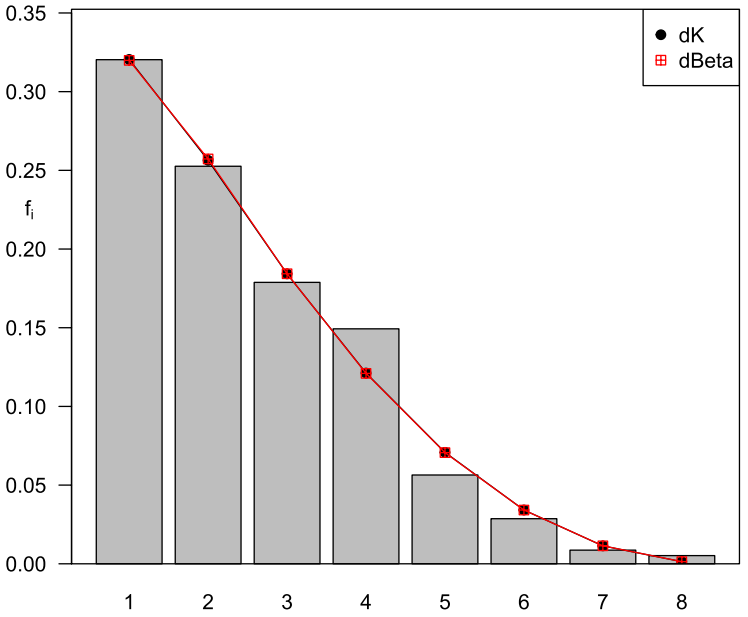

Fig. 11 Empirical distribution of the variable `IND` of the data set `Utah`. The heights of the bars represent the sample proportions of the $k = 8$ ranks, with circles and crosshatched squares overlaid to indicate the fitted probabilities of the discrete Kumaraswamy (dK) and discrete Beta (dBeta) distributions, respectively

Section 3.4, for the selected model. When examining the question of responsibility for air pollution, the options ranked in the top positions are generally viewed as the most accountable, while those ranked lower are perceived as the least accountable. In this context, the liking indicator η can be interpreted in a manner similar to a rating framework, as it highlights aversion patterns toward the former options. The accuracy indicator γ still quantifies the extent of consensus in these perceptions among various demographic groups.

The intercept value of -1.269 related to η applies to individuals who identify as Democrats with an annual income below $25,000, and it is statistically significant. The estimated ratio $\hat{a}/(\hat{a}+\hat{b}) = 0.219$ is significantly lower than 0.5, which indicates that this group tends to place more responsibility for air pollution on major industries compared to other causes. Concerning the impact of other political orientations, individuals who do not identify with any political party show less opposition to major industries than Democrats. This difference is evident in the ratio $\hat{a}/(\hat{a}+\hat{b})$, which increases by 0.230 on the logit scale for individuals with no party preference. This implies that these individuals tend to rank major industries lower compared to Democrats, indicating that they perceive them as less responsible for air pollution. Similarly, a pattern emerges among individuals declaring an annual income between 50,0000 and 75,000, where the ratio $\hat{a}/(\hat{a}+\hat{b})$ decreases by 0.391 on a logit scale.

Table 7 Estimates, standard errors (SE), and p-values of the model coefficients for the η and γ parameters using Utah survey data, with "Major Industries" as response variable and income and political orientation as covariates

Parameter	Estimate	SE	p-value
η			
Intercept	−1.269	0.092	0.000***
25,000 to 34,999	0.176	0.110	0.111
35,000 to 49,999	−0.009	0.123	0.940
50,000 to 74,999	−0.391	0.117	0.001***
75,000 to 99,999	0.159	0.111	0.150
100,000 to 149,999	0.172	0.117	0.141
Greater than 150,000	0.081	0.169	0.633
Prefer not to answer	−0.020	0.177	0.909
Independent	0.066	0.102	0.514
Libertarian	−0.125	0.226	0.580
Republican	0.123	0.087	0.157
Otherara>	0.337	0.183	0.065
No preference	0.230	0.108	0.034*
γ			
Intercept	1.444	0.160	0.000***
25,000 to 34,999	−0.046	0.188	0.806
35,000 to 49,999	0.296	0.203	0.146
50,000 to 74,999	0.720	0.190	0.000***
75,000 to 99,999	−0.007	0.180	0.971
100,000 to 149,999	−0.071	0.193	0.714
Greater than 150,000	0.371	0.258	0.151
Prefer not to answer	0.329	0.280	0.240
Independent	−0.127	0.164	0.441
Libertarian	−0.19	0.393	0.615
Republican	−0.020	0.139	0.883
Other	−0.340	0.310	0.273
No preference	−0.417	0.179	0.020*

*$p = 0.01$–0.05; **$p = 0.001$–0.01; ***$p < 0.001$

Consequently, those with incomes in (50,0000; 75,000) per year are more likely to perceive major industries as the primary culprits responsible for air pollution.

As regards the accuracy indicator, it measures the level of agreement among respondents in ranking major industries as a cause of air pollution. The accuracy coefficient for a person who identifies as a Democrat with an annual income of less than 25,000, is 1.444, which indicates a high level of agreement within this demographic group regarding major industries' responsibility for air pollution in Utah. Moreover, the agreement between respondents significantly increases moving from the "baseline" group earning less than 25,000 to the group of those earning between 50,000 and 75,000. In this transition, the agreement indicator experiences a notable increase on the log scale of 0.720 for $\hat{a} + \hat{b}$. The agreement between respondents decreases statistical jargon when considering individuals with no political preference (-0.417 on a log scale for $\hat{a} + \hat{b}$) with respect to those perceiving themselves as Democrats.

The AIC of this model is 3780.976, which is smaller than the value for the model without covariates, thus indicating a better relative fit, and is smaller than the AIC for the discrete Beta model fitted to the same data set in Sciandra et al. (2024). In that

study, however, the category Other for the variable Political Orientation was likely merged with the category No preference.

5 Conclusion

We introduced a novel discrete probability distribution, constructed as a counterpart to a two-parameter continuous probability distribution supported on the unit interval and designed specifically for modeling ordinal data. This model offers several attractive features: its probability mass function, cumulative distribution function, and quantile function are all expressed in closed-form. This enables straightforward pseudo-random simulation, setting it apart from the discrete Beta distribution. Parameter estimation, although it does not generally yield analytic expressions for the point estimators, is still simple to implement using standard procedures. The discrete Kumaraswamy distribution can be fitted in settings in which the distribution depends upon predictor variables, where this dependence takes a linear form, by appropriately reparametrizing the probability mass function in terms of two new parameters: one that is positively related to the expectation, and the other inversely related to the variance. Several examples, where the proposed distribution has been fitted to real data on individuals' attitudes towards environmental issues, empirically prove its usefulness, establishing it as a valid competitor to existing models, such as the discrete Beta, CUB, and Beta-Binomial distributions. Parameter estimation via the maximum likelihood method in presence of covariates has proved to be straightforward by using well-established routines within the `R` statistical environment. Although the proposed model performs well in the analyzed cases, further research is required to assess its general applicability and possible extensions. For instance, one could explore extensions of this discrete distribution that account for potential overdispersion in ordinal data, analogous to how the CUBE model extends the CUB model by introducing an additional parameter (Iannario 2014).

A.1 Appendix: Alternative estimation methods for the discrete Kumaraswamy distribution

Method of proportions

Specializing the expression for the pmf (6) for $i = 1$ and $i = k$, we obtain

$$\begin{cases} p_1 &= 1 - (1 - (1/k)^a)^b \\ p_k &= [1 - ((k-1)/k)^a]^b. \end{cases}$$

From the former we derive

$$a = \frac{\ln[1-(1-p_1)^{1/b}]}{-\ln k}; \tag{A1}$$

from the latter

$$a = \frac{\ln(1-p_k^{1/b})}{\ln(k-1)-\ln k}. \tag{A2}$$

By equating the two expressions for a above, we obtain

$$1-(1-p_1)^{1/b} = (1-p_k^{1/b})^{-\ln k/(\ln(k-1)-\ln k)}, \tag{A3}$$

that can be solved for b, numerically, after substituting to p_1 and p_k the corresponding sample frequencies of 1's and k's, say, f_1 and f_k; then one can obtain a through either (15) or (16). The two values found, say $\hat{a}_P$ and $\hat{b}_P$, can be regarded as the estimates given by the method of proportions. Such estimates can be also used as starting values for the optimization routine involved by the maximum likelihood method. We highlight that this method is applicable to the discrete Kumaraswamy distribution, as its pmf is available in closed-form. This is not the case for the discrete Beta distribution in Sciandra et al. (2024), where the pmf generally needs to be computed numerically based on the cdf of the underlying continuous Beta distribution. Obviously, for the method to work, the sample must include at least one 1 and at least one k. Equally evident is that this method does not efficiently utilize all the available sample information, as it only considers the frequencies and probabilities of the two extreme categories.

A.2 Method of moments

The first two non-central moments of the discrete Kumaraswamy distribution can be written as a finite sum of simple functions, see (9) and (10). If applied to this distribution, the method of moments, which equates the two sample moments to the two theoretical moments and consequently derives the values of a and b, however, cannot return analytic expressions for the estimates of its parameters. However, proceeding as in Khan et al. (1989, Section 4), one can think of minimizing the following quadratic loss function, which is the sum of the squared differences between the two sample moments and the corresponding theoretical moments:

$$\mathcal{L}(a,b;x_1,\ldots,x_n) = (\bar{x}-\mathbb{E}(X;a,b))^2 + (\hat{\mu}_2-\mathbb{E}(X;a,b))^2,$$

where $\bar{x} = \sum_{i=1}^n x_i/n$ and $\hat{\mu}_2 = \sum_{i=1}^n x_i^2/n$. The optimal solution, $(\hat{a}_M, \hat{b}_M)$, corresponding to the global minimum point of $\mathcal{L}(a,b)$ (where the function is expected to be zero, if the two sample moments belong to the feasible region discussed in Section 3.2), can be regarded as the method of moments' estimate. This formulation of the method of moments also overcomes situations when the "standard" solution (obtained by directly equating the sample moments to the theoretical moments) does not exist (i.e., the pair $(\bar{x}, \hat{\mu}_2)$ does not fall in the feasible region).

A.3 Method of minimum chi-square

A method closely related to the maximum likelihood method, and conceptually similar to the method of proportions in that it works with discrete or discretized variables, is the minimum Chi-square method (see, e.g., Barbiero 2017, for an application to the discrete Weibull distribution). For all categories of a discrete rv, the sample relative frequencies f_i can be compared with the corresponding probabilities $p_i = P(X = i; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a p-variate parameter vector indexing the distribution. Alternatively, the sample frequencies n_i can be compared with the corresponding expected frequencies np_i, thus defining the following discrepancy function:

$$\chi^2 = \sum_{i=1}^{k} \frac{(np_i - n_i)^2}{np_i} = n \sum_{i=1}^{k} \frac{(p_i - f_i)^2}{p_i}. \tag{A4}$$

The global minimum point $\boldsymbol{\theta}_{MCS}$ of χ^2 represents the minimum Chi-square estimate of $\boldsymbol{\theta}$. Minimizing (18) with respect to the component θ_j, $j = 1 \ldots, p$, we obtain the following p equations:

$$\begin{aligned} \frac{\partial \chi^2}{\partial \theta_j} &= n \sum_{i=1}^{k} \frac{2(p_i - f_i)\frac{\partial p_i}{\partial \theta_j} p_i - (p_i - f_i)^2 \frac{\partial p_i}{\partial \theta_j}}{p_i^2} \\ &= n \sum_{i=1}^{k} \frac{\frac{\partial p_i}{\partial \theta_j}[2(p_i - f_i)p_i - (p_i - f_i)^2]}{p_i^2} = n \sum_{i=1}^{k} \frac{\partial p_i}{\partial \theta_j}(1 - f_i^2/p_i^2), \end{aligned} \tag{A5}$$

which, set equal to zero, form a system of p equations in p unknowns that typically must be solved numerically, as is the case for the discrete Kumaraswamy distribution, where $\boldsymbol{\theta} = (a, b)$. It is worth recalling that by properly approximating the general term of the sum in (18), the equations obtained by setting the partial derivatives of χ^2 equal to zero can be made to coincide with the score equations, whence the method of minimum Chi-square and maximum likelihood become equivalent (Landenna et al. 1996).

Examples of estimation

We consider two samples of size $n = 100$, with their empirical distributions shown in Table 8. Both distributions are unimodal and symmetric around the central category (3 for the first sample and 4 for the second sample). This symmetry presents a potential issue for the discrete Kumaraswamy distribution, which is not designed to model symmetric and unimodal distributions effectively, as highlighted in Section 3.

For the two samples, the parameters a and b are estimated using the four methods described in this appendix and in Section 3.3: maximum likelihood, method of moments, method of proportions, and minimum Chi-square. The results are presented

Table 8 Two sample distributions

1st sample, $k = 5$							
	i	1	2	3	4	5	
	f_i	0.1	0.2	0.4	0.2	0.1	
2nd sample, $k = 7$							
i	1	2	3	4	5	6	7
f_i	0.05	0.1	0.2	0.3	0.2	0.1	0.05

Table 9 Parameter estimation for the samples whose distributions are displayed in Table 8

	First sample, $k = 5$					Second sample, $k = 7$				
Par	ML	MM	MP	MCS	AIC	ML	MM	MP	MCS	AIC
a	2.160 (.267)	2.227	1.892	2.213	298.03	2.137(.246)	2.198	1.941	2.123	361.27
b	2.697 (.520)	2.840	2.160	2.783		2.645(.466)	2.778	2.215	2.598	
α	2.483 (.406)	-	-	-	298.14	2.448(.369)	-	-	-	361.32
β	2.483 (.406)	-	-	-		2.448(.369)	-	-	-	

in Table 9. While the maximum likelihood method and the methods of moments and minimum Chi-square provide similar estimates for a and b across both samples, the estimates from the method of proportions show notable deviations. We then fit the discrete Beta distribution to the same samples using the maximum likelihood method: the MLEs of α and β, along with the corresponding standard errors between brackets, are also displayed in the same table. The values of the Akaike Information Criterion (AIC) for both probability distributions across the two samples are presented as well. They indicate a slightly better fit for the discrete Kumaraswamy distribution, which shows a slightly smaller AIC for both samples. These results suggest that, although the proposed discrete model does not perfectly fit the two symmetric distributions, its performance is comparable to that of the discrete Beta model and is not outperformed by it in terms of AIC.

A.4 Monte Carlo assessment of estimation methods

In this section, a Monte Carlo simulation study is described, which was conducted to compare the performance of the different estimation procedures discussed in this appendix and in Section 3.3 for the proposed discrete Kumaraswamy model. In this study, we generated $S = 10,000$ random samples $(x_1, x_2, \ldots, x_n)$ of size $n \in \{25, 50, 100\}$ from the discrete Kumaraswamy rv with parameters $k \in \{5, 7\}$, $a \in \{0.5, 0.75, 1, 1.5, 2.5\}$, and $b \in \{0.5, 0.75, 1, 1.5, 2.5\}$. For each setting, on each sample, we computed the estimates of the two parameters a and b (with k assumed to be known a priori) using the method of proportions, the method of moments, the method of minimum Chi-square, and the maximum likelihood method. We then compared the performance of these estimators in terms of their Monte Carlo biases and root mean square errors (RMSEs) over the S realizations of each setting. For an estimator $\hat{\theta}$ of a parameter θ, the Monte Carlo bias and RMSE are defined as follows:

$$\text{bias}_{MC}(\hat{\theta}) = \frac{1}{S}\sum_{s=1}^{S}\hat{\theta}_s - \theta$$

$$\text{RMSE}_{MC}(\hat{\theta}) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}(\hat{\theta}_s - \theta)^2},$$

where $\hat{\theta}_s$ is the estimate of θ computed on the s-th sample, $s = 1, \ldots, S$. For each examined setting, we also computed the 95% confidence intervals (CIs) for a and b based on the likelihood profile for each sample, along with the average confidence interval lengths and the coverage probabilities across the S replications. For the method of proportions, we recorded, for each setting, the fraction of samples for which it was not possible to provide valid estimates for a and b; these are the samples that do not contain any 1's or k's.

Since the Monte Carlo RMSE is the square root of the sum of the square of the Monte Carlo bias and the Monte Carlo variance of the estimator, we will use it for comparing the performance of the different estimators of a and b. Figures 12 and 13 display the RMSEs of the four estimators of a and b, respectively, in each of the settings previously defined, when $k = 7$ and $n = 100$.

From Figure 12, it can be observed that, for the parameter a, no estimator performs uniformly better than the others. However, quite surprisingly, the MCS emerges as the overall best performer, although it is closely followed by the ML estimator, which shows visibly larger RMSE when $a = 2.5$ for all the examined values of b. The MM and MP estimators perform slightly worse than the former; in particular, the MP estimator exhibits a significantly larger RMSE compared to the others when $a = 0.5$ and $b = 2.5$, as well as when $a = 2.5$ and $b = 0.5$.

From Figure 13, it can be observed that even for the parameter b, no estimator performs uniformly better than the others. However, the MCS estimator again emerges as the overall best performer, although it is closely followed by the ML estimator, which exhibits a larger RMSE when $b = 2.5$ for all the examined values of a. The MM and MP estimators perform slightly worse than the former.

The apparent paradox whereby the MCS estimator achieves a smaller RMSE than the MLE, although the latter is theoretically expected to be asymptotically the most efficient, can be understood in light of finite-sample considerations. Asymptotic efficiency results are guaranteed only in the limit as the sample size tends to infinity, while for moderate sample sizes such as $n = 100$ the asymptotic properties of the MLE may not yet be fully realized. Furthermore, the Monte Carlo experiments themselves introduce an additional source of variability due to the finite number of replications, and occasional convergence issues or suboptimal solutions in the numerical optimization of the likelihood function may produce poor estimates that disproportionately increase the empirical RMSE.

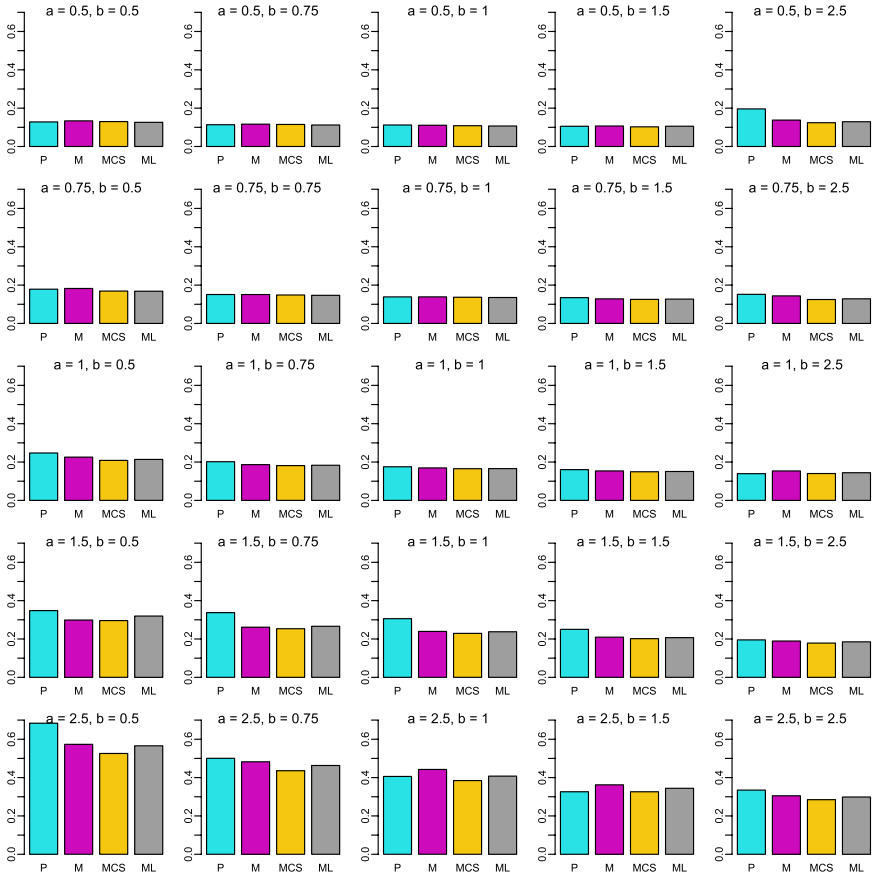

Fig. 12 Results on RMSE for the estimators of the parameter a with $k = 7$ and $n = 100$

However, to further investigate the issue, we conducted an additional ad hoc simulation study. Table 10 shows the values of RMSE values of the estimator $\hat{b}$ when $a = 1.5$, $b = 0.75$, and $k = 7$ across sample sizes $n \in \{50, 100, 200, 500\}$, along with the relative efficiency of the MLE with respect to the MCS estimator, defined as the squared ratio of the RMSE of the latter to that of the former. As can be seen, for smaller sample sizes ($n = 50$ and 100) the MCS estimator outperforms the MLE, whereas for larger sample sizes ($n = 200$ and 500) the MLE emerge as slightly more efficient. The relative efficiency is only marginally above 1, confirming that asymptotically the MLE is the most efficient estimator, but it is also equivalent to the MCS estimator.

For the setting $(a = 2.5, b = 2.5, k = 7)$, Figure 14 displays the scatterplots of $(\hat{a}_{MCS}, \hat{a}_{ML})$ – left panel – and $(\hat{b}_{MCS}, \hat{b}_{ML})$ – right panel – for $n = 100$. It is

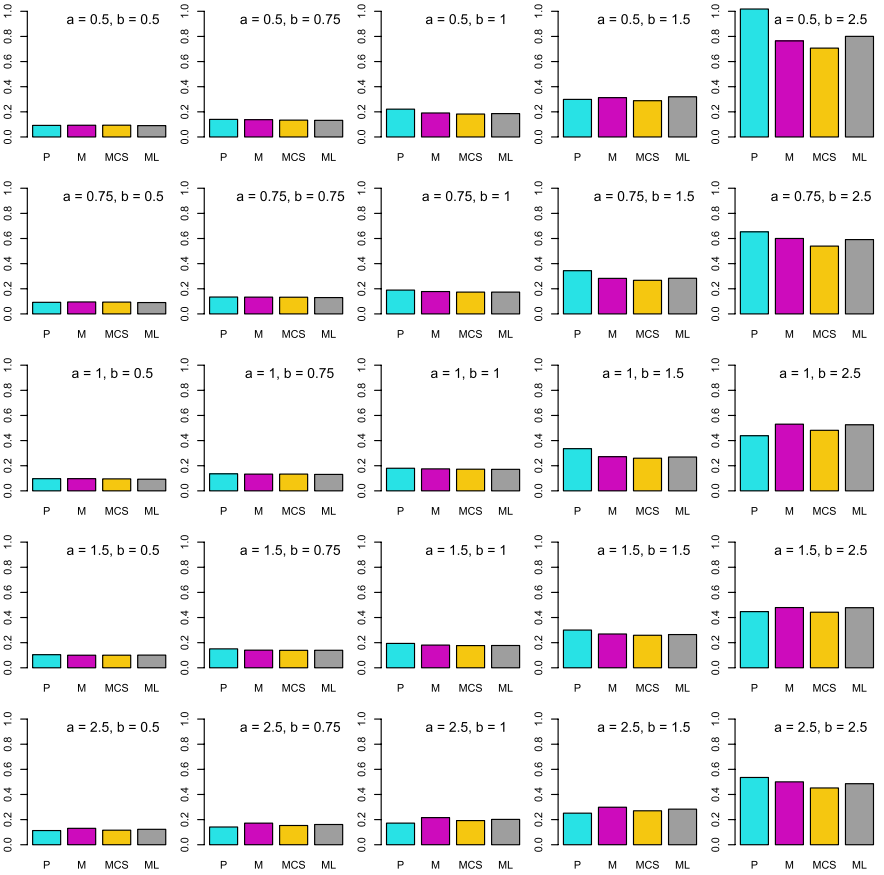

Fig. 13 Results on RMSE for the estimators of the parameter b with $k = 7$ and $n = 100$

Table 10 RMSEs of the MLE and MCS estimator of b, and relative efficiency of the MLE, for $a = 1.5$, $b = 0.75$, and $k = 7$, across different sample sizes

Estimator, n	50	100	200	500
MCS	0.2049	0.1391	0.0948	0.0587
ML	0.2075	0.1394	0.0947	0.0586
rel.eff	0.9748	0.9953	1.0018	1.0011

readily apparent that the MCS and ML estimators return two similar values on each sample, for both the shape parameters; all the points in each graph lie in fact very close to the bisector of the first and third quadrants. Therefore, the two estimators not only exhibit very similar marginal distributions and RMSE values, but also show a high level of agreement in the estimates provided for each possible sample.

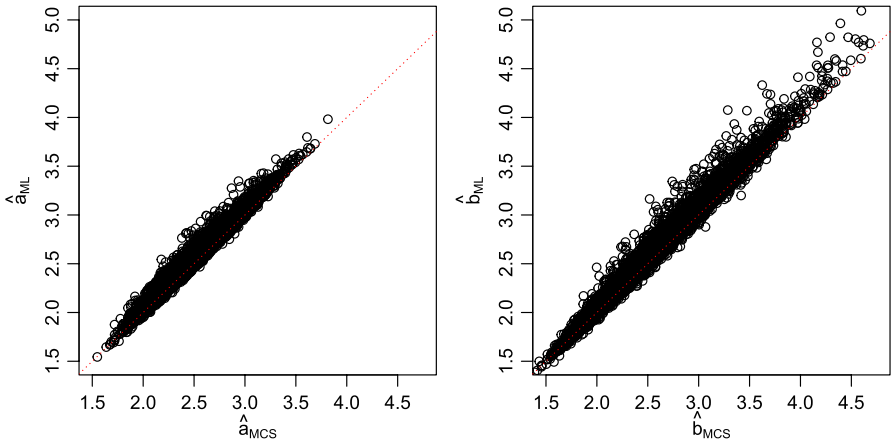

Fig. 14 Joint distribution of the minimum Chi-square and maximum likelihood estimators of the shape parameters of the discrete Kumaraswamy distribution, when $a = b = 2.5$, $k = 7$, and $n = 100$. On the left panel, the scatterplot of $(\hat{a}_{MCS}, \hat{a}_{ML})$; on the right panel, the scatterplot of $(\hat{b}_{MCS}, \hat{b}_{ML})$. The red dotted line represents the 1st-3rd quadrants' bisector. The scale on the x and y axes are the same across the two graphs

Acknowledgements We thank the Editor, the Associate Editor, and the anonymous Referee for their thorough review of the paper and constructive comments, which helped us improve the clarity and quality of the early draft of the manuscript.

Author contributions Conceptualization: A.B., A.H.; Methodology: A.B., A.H.; Formal analysis and investigation: A.B., A.H.; Writing - original draft preparation: A.B.; Writing - review and editing: A.B., A.H.; Resources: A.B., A.H.; Supervision: A.B., A.H.

Funding Open access funding provided by Università degli Studi di Milano within the CRUI-CARE Agreement. The corresponding author acknowledges financial support by the PRIN2022 project 'The effects of climate change in the evaluation of financial instruments' financed by the Italian 'Ministero dell'Università e della Ricerca' with grant number 20225PC98R, CUP Code: G53D23001960006.

Data availability Nodatasets weregenerated during the current study. R code implementing the proposed distribution and the real data analyses are available on GitHub: https://github.com/alessandro-barbiero/discreteKumaraswamy.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Consent to publication All authors gave explicit consent to publish this manuscript.

References

Agresti A (2010) Analysis of ordinal categorical data. Wiley, Hoboken

An MY (1997) Log-concave probability distributions: Theory and statistical testing. Duke University Dept of Economics Working Paper **95**(03)

Barbiero A (2017) Least-squares and minimum chi-square estimation in a discrete Weibull model. Commun Stat-Simul Comput 46(10):8028–8048

Barbiero A, Hitaj A (2023) Discrete approximations of continuous probability distributions obtained by minimizing Cramér-von Mises-type distances. Stat Pap 64(5):1669–1697

Benney T, Chaney R, Singer P, Sloan C (2020) Utah air quality risk and behavioral action survey. Technical report, Interuniversity Consortium for Political and Social Research [distributor], Ann Arbor. https://doi.org/10.3886/E117904V1

Bolker B (2023) R Development Core Team: Bbmle: tools for general maximum likelihood estimation. R package version 1.0.25.1. https://CRAN.R-project.org/package=bbmle

Chyung SY, Hutchinson D, Shamsy JA (2020) Evidence-based survey design: ceiling effects associated with response scales. Perform Improv 59(6):6–13

Corduas M, Iannario M, Piccolo D (2009) A class of statistical models for evaluating services and performances. Statistical methods for the evaluation of educational services and quality of products. Physica-Verlag, Heidelberg, pp 99–117

D'Elia A, Piccolo D (2005) A mixture model for preferences data analysis. Comput Stat Data Anal 49(3):917–934

Dey S, Mazucheli J, Nadarajah S (2018) Kumaraswamy distribution: different methods of estimation. Comput Appl Math 37:2094–2111

Fasola S, Sciandra M (2015) New flexible probability distributions for ranking data. Adv Stat Models Data Anal. Springer, Cham, pp 117–124

Flowerday CE, Thalman R, Hansen JC (2023) Twenty-year review of outdoor air quality in Utah, USA. Atmosphere 14(10):1496

Flury BA (1990) Principal points. Biometrika 77(1):33–41

Ghosh SK, Burns CB, Prager DL, Zhang L, Hui G (2018) On nonparametric estimation of the latent distribution for ordinal data. Comput Stat Data Anal 119:86–98

Golub GH, Welsch JH (1969) Calculation of Gauss quadrature rules. Math Comput 23(106):221–230

Hatzinger R, Dittrich R (2012) prefmod: An R package for modeling preferences based on paired comparisons, rankings, or ratings. J Stat Softw 48:1–31

Hatzinger R, Maier MJ (2023) Prefmod: utilities to fit paired comparison models for preferences. R package version 0.8-36. https://CRAN.R-project.org/package=prefmod

Iannario M (2014) Modelling uncertainty and overdispersion in ordinal data. Commun Stat-Theory Methods 43(4):771–786

Johnson NL, Kotz S, Balakrishnan N (1994) Continuous univariate distributions, vol 2. Wiley, New York

Jones M (2009) Kumaraswamy's distribution: a beta-type distribution with some tractability advantages. Stat Methodol 6(1):70–81

Khan MSA, Khalique A, Abouammoh AM (1989) On estimating parameters in a discrete Weibull distribution. IEEE Trans Reliab 38(3):348–350

Kumaraswamy P (1980) A generalized probability density function for double-bounded random processes. J Hydrol 46(1–2):79–88

Landenna G, Marasini D, Ferrari PA (1996) Teoria della Stima. Il Mulino, Bologna

Leti G (1983) Statistica descrittiva. Il Mulino, Bologna

Mitnik PA (2013) New properties of the Kumaraswamy distribution. Commun Stat-Theory Methods 42(5):741–755

Mitnik PA, Baek S (2013) The Kumaraswamy distribution: median-dispersion re-parameterizations for regression modeling and simulation-based estimation. Stat Pap 54:177–192

Punzo A (2010) Discrete beta-type models. In: Classification as a tool for research: proceedings of the 11th IFCS Biennial Conference and 33rd annual conference of the Gesellschaft Für Klassifikation e.V. Springer, pp 253–261

Punzo A, Zini A (2008) Discrete approximations of continuous and mixed measures on a closed interval. Technical Report 160, Università di Milano-Bicocca, Dipartimento di Metodi Quantitativi per le Scienze Economiche e Aziendali

R Core Team (2024) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. R Foundation for Statistical Computing. https://www.R-project.org/
Sciandra M, Fasola S, Albano A, Di Maria C, Plaia A (2024) Discrete beta and shifted beta-binomial models for rating and ranking data. Environ Ecol Stat 31:317–338
Turner R (2021a) dbd: Discretised Beta Distribution. R package version 0.0-22. https://CRAN.R-project.org/package=dbd
Turner R (2021b) A new versatile discrete distribution. R J 13(2):485–506