Advances in Data Analysis and Classification (2023) 17:659–699
https://doi.org/10.1007/s11634-022-00520-8

REGULAR ARTICLE

New models for symbolic data analysis

Boris Beranger[1] · Huan Lin[1] · Scott Sisson[1]

Received: 7 April 2020 / Revised: 23 August 2022 / Accepted: 30 August 2022 /
Published online: 19 September 2022

Abstract

Symbolic data analysis (SDA) is an emerging area of statistics concerned with understanding and modelling data that takes distributional form (i.e. *symbols*), such as random lists, intervals and histograms. It was developed under the premise that the statistical unit of interest is the symbol, and that inference is required at this level. Here we consider a different perspective, which opens a new research direction in the field of SDA. We assume that, as with a standard statistical analysis, inference is required at the level of individual-level data. However, the individual-level data are unobserved, and are aggregated into observed symbols—group-based distributional-valued summaries—prior to the analysis. We introduce a novel general method for constructing likelihood functions for symbolic data based on a desired probability model for the underlying measurement-level data, while only observing the distributional summaries. This approach opens the door for new classes of symbol design and construction, in addition to developing SDA as a viable tool to enable and improve upon classical data analyses, particularly for very large and complex datasets. We illustrate this new direction for SDA research through several real and simulated data analyses, including a study of novel classes of multivariate symbol construction techniques.

Keywords Binned data · Interval data · Likelihoods · Summary statistics · Symbol design

Mathematics Subject Classification 62H86 · 62R07

1 Introduction

Symbolic data analysis (SDA) is an emerging area of statistics that has immense potential to become a standard inferential technique in the near future (Billard and Diday

✉ Boris Beranger
b.beranger@unsw.edu.au

[1] School of Mathematics and Statistics, UNSW Data Science Hub (uDASH), University of New South Wales, Sydney, Australia

2003). At its core, it builds on the notion that exploratory analyses and statistical inferences are commonly required at a group level rather than at an individual level (Diday 1988; Billard 2011; Billard and Diday 2006). This is the familiar notion behind hierarchical modelling (e.g. Gelman et al. 2013, Chapter 5). For example, the performance of school and higher level units in standardised testing exams is usually of interest rather than the performance of the individual students (Rodrigues et al. 2016; Rubin 1981).

SDA explicitly embraces this idea by considering group level distributional summaries (i.e. *symbols*) as the statistical unit of interest, and then analysing the data at this summary level (Billard 2011; Billard and Diday 2006). The most common choice of these summaries is the random interval (or the d-dimensional equivalent, the random rectangle; throughout we use the term 'random rectangle' to include d-dimensional *hyper* rectangles). Here, for individual-level observations $X_1, \ldots, X_n \in \mathbb{R}$, the random interval is typically constructed as $S = (\min_i X_i, \max_i X_i) \subseteq \mathbb{R}$. Quantile-based intervals have only received little attention (e.g. Hron et al. 2017). Other common symbol types include random histograms (Dias and Brito 2015; Le-Rademacher and Billard 2013) and categorical multi-valued variables (Billard and Diday 2006). Under the SDA framework, the collection of group-level data summaries $S_1, \ldots, S_m \in \mathcal{S}$ are considered the new data "points", whereby each datum is a distribution of some kind with an internal distributional structure. Statistical inference is then performed at the level of the symbols directly, with reference to their distributional forms, and without any further reference to the underlying measurement-level data. See e.g. Noirhomme-Fraiture and Brito (2011), Billard (2011) and Billard and Diday (2003) for a comprehensive overview of symbolic data types and their analysis.

This approach is potentially extremely attractive given present technological trends requiring the analysis of increasingly large and complex datasets. SDA effectively states that for many analyses, the high level of computation required for e.g. divide-and-recombine techniques (e.g. Guha et al. 2012; Jordan et al. 2019; Vono et al. 2019; Rendell et al. 2020) or subsampling-based techniques (Quiroz et al. 2018; Bardenet et al. 2014; Quiroz et al. 2019), is not necessary to make inference at the group level.

By aggregating the individual-level data to a much smaller number of group level symbols m (where $m \ll n$), 'big data' analyses can be performed cheaply and effectively on low-end computing devices. Recent work by Whitaker et al. (2021) has shown that SDA can outperform bespoke subsampling techniques for logistic regression models, in terms of much lower computational overheads for the same predictive accuracy. Beyond data aggregation, distributional-valued observations can arise naturally through the data recording process, representing underlying variability. This can include e.g. observational rounding or truncation, which results in imprecise data known to lie within some interval (Heitjan and Rubin 1991; Vardeman and Lee 2005), and the elicitation of distributions from experts thought to contain quantities of interest (Fisher et al. 2015; Lin et al. 2022). In this sense, Schweizer (1984)'s often-quoted statement that "distributions are the numbers of the future" seems remarkably prescient.

Many SDA techniques for analysing distributional-valued random variables have been developed (and here we take 'distributional-valued' random variables to include random intervals or random rectangles that have no specific distributional form aside from the specified quantiles). These include regression models (Irpino and Verde 2015; Le-Rademacher and Billard 2013), time series (Lin and González-Rivera 2016), clustering and classification (Whitaker et al. 2021), discriminant analysis (Duarte Silva and Brito 2015) and Bayesian hierarchical modelling (Lin et al. 2022). Likelihood-based inference was introduced by Le-Rademacher and Billard (2011) and Brito and Duarte Silva (2012) with further development and application by Zhang et al. (2020), Rahman et al. (2022), Lin et al. (2022).

While there have been many successes in the analysis of symbolic data, from a statistical perspective there are many opportunities for methodological improvement. Some of these opportunities relate to existing SDA approaches, under which the statistical unit of interest is the symbol, and where inference is required at this level (either exploratory or statistical inference). For example, the large majority of SDA techniques are descriptive and do not permit statistical inference on model parameters. E.g., regression models tend to be fitted by symbolic variants of least squares. Other opportunities arise, as with the present work, by re-imagining how the ideas behind SDA can be used to solve modern statistical challenges. Here we assume that, as with a standard statistical analysis, inference is required at the level of the individual-level data, but where we deliberately aggregate the individual-level data into symbols prior to the analysis. Hence, if we can develop a way to perform statistical inference on the individual-level data when only given the group-level summaries, then we can potentially perform standard statistical inference for large and complex datasets more efficiently via these distributional summaries than when directly using the original data. This alternative perspective on the ideas underlying SDA methodology opens up a new research direction in the field of SDA. Here, we focus on likelihood-based inference.

The likelihood approach of Le-Rademacher and Billard (2011), Brito and Duarte Silva (2012) maps each symbol to a random vector that uniquely defines the symbol, and then models this via a standard likelihood model. E.g., suppose that $X_{ij} \in \mathbb{R}$ is the value of some process recorded on the i-th second, $i = 1, \ldots, n = 86{,}400$, of the j-th day, $j = 1, \ldots, m$. If interest is in modelling these data as, say, i.i.d draws from a skew-normal distribution $X_{ij} \sim SN(\mu_0, \sigma_0, \alpha_0)$, the likelihood function $L(x|\theta)$, $\theta \in \Theta$, may then be easily constructed. However, suppose that interval symbols are now constructed so that $S_j = (\min_i X_{ij}, \max_i X_{ij}) \subseteq \mathbb{R}$ is the random interval describing the observed process range on day j. Due to the equivalence of representing continuous subsets of $\mathbb{R}$ by the associated bivariate vector in this setting (Zhang et al. 2020), the approach of Le-Rademacher and Billard (2011), Brito and Duarte Silva (2012) constructs a model for the vectorised symbols $S_1, \ldots, S_m$, perhaps after a reparameterisation. For example,

$$S_j \sim SN_2(\mu, \Sigma, \alpha) \qquad \text{or} \qquad \tilde{S}_j \sim SN_2(\mu, \Sigma, \alpha),$$

where $\tilde{S}_j = ((a+b)/2, \log(b-a))$ is a typical reparameterisation of $S_j = (a, b)$ into a function of interval mid-point and log range (Brito and Duarte Silva 2012). While there is inferential value in models of these kind (e.g. Brito and Duarte Silva 2012; Lin et al. 2022), it is clear that if there is interest in modelling the underlying X_{ij} as skew-normal, it is difficult to construct even a loosely equivalent model at the level of the symbol S_j (or $\tilde{S}_j$). That is, while the analyst may intuitively construct complex statistical models at the level of the individual-level data, it is less obvious how to construct models at the symbolic level and for different symbolic forms.

By design, modelling symbols directly, without specifying a probabilistic model for the underlying micro-data, only permits inference and predictions at the symbol level. This is unsatisfactory because predictive inference for the underlying micro-data is often of interest, even if primary focus is on group-level analysis, and as we demonstrate in Sect. 3.3, ignoring the structure of the micro-data can result in symbolic-level analyses producing poorer inferential outcomes. Another clear and acknowledged problem (Kosmelj et al. 2014; Cariou and Billard 2015) is that even though existing SDA techniques do not focus on the individual-level data, the distribution of this data within random intervals/rectangles and within histogram bins is typically assumed to be uniform. Alternatives include the triangular distribution (Le-Rademacher and Billard 2011; Dias and Brito 2017). When considering that random intervals are typically constructed by specifying $S_j = (\min_i X_{ij}, \max_i X_{ij})$, it is almost certain that the distribution of the underlying data within S_j is non-uniform. This implies that any inferential procedure built on the uniformity assumption (i.e. almost all current SDA methods) is likely to produce questionable results.

One principled difference between SDA and regular statistical analyses is that the analysed symbolic data can be constructed by the analyst. This raises the question of how this should be undertaken. Intuitively, if looking to design, say, a random interval S_j to maximise information about a location parameter, using the sample maximum and minimum is likely a poor choice as these statistics are highly variable. A more useful alternative could use e.g. sample quantiles to define the interval. While sample quantiles have been considered in SDA methods, they have only been used as a robust method to avoid outliers that would otherwise dominate the size of a random interval (Hron et al. 2017). In general, little consideration has been given to the design of informative symbols.

In this paper we introduce a novel general method for constructing likelihood functions for symbolic data based on specifying a standard statistical model $L(X|\theta)$ for the underlying measurement-level data and then deriving the implied model $L(S|\theta)$ at the symbolic level by considering how S is constructed from x. This construction assumes that we are in the setting where the symbolic data are created through a data aggregation process. This provides a way to fit the measurement-level data model $L(X|\theta)$ while only observing the symbol level data, S. It provides both a natural way of specifying models for symbolic data, while also opening up SDA methods as a mainstream technique for the fast analysis of large and complex datasets. This approach naturally avoids making the likely invalid assumption of within-symbol uniformity, allows inference and predictions at both the measurement data and symbolic data levels, permits symbolic inference using multivariate symbols (a majority of symbolic analyses are based on vectors of univariate symbols), and can provide a higher quality of inference than standard SDA techniques. The method recovers some known models in the statistical literature, as well as introducing several new ones, and reduces to standard likelihood-based inference for the measurement-level data (so that $L(S|\theta) \to L(X|\theta)$) when $S \to X$.

As a result we demonstrate some weaknesses of current symbol construction techniques. In particular we establish informational limits on random rectangles constructed from marginal minima/maxima or quantiles, and introduce a new class of quantile-based random rectangles. These alternative symbol variations produce more efficient analyses than existing symbol constructions, and permit the estimation of within-symbol multivariate dependencies that were not previously estimable.

The new symbolic likelihood function is presented in Sect. 2 with specific results for random rectangles and histograms. All derivations are relegated to the Appendix. The performance of these models is demonstrated in Sect. 3 through a meta-analysis of univariate histograms, a simulation study of the inferential performance of an alternative class of multivariate random rectangle constructions, and an analysis of a large loan dataset. In all cases, the existing state-of-the-art models and symbolic constructions are outperformed by the new symbolic model. Section 4 concludes with a discussion.

Throughout this manuscript we adopt the convention that upper case letters X denote (vector or scalar) random variables, whereas lower case letters x denote their observed values. We also write matrices (as vectors of random variables) $\boldsymbol{X}, \boldsymbol{x}$ in bold font.

2 A general construction tool for symbolic likelihoods

2.1 Symbolic likelihood functions

Consider that Ω is a sample space defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and that each element of Ω is described by a measurable random variable X, defined by $X : \Omega \rightarrow \mathcal{X}$, and $\mathbb{P}(X \in \mathcal{Y}) = \mathbb{P}(\omega \in \Omega | X(\omega) \in \mathcal{Y})$, for $\mathcal{Y} \subset \mathcal{X}$. We follow the standard SDA construction of a *class* (Billard and Diday 2003) and let the random variable $C : \Omega \rightarrow \mathcal{C}$ denote the class to which an individual belongs. For simplicity we assume that $\mathcal{C} = \{1, \ldots, m\}$ is finite. Consequently let $\Omega_c = \{\omega \in \Omega \text{ s.t. } C(\omega) = c\} \subseteq \Omega$ be the set of all possible outcomes that belong to class $c \in \mathcal{C}$, and define $X_c : \Omega_c \rightarrow \mathcal{X}_c \subseteq \mathcal{X}$ as the random variable that describes them. No assumptions about the countability of Ω or Ω_c are made since the results presented in this manuscript hold for both discrete and continuous random variables. We assume that we have a population of size N for which the variables $X_1, \ldots, X_N$ are measured and given in vector form by $\boldsymbol{X} = (X_1, \ldots, X_N)$. The population is itself decomposed into m sub-populations of size N_c such that $\sum_{c \in \mathcal{C}} N_c = N$, corresponding to the groups of individuals belonging to class c. For each class $c \in \mathcal{C}$, we define by $\boldsymbol{X}_c = (X_{1,c}, \ldots, X_{N_c,c})$ the vector of random variables describing the population of class c.

We now consider an interpretation of SDA, where the symbolic random variable S_c for class $c \in \mathcal{C}$ is assumed to be the result of the aggregation of the random vector $\boldsymbol{X}_c$ via some aggregation function π_c, so that $S_c = \pi_c(\boldsymbol{X}_c) : [\mathcal{X}_c]^{N_c} \rightarrow \mathcal{S}_c$ and $\boldsymbol{x}_c \mapsto \pi(\boldsymbol{x}_c)$. That is, a symbolic random variable is a statistic which represents a summary of the information brought by measurement over individuals. The choice of this summary (and thus of the aggregation function) is critical and we explore this in later sections. In the following we refer to random variables of the measurement-level data X as *classical* random variables. By construction symbolic random variables require knowledge of the underlying classical random variables. Accordingly, this should also be true when dealing with likelihood functions, particularly if inference is required at both classical and symbolic levels, but when only information at the symbolic level is observed.

To construct a symbolic likelihood function, suppose that the classical random variable X has probability density and distribution functions $g_X(\cdot\,;\theta)$ and $G_X(\cdot\,;\theta)$ respectively, where $\theta \in \Theta$. Consider a random classical data sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ of size $n < N$ from the population, and denote by $\boldsymbol{x}_c = (x_{1,c}, \ldots, x_{n_c,c})$, the collection of those in class c, where $\sum_{c\in\mathcal{C}} n_c = n$. Similarly let $s_c = \pi_c(\boldsymbol{x}_c)$ be the resulting observed symbol obtained through the aggregate function π_c and define the symbolic dataset to be the collection of symbols $\boldsymbol{s} = (s_c; c \in \mathcal{C})$.

Proposition 1 *For the subset $\boldsymbol{x}_c$ of $\boldsymbol{x}$ associated with class $c \in \mathcal{C}$, the likelihood function of the corresponding symbolic observation $s_c = \pi_c(\boldsymbol{x}_c)$ is given by*

$$L(s_c; \vartheta, \theta) \propto \int_{\mathcal{X}^n} f_{S_c|\boldsymbol{X}_c=\boldsymbol{z}_c}(s_c; \vartheta) g_{\boldsymbol{X}}(\boldsymbol{z}; \theta) \mathrm{d}\boldsymbol{z}, \quad \forall c \in \mathcal{C}, \tag{1}$$

where $\boldsymbol{z}_c \in \mathcal{X}_c^{n_c}$ is a subset of $\boldsymbol{z} \in \mathcal{X}^n$, $f_{S_c|\boldsymbol{X}_c}(\cdot; \vartheta)$ is the conditional density of S_c given $\boldsymbol{X}_c$ and $g_{\boldsymbol{X}}(\cdot\,;\theta)$ is the joint density of $\boldsymbol{X}$.

We refer to $L(s_c; \vartheta, \theta)$ given in (1) as the symbolic likelihood function. A discrete version of (1) is easily constructed. Note that by writing the joint density $g_{\boldsymbol{X}}(\cdot\,;\theta) = g_{\boldsymbol{X}_c}(\cdot\,;\theta) g_{\boldsymbol{X}_{-c}|\boldsymbol{X}_c}(\cdot\,;\theta)$, where $\boldsymbol{X}_{-c} = \boldsymbol{X}\backslash\boldsymbol{X}_c$, then after integration with respect to $\boldsymbol{x}_{-c} = \boldsymbol{x}\backslash\boldsymbol{x}_c$, equation (1) becomes

$$L(s_c; \vartheta, \theta) \propto \int_{\mathcal{X}_c^{n_c}} f_{S_c|\boldsymbol{X}_c=\boldsymbol{z}_c}(s_c; \vartheta) g_{\boldsymbol{X}_c}(\boldsymbol{z}_c; \theta) \mathrm{d}\boldsymbol{z}_c.$$

This construction method can easily be interpreted: the probability of observing a symbol s_c is equal to the probability of generating a classical dataset under the classical data model that produces the observed symbol under the aggregation function π_c. That is, we have established a direct link between the user-specified classical likelihood function $L(\boldsymbol{x}|\theta) \propto g_{\boldsymbol{X}}(\boldsymbol{x}; \theta)$ and the resulting probabilistic model on the derived symbolic data. As a result we may directly estimate the parameters θ of the underlying classical data model, based only on observing the symbols $\boldsymbol{s}$.

In the case where there is no aggregation of $\boldsymbol{x}_c$ into a symbol, so that $\pi(\boldsymbol{x}_c) = \boldsymbol{x}_c$ and $\mathcal{S}_c = [\mathcal{X}_c]^{N_c}$, then $f_{S_c|\boldsymbol{X}_c=\boldsymbol{z}_c}(s_c) \equiv f_{\pi(\boldsymbol{X}_c)|\boldsymbol{X}_c=\boldsymbol{z}_c}(\pi(\boldsymbol{x}_c)) = f_{\boldsymbol{X}_c|\boldsymbol{X}_c=\boldsymbol{z}_c}(\boldsymbol{x}_c) = \delta_{\boldsymbol{z}_c}(\boldsymbol{x}_c)$, where $\delta_{\boldsymbol{z}_c}(\boldsymbol{x}_c)$ is the Dirac delta function, taking the value 1 if $\boldsymbol{z}_c = \boldsymbol{x}_c$ and 0 otherwise. As a result the symbolic likelihood function reduces to $g_{\boldsymbol{X}_c}(\boldsymbol{x}_c; \theta)$, the classical likelihood contribution of class c. Under the assumption that the classical data are independently distributed between classes, so that $g_{\boldsymbol{X}}(\cdot\,;\theta) = \prod_{c\in\mathcal{C}} g_{\boldsymbol{X}_c}(\cdot\,;\theta)$, the associated symbols are also independent and the likelihood of the symbolic dataset $\boldsymbol{s}$ is given by

$$L(\boldsymbol{s}; \vartheta, \theta) = \prod_{c\in\mathcal{C}} L(s_c; \vartheta, \theta) \propto \prod_{c\in\mathcal{C}} \int_{\mathcal{X}_c^{n_c}} f_{S_c|\boldsymbol{X}_c=\boldsymbol{z}_c}(s_c; \vartheta) g_{\boldsymbol{X}_c}(\boldsymbol{z}_c; \theta) \mathrm{d}\boldsymbol{z}_c.$$

If, further, the observations within a class $c \in \mathcal{C}$ are independent and identically distributed, then in the scenario where $\pi(\boldsymbol{x}_c) = \boldsymbol{x}_c$ we have $L(\theta) = \prod_{i=1}^{n} g_X(x_i; \theta)$. Because $\text{Card}(\mathcal{C}) = m$ and typically $m \ll n$, this implies that large computational savings can be made through the analysis of symbolic rather than classical data, depending on the complexity of the classical data likelihood function. The method established in Proposition 1 specifies a probability model for the micro-data which, combined with knowledge of the aggregation process π, induces a likelihood function at the aggregate level. In contrast, the likelihood function defined by Le-Rademacher and Billard (2011), Brito and Duarte Silva (2012) specifies a probability model directly on the symbols.

While we do not pursue this further here, we note that the function $f_{S_c|\boldsymbol{X}_c}(\cdot; \vartheta)$ is not constrained to be constructed from Dirac functions (such as when S_c is fully determined by $\boldsymbol{X}_c$), and may be a full probability function. This allows for the incorporation of randomness in construction of the symbols from the micro-data, such as the random allocation of the micro-data to different symbols. The parameters ϑ are fixed quantities that determine the structure of how a symbol will be constructed, e.g., the locations of bins for histogram symbols. While we explore this in Sect. 3 where we introduce a number of new ideas in symbol construction techniques, there is much scope, beyond this paper, to explore these ideas further.

In the following subsections, we establish analytical expressions of the symbolic likelihood function based on various choices of the aggregation function π, which leads to different symbol types. The performance of each of these models will be examined in Sect. 3. For clarity of presentation the class index c is omitted in the remainder of this section as the results presented are class specific.

2.2 Modelling random intervals

The univariate random interval is the most common symbolic form, and is typically constructed as the range of the underlying classical data e.g. $S = (\min_i X_i, \max_i X_i)$. Here we generalise this to order statistics $S = (X_{(l)}, X_{(u)})$ for indices $l \leq u$ given their higher information content. We define an interval-valued symbolic random variable to be constructed by the aggregation function π where

$$S = \pi(\boldsymbol{X}) : \mathbb{R}^N \to \mathcal{S} = \{(a_1, a_2) \in \mathbb{R}^2 : a_1 \leq a_2\} \times \mathbb{N} \tag{2}$$

so that $\boldsymbol{x} \mapsto (x_{(l)}, x_{(u)}, N)$, where $x_{(k)}$ is the k-th order statistic of $\boldsymbol{x}$ and $l, u \in \{1, \ldots, N\}$, $l \leq u$ are fixed. Taking $l = 1$, $u = N$ corresponds to determining the range of the data. (Note that modelling an interval $(a_1, a_2) \in \mathbb{R}^2$ as a bivariate random vector is mathematically equivalent to modelling it as a subset of the real line $(a_1, a_2) \subseteq R$. See e.g. Zhang et al. 2020). Note that this construction explicitly includes the number of underlying datapoints N in the interval as part of the symbol, in direct contrast to almost all existing SDA techniques. This allows random intervals constructed using different numbers of underlying classical datapoints to contribute to the likelihood function in relation to the size of the data that they represent. This is not available in the construction of Le-Rademacher and Billard (2011), Brito and Duarte Silva (2012).

Lemma 1 *Consider a univariate interval-valued random variable $S = (s_l, s_u, n) \in \mathcal{S}$, obtained through (2) and assume that $g_{\boldsymbol{X}}(\boldsymbol{x}; \theta) = \prod_{i=1}^n g_X(x_i; \theta)$, $\boldsymbol{x} \in \mathbb{R}^n$, where g_X is a continuous density function with unbounded support. The corresponding symbolic likelihood function is then given by*

$$L(s_l, s_u, n; \theta) = \frac{n!}{(l-1)!(u-l-1)!(n-u)!} \left[G_X(s_l; \theta)\right]^{l-1} \\ \times \left[G_X(s_u; \theta) - G_X(s_l; \theta)\right]^{u-l-1} \left[1 - G_X(s_u; \theta)\right]^{n-u} g_X(s_l; \theta) g_X(s_u; \theta).$$

It is worth noting that this expression can also be obtained by evaluating $\mathbb{P}(S_l \leq s_l, S_u \leq s_u) = \mathbb{P}(X_{(l)} \leq s_l, X_{(u)} \leq s_u)$ and then taking derivatives with respect to s_l and s_u, and corresponds to the joint distribution of order two statistics. This model was previously established by Zhang et al. (2020) as a generative model for random intervals built from i.i.d. random variables.

2.3 Modelling random rectangles

The typical method of constructing multivariate random rectangles from underlying d-dimensional data $X \in \mathbb{R}^d$, $d \in \mathbb{N}$ is by taking the cross product of each d univariate random interval described by their marginal minima and maxima (e.g. Neto et al. 2011; Ichino 2011). The number of datapoints underlying this rectangle is often not used. We improve on this scheme by making use of additional information available at the time of rectangle construction (Sect. 2.3.1), and then develop several alternative constructions for random rectangles based on marginal order statistics (Sect. 2.3.2).

2.3.1 Using marginal maxima and minima

While it is in principle possible to identify a small amount of information about the dependence between two variables summarised by a marginally constructed bounding box, this information content is very weak, and the direction of dependence is not identifiable (Zhang et al. 2020). E.g. if n datapoints are generated from a multivariate distribution and the marginal minimum and maximum values recorded, what can be said about the correlation strength and direction? We propose that dependence information can be obtained if the locations of those datapoints involved in construction of the bounding rectangle, and the total number of points are known. For the examples in Fig. 1 (top), if the rectangle is generated from only two points (left panel) one can surmise stronger dependence than if three points are used (centre), with rectangle construction using four points (right) producing the weakest dependence. The exact locations of these bounding points is informative of dependence direction. (We note that the data points used to construct a multivariate random rectangle are immediately obtained when constructing the random rectangle in the usual way.)

As such, we define the aggregation function π to incorporate these construction points (where available) into the definition of the random rectangle as

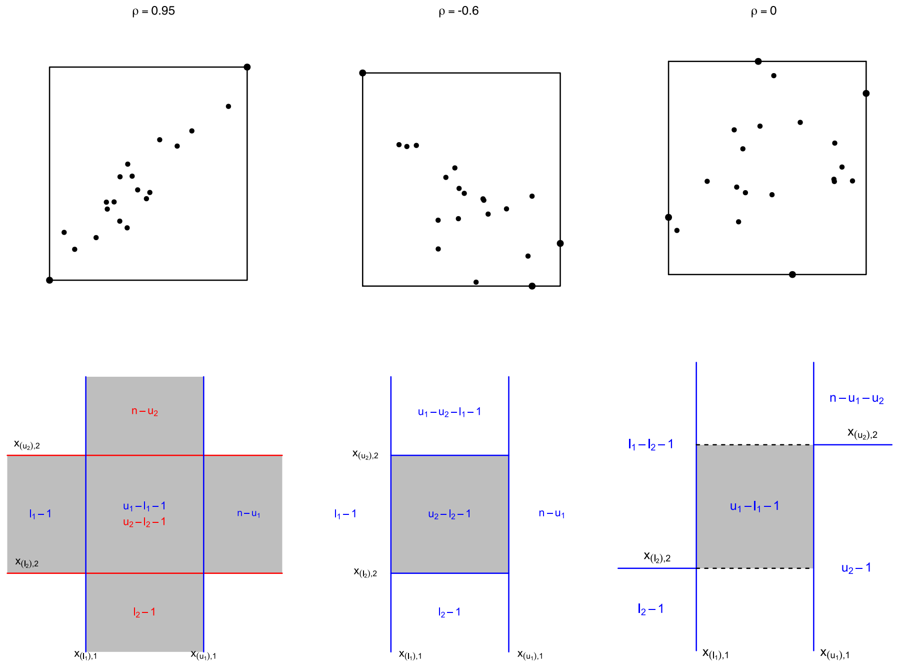

Fig. 1 Construction methods for bivariate intervals using marginal minima/maxima (top panels) or marginal order statistics (bottom) Top panels: Illustrative random rectangles constructed from 2 points (high correlation), 3 points (moderate correlation) and 4 points (low/no correlation). Bottom panels: Three alternative construction methods: marginal only (left panel), sequential nesting (centre; equation (9)) and iterative segmentation (right; equation (11)). Values in blue (red) denote the number of observations in the area bounded by blue (red) lines (colour figure online)

$$
\begin{aligned}
S = \pi(\boldsymbol{X}) &: \mathbb{R}^{d\times N} \to \mathcal{S} \\
&= \{(a_1, a_2) \in \mathbb{R}^2 : a_1 \le a_2\}^d \times \{2, \ldots, \min(2d, n)\} \times \mathcal{T} \times \mathbb{N}
\end{aligned} \tag{3}
$$

so that $\boldsymbol{x} \mapsto \left((x_{(1),i}, x_{(n),i})_{i=1,\ldots,d}, p, I(p), N\right)$, where $\boldsymbol{x} = (x_1, \ldots, x_n)$, $x_j = (x_{j,1}, \ldots, x_{j,d})^\top$ and $x_{(k),i}$ is the k-th order statistic of the i-th marginal component of $\boldsymbol{x}$. The quantity p represents the number of unique points involved in constructing the rectangle. The quantity $I(p)$ contains those points that reside in any marginal vertex (in 2 dimensions or higher) in the d-dimensional rectangle. (A 'marginal vertex' is a vertex of a lower dimensional marginal rectangle.) In this context a symbol is written as $S = (S_{\min}, S_{\max}, S_p, S_{I_p}, N)$, where $S_{\min}$ and $S_{\max}$ are respectively the d-vectors corresponding to the marginal minima and maxima.

Lemma 2 *Consider a multivariate random rectangle $S \in \mathcal{S}$, obtained through (3) and assume that $g_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \prod_{i=1}^{n} g_X(x_i;\theta)$, $\boldsymbol{x} \in \mathbb{R}^{n\times d}$, where g_X is a continuous density function with unbounded support. Then the symbolic likelihood function is given by*

$$L(s;\theta) = \frac{n!}{(n-s_p)!}\left[\int_{s_{\min}}^{s_{\max}} g_X(z;\theta)\mathrm{d}z\right]^{n-s_p} \times \ell_{s_p}, \tag{4}$$

where the multivariate integral is taken over the rectangular region defined by $s_{\min}$ *and* $s_{\max}$*, and where* ℓ_{s_p} *is defined as follows. If* $s_p = 2$ *then* $s_{I_p} = (s_a, s_b)$ *is the two co-ordinates of* d*-dimensional space which define the bounding rectangle, and* $\ell_2 = g_X(s_a;\theta)g_X(s_b;\theta)$*. If* $s_p = 2d$ *then* $s_{I_p} = \emptyset$ *and*

$$\begin{aligned}\ell_{2d} = &\prod_{i=1}^{d}\left[G_{X_{-i}|X_i=s_{\min,i}}(s_{\max,-i};\theta) - G_{X_{-i}|X_i=s_{\min,i}}(s_{\min,-i};\theta)\right] g_{X_i}(s_{\min,i}) \\ &\times \prod_{i=1}^{d}\left[G_{X_{-i}|X_i=s_{\max,i}}(s_{\max,-i};\theta) - G_{X_{-i}|X_i=s_{\max,i}}(s_{\min,-i};\theta)\right] g_{X_i}(s_{\max,i}),\end{aligned} \tag{5}$$

where X_i *is the* i*-th component of* X*,* $X_{-i} = X\backslash X_i$ *and similarly for* $s_{\min,-i}$*,* $s_{\max,-i}$*,* $s_{\min,i}$ *and* $s_{\max,i}$*, and* $G_{X_{-i}|X_i}$ *is the conditional distribution function of* X_{-i} *given* X_i*.*

In (5) the product terms represent the joint distributions of X_{-i} being between $s_{\min,-i}$ and $s_{\max,-i}$ given that X_i is equal to $s_{\min,i}$ or $s_{\max,i}$. When $s_p = 2$, (5) reduces to $\ell_{2d} = \ell_2$. General expressions for ℓ_{s_p} for $p \neq 2$ or $2d$ can be complex. Simple expressions are available for $s_p = 3$ when $d = 2$.

Corollary 1 *For a bivariate random rectangle, if* $s_p = 3$ *then* $S_{I_p} = s_c \in \mathbb{R}^2$ *is the co-ordinate of the point defining the bottom-left, top-left, top-right or bottom-right corner of the rectangle.*

In this case, if $\bar{s}_c$ *is the element-wise complement of* s_c*, i.e.* $\bar{s}_{c,i} = \{s_{\min,i}, s_{\max,i}\}\backslash\{s_{c,i}\}$*,* $i = 1, 2$*, then*

$$\begin{aligned}\ell_3 = g_X(s_c;\theta) \times &\prod_{i=1}^{2}\left[G_{X_{-i}|X_i=\bar{s}_{c,i}}(s_{\max,-i};\theta) - G_{X_{-i}|X_i=\bar{s}_{c,i}}(s_{\min,-i};\theta)\right] \\ &\times g_{X_i}(\bar{s}_{c,i};\theta).\end{aligned} \tag{6}$$

E.g. if $s_c = (s_{\min,1}, s_{\min,2})$ *is in the bottom-left corner, then* $\bar{s}_c = (s_{\max,1}, s_{\max,2})$*.*

The first term in (6) is the density of the point in the corner of the rectangle, and the other terms are the probabilities of the two points on the edges being between two interval values given that the other component is fixed. Qualitatively similar expressions can be derived for d-dimensional random rectangles in the cases where $s_p \neq 2$ or $2d$, although there is no simple general expression.

2.3.2 Using marginal order statistics

As order statistics are defined in the univariate setting, there are a number of methods to use fixed vectors of lower $l = (l_1, \ldots, l_d)^\top$ and upper $u = (u_1, \ldots, u_d)^\top$ order statistic values, with $1 \leq l_i < u_i \leq N$, to define a d-dimensional random rectangle. The simplest takes the cross product of the d-univariate marginal quantiles as suggested by Neto et al. (2011), where the authors indicate that the lower and upper values can take any pair of interval feature possible. Here the aggregation function π is

$$S = \pi(\boldsymbol{X}) : \mathbb{R}^{d \times N} \rightarrow \mathcal{S} = \{(a_1, a_2) \in \mathbb{R}^2 : a_1 \leq a_2\}^d \times \mathbb{N} \tag{7}$$

$$\boldsymbol{x} \mapsto \left(\left(x_{(l_i),i}, x_{(u_i),i} \right)_{i=1,\ldots,d}, N \right). \tag{8}$$

In this context the symbol is written as $S = (S_l, S_u, N)$, where S_l and S_u are respectively the d-vectors corresponding to the marginal lower and upper order statistics. This process is illustrated in Fig. 1 (bottom left panel) in the $d = 2$ setting. For fixed l and u, the observed counts in each region are then known as a function of the construction (8). The resulting symbolic likelihood function is then

$$L(s; \theta) = \prod_{i=1}^{d} L(s_{l_i}, s_{u_i}, n; \theta_i)$$

where $L(s_{l_i}, s_{u_i}, n; \theta_i)$ is as obtained in Lemma 1 using the i-th marginal distribution with parameter $\theta_i \in \Theta$. However, as the construction (8) only contains marginal information, such a symbol will fail to adequately capture dependence between variables. As an alternative, we introduce two new order-statistic based representations of random rectangles that do account for such dependence.

The first, *sequential nesting* (Fig. 1, bottom centre panel), constructs the order statistics within dimension i conditionally on already being within the random rectangle in dimensions $j < i$. The aggregation function π is given by (7) as before, but where now

$$\boldsymbol{x} \mapsto \left(\left(\left(x_{(l_i),i}, x_{(u_i),i} \right) | \{ x_{(l_j),j} < x_j < x_{(u_j),j}; j < i \} \right)_{i=1,\ldots,d}, N \right). \tag{9}$$

As before, $S = (S_l, S_u, N)$, but where the known observed counts now lie in different regions (Fig. 1), and with the additional constraints of $2 \leq u_{i+1} \leq u_i - l_i - 1$.

Lemma 3 *Consider a multivariate random rectangle $S \in \mathcal{S}$, constructed via (9) and suppose that $g_{\boldsymbol{X}}(\boldsymbol{x}; \theta) = \prod_{i=1}^{n} g_X(x_i; \theta)$, $\boldsymbol{x} \in \mathbb{R}^{n \times d}$, where g_X is a continuous density function with unbounded support. The symbolic likelihood function is then given by*

$$L(s; \theta) \propto \mathbb{P}(s_{\boldsymbol{l}} < X < s_{\boldsymbol{u}})^{u_d - l_d - 1} \mathrm{d}\mathbb{P}(X_1 < s_{l,1}) \mathrm{d}\mathbb{P}(X_1 < s_{u,1}) \prod_{i=1}^{d} p_i(s_{\boldsymbol{l}}) q_i(s_{\boldsymbol{u}}), \tag{10}$$

where $p_1(s_{\boldsymbol{l}}) = \mathbb{P}(X_1 < s_{l,1})^{l_1 - 1}$, $q_1(s_{\boldsymbol{u}}) = \mathbb{P}(X_1 > s_{u,1})^{n - u_1}$ and for $i = 2, \ldots, d$,

$$p_i(s_{\boldsymbol{l}}) = \mathbb{P}(s_{l,j} < X_j < s_{u,j}; j < i | X_i = s_{l,i}) \mathrm{d}\mathbb{P}(X_i < s_{l,i})$$
$$\times \mathbb{P}(X_i < s_{l,i} | s_{l,j} < X_j < s_{u,j}; j < i)^{l_i - 1}$$

$$\begin{aligned} q_i(s_{\boldsymbol{u}}) = &\ \mathbb{P}(s_{l,j} < X_j < s_{u,j};\ j < i | X_i = s_{u,i}) \mathrm{d}\mathbb{P}(X_i < s_{u,i}) \\ &\times \mathbb{P}(X_i > s_{u,i} | s_{l,j} < X_j < s_{u,j};\ j < i)^{u_{i-1}-u_i-l_{i-1}-1}. \end{aligned}$$

Corollary 2 *With $d = 2$, the symbolic likelihood function in Lemma 3 is given by*

$$\begin{aligned} L(s;\theta) \propto &\left(G_X(s_{\boldsymbol{u}}) - G_X(s_{\boldsymbol{l}})\right)^{u_2-l_2-1} g_{X_1}(s_{l,1}) g_{X_1}(s_{u,1}) g_{X_2}(s_{l,2}) g_{X_2}(s_{u,2}) \\ &\times G_{X_1}(s_{l,1})^{l_1-1}\left[1 - G_{X_1}(s_{u,1})\right]^{n-u_1}\left[G_{X_1|X_2=s_{l,2}}(s_{u,1}) - G_{X_1|X_2=s_{l,2}}(s_{l,1})\right] \\ &\times \left[G_{X_1|X_2=s_{u,2}}(s_{u,1}) - G_{X_1|X_2=s_{u,2}}(s_{l,1})\right]\left[G_X((s_{u,1}, s_{l,2})) - G_X(s_{\boldsymbol{l}})\right]^{l_2-1} \\ &\times \left[G_{X_1}(s_{u,1}) - G_X(s_{\boldsymbol{u}}) - G_{X_1}(s_{l,1}) + G_X((s_{l,1}, s_{u,2}))\right]^{u_1-u_2-l_1-1}, \end{aligned}$$

where $G_{X_i}(\cdot) \equiv G_{X_i}(\,\cdot\,;\theta)$ and $G_{X_i|X_j}(\cdot) \equiv G_{X_i|X_j}(\,\cdot\,;\theta)$; $i \neq j$ respectively denote the marginal and conditional distribution functions of $g_{\boldsymbol{X}}(\boldsymbol{x};\theta)$.

An alternative to sequential nesting is an *iterative segmentation* construction (Fig. 1, bottom right). As before, for fixed vectors l and u, the aggregation function π is given by (7) but where

$$\boldsymbol{x} \mapsto \Big(\big(x_{(l_i),i} | \{x_j < x_{(l_j),j};\ j < i\}, x_{(u_i),i} | \{x_j > x_{(u_j),j};\ j < i\}\big)_{i=1,\ldots,d}, N\Big). \quad (11)$$

Again $S = (S_l, S_u, N)$, but now where $S_{l,i}$, the l_i-th order statistic of the i-th margin, is restricted to the area where the previous margins $j < i$ are all below their respective lower (l_j-th) order statistic. Similarly, $S_{u,i}$ is restricted to the area where the previous margins $j < i$ are all above their respective upper order statistic. For fixed l and u the observed counts are then known (Fig. 1, bottom right) but are attributed to different regions than for sequential nesting. Iterative segmentation implies the additional constraints $l_{i+1} < l_i - 1$ and $u_{i+1} < N - \sum_{j=1}^{i} u_j$ for $i = 1, \ldots, d-1$.

Lemma 4 *Consider a multivariate random rectangle $S \in \mathcal{S}$, constructed via (11) and suppose that $g_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \prod_{i=1}^{n} g_X(x_i;\theta)$, $\boldsymbol{x} \in \mathbb{R}^{n\times d}$, where g_X is a continuous density function with unbounded support. The symbolic likelihood function is then given by*

$$L(s;\theta) \propto \mathbb{P}(s_{l,1} < X_1 < s_{u,1})^{u_1-l_1-1} \mathrm{d}\mathbb{P}(X_1 < s_{l,1}) \mathrm{d}\mathbb{P}(X_1 < s_{u,1}) \prod_{i=2}^{d+1} p_i(s_{\boldsymbol{l}}) q_i(s_{\boldsymbol{u}}), \quad (12)$$

where $p_{d+1}(s_{\boldsymbol{l}}) = \mathbb{P}(X_1 < s_{l,1}, \ldots, X_d < s_{l,d})^{l_d-1}$, $q_{d+1}(s_{\boldsymbol{u}}) = \mathbb{P}(X_1 > s_{u,1}, \ldots, X_d > s_{u,d})^{n-\sum_{i=1}^{d} u_i}$ and for $i = 2, \ldots, d$

$$\begin{aligned} p_i(s_{\boldsymbol{l}}) = &\ \mathbb{P}(X_j < s_{l,j};\ j < i | X_i = s_{l,i}) \mathrm{d}\mathbb{P}(X_i < s_{l,i}) \\ &\times \left[\mathbb{P}(X_j < s_{l,j};\ j < i) - \mathbb{P}(X_j < s_{l,j};\ j \leq i)\right]^{l_i-l_{i-1}-1} \end{aligned}$$

$$q_i(s_{\boldsymbol{u}}) = \mathbb{P}(X_j > s_{u,j};\, j < i | X_i = s_{u,i}) \mathrm{d}\mathbb{P}(X_i < s_{u,i}) \\ \times \left[\mathbb{P}(X_j > s_{u,j};\, j < i) - \mathbb{P}(X_j > s_{u,j};\, j \leq i)\right]^{u_i - 1}.$$

Corollary 3 *With $d = 2$, the symbolic likelihood function in Lemma 4 is given by*

$$\begin{aligned} L(s;\theta) \propto &\left(G_{X_1}(s_{u,1}) - G_{X_1}(s_{l,1})\right)^{u_1 - l_1 - 1} g_{X_1}(s_{l,1}) g_{X_1}(s_{u,1}) g_{X_2}(s_{l,2}) g_{X_2}(s_{u,2}) \\ &\times G_{X_1|X_2 = s_{l,2}}(s_{l,1})(1 - G_{X_1|X_2=s_{u,2}}(s_{u,1})) \left[G_{X_1}(s_{l,1}) - G_{\boldsymbol{X}}(s_{\boldsymbol{l}})\right]^{l_2 - l_1 - 1} \\ &\times \left[G_{X_2}(s_{u,2}) - G_{\boldsymbol{X}}(s_{\boldsymbol{u}})\right]^{u_2 - 1} G_{\boldsymbol{X}}(s_{\boldsymbol{l}})^{l_2 - 1} \\ &\times \left(1 - G_{X_1}(s_{u,1}) - G_{X_2}(s_{u,2}) - G_{\boldsymbol{X}}(s_{\boldsymbol{u}})\right)^{n - u_1 - u_2}, \end{aligned}$$

where $G_{X_i}(\cdot) \equiv G_{X_i}(\,\cdot\,;\theta)$ and $G_{X_i|X_j}(\cdot) \equiv G_{X_i|X_j}(\,\cdot\,;\theta)$; $i \neq j$ respectively denote the marginal and conditional distribution functions of $g_{\boldsymbol{X}}(\boldsymbol{x};\theta)$.

When $l_1 = \cdots = l_d = 1$ and $u_i = n - 2(i-1)$ (so that the marginal minima and maxima are selected), the sequential nesting random interval construction (9) approximately reduces to the rectangle construction (3) based on univariate marginal maxima and minima, indicating some degree of construction consistency. That is, $S = (S_l, S_u, N)$ contains nearly the same information as the symbol $S = (S_{\min}, S_{\max}, S_p, S_{I_p}, N)$ when $S_p = 2d$, and so the symbolic likelihood function (10) approximately reduces to (4). For highly correlated data $S = (S_l, S_u, N)$ is slightly more informative as the lower and upper bounds of each dimension i are calculated on a set from which the $(i-1)$ lowest and largest observations are removed. The approximation improves as the correlation decreases until both symbols become identical when the random variables are completely independent. A similar reduction cannot be obtained for the iterative segmentation construction.

2.4 Modelling histograms with random counts

Histograms are a popular and typically univariate SDA tool to represent the distribution of continuous data. They are commonly constructed as a set of fixed consecutive intervals for which random relative frequencies (or counts) are reported (e.g. Bock and Diday 2000; Billard and Diday 2006). Following Le-Rademacher and Billard (2011), a (d-dimensional) histogram-valued random variable may be defined as a set of counts associated with a deterministic partition of the domain $\mathcal{X} = \mathbb{R}^d$. Suppose that the i-th margin of $\mathcal{X}$ is partitioned into B^i bins, so that $B^1 \times \cdots \times B^d$ bins are created in $\mathcal{X}$ through the d-dimensional intersections of each marginal bin. Index each bin by $\boldsymbol{b} = (b_1, \ldots, b_d)$, $b_j = 1, \ldots, B^j$ as the vector of co-ordinates of each bin in the histogram. Each bin $\boldsymbol{b}$ may then be constructed as

$$\mathcal{B}_{\boldsymbol{b}} = \mathcal{B}^1_{b_1} \times \cdots \times \mathcal{B}^d_{b_d} \quad \text{where} \quad \mathcal{B}^j_{b_j} = (y^j_{b_j - 1}, y^j_{b_j}],\ j = 1, \ldots, d,$$

where for each j, the marginal sequences $-\infty < y^j_0 < y^j_1 < \cdots < y^j_{B^j} < \infty$ are fixed. We assume that all data counts outside of the constructed histogram are zero. A d-dimensional histogram-valued random variable is constructed through the aggregation function π where

$$\begin{aligned} S = \pi(\boldsymbol{X}) : \mathbb{R}^{d\times N} &\to \mathcal{S} = \{0, \ldots, N\}^{B^1\times\cdots\times B^d} \\ \boldsymbol{x} &\mapsto \left(\sum_{i=1}^n \mathbb{1}\{x_i \in \mathcal{B}_{\mathbf{1}}\}, \ldots, \sum_{i=1}^n \mathbb{1}\{x_i \in \mathcal{B}_{\boldsymbol{B}}\}\right), \end{aligned} \tag{13}$$

where $\mathbf{1} = (1, \ldots, 1)$ and $\boldsymbol{B} = (B^1, \ldots, B^d)$, and $\mathbb{1}$ is the indicator function. The symbol $S = (S_{\mathbf{1}}, \ldots, S_{\boldsymbol{B}})$ is a vector of counts, $\sum_{\boldsymbol{b}} S_{\boldsymbol{b}} = N$, where $S_{\boldsymbol{b}}$ denotes the frequency of data in bin $\mathcal{B}_{\boldsymbol{b}}$.

Lemma 5 *Consider a multivariate histogram-valued random variable $S \in \mathcal{S}$, constructed via (13) and suppose that $g_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \prod_{i=1}^n g_X(x_i;\theta)$, $\boldsymbol{x} \in \mathbb{R}^{n\times d}$, where g_X is a continuous density function with unbounded support over the region defined by the histogram bins. The symbolic likelihood function is given by*

$$L(s;\theta) = \frac{n!}{s_{\mathbf{1}}!\cdots s_{\boldsymbol{B}}!}\prod_{\boldsymbol{b}}\left(\int_{\mathcal{B}_b} g_X(z;\theta)\mathrm{d}z\right)^{s_b}, \tag{14}$$

where the integral denotes the probability that data $x \in \mathcal{X}$ falls in bin $\mathcal{B}_{\boldsymbol{b}}$ under the model.

In the univariate setting, this multinomial likelihood coincides with the likelihood function for binned and truncated data introduced by McLachlan and Jones (1988). It also extends the method of Heitjan and Rubin (1991) who build corrected likelihood functions for coarsened data, where the authors highlight the need to account for both the grouping and the stochastic nature of the coarsening.

In the limit as the histogram is reduced to its underlying classical data, the likelihood (14) reduces to the classical data likelihood. As the number of bins becomes large each bin of the histogram reduces in size and approaches a single point $\mathcal{B}_{\boldsymbol{b}} \to x^*_{\boldsymbol{b}} = (x^*_{b_1}, \ldots, x^*_{b_d}) \in \mathbb{R}^d$. More precisely, this means that, as the number of bins gets large, for each marginal component j, $j = 1, \ldots, d$, the lower bound $y^j_{b_j-1}$ of $\mathcal{B}^j_{b_j}$ approaches $x^*_{b_j}$ from below while the upper bound $y^j_{b_j}$ approaches $x^*_{b_j}$ from above. In the limit as the number of bins $\to \infty$, only those n bins for which $x^*_{\boldsymbol{b}}$ coincides with the underlying micro data will have a count of 1, while the others will have a count of 0 removing their contribution to the symbolic likelihood function.

Now, since the density g_X is assumed continuous we can use the fact that

$$\frac{1}{|\mathcal{B}_{\boldsymbol{b}}|}\int_{\mathcal{B}_b} g_X(z;\theta)dz \to g_X(x_{\boldsymbol{b}}),$$

as the number of bins gets large, implying that the likelihood contribution of the non-empty bins $\mathcal{B}_{\boldsymbol{b}}$ is then proportional to $g_X(x_b;\theta)$. This is equivalent to specifying $f_{S|\boldsymbol{X}=\boldsymbol{z}}(s;\vartheta) = \prod_{i=1}^n \delta_{z_i}(x_i)$ in (1). Consequently $L(\boldsymbol{s};\theta) \propto \prod_{i=1}^n g_X(x_i;\theta)$ reduces to the classical data likelihood function.

Finally, note that when the classical data are only observed on a subset of the domain $\mathcal{X}$, $g_{\boldsymbol{X}}(\boldsymbol{x};\theta)$ should be truncated and rescaled over the same subdomain.

2.5 Modelling histograms with random bins

A common alternative to histograms with random counts over fixed bins is constructing histograms with fixed counts within random bins (e.g. Mousavi and Zaniolo 2011; Ioannidis 2013). Such random histograms can be seen as a generalisation of interval-valued random variables (Sects. 2.2–2.3). In particular, random intervals can be viewed as histograms with the number of bins ranging from 1 (all margins are intervals calculated from sample minima and maxima; Fig. 1, top) to $3d$ (all margins are intervals calculated from order statistics $l > 1$ and $u < n$; Fig. 1 bottom left). In the following we focus on the univariate setting ($\mathcal{X} = \mathbb{R}$) since extension to d-dimensions is challenging. E.g. given a matrix of counts, then a simply constructed grid matching these counts does not necessarily exist.

For a vector of orders $k = (k_1, \ldots, k_B)$, such that $1 \le k_1 \le \cdots \le k_B \le N$, a univariate random histogram is constructed through the aggregation function π where

$$\begin{aligned} S = \pi(\boldsymbol{X}) : \mathbb{R}^N &\to \mathcal{S} = \{(a_1, \ldots, a_B) \in \mathbb{R}^B : a_1 \le \cdots \le a_B\} \times \mathbb{N} \\ \boldsymbol{x} &\mapsto (x_{(k_1)}, \ldots, x_{(k_B)}, N). \end{aligned} \tag{15}$$

This defines a histogram with bin b located at $(s_{b-1}, s_b]$ with fixed count $k_b - k_{b-1}$, for $b = 1, \ldots, B + 1$, where $s_0 = -\infty$, $s_{B+1} = +\infty$, $k_0 = 0$ and $k_{B+1} = N + 1$, and knowledge that there is an observed data point located at each s_b, $b = 1, \ldots, B$. The symbol $S = ((S_1, \ldots, S_B), N)$ is a B-vector of order statistics, plus N.

Lemma 6 *Consider a univariate random histogram $S \in \mathcal{S}$, obtained through (15) and assume that $g_{\boldsymbol{X}}(\boldsymbol{x}; \theta) = \prod_{i=1}^n g_X(x_i)$, $\boldsymbol{x} \in \mathbb{R}^{n \times d}$. Then the symbolic likelihood function is given by*

$$L(s; \theta) = n! \prod_{b=1}^{B} g_X(s_b; \theta) \prod_{b=1}^{B+1} \frac{(G_X(s_b; \theta) - G_X(s_{b-1}; \theta))^{k_b - k_{b-1} - 1}}{(k_b - k_{b-1} - 1)!}. \tag{16}$$

When $B = 2$, $k_1 = l$ and $k_2 = u$ with $l, u = 1, \ldots, n$; $l < u$, then (16) reduces to the likelihood function in Lemma 1 (see Appendix A.4). Further, under this construction it is straightforward to show that if $B = N$ then the symbolic likelihood (16) recovers the classical data likelihood. Specifically this implies $k_b = b$ for all $b = 1, \ldots, B$ so that the aggregation function (15) is $S = \pi(\boldsymbol{X}) = ((X_{(1)}, \ldots, X_{(n)}), N)$, $k_b - k_{b-1} = 1$ for all b and so $L(s; \theta) \propto \prod_{b=1}^N g_X(x_b; \theta)$.

3 Illustrative analyses

Our symbolic likelihood function resolves many of the conceptual and practical issues with current SDA methods, opens the door for new classes of symbol design and construction, and positions SDA as a viable tool to enable and improve upon classical data analyses. In Sect. 3.1 we demonstrate that the proposed symbolic likelihood function is able to outperform bespoke statistical techniques for analysing datasets used in medical research. In Sect. 3.2 we explore the ability of random rectangles to contain information about the *dependence structure* between the data margins. We also propose two novel methods of constructing random rectangles that are able to incorporate this information, and favourably compare these to standard SDA random rectangle construction techniques which can only provide marginal information. Finally in Sect. 3.3 we provide a direct comparison between the developed methods and the likelihood-based approach of Le-Rademacher and Billard (2011). Here we demonstrate that, as might be expected, when one is willing to specify a model at the level of the micro-data, improved statistical performance can be achieved in comparison to methods that only specify models at the level of the symbol.

3.1 Effect reconstruction for meta-analyses

In medical research, meta-analyses are often implemented to systematically examine the clinical effects of certain treatments, and typically use the effect sample mean and standard deviation from the dataset in each individual study. However it is common practice that such studies only report certain quantile statistics, namely the sample minimum (q_0), maximum (q_4) and the sample quartiles (q_1, q_2, q_3), rather than the dataset mean and standard deviation required to perform the meta-analysis. As a result, we have the problem of trying to estimate the sample mean and standard deviation of a dataset from observed quantiles.

The most sophisticated practiced method to estimate the sample mean was developed by Luo et al. (2018) based on previous work by Hozo et al. (2005) and Wan et al. (2014), whereby

$$\hat{\hat{x}}_L = w_1 \left(\frac{q_0 + q_4}{2} \right) + w_2 \left(\frac{q_1 + q_3}{2} \right) + (1 - w_1 - w_2) q_2, \tag{17}$$

with $w_1 = 2.2/(2.2 + n^{0.75})$ and $w_2 = 0.7 - 0.72/n^{0.55}$. Based on previous work by Hozo et al. (2005) and Bland (2015) the best performing estimators of the sample standard deviation are due to Wan et al. (2014) and Shi et al. (2018), which are respectively given by

$$\hat{s}_W = \frac{1}{2} \left(\frac{q_4 - q_0}{\zeta(n)} + \frac{q_3 - q_1}{\eta(n)} \right) \quad \text{and} \quad \hat{s}_S = \frac{q_4 - q_0}{\theta_1(n)} + \frac{q_3 - q_1}{\theta_2(n)}, \tag{18}$$

where $\zeta(n) = 2\Phi^{-1}\left(\frac{n-0.375}{n+0.25}\right)$, $\eta(n) = 2\Phi^{-1}\left(\frac{0.75n-0.125}{n+0.25}\right)$, $\theta_1(n) = (2 + 0.14n^{0.6})\Phi^{-1}(\frac{n-0.375}{n+0.25})$, $\theta_2(n) = (2 + \frac{2}{0.07n^{0.6}})\Phi^{-1}(\frac{0.75n-0.125}{n+0.25})$, and $\Phi^{-1}(\cdot)$ is the inverse of the standard normal c.d.f. Each estimator in (17) and (18) assumes the underlying data are normally distributed.

In the context of the symbolic random variables developed in Sect. 2, this setting corresponds to constructing the symbolic variable S defined through (15) with $n = 4Q + 1$, $Q \in \mathbb{N}$ where $k = (1, Q + 1, 2Q + 1, 3Q + 1, n)$ i.e. a histogram with $B = 4$ random bins and equal counts. If we make the same assumption of i.i.d. normality of the underlying data, then maximising the symbolic likelihood (16) with $g_X(x; \theta) = \phi(x; \mu, \sigma)$ will yield maximum likelihood estimators $\hat{\theta} = (\hat{\mu}, \hat{\sigma}) \approx$

$(\bar{x}, \sqrt{(n-1)/ns})$ which provide direct estimates $(\hat{\bar{x}}_*, \hat{s}_*) = (\hat{\mu}, \sqrt{n/(n-1)}\hat{\sigma})$ of the sample mean $\bar{x}$ and standard deviation s of the underlying data. Of course, other distributional assumptions can easily be made.

Figure 2 illustrates the performance of each estimator compared to the true sample values (i.e. $(\hat{\bar{x}} - \bar{x}_0)$ and $(\hat{s} - s_0)$) based on data generated from normal (top panels) and lognormal (bottom) distributions, averaged over 10,000 replicates, and for a range of sample sizes n. For normal data, the sample mean estimator $\hat{\bar{x}}_L$ by Luo et al. (2018) (red) and the symbolic likelihood-based estimator (green) perform comparably (top left). Identifying performance differences of the sample standard deviation estimators is much clearer (top right), with the symbolic estimator strongly outperforming the

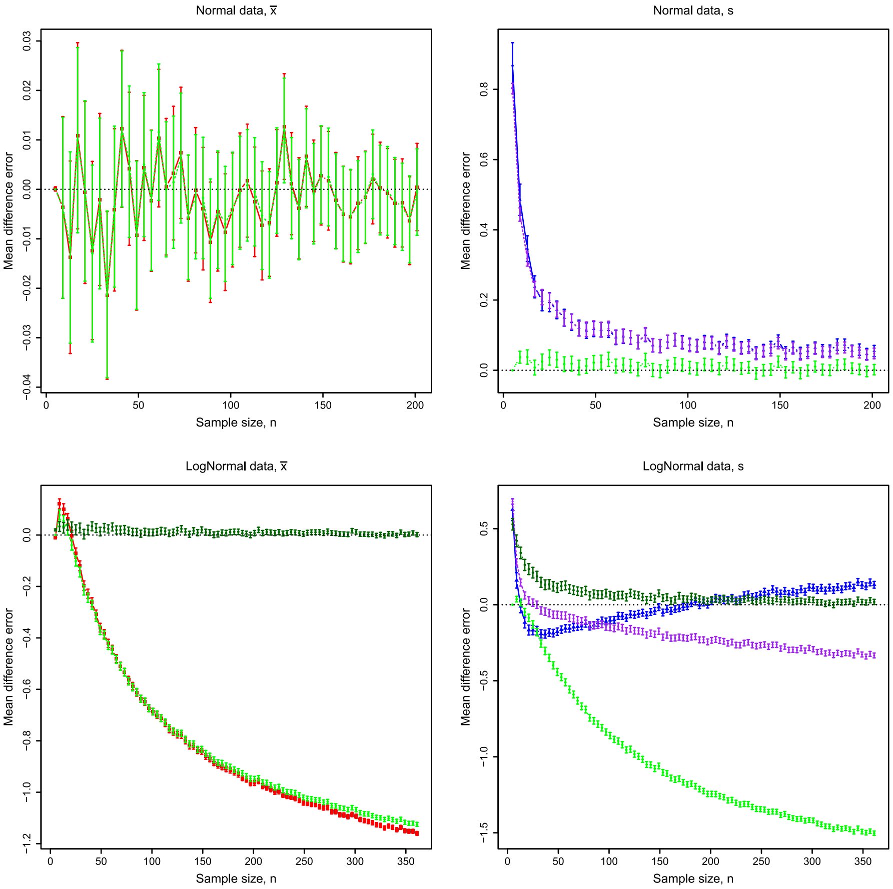

Fig. 2 Mean difference errors, $(\hat{\bar{x}} - \bar{x}_0)$ and $(\hat{s} - s_0)$, of various estimates of the sample mean (left panels) and standard deviation (right) as a function of sample size $n = 4Q + 1$, $Q = 1, \ldots, 50$ or 90, for both normally (top panels) and log-normally (bottom) distributed data. $\bar{x}_0$ and s_0 denote the true sample mean and standard deviation for each dataset. Errors are averaged over $T = 10{,}000$ dataset replicates generated from $\theta_0 = (\mu_0, \sigma_0) = (50, 17)$ (normal data) and $\theta_0 = (\mu_0, \sigma_0) = (4, 0.3)$ following Hozo et al. (2005) and Luo et al. (2018). Colouring indicates the SDA estimates (light and dark green), $\hat{\bar{x}}_L$ (red), $\hat{s}_W$ (blue) and $\hat{s}_S$ (purple). Confidence intervals indicate ±1.96 standard errors (colour figure online)

discipline-standard estimators of Wan et al. (2014) and Shi et al. (2018) (blue and purple, respectively). The differences are particularly stark for low n. As $\hat{s}_W$ and $\hat{s}_S$ substantially overestimate the true standard deviation, their usage will systematically undervalue the contribution of each study in any larger analysis, potentially weakening the power of the meta-analysis to detect significant clinical effects. Note that for $n = 5$, the symbolic estimator of the sample standard deviation is exact (i.e. zero error) as the symbolic likelihood (16) reduces to the classical likelihood in this case.

When the sample data are lognormal (bottom panels), both symbolic (light green) and the industry-standard estimators perform poorly. This is unsurprising given the common normality assumption. While estimators equivalent to those in (17) and (18) but for lognormally distributed data could in principle be derived, it is trivial to achieve this for the symbolic estimator by substituting the lognormal density (or any other distribution) for $g_X(\cdot\,;\theta)$ in (16). The resulting sample mean and standard deviation estimators assuming the lognormal distribution are illustrated in dark green. The lognormal-based symbolic likelihood estimator performance is clearly excellent in comparison.

One factor influencing the efficiency of the symbolic maximum likelihood estimate (MLE) is the form and specification of the symbol as a summary representation of the underlying data. While a histogram with more bins should be more informative than one with less, for a fixed number of bins, sensible choice of location can result in increased MLE performance. This idea of *symbol design* has been largely ignored in the SDA literature e.g. with random intervals routinely constructed from sample minima and maxima.

Consider the simplified setting of the univariate random interval $S = (s_l, s_u, n)$ defined in Lemma 1 constructed using symmetric upper and lower order statistics, and the associated 2-bin random histogram (15) that results by additionally including the sample median, q_2. I.e. for sample sizes $n = 4Q + 1$, $Q \in \mathbb{N}$ we have $l = i, u = n+1-i$ for the interval and $k = (i, 2Q+1, n+1-i)$ for the histogram. We examine the efficiency of the symbolic MLE for the symbols defined by $i = 1, \ldots, 2Q$. For each of $t = 1, \ldots, T = 10{,}000$ replicate datasets of size $n = 21$, 81 and 201 (i.e. $Q = 5, 20, 50$) drawn from a $N(\mu_0, \sigma_0)$ distribution with $(\mu_0, \sigma_0) = (50, 17)$, we compute the rescaled symbolic MLE $(\hat{\mu}_t, \tilde{\sigma}_t)$ where $\tilde{\sigma}_t = \sqrt{n/(n-1)}\hat{\sigma}_t$, and calculate the relative mean square errors (RMSE) defined by

$$\text{RMSE}_{\hat{\mu}} = \frac{\sum_{t=1}^{T}(\hat{\mu}_t - \mu_0)^2}{\sum_{t=1}^{T}(\bar{x}_t - \mu_0)^2} \quad \text{and} \quad \text{RMSE}_{\tilde{\sigma}} = \frac{\sum_{t=1}^{T}(\tilde{\sigma}_t - \sigma_0)^2}{\sum_{t=1}^{T}(s_t - \sigma_0)^2},$$

where $\bar{x}_t$ and s_t denote the sample mean and standard deviation of the t-th replicate.

Figure 3 shows the RMSEs as function of the quantile $q = (n + 1 - i)/n$ used to construct the symbol. As expected, using a histogram (dark lines) provides more information about μ than the associated random interval (grey), as the extra information contained in the median is informative for this parameter. In contrast, the median provides no information about σ in addition to the two bounding quantiles, for the normal distribution. Including alternative quantiles would be informative.

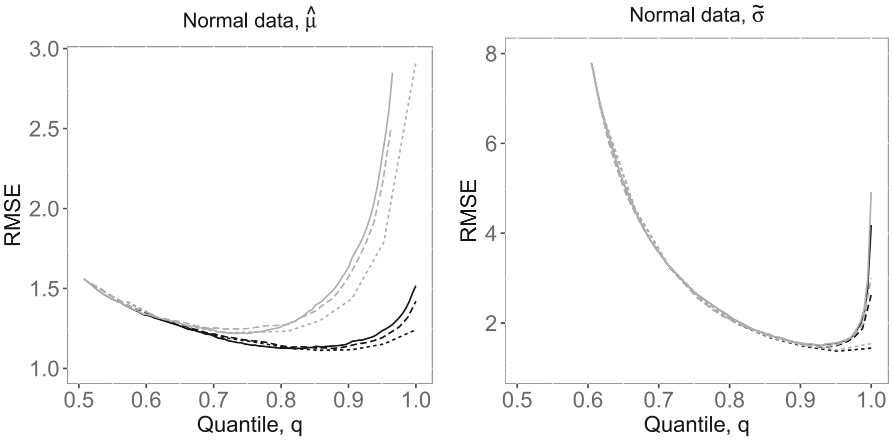

Fig. 3 $\mathrm{RMSE}_{\hat{\mu}}$ (left) and $\mathrm{RMSE}_{\tilde{\sigma}}$ (right) as a function of quantile $q = (n+1-i)/n$ for $i = 1, \ldots, (n+1)/2$. Grey and black lines respectively denote random intervals and histograms. Short-dashed, long-dashed and solid lines indicate samples of size $n = 21$, 81 and 201 respectively

The convex shape of each RMSE curve indicates that the prevailing SDA practice of constructing intervals from sample minima and maxima ($i = 1, q = n$) is highly inefficient for parameter estimation. Greater precision for both location and scale parameters is achieved by using less extreme quantiles, in this setting around the $q = 0.85$-0.90 range (balancing optimal minimum RMSE values between the two parameters). There is also a severe penalty for using too low quantiles when estimating σ, as the data scale is not easily estimated using overly central quantities. Estimating μ is less sensitive in this regard. These conclusions are robust to sample size, n. Overall this analysis indicates that substantial efficiency gains should be possible in standard SDA with more informed symbol design.

3.2 Information content in multivariate random rectangles

In Sects. 2.3.1 and 2.3.2 we introduced two new symbolic constructions to increase the information content within multivariate random rectangles. We now examine the performance of each of these representations and contrast them with standard SDA constructions. While we focus on bivariate intervals for clarity, extension of the results to higher dimensions is immediate.

When constructing random rectangles from marginal minima and maxima, Lemma 2 and Corollary 1 provide an expression for the symbolic likelihood that incorporates full knowledge of the number and location of unique points from which the interval is constructed (e.g. Fig. 1). We denote the resulting likelihood function (4) by $L_{\mathrm{full}}(s; \theta)$. Existing SDA definitions of random rectangles do not use this information. In its absence, the best likelihood model that can be constructed is by averaging the likelihood L_{full} over all possible combinations of the unique point constructions, weighted according to the probability of that configuration arising under the classical data model. That is,

$$L_{\emptyset}(s;\theta) = \sum_{t_p}\sum_{t_{I_p}} L_{\text{full}}((s_{\min}, s_{\max}, t_p, t_{I_p}, n);\theta)\mathbb{P}(S_p = t_p, S_{I_p} = t_{I_p};\theta),$$

where

$$\mathbb{P}(S_p = t_p, S_{I_p} = t_{I_p};\theta) = \int\int L_{\text{full}}((a, b, t_p, t_{I_p}, n);\theta)\prod_{i=1}^{d} I(a_i \leq b_i)\,\mathrm{d}a\mathrm{d}b, \tag{19}$$

where $a = (a_1, \dots, a_d)$ and $b = (b_1, \dots, b_d)$. While not generally viable, below we estimate the probabilities (19) to high accuracy using Monte Carlo with a large number of samples, each time $L_{\emptyset}$ is evaluated. One alternative is to assume each random rectangle is constructed by the maximum number of unique points ($2d$), which is perhaps realistic when the number of points n_c underlying a symbol is large and the dependence between the variables not too strong. We denote the particular case of L_{full} with $S_p = 2d$ as $L_{2d}(s;\theta)$. Here, L_{2d} effectively represents the current state-of-the-art in SDA methods, $L_{\emptyset}$ represents the best that can likely be done with the existing constructions of random rectangles in the SDA literature (although it is likely impractical), and L_{full} is our construction.

We assume $m = 20, 50$ classes, for each of which a random sample of size $n_c = 5, 10, 50, 100$ is drawn from a $N_2(\mu_0, \Sigma_0)$ distribution ($d = 2$) with $\mu_0 = (2, 5)^\top$, $\text{diag}(\Sigma_0) = (\sigma^2_{0,1}, \sigma^2_{0,2}) = (0.5, 0.5)$ and correlation $\rho_0 = 0, 0.3, 0.5, 0.7, 0.9$. The m random rectangles are then constructed, retaining the information (s_p, s_{I_p}) required to maximise L_{full} but which is ignored when maximising $L_{\emptyset}$ and L_4. For each of $T = 100$ replicate datasets, the symbolic MLE $\hat{\theta} = (\hat{\mu}, \hat{\Sigma})$ is computed.

Table 1 reports the mean and standard deviation of $\hat{\rho}$ over the replicate datasets under each likelihood. The marginal parameters (μ, σ_1 and σ_2) are well estimated in each case (see Supplementary Material B.1). The main conclusion from Table 1 is that only L_{full}, which incorporates full information of the number and location of the unique points that define the random rectangle, is able to accurately estimate dependence between the variables. For L_4 and $L_{\emptyset}$ the MLEs are either zero (no dependence can be estimated) or they are biased upwards. Note that for L_{full}, variability of the MLE mostly increases as n_c increases, and is more variable for lower correlation values. This can be explained as dependence information is contained in the proportion of rectangles constructed from 2 and 3 unique points (and their locations). For a fixed correlation, as n_c gets large it is increasingly likely that the rectangles will be generated by 4 unique points, thereby weakening the dependence information that the sample of random rectangles can contain. This weakening naturally occurs more slowly for higher correlations, and so the correlation MLE has greater accuracy and precision for stronger dependence. I.e. Depending on the number of points, n_c, in the random rectangle, the tendency of the correlation to be underestimated is lower when the strength of the correlation is higher. As $n_c \to \infty$ all rectangles will be generated from 4 unique points, and it will not be possible to accurately estimate within-rectangle dependence. This effect can be seen for $n_c = 1{,}000$ and 100,000 for $\rho = 0.3, 0.5, 0.7$ but not yet for $\rho = 0.9$.

Table 1 Mean (and standard deviation) of the symbolic maximum likelihood estimate of the correlation, ρ, over $T = 100$ replicate bivariate random rectangle datasets. The symbolic datasets vary in the number of symbols (m), the number of classical datapoints per symbol (n_c), and the strength of the correlation between the two variables (ρ_0). Estimates maximise the three symbolic likelihoods L_{full}, $L_{\emptyset}$ and L_4

		$m = 20$					$m = 50$				
n_c		5	10	100	1,000	100,000	5	10	100	1,000	100,000
$\rho_0 = 0.0$	L_4	0.004	−0.001	0.002	0.000	0.000	−0.004	−0.003	0.001	0.000	0.000
		(0.071)	(0.054)	(0.020)	(0.012)	(0.004)	(0.056)	(0.032)	(0.015)	(0.008)	(0.004)
	$L_{\emptyset}$	−0.018	−0.017	−0.006	−0.001	–[a]	−0.055	−0.018	−0.009	−0.005	–[a]
		(0.476)	(0.059)	(0.017)	(0.009)	–[a]	(0.399)	(0.029)	(0.016)	(0.008)	–[a]
	L_{full}	−0.001	0.015	0.006	−0.003	0.000	−0.009	0.001	−0.001	0.011	0.000
		(0.126)	(0.123)	(0.146)	(0.068)	(0.004)	(0.087)	(0.082)	(0.100)	(0.108)	(0.004)
0.3	L_4	0.082	0.034	0.006	−0.002	−0.001	0.089	0.041	0.007	0.000	−0.002
		(0.080)	(0.055)	(0.025)	(0.012)	(0.015)	(0.046)	(0.035)	(0.015)	(0.008)	(0.016))
	$L_{\emptyset}$	0.499	0.014	−0.005	0.002	–[a]	0.523	0.034	−0.004	0.000	–[a]
		(0.281)	(0.059)	(0.019)	(0.013)	–[a]	(0.115)	(0.044)	(0.018)	(0.011)	–[a]
	L_{full}	0.304	0.297	0.273	0.168	0.011	0.306	0.303	0.289	0.249	0.041
		(0.112)	(0.129)	(0.160)	(0.217)	(0.088)	(0.067)	(0.066)	(0.100)	(0.152)	(0.143)
0.5	L_4	0.147	0.073	0.014	0.001	0.023	0.157	0.082	0.015	0.002	0.029
		(0.081)	(0.060)	(0.026)	(0.012)	(0.021)	(0.048)	(0.038)	(0.016)	(0.009)	(0.021)
	$L_{\emptyset}$	0.687	0.053	−0.001	0.001	–[a]	0.677	0.077	0.003	−0.001	–[a]
		(0.117)	(0.069)	(0.018)	(0.013)	–[a]	(0.067)	(0.049)	(0.017)	(0.012)	–[a]
	L_{full}	0.505	0.499	0.490	0.426	0.224	0.508	0.503	0.494	0.488	0.327
		(0.094)	(0.105)	(0.134)	(0.204)	(0.290)	(0.058)	(0.055)	(0.083)	(0.076)	(0.259)

Table 1 continued

		$m = 20$					$m = 50$				
n_c		5	10	100	1,000	100,000	5	10	100	1,000	100,000
0.7	L_4	0.239	0.134	0.034	0.009	0.135	0.252	0.148	0.034	0.010	0.129
		(0.083)	(0.071)	(0.031)	(0.014)	(0.047)	(0.051)	(0.044)	(0.019)	(0.010)	(0.050)
	$L_\emptyset$	0.821	0.202	0.007	0.009	–[a]	0.819	0.155	0.011	0.005	–[a]
		(0.056)	(0.262)	(0.019)	(0.017)	–[a]	(0.035)	(0.113)	(0.013)	(0.015)	–[a]
	L_{full}	0.701	0.700	0.696	0.692	0.655	0.706	0.702	0.701	0.702	0.695
		(0.077)	(0.074)	(0.079)	(0.081)	(0.194)	(0.044)	(0.039)	(0.047)	(0.045)	(0.055)
0.9	L_4	0.408	0.267	0.098	0.038	0.029	0.425	0.290	0.095	0.036	0.016
		(0.093)	(0.096)	(0.060)	(0.033)	(0.074)	(0.055)	(0.060)	(0.034)	(0.016)	(0.036)
	$L_\emptyset$	0.936	0.933	0.282	0.023	–[a]	0.935	0.937	0.188	0.025	–[a]
		(0.017)	(0.036)	(0.420)	(0.024)	–[a]	(0.010)	(0.010)	(0.355)	(0.020)	–[a]
	L_{full}	0.901	0.899	0.901	0.901	0.903	0.902	0.901	0.900	0.900	0.902
		(0.029)	(0.026)	(0.025)	(0.028)	(0.023)	(0.017)	(0.014)	(0.016)	(0.016)	(0.015)

[a] indicates that computation times were too high

This insight identifies clear limits on the dependence information content that this (discipline standard) interval construction can possess.

Given the statistical inefficiency of intervals constructed from minima and maxima (Fig. 3) and their informational limits, a sensible alternative is to construct random rectangles using marginal order statistics (Sect. 2.3.2), which should be robust to these limitations. Given that such intervals constructed from independent marginal quantiles ((7) and (8)) will not contain dependence information, we examine the performance of the sequential nesting (9) and iterative segmentation (11) constructions, for which we denote the respective likelihood functions as $L_{\text{sn}}(s;\theta)$ and $L_{\text{is}}(s;\theta)$.

For each of $T = 100$ replicate datasets, we generate $m = 20$ classes, each constructed from $n = 60$ and 300 draws from a bivariate ($d = 2$) $N_2(\mu_0, \Sigma_0)$ distribution with $\mu_0 = (2, 5)^\top$, $\sigma_{0,1} = \sigma_{0,2} = 0.5$ and correlation $\rho_0 = -0.7, 0, 0.7$. The symbols are constructed in four ways: $L_{\text{sn},x}$ using sequential nesting (9); $L_{\text{sn},y}$ using sequential nesting but by exchanging the conditioning order of the x and y margins for symbol construction; $L_{\text{is},x}$ using iterative segmentation (11); $L_{\text{is},y}$ using iterative segmentation but again by exchanging the conditioning order of the x and y margins.

Table 2 reports the mean (and standard deviation) of σ_1, σ_2, ρ under each experimental setup when $\rho_0 = 0.7$ (results for $\rho_0 = -0.7$ and 0 are in Supplementary Material B.2). Estimates of σ_1 and σ_2 are unbiased for any rectangle configuration. However the standard deviations of the estimates are smaller for components which are conditioned on first in the symbol construction e.g. σ_1 is more precisely estimated by $L_{sn,x}$ and $L_{is,x}$, and σ_2 by $L_{sn,y}$ and $L_{is,y}$. Constructing intervals using iterative segmentation produces more precise estimates of the correlation ρ than using sequential nesting. This is because iterative segmentation provides more information about joint upper and lower values of the margins than nested segmentation, which provides stronger information about the centre of the marginal distributions (Fig. 1). Different axis constructions ($L_{\cdot,x}$ or $L_{\cdot,y}$) have little effect on the estimates in this case, due to the symmetry of the underlying Gaussian distribution. As expected, increasing the amount of data per symbol, n_c, leads to more precise estimates of all parameters.

All estimates of ρ are more precise than that obtained using marginal minima and maxima, which gave a MLE standard deviation of 0.0720 (for $n_c = 50, m = 20, \rho_0 = 0.7$ and using L_{full} in Table 1). Similar to Fig. 3, within any method of symbol construction, the choice of order statistics has an impact on the performance of the MLE. Clearly there is an important optimal symbol design question to be addressed, that goes beyond the scope of this paper. However, the iterative segmentation approach appears to be more informative for all parameters, for reasons described above. It is likely that there are other random rectangle constructions that would be even more informative.

3.3 Peer-to-peer loan data analysis

We analyse data from the U.S. peer-to-peer lending company *LendingClub* available from the Kaggle platform (https://www.kaggle.com/husainsb/lendingclub-issued-loans). After removing missing values, it comprises 887,373 unsecured personal loans of between \$1k–\$40k, issued by individual investors during 2007–2015, each with an

Table 2 Mean (and standard deviation) of the symbolic maximum likelihood estimate of σ_1, ρ and σ_2, over $T = 100$ replicate bivariate random rectangle datasets containing $m = 20$ symbols. The symbolic datasets vary in the number of classical datapoints per symbol (n_c), the type of symbol construction (sn = sequential nesting; is = iterative segmentation), which axis is used first in the symbol construction (x or y), and the vectors of lower (l) and upper (u) order statistics used. True parameter values are $\sigma_{0,1} = \sigma_{0,2} = 0.5$ and $\rho_0 = 0.7$. For $L_{sn,x}$, orders $(l, u) = ((6, 5), (55, 35))$ mean firstly take the (6,55) lower/upper order statistics on the x-axis, and then the $(5, 35)$ y-order statistics of the remaining $n_c - 12$ observations in the central x range (Fig. 1, bottom centre). For $L_{is,x}$, orders $(l, u) = ((6, 3), (55, 3))$ mean firstly take the $(6, 55)$ lower/upper order statistics on the x-axis, the 3-rd y-order statistic of the remaining 5 observations below the lower x quantile, and the 3-rd y-order statistic of the remaining 5 observations above the upper x order statistic (Fig. 1, bottom right). For $L_{\cdot,y}$ the procedure is the same as for $L_{\cdot,x}$ but starting with the y-quantiles (the resulting 3 bivariate intervals for e.g. $L_{sn,x}$ are identical to those for $L_{sn,y}$). The orders shown are for $n_c = 60$. For $n_c = 300$ the utilised orders are multiplied by 5 so that the intervals are directly comparable

		$n_c = 60$			$n_c = 300$		
	Orders (l,u)	σ_1	ρ	σ_2	σ_2	ρ_0	σ_2
$L_{\text{sn},x}$	((6, 5), (55, 35))	0.4992	0.6933	0.5050	0.4984	0.6772	0.5075
		(0.0019)	(0.0255)	(0.0054)	(0.0004)	(0.0146)	(0.0024)
	((16,6), (45,24))	0.4981	0.6402	0.5043	0.4985	0.6739	0.5177
		(0.0021)	(0.0273)	(0.0107)	(0.0005)	(0.0115)	(0.0048)
	((20, 5), (41, 16))	0.4991	0.6396	0.5054	0.4981	0.6451	0.5141
		(0.0027)	(0.0256)	(0.0129)	(0.0006)	(0.0127)	(0.0059)
$L_{\text{sn},y}$	((5,6), (35, 55))	0.5106	0.6912	0.4974	0.5082	0.6774	0.4998
		(0.0061)	(0.0339)	(0.0016)	(0.0024)	(0.0156)	(0.0004)
	((6, 16), (24, 45))	0.5289	0.6933	0.4986	0.5088	0.6453	0.4994
		(0.0123)	(0.0239)	(0.0021)	(0.0049)	(0.0129)	(0.0004)
	((5,20), (16, 41))	0.5231	0.6699	0.5004	0.5154	0.6702	0.4992
		(0.0127)	(0.0253)	(0.0024)	(0.0053)	(0.0106)	(0.0005)
$L_{\text{is},x}$	((6, 3), (55, 3))	0.4993	0.7130	0.4900	0.4984	0.7124	0.4932
		(0.0019)	(0.0067)	(0.0037)	(0.0004)	(0.0032)	(0.0019)
	((16,10), (45, 2))	0.4981	0.7037	0.4806	0.4985	0.7051	0.4866
		(0.0021)	(0.0039)	(0.0064)	(0.0005)	(0.0011)	(0.0025)
	((20,7), (41,14))	0.4993	0.7465	0.4871	0.4981	0.7169	0.4979
		(0.0027)	(0.0128)	(0.0037)	(0.0006)	(0.0051)	(0.0013)
$L_{\text{is},y}$	((3,6), (3, 55))	0.4929	0.7133	0.4975	0.4896	0.7151	0.4998
		(0.0051)	(0.0064)	(0.0016)	(0.0018)	(0.0032)	(0.0004)
	((10,16), (2, 45))	0.4868	0.7053	0.4986	0.4848	0.7066	0.4993
		(0.0068)	(0.0035)	(0.0021)	(0.0026)	(0.0011)	(0.0004)
	((7,20), (14, 41))	0.4933	0.7311	0.5004	0.4947	0.7268	0.4993
		(0.0040)	(0.0115)	(0.0023)	(0.0016)	(0.0057)	(0.0005)

associated grade, from A1 (least risky) to G5 (most risky), based on risk and market conditions, which defines the interest rate. To properly balance their investment portfolio, individual investors need to understand how borrower characteristics relate to their ability to repay the loan plus interest, as the interest is where the investor makes their profit. We examine the link between the borrower's log annual income (in \$US)—taken as an indicator of the borrower's ability to repay a loan—and loan grade (i.e. the investment risk) via a highly computational analysis of the full dataset (providing a gold standard), a symbolic analysis using (13) based on aggregating the income data in each risk group into a 5-bin histogram, and a reference SDA analysis following Le-Rademacher and Billard (2011) (denoted LRB).

Denoting X_{ij} as the log-income for individual j in grade $i = 1, \dots, 35$, we consider both normal $X_{ij} \sim N(\mu_i, \sigma_i^2)$ and skew-normal $X_{ij} \sim SN(\mu_i, \sigma_i^2, \gamma_i)$ models for each grade (with the skew-normal parameterised in terms of mean μ_i and variance σ_i^2), given that standard likelihood ratio tests identify the presence of asymmetry in 34/35 groups ($\alpha = 0.05$). Within-grade sample sizes range from 576 (G5) to 56,323 (B3). Coding the ordered grades A1–G5 as the numbers 1–35, each model specifies

$$\mu_i \sim T_3(c_0 + c_1 i + c_2 i^2, \tau^2) \quad \text{and} \quad \sigma_i^2 \sim IG(\alpha, \beta), \tag{20}$$

where $T_\nu(m, v)$ denotes a t-distribution with mean m, variance v and ν degrees of freedom, and $IG(\alpha, \beta)$ the inverse-Gamma distribution with shape α and scale β. For the skew-normal model we additionally specify $\gamma_i \sim N(\eta, \epsilon)$ for the skewness parameter. For the reference LRB analysis, we implement the model (20) where μ_i and σ_i^2 correspond to the mean and variance of the histogram of the i-th group (Le-Rademacher and Billard 2011, section 2.3).

Figure 4 presents the fitted group means and variances obtained through each method. The grade specific means under the Normal (top row) model are uniformly well estimated. Our symbolic model produces standard errors only slightly larger than the classical ones, while those from the LRB model are about twice as large. The means under the skew-Normal (bottom) model are less well estimated, but remain, for the majority, within the classical 95% confidence band. The right panels highlight the inability of the LRB method to correctly replicate the classical analysis. This is essentially because the LRB approach models the variances of a histogram generated from the underlying data (assuming uniformity within bins), rather than modelling the variance of the underlying data. However, given the same histograms, our symbolic approach approximates the classical analysis well.

By design, the LRB approach cannot discriminate between normal and skew-normal models (the LRB fits in Fig. 4 for both models are the same), unlike our symbolic analysis which approximates the full classical analysis. This means that we are able to make inference at both the level of the underlying data as well as the symbol level (LRB is restricted to the latter). This is illustrated for the distribution of loan grade C3 in Fig. 5. Qq-plots for both models (left panel) suggest the skew-normal model appears to be a better fit in the upper tail of the log income distribution, and slightly worse in the lower tail. This could be tested formally. As the LRB approach cannot make such judgements of model adequacy, it is confined to predictions about the mean and variance of (histograms constructed from data generated by) this underlying process.

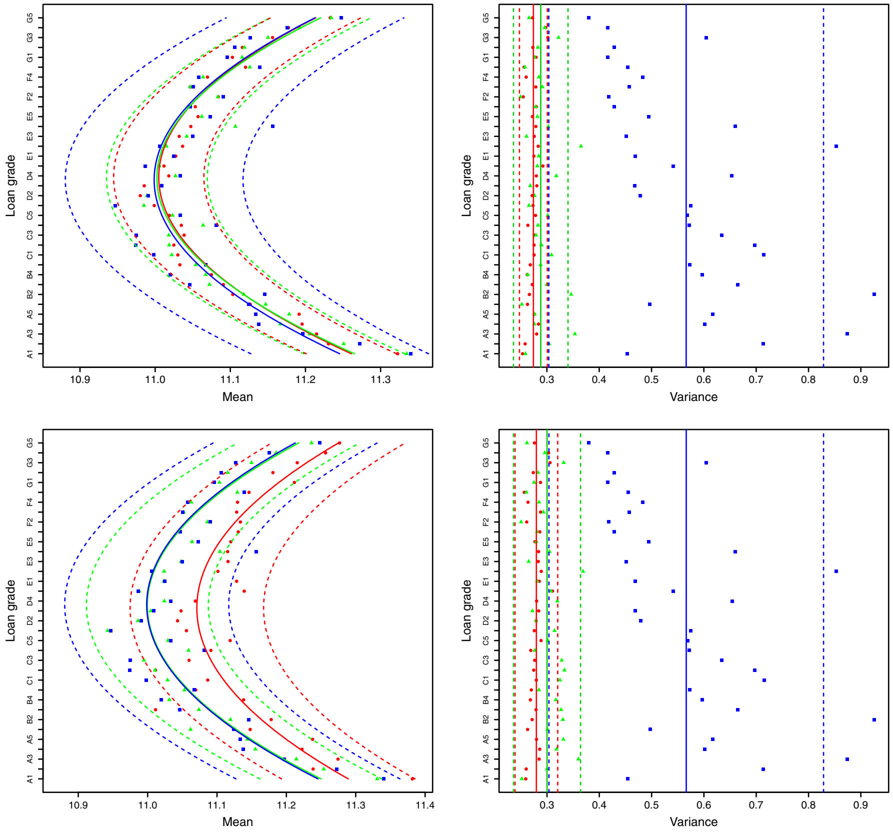

Fig. 4 Fitted group means and variances (solid lines) when the underlying distribution is Normal (top) and skew-Normal (bottom), using the classical (red) and symbolic (green) likelihoods and the LRB model (blue). Dashed lines indicate pointwise 95% confidence intervals. Points denote $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ under the classical and symbolic models, and the sample mean and variance of each grade histogram for the LRB model (colour figure online)

These predictive distributions are shown in the right panel. Even when considered on its own terms, the LRB method produces less accurate and precise predictions than our symbolic approach (the dot indicates the observed histogram mean/variance). Beyond this, our symbolic approach can produce the equivalent predictive distributions for the sample mean and variance of the underlying predicted data, without first producing histograms (centre), which is perhaps more useful in an analysis as it captures knowledge of the underlying data generation process. The LRB method cannot produce these predictions.

Finally, Table 3 provides the mean time to evaluate each likelihood function under each model, averaged over 1,000 randomly generated parameter vectors. The LRB analysis is most efficient as it is based on a likelihood with 35 bivariate points. The classical analysis is efficient for the normal model given the available sufficient statistic for each loan grade, however the the skew-normal likelihood requires iteration over all

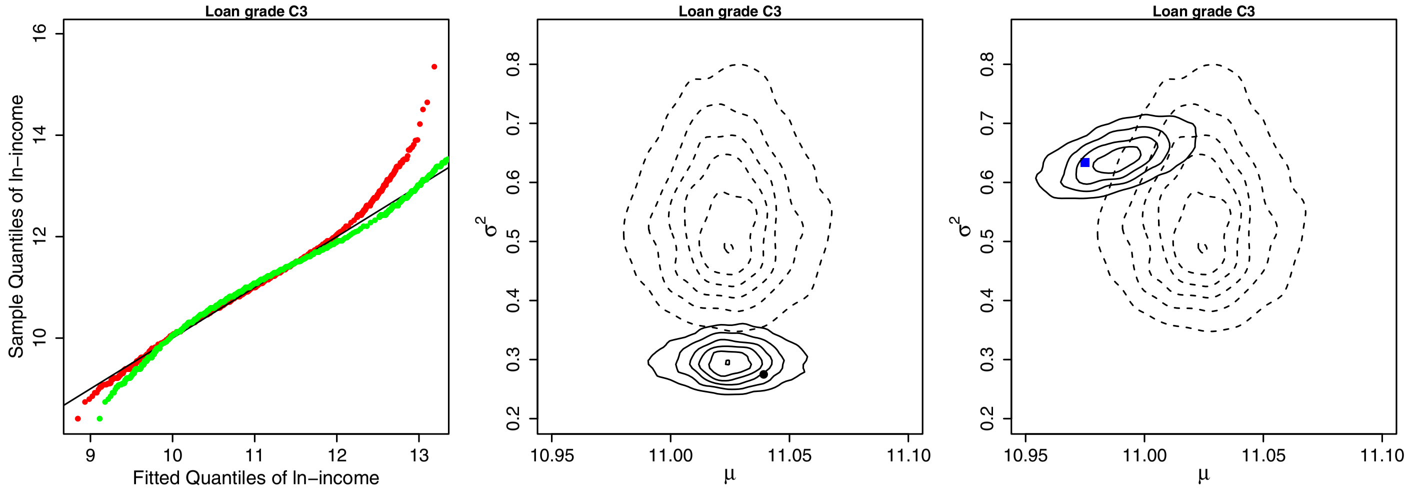

Fig. 5 Predictive inference for loan grade C3 ($n_{C3} = 50, 161$). Left panel: qq-plot of fitted versus empirical quantiles for our symbolic likelihood under Normal (red) and skew-Normal (green) distributions. Predictive distributions of the group mean (μ_i) and variance (σ_i^2) at the underlying data (centre panel; obtained by computing sample means and variances of data generated from the fitted model) and symbol (right panel; obtained by computing means and variances of histograms constructed from data generated from the fitted model) levels. Solid and dashed contours respectively denote the symbolic and LRB predictive distributions. The LRB model predicts directly at the symbol level (not via underlying data). Black dot and blue square denote the underlying-data-based and histogram-based mean and variance for group $C3$ (colour figure online)

Table 3 Mean (s.e.) likelihood evaluation times (seconds $\times 10^{-3}$) over $T = 1{,}000$ parameter vector replicates, using the loan dataset (first two columns) and simulated data with 35 groups of $n = 1{,}000{,}000$ observations

	Normal	Skew-normal	Skew-normal ($n = 1{,}000{,}000$)
Classical	3.886 (0.478)	90.754 (0.097)	3533.900 (2.472)
New symbolic	1.551 (0.045)	12.721 (0.034)	11.487 (0.030)
LRB	0.498 (0.001)	0.476 (0.001)	0.457 (0.001)

887,373 records. In contrast, our symbolic likelihood (14) requires 6 cdf evaluations per loan grade. It is slower than the normal classical analysis, but 14 times faster than the skew-normal classical analysis, with comparable model fit. The symbolic computational times will remain roughly constant as the dataset size increases (right column), generating increasingly large computational savings for the skew-normal model compared to the classical analysis.

This analysis highlights the the differences and similarities between the LRB approach and our own. The former makes model assumptions and inference at the symbol level, with no assumptions at the level of the micro-data. Our own approach makes model assumptions at the level of the micro-data and so can make inference at both micro-data and symbol levels. When the micro-data model assumptions are adequate, then our approach will naturally outperform that of LRB. Conversely, if the micro-data model assumptions are incorrect then using the LRB method may be the preferred approach. Clearly these two methods are complementary in nature.

4 Discussion

In this article we have introduced a novel framework for the analysis of data that have been summarised into distributional forms. For the general statistical analyst, this method opens up the use of SDA as a broadly applicable statistical technique for analysing large and complex datasets with the potential for large data-storage and computational savings. Within the SDA setting, the fundamentally different approach taken—that of specifying probability models for the data underlying a symbol and deriving the resulting model at the symbolic level, rather than direct model specification at the symbolic level—has introduced a new research direction in the field of SDA. The proposed framework resolves open and new problems including the difficulty of specifying meaningful models at the symbolic level, avoidance of the routinely violated uniformity-within-symbols assumption, the ability to perform accurate inference at the level of the underlying data, including model choice, and providing a means to construct and analyse multivariate symbols. We have exposed some weaknesses of current symbol design, and have introduced several alternative, more efficient symbol constructions.

When the aim of a SDA practitioner is to understand the behaviour of some underlying process, we have demonstrated how the design of the aggregation function can affect the outcome of an analysis. Regardless of whether the symbols were computed from micro-data or directly collected, we have shown that keeping details about the aggregation procedure is of absolute necessity for inference.

While providing a step forwards, our approach is not without some caveats. Most obviously, the symbolic likelihood function (1) requires enumeration of the integral over the underlying data space, which may be problematic in high dimensions. For many standard classes of models, distribution functions $G_X(\boldsymbol{x}; \theta)$ are available in closed form. In other cases, numerical or approximate methods may be required, such as quadrature, Monte Carlo techniques (Andrieu and Roberts 2009), or factorisation of $g_X(\boldsymbol{x}; \theta)$ to reduce the dimension of the integral. Alternatively, composite likelihood solutions (Whitaker et al. 2020) can be considered.

The symbolic likelihood is clearly an approximation of the classical likelihood as it is based on summary data, and so there will be some information loss. While the accuracy of the classical data model can be approached by letting the symbols approach the classical data (e.g. by letting the number of histogram bins $B \rightarrow \infty$), this may not be viable in practice, and in the extreme (e.g. with huge numbers of bins) the computational costs could exceed those of the classical analysis. It is therefore of interest, and the subject of future research, to understand the quality of the approximation. It is possible that some of the theory supporting approximate Bayesian computation (e.g. Sisson et al. 2018), which is also based on computation via summary statistics, could be useful here.

Within this context there is immense scope for optimum symbol design, whereby symbols are constructed to provide maximal information for a specific or more general analyses that may be performed in the future. Different symbolic types could be developed such as continuous distribution-based symbols, which may additionally enable direct integration of the integral in (1) through conjugacy.

Since Schweizer (1984)'s 35-year old prediction that "distributions are the numbers of the future", the explosive emergence of the data-rich biome—the *infome*—in which we now reside, clearly substantiates the potential for symbolic data analysis to become a powerful everyday tool for the statistical analyst. Schweizer (1984)'s future is very much here.

Acknowledgements The authors are supported by the Australian Centre of Excellence for Math- ematical and Statistical Frontiers (ACEMS, CE140100049). BB and SAS are also supported by the Australian Research Council Discovery Project Scheme (DP160102544 and FT170100079). The authors thank Felix Camirand-Lemyre and Tom Whitaker for many useful conversations.

Author Contributions Not applicable.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Availability of data and material Data is publicly available, see reference in the manuscript.

Code Availability Code for the analyses in this manuscript can be found at: https://www.borisberanger.com/zip/BLS.zip.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

A Proofs

A.1 Univariate intervals: Proof of Lemma 1

Based on the aggregation function (2) we have $S = (S_l, S_u, N) = (X_{(l)}, X_{(u)}, N)$ and thus the role of $f_{S|\boldsymbol{X}=\boldsymbol{z}}(s; \vartheta)$ in Proposition 1 is to ensure that $s_l = z_{(l)}$ and $s_u = z_{(u)}$. Consequently we can write

$$f_{S|\boldsymbol{X}=\boldsymbol{z}}(s; \vartheta) = \delta_{z_{(l)}, z_{(u)}}(s_l, s_u) = \delta_{z_{(l)}}(s_l)\, \delta_{z_{(u)}}(s_u),$$

so that $l-1$ points of $\boldsymbol{z}$ belong to $(-\infty, s_l)$, one is at s_l, $u-l-1$ belong to (s_l, s_u), one is at s_u and $n-u$ belong to (s_u, ∞). As there are $n!/((l-1)!(u-l-1)!(n-u)!)$ possible combinations to arrange n points in such a way, the likelihood function can then be written as

$$\begin{aligned} &L(s_l, s_u, n; \theta)\\ &= \frac{n!}{(l-1)!(u-l-1)!(n-u)!} \left(\int_{-\infty}^{s_l} g_X(z; \theta) \mathrm{d}z\right)^{l-1} \int_{-\infty}^{+\infty} g_X(z; \theta) \delta_z(s_l) \mathrm{d}z\\ &\quad \times \left(\int_{s_l}^{s_u} g_X(z; \theta) \mathrm{d}z\right)^{u-l-1} \int_{-\infty}^{+\infty} g_X(z; \theta) \delta_z(s_u) \mathrm{d}z \left(\int_{s_u}^{\infty} g_X(z; \theta) \mathrm{d}z\right)^{n-u}\\ &= \frac{n!}{(l-1)!(u-l-1)!(n-u)!} \left[G_X(s_l; \theta)\right]^{l-1} \left[G_X(s_u; \theta) - G_X(s_l; \theta)\right]^{u-l-1}\\ &\quad \times \left[1 - G_X(s_u; \theta)\right]^{n-u} g_X(s_l; \theta) g_X(s_u; \theta), \end{aligned}$$

using the independence between the n replicates $X_1, \ldots, X_n$.

A.2 Multivariate intervals: Proof of Lemma 2 and Corollary 1

Consider bivariate intervals for simplicity (with identical arguments providing a full multivariate extension), so that X is a bivariate random vector with pdf $g_X(\cdot\,; \theta)$ and marginal and conditional pdfs respectively denoted by $g_{X_i}(\cdot\,; \theta)$, $i = 1, 2$ and $g_{X_i|X_j}(\cdot\,; \theta)$, $i, j = 1, 2$; $i \neq j$. The conditional distribution of S given $\boldsymbol{X} = \boldsymbol{z} \in \mathbb{R}^2$ is obtained from the aggregation function (3). When $S_p = 2$, $S_{I_p} = (s_a, s_b)$. Now $s_a = (s_{a_1}, s_{a_2})$ and $s_b = (s_{b_1}, s_{b_2})$ which take values $s_a = (s_{\min,1}, s_{\min,2})$ and $s_b = (s_{\max,1}, s_{\max,2})$ if the rectangle constructed from top right and bottom left points, or

$s_a = (s_{\min,1}, s_{\max,2})$ and $s_b = (s_{\max,1}, s_{\min,2})$ if from the top left, bottom right points. Then

$$f_{S|\boldsymbol{X}=\boldsymbol{z}}(s;\vartheta) = \begin{cases} \delta_{z_{(1),1}, z_{(1),2}, z_{(n),1}, z_{(n),2}}\left(s_{a_1}, s_{a_2}, s_{b_1}, s_{b_2}\right) \\ \text{or} \\ \delta_{z_{(1),1}, z_{(n),2}, z_{(n),1}, z_{(1),2}}\left(s_{a_1}, s_{a_2}, s_{b_1}, s_{b_2}\right). \end{cases}$$

Straightforwardly, this ensures that two points give the marginal minima and maxima and the remaining points are within the interval. There are $n(n-1)$ possible combinations to arrange n points in such a way and so the likelihood function is

$$\begin{aligned} &L(s;\theta) \\ &\quad = n(n-1)\left(\int_{s_{\min}}^{s_{\max}} g_X(z;\theta)\mathrm{d}z\right)^{n-2} \int_{\mathbb{R}^2} g_X(z;\theta)\delta_{s_a}(z)\mathrm{d}z \int_{\mathbb{R}^2} g_X(z;\theta)\delta_{s_b}(z)\mathrm{d}z \\ &\quad = n(n-1)\left(\int_{s_{\min}}^{s_{\max}} g_X(z;\theta)\mathrm{d}z\right)^{n-2} g_X(s_a;\theta)g_X(s_b;\theta). \end{aligned}$$

When $S_p = 3$ so that a single point $S_{I_p} = s_c = s_{\min}$ defines the bottom left rectangle corner, then

$$\begin{aligned} &f_{S|\boldsymbol{X}=\boldsymbol{z}}(s;\vartheta) \\ &\quad = \delta_{z_{(1),1}, z_{(1),2}}(s_c)\delta_{(s_{\min,1}, s_{\max,1}),(s_{\min,2}, s_{\max,2})}\left(z_{j,1}|z_{j,2} = s_{\max,2}, z_{j,2}|z_{j,1} = s_{\max,1}\right). \end{aligned}$$

There are $n(n-1)(n-2)$ possible combinations to arrange n points such that one is at a corner, two are on two different edges and the rest are inside the interval. The likelihood is then

$$\begin{aligned} L(s;\theta) = {} & n(n-1)(n-2)\int_{\mathbb{R}^2} g_X(z;\theta)\delta_{s_{\min}}(z)\mathrm{d}z \left(\int_{s_{\min,1}}^{s_{\max,1}} g_{X_1|X_2=s_{\max,2}}(z_1;\theta)\mathrm{d}z_1\right) \\ & \times g_{X_2}(s_{\max,2};\theta) \\ & \times \left(\int_{s_{\min,2}}^{s_{\max,2}} g_{X_2|X_1=s_{\max,1}}(z_2;\theta)\mathrm{d}z_2\right) g_{X_1}(s_{\max,1};\theta) \left(\int_{s_{\min}}^{s_{\max}} g_X(z;\theta)\mathrm{d}z\right)^{n-3} \\ = {} & n(n-1)(n-2)g_X(s_{\min};\theta)\left(\int_{s_{\min}}^{s_{\max}} g_X(z;\theta)\mathrm{d}z\right)^{n-3} \\ & \times \left[G_{X_1|X_2=s_{\max,2}}(s_{\max,1};\theta) - G_{X_1|X_2=s_{\max,2}}(s_{\min,1};\theta)\right] g_{X_2}(s_{\max,2};\theta) \\ & \times \left[G_{X_2|X_1=s_{\max,1}}(s_{\max,2};\theta) - G_{X_2|X_1=s_{\max,1}}(s_{\min,2};\theta)\right] g_{X_1}(s_{\max,1};\theta). \end{aligned}$$

Finally when $S_p = 4$ then

$$\begin{aligned} f_{S|\boldsymbol{X}=\boldsymbol{z}}(s;\vartheta) = {} & \delta_{(s_{\min,1}, s_{\max,1}),(s_{\min,1}, s_{\max,1})}\left(z_{j,1}|z_{j,2} = s_{\min,2}, z_{j,1}|z_{j,2} = z_{\max,2}\right) \\ & \times \delta_{(s_{\min,2}, s_{\max,2}),(s_{\min,2}, s_{\max,2})}\left(z_{j,2}|z_{j,1} = s_{\min,1}, z_{j,2}|z_{j,1} = s_{\max,1}\right), \end{aligned}$$

and there are $n(n-1)(n-2)(n-3)$ possible combinations to arrange four points on different edges and the rest inside the interval. The likelihood is then

$$
\begin{aligned}
L(s;\theta) = n(n-1)(n-2)(n-3) & \left(\int_{s_{\min}}^{s_{\max}} g_X(z;\theta)\mathrm{d}z\right)^{n-4} \\
& \times \left(\int_{s_{\min,1}}^{s_{\max,1}} g_{X_1|X_2=s_{\min,2}}(z_1;\theta)\mathrm{d}z_1\right) g_{X_2}(s_{\min,2};\theta) \\
& \times \left(\int_{s_{\min,1}}^{s_{\max,1}} g_{X_1|X_2=s_{\max,2}}(z_1;\theta)\mathrm{d}z_1\right) g_{X_2}(s_{\max,2};\theta) \\
& \times \left(\int_{s_{\min,2}}^{s_{\max,2}} g_{X_2|X_1=s_{\min,1}}(z_2;\theta)\mathrm{d}z_2\right) g_{X_1}(s_{\min,1};\theta) \\
& \times \left(\int_{s_{\min,2}}^{s_{\max,2}} g_{X_2|X_1=s_{\max,1}}(z_2;\theta)\mathrm{d}z_2\right) g_{X_1}(s_{\max,1};\theta) \\
= n(n-1)(n-2)(n-3) & \left(\int_{s_{\min}}^{s_{\max}} g_X(z;\theta)\mathrm{d}z\right)^{n-4} \\
& \times \left[G_{X_1|X_2=s_{\min,2}}(s_{\max,1};\theta) - G_{X_1|X_2=s_{\min,2}}(s_{\min,1};\theta)\right] g_{X_2}(s_{\min,2};\theta) \\
& \times \left[G_{X_1|X_2=s_{\max,2}}(s_{\max,1};\theta) - G_{X_1|X_2=s_{\max,2}}(s_{\min,1};\theta)\right] g_{X_2}(s_{\max,2};\theta) \\
& \times \left[G_{X_2|X_1=s_{\min,1}}(s_{\max,2};\theta) - G_{X_2|X_1=s_{\min,1}}(s_{\min,2};\theta)\right] g_{X_1}(s_{\min,1};\theta) \\
& \times \left[G_{X_2|X_1=s_{\max,1}}(s_{\max,2};\theta) - G_{X_2|X_1=s_{\max,1}}(s_{\min,2};\theta)\right] g_{X_1}(s_{\max,1};\theta).
\end{aligned}
$$

A.3 Multivariate histograms with fixed bins: Proof of Lemma 5

As $S = \pi(\boldsymbol{X})$ is given by (13), then $s_{\boldsymbol{b}} = \sum_{i=1}^{n} \mathbb{1}\{z_i \in \mathcal{B}_{\boldsymbol{b}}\}$ for $\boldsymbol{b} = \mathbf{1}, \ldots, \boldsymbol{B}$, which is equivalent to

$$
f_{S|\boldsymbol{X}=\boldsymbol{z}}(s;\vartheta) = \prod_{\boldsymbol{b}=\mathbf{1}}^{\boldsymbol{B}} \delta_{\sum_{i=1}^{n} \mathbb{1}\{z_i \in \mathcal{B}_{\boldsymbol{b}}\}}(s_{\boldsymbol{b}})\,.
$$

The number of combinations to arrange $z_1, \ldots, z_n$ into the $B_1 \times \cdots \times B_{\boldsymbol{B}}$ bins is the multinomial coefficient $n!/\prod_{\boldsymbol{b}} s_{\boldsymbol{b}}!$, and sp the likelihood function (1) becomes

$$
\begin{aligned}
& L(s;\theta) \\
& \quad = \frac{n!}{\boldsymbol{s}_{\mathbf{1}}!\cdots \boldsymbol{s}_{\boldsymbol{B}}} \int_{\mathbb{R}^{n\times d}} \delta_{z_1}(\mathcal{B}_{\mathbf{1}}) \cdots \delta_{z_{s_{\mathbf{1}}}}(\mathcal{B}_{\mathbf{1}}) \cdots \delta_{z_{n-s_{\boldsymbol{B}}+1}}(\mathcal{B}_{\boldsymbol{B}}) \cdots \delta_{z_n}(\mathcal{B}_{\boldsymbol{B}}) \prod_{i=1}^{n} g_X(z_i;\theta)\mathrm{d}\boldsymbol{z} \\
& \quad = \frac{n!}{\boldsymbol{s}_{\mathbf{1}}!\cdots \boldsymbol{s}_{\boldsymbol{B}}} \left(\int_{\mathbb{R}^d} g_X(z;\theta)\delta_z(\mathcal{B}_{\mathbf{1}})\mathrm{d}z\right)^{s_{\mathbf{1}}} \cdots \left(\int_{\mathbb{R}^d} g_X(z;\theta)\delta_z(\mathcal{B}_{\boldsymbol{B}})\mathrm{d}z\right)^{s_{\boldsymbol{B}}} \\
& \quad = \frac{n!}{\boldsymbol{s}_{\mathbf{1}}!\cdots \boldsymbol{s}_{\boldsymbol{B}}} \prod_{\boldsymbol{b}=\mathbf{1}}^{\boldsymbol{B}} \left(\int_{\mathcal{B}_{\boldsymbol{b}}} g_X(z;\theta)\mathrm{d}z\right)^{s_{\boldsymbol{b}}}.
\end{aligned}
$$

A.4 Histograms with fixed counts: Lemma 6

The aggregation function (15) ensures that the B bins are defined as order statistics, so that the symbol S provides the location of B out of n points, with the number of points between these fixed and known through $k = (k_1, \dots, k_B)$. As a consequence the conditional density $f_{S|\boldsymbol{X}=\boldsymbol{z}}(s; \vartheta)$ is

$$f_{S|\boldsymbol{X}=\boldsymbol{z}}(s; \vartheta) = \prod_{b=1}^{B} \delta_{z_{(k_b)}}(s_b) \prod_{b=1}^{B+1} \prod_{j=k_{b-1}}^{k_b-1} \delta_{z_{(j)}}((s_{b-1}, s_b)),$$

for which there are $n!/\prod_{b=1}^{B+1}(k_b - k_{b-1} - 1)!$ possible combinations to arrange n points. Hence

$$\begin{aligned}
\mathcal{L}(s; \theta) &= \frac{n!}{\prod_{b=1}^{B+1}(k_b - k_{b-1} - 1)!} \int_{\mathbb{R}^n} \left(\prod_{b=1}^{B} \delta_{z_{(k_b)}}(s_b) \right) \prod_{b=1}^{B+1} \\
&\quad \times \left(\prod_{j=k_{b-1}}^{k_b-1} \delta_{z_{(j)}}((s_{b-1}, s_b)) \right) \prod_{i=1}^{n} g_X(z_i; \theta) \mathrm{d}\boldsymbol{z} \\
&= \frac{n!}{\prod_{b=1}^{B+1}(k_b - k_{b-1} - 1)!} \prod_{b=1}^{B} \\
&\quad \times \left(\int_{\mathbb{R}} \delta_z(s_b)\, g_X(z; \theta) \mathrm{d}z \right) \prod_{b=1}^{B+1} \left(\int_{s_{b-1}}^{s_b} g_X(z; \theta) \mathrm{d}z \right)^{k_b - k_{b-1} - 1} \\
&= \frac{n!}{\prod_{b=1}^{B+1}(k_b - k_{b-1} - 1)!} \\
&\quad \times \prod_{b=1}^{B} g_X(s_b; \theta) \prod_{b=1}^{B+1} (G_X(s_b; \theta) - G_X(s_{b-1}; \theta))^{k_b - k_{b-1} - 1}.
\end{aligned}$$

B Supplementary Material

B.1 Estimates of the μ_1, μ_2, σ_1 and σ_2, from Section 3.2

See Tables 4, 5, 6 and 7.

Table 4 As for Table 1 but for estimates of the mean μ_1

		$m = 20$					$m = 50$				
n_c		5	10	100	1,000	100,000	5	10	100	1,000	100,000
$\rho_0 = 0.0$	L_4	1.999	2.004	2.006	1.998	2.003	1.999	2.000	2.001	1.999	2.002
		(0.051)	(0.045)	(0.035)	(0.025)	(0.018)	(0.031)	(0.027)	(0.021)	(0.015)	(0.012)
	$L_\emptyset$	1.999	2.004	2.006	1.998	–[a]	1.999	2.000	2.001	1.999	–[a]
		(0.051)	(0.045)	(0.035)	(0.025)	–[a]	(0.031)	(0.027)	(0.021)	(0.015)	–[a]
	L_{full}	1.999	2.004	2.006	1.998	2.003	1.999	2.000	2.001	1.999	2.002
		(0.051)	(0.045)	(0.035)	(0.025)	(0.018)	(0.031)	(0.027)	(0.021)	(0.015)	(0.012)
0.3	L_4	1.995	1.996	1.998	2.000	2.001	1.996	1.999	1.996	1.998	2.001
		(0.052)	(0.044)	(0.034)	(0.024)	(0.016)	(0.034)	(0.028)	(0.020)	(0.016)	(0.011)
	$L_\emptyset$	1.995	1.996	1.998	2.000	–[a]	1.996	1.999	1.996	1.998	–[a]
		(0.053)	(0.044)	(0.034)	(0.024)	–[a]	(0.034)	(0.028)	(0.020)	(0.015)	–[a]
	L_{full}	1.995	1.996	1.998	2.000	2.001	1.996	2.000	1.996	1.998	2.001
		(0.053)	(0.044)	(0.034)	(0.024)	(0.016)	(0.034)	(0.028)	(0.020)	(0.016)	(0.011)
0.5	L_4	1.995	1.995	1.998	2.000	2.000	1.996	1.999	1.996	1.998	2.001
		(0.053)	(0.044)	(0.034)	(0.024)	(0.016)	(0.035)	(0.028)	(0.021)	(0.016)	(0.012)
	$L_\emptyset$	1.995	1.995	1.998	2.000	–[a]	1.996	1.999	1.996	1.998	–[a]
		(0.054)	(0.044)	(0.034)	(0.024)	–[a]	(0.034)	(0.028)	(0.021)	(0.016)	–[a]
	L_{full}	1.996	1.995	1.998	2.000	2.000	1.996	2.000	1.996	1.998	2.001
		(0.054)	(0.044)	(0.034)	(0.024)	(0.016)	(0.034)	(0.028)	(0.021)	(0.016)	(0.012)
0.7	L_4	1.994	1.996	1.997	2.001	2.000	1.997	1.999	1.995	1.999	2.001
		(0.054)	(0.043)	(0.034)	(0.023)	(0.017)	(0.035)	(0.027)	(0.020)	(0.017)	(0.012)
	$L_\emptyset$	1.994	1.996	1.997	2.001	–[a]	1.997	1.999	1.995	1.998	–[a]
		(0.055)	(0.043)	(0.034)	(0.023)	–[a]	(0.035)	(0.028)	(0.020)	(0.017)	–[a]
	L_{full}	1.995	1.995	1.998	2.000	2.000	1.997	2.000	1.995	1.998	2.001
		(0.054)	(0.043)	(0.033)	(0.023)	(0.017)	(0.034)	(0.027)	(0.020)	(0.017)	(0.012)
0.9	L_4	1.993	1.998	1.996	2.000	1.999	1.997	1.999	1.995	1.999	2.001
		(0.054)	(0.043)	(0.033)	(0.024)	(0.019)	(0.035)	(0.027)	(0.020)	(0.018)	(0.012)
	$L_\emptyset$	1.993	1.997	1.996	2.000	–[a]	1.997	1.999	1.996	1.999	–[a]
		(0.055)	(0.044)	(0.033)	(0.024)	–[a]	(0.035)	(0.027)	(0.019)	(0.018)	–[a]
	L_{full}	1.994	1.998	1.998	1.999	2.000	1.997	2.001	1.996	1.999	2.001
		(0.054)	(0.041)	(0.032)	(0.022)	(0.018)	(0.035)	(0.026)	(0.019)	(0.017)	(0.011)

Table 5 As for Table 1 but for estimates of the mean μ_2

		$m = 20$					$m = 50$				
n_c		5	10	100	1,000	100,000	5	10	100	1,000	100,000
$\rho_0 = 0.0$	L_4	4.993	5.000	4.997	5.002	4.998	4.998	5.000	4.996	5.001	5.000
		(0.053)	(0.045)	(0.032)	(0.027)	(0.018)	(0.034)	(0.026)	(0.018)	(0.019)	(0.012)

Table 5 continued

		$m = 20$					$m = 50$				
n_c		5	10	100	1,000	100,000	5	10	100	1,000	100,000
	$L_\emptyset$	4.993	5.001	4.997	5.002	–[a]	4.998	5.000	4.996	5.001	–[a]
		(0.053)	(0.045)	(0.032)	(0.027)	–[a]	(0.035)	(0.026)	(0.018)	(0.019)	–[a]
	L_{full}	4.993	5.001	4.997	5.002	4.998	4.998	5.000	4.997	5.001	5.000
		(0.053)	(0.046)	(0.032)	(0.027)	(0.018)	(0.034)	(0.026)	(0.018)	(0.019)	(0.012)
0.3	L_4	4.999	5.001	5.004	4.997	4.999	5.000	5.002	5.001	5.000	5.001
		(0.053)	(0.045)	(0.032)	(0.023)	(0.018)	(0.033)	(0.028)	(0.017)	(0.017)	(0.011)
	$L_\emptyset$	4.999	5.001	5.004	4.997	–[a]	5.000	5.002	5.002	5.000	–[a]
		(0.053)	(0.045)	(0.032)	(0.023)	–[a]	(0.033)	(0.027)	(0.017)	(0.017)	–[a]
	L_{full}	4.999	5.001	5.004	4.997	4.999	5.000	5.003	5.001	5.000	5.001
		(0.053)	(0.045)	(0.032)	(0.023)	(0.018)	(0.033)	(0.027)	(0.017)	(0.017)	(0.011)
0.5	L_4	4.998	5.001	5.004	4.996	4.999	5.000	5.002	5.001	5.001	5.000
		(0.053)	(0.047)	(0.032)	(0.024)	(0.018)	(0.033)	(0.028)	(0.018)	(0.017)	(0.012)
	$L_\emptyset$	4.998	5.001	5.004	4.996	–[a]	5.000	5.002	5.001	5.001	–[a]
		(0.053)	(0.046)	(0.032)	(0.024)	–[a]	(0.033)	(0.027)	(0.018)	(0.017)	–[a]
	L_{full}	4.999	5.001	5.005	4.996	4.999	5.000	5.002	5.002	5.001	5.000
		(0.052)	(0.046)	(0.032)	(0.024)	(0.018)	(0.032)	(0.027)	(0.018)	(0.017)	(0.012)
0.7	L_4	4.997	5.001	5.003	4.996	4.998	5.000	5.001	5.001	5.001	5.000
		(0.053)	(0.047)	(0.032)	(0.024)	(0.018)	(0.033)	(0.028)	(0.018)	(0.018)	(0.012)
	$L_\emptyset$	4.997	5.001	5.003	4.996	–[a]	5.000	5.001	5.001	5.001	–[a]
		(0.053)	(0.047)	(0.032)	(0.024)	–[a]	(0.033)	(0.028)	(0.018)	(0.018)	–[a]
	L_{full}	4.998	5.001	5.004	4.996	4.998	5.000	5.002	5.001	5.001	5.000
		(0.052)	(0.046)	(0.031)	(0.024)	(0.018)	(0.032)	(0.027)	(0.018)	(0.017)	(0.012)
0.9	L_4	4.995	5.001	5.000	5.000	4.998	4.999	5.000	5.000	5.002	4.999
		(0.053)	(0.047)	(0.031)	(0.024)	(0.018)	(0.034)	(0.028)	(0.017)	(0.017)	(0.011)
	$L_\emptyset$	4.994	5.001	5.000	5.000	–[a]	4.999	5.000	5.000	5.002	–[a]
		(0.054)	(0.048)	(0.032)	(0.024)	–[a]	(0.034)	(0.028)	(0.017)	(0.017)	–[a]
	L_{full}	4.995	5.002	5.002	4.999	4.999	4.999	5.002	5.000	5.002	5.000
		(0.052)	(0.045)	(0.029)	(0.021)	(0.018)	(0.033)	(0.027)	(0.018)	(0.016)	(0.011)

Table 6 As for Table 1 but for estimates of the standard deviation σ_1

		$m = 20$					$m = 50$				
n_c		5	10	100	1,000	100,000	5	10	100	1,000	100,000
$\rho_0 = 0.0$	L_4	0.247	0.247	0.250	0.250	0.251	0.247	0.247	0.249	0.250	0.250
		(0.039)	(0.029)	(0.015)	(0.008)	(0.004)	(0.024)	(0.018)	(0.009)	(0.006)	(0.002)
	$L_\emptyset$	0.249	0.247	0.250	0.250	–[a]	0.248	0.247	0.249	0.250	–[a]
		(0.039)	(0.029)	(0.015)	(0.008)	–[a]	(0.024)	(0.018)	(0.009)	(0.006)	–[a]

Table 6 continued

		$m = 20$					$m = 50$				
n_c		5	10	100	1,000	100,000	5	10	100	1,000	100,000
	L_{full}	0.247	0.247	0.250	0.250	0.251	0.247	0.247	0.249	0.250	0.250
		(0.039)	(0.029)	(0.015)	(0.008)	(0.004)	(0.024)	(0.018)	(0.009)	(0.006)	(0.002)
0.3	L_4	0.250	0.255	0.248	0.250	0.250	0.246	0.251	0.249	0.250	0.250
		(0.044)	(0.031)	(0.014)	(0.009)	(0.005)	(0.026)	(0.018)	(0.008)	(0.005)	(0.003)
	$L_{\emptyset}$	0.253	0.255	0.248	0.250	–^{a}	0.249	0.251	0.249	0.250	–^{a}
		(0.044)	(0.031)	(0.014)	(0.009)	–^{a}	(0.026)	(0.018)	(0.008)	(0.005)	–^{a}
	L_{full}	0.250	0.255	0.248	0.250	0.250	0.246	0.251	0.249	0.250	0.250
		(0.043)	(0.031)	(0.014)	(0.009)	(0.005)	(0.026)	(0.018)	(0.008)	(0.005)	(0.003)
0.5	L_4	0.251	0.256	0.248	0.251	0.250	0.248	0.252	0.249	0.250	0.250
		(0.043)	(0.031)	(0.013)	(0.008)	(0.004)	(0.025)	(0.018)	(0.008)	(0.005)	(0.003)
	$L_{\emptyset}$	0.257	0.255	0.248	0.251	–^{a}	0.252	0.252	0.249	0.250	–^{a}
		(0.045)	(0.030)	(0.013)	(0.008)	–^{a}	(0.026)	(0.018)	(0.008)	(0.005)	–^{a}
	L_{full}	0.250	0.255	0.248	0.251	0.250	0.247	0.251	0.249	0.250	0.250
		(0.042)	(0.030)	(0.013)	(0.008)	(0.004)	(0.025)	(0.018)	(0.008)	(0.005)	(0.003)
0.7	L_4	0.253	0.256	0.248	0.251	0.251	0.250	0.253	0.249	0.250	0.250
		(0.042)	(0.029)	(0.013)	(0.008)	(0.004)	(0.025)	(0.019)	(0.008)	(0.004)	(0.003)
	$L_{\emptyset}$	0.259	0.255	0.248	0.251	–^{a}	0.258	0.252	0.249	0.250	–^{a}
		(0.043)	(0.029)	(0.013)	(0.008)	–^{a}	(0.025)	(0.018)	(0.008)	(0.004)	–^{a}
	L_{full}	0.249	0.255	0.248	0.251	0.251	0.247	0.251	0.249	0.250	0.250
		(0.041)	(0.029)	(0.012)	(0.008)	(0.004)	(0.024)	(0.018)	(0.008)	(0.004)	(0.003)
0.9	L_4	0.260	0.258	0.248	0.251	0.251	0.258	0.256	0.249	0.251	0.250
		(0.041)	(0.029)	(0.013)	(0.008)	(0.004)	(0.024)	(0.020)	(0.008)	(0.005)	(0.003)
	$L_{\emptyset}$	0.257	0.253	0.247	0.251	–^{a}	0.253	0.251	0.249	0.251	–^{a}
		(0.040)	(0.027)	(0.013)	(0.008)	–^{a}	(0.022)	(0.018)	(0.008)	(0.005)	–^{a}
	L_{full}	0.249	0.253	0.248	0.251	0.251	0.248	0.251	0.249	0.251	0.250
		(0.039)	(0.027)	(0.012)	(0.008)	(0.004)	(0.023)	(0.018)	(0.008)	(0.004)	(0.003)

Table 7 As for Table 1 but for estimates of the standard deviation σ_2

		$m = 20$					$m = 50$				
n_c		5	10	100	1,000	100,000	5	10	100	1,000	100,000
$\rho_0 = 0.0$	L_4	0.251	0.251	0.250	0.251	0.251	0.251	0.251	0.250	0.250	0.251
		(0.038)	(0.029)	(0.013)	(0.009)	(0.005)	(0.023)	(0.019)	(0.008)	(0.005)	(0.003)
	$L_{\emptyset}$	0.253	0.251	0.250	0.251	–^{a}	0.253	0.251	0.250	0.250	–^{a}
		(0.039)	(0.029)	(0.013)	(0.009)	–^{a}	(0.023)	(0.019)	(0.008)	(0.005)	–^{a}
	L_{full}	0.251	0.251	0.250	0.251	0.251	0.251	0.251	0.250	0.250	0.251
		(0.038)	(0.028)	(0.013)	(0.009)	(0.005)	(0.023)	(0.019)	(0.008)	(0.005)	(0.003)
0.3	L_4	0.250	0.247	0.250	0.251	0.251	0.254	0.250	0.251	0.251	0.250
		(0.032)	(0.026)	(0.011)	(0.009)	(0.004)	(0.024)	(0.019)	(0.007)	(0.005)	(0.003)

Table 7 continued

		$m = 20$					$m = 50$				
n_c		5	10	100	1,000	100,000	5	10	100	1,000	100,000
	$L_{\emptyset}$	0.254	0.247	0.250	0.251	–[a]	0.257	0.250	0.251	0.251	–[a]
		(0.033)	(0.026)	(0.011)	(0.009)	–[a]	(0.025)	(0.020)	(0.007)	(0.005)	–[a]
	L_{full}	0.250	0.246	0.250	0.251	0.251	0.253	0.250	0.251	0.251	0.250
		(0.033)	(0.026)	(0.011)	(0.009)	(0.004)	(0.024)	(0.019)	(0.007)	(0.005)	(0.003)
0.5	L_4	0.252	0.247	0.250	0.251	0.251	0.255	0.250	0.251	0.251	0.250
		(0.033)	(0.027)	(0.011)	(0.009)	(0.004)	(0.024)	(0.020)	(0.007)	(0.005)	(0.002)
	$L_{\emptyset}$	0.257	0.247	0.250	0.251	–[a]	0.259	0.250	0.251	0.251	–[a]
		(0.034)	(0.027)	(0.011)	(0.009)	–[a]	(0.025)	(0.020)	(0.007)	(0.005)	–[a]
	L_{full}	0.250	0.247	0.250	0.251	0.251	0.253	0.250	0.251	0.251	0.250
		(0.033)	(0.027)	(0.011)	(0.009)	(0.004)	(0.024)	(0.020)	(0.007)	(0.005)	(0.002)
0.7	L_4	0.254	0.249	0.250	0.251	0.251	0.257	0.252	0.251	0.251	0.251
		(0.033)	(0.028)	(0.011)	(0.008)	(0.004)	(0.024)	(0.020)	(0.007)	(0.005)	(0.003)
	$L_{\emptyset}$	0.260	0.248	0.250	0.251	–[a]	0.264	0.251	0.251	0.251	–[a]
		(0.035)	(0.028)	(0.011)	(0.008)	–[a]	(0.024)	(0.020)	(0.007)	(0.005)	–[a]
	L_{full}	0.250	0.247	0.250	0.251	0.251	0.253	0.250	0.251	0.251	0.251
		(0.033)	(0.027)	(0.011)	(0.008)	(0.004)	(0.024)	(0.020)	(0.007)	(0.005)	(0.003)
0.9	L_4	0.260	0.253	0.251	0.252	0.251	0.262	0.255	0.251	0.251	0.251
		(0.036)	(0.030)	(0.011)	(0.008)	(0.004)	(0.024)	(0.021)	(0.007)	(0.005)	(0.003)
	$L_{\emptyset}$	0.257	0.248	0.249	0.252	–[a]	0.257	0.250	0.250	0.251	–[a]
		(0.035)	(0.028)	(0.011)	(0.008)	–[a]	(0.022)	(0.019)	(0.007)	(0.005)	–[a]
	L_{full}	0.249	0.248	0.250	0.252	0.251	0.252	0.250	0.251	0.251	0.251
		(0.034)	(0.028)	(0.011)	(0.008)	(0.004)	(0.023)	(0.019)	(0.007)	(0.005)	(0.003)

B.2 Estimates of $(\sigma_1, \rho, \sigma_2)$, from Section 3.2

See Tables 8 and 9.

Table 8 As for Table 2 but with $\rho_0 = -0.7$

		$n_c = 60$			$n_c = 300$		
	Orders (l,u)	σ_1	ρ	σ_2	σ_1	ρ	σ_2
$L_{sn,x}$	((6, 5), (55, 35))	0.4974	−0.6912	0.5106	0.4998	−0.6596	0.5040
		(0.0124)	(0.2625)	(0.0472)	(0.0060)	(0.2790)	(0.0410)
	((16,6), (45,24))	0.4986	−0.6933	0.5289	0.4994	−0.6606	0.5144
		(0.0164)	(0.1854)	(0.0949)	(0.0075)	(0.2146)	(0.0856)
	((20,5), (41,16))	0.5004	−0.6699	0.5231	0.4993	−0.6790	0.5201
		(0.0184)	(0.1963)	(0.0987)	(0.0080)	(0.1753)	(0.0919)

Table 8 continued

		$n_c = 60$			$n_c = 300$		
	Orders (l,u)	σ_1	ρ	σ_2	σ_1	ρ	σ_2
$L_{sn,y}$	((5,6), (35, 55))	0.4979	−0.6423	0.4993	0.5006	−0.6405	0.4984
		(0.0394)	(0.2486)	(0.0148)	(0.0364)	(0.2746)	(0.0061)
	((6, 16), (24, 45))	0.5060	−0.6447	0.4981	0.5223	−0.6726	0.4985
		(0.0859)	(0.2168)	(0.0162)	(0.0910)	(0.2231)	(0.0078)
	(5, 20), (16, 41))	0.5054	−0.6396	0.4991	0.5141	−0.6451	0.4981
		(0.0999)	(0.1981)	(0.0206)	(0.1018)	(0.2205)	(0.0101)
$L_{is,x}$	((6,3), (55, 3))	0.4975	−0.7133	0.4929	0.4999	−0.7133	0.4896
		(0.0121)	(0.0497)	(0.0393)	(0.0060)	(0.0472)	(0.0320)
	((16, 10), (45, 2))	0.4987	−0.7325	0.4966	0.4994	−0.7215	0.4983
		(0.0162)	(0.0932)	(0.0277)	(0.0075)	(0.1051)	(0.0248)
	((20, 7), (41, 14))	0.5004	−0.7108	0.4869	0.4993	−0.7128	0.4771
		(0.0180)	(0.0363)	(0.0444)	(0.0080)	(0.0275)	(0.0453)
$L_{is,y}$	((3, 6), (3, 55))	0.4900	−0.7130	0.4993	0.4915	−0.7127	0.4984
		(0.0288)	(0.0517)	(0.0147)	(0.0326)	(0.0447)	(0.0061)
	((10, 16), (2, 45))	0.4915	−0.7327	0.4982	0.4955	−0.7284	0.4985
		(0.0228)	(0.1020)	(0.0163)	(0.0238)	(0.0999)	(0.0077)
	((7, 20), (14, 41))	0.4802	−0.7155	0.4990	0.4850	−0.7096	0.4981
		(0.0424)	(0.0335)	(0.0205)	(0.0401)	(0.0253)	(0.0101)

Table 9 As for Table 2 but with $\rho_0 = 0$

		$n_c = 60$			$n_c = 300$		
	Orders (l, u)	σ_1	ρ	σ_2	σ_1	ρ	σ_2
$L_{sn,x}$	((6,5), (55, 35))	0.4980	0.0183	0.5235	0.4998	−0.0191	0.5216
		(0.0126)	(0.4156)	(0.0322)	(0.0059)	(0.3888)	(0.0301)
	((16,6), (45, 24))	0.4968	0.0670	0.5329	0.5001	−0.0172	0.5307
		(0.0157)	(0.3490)	(0.0612)	(0.0076)	(0.3375)	(0.0572)
	((20,5), (41, 16))	0.4957	0.0847	0.5394	0.4990	−0.0018	0.5355
		(0.0186)	(0.3747)	(0.0551)	(0.0099)	(0.3671)	(0.0508)
$L_{sn,y}$	((5,6), (35, 55))	0.5252	0.0024	0.4983	0.5235	0.0261	0.4995
		(0.0412)	(0.4303)	(0.0142)	(0.0306)	(0.4018)	(0.0058)
	((6, 16), (24, 45))	0.5382	−0.0048	0.4983	0.5359	0.0240	0.4986
		(0.0532)	(0.3863)	(0.0151)	(0.0558)	(0.3647)	(0.0063)
	((5, 20), (16, 41))	0.5343	−0.0024	0.5005	0.5434	0.0080	0.4984
		(0.0586)	(0.3645)	(0.0174)	(0.0569)	(0.3855)	(0.0089)

Table 9 continued

		$n_c = 60$			$n_c = 300$		
	Orders (l, u)	σ_1	ρ	σ_2	σ_1	ρ	σ_2
$L_{is,x}$	((6,3), (55, 3))	0.4980	−0.0048	0.4856	0.4998	0.0008	0.4838
		(0.0126)	(0.0519)	(0.0546)	(0.0059)	(0.0205)	(0.0584)
	((16, 10), (45, 2))	0.4968	−0.0353	0.4777	0.5001	−0.0260	0.4881
		(0.0157)	(0.0828)	(0.0520)	(0.0076)	(0.0653)	(0.0514)
	((20,7), (41, 14))	0.4957	0.0214	0.4846	0.4990	0.0184	0.4829
		(0.0186)	(0.0618)	(0.0547)	(0.0099)	(0.0566)	(0.0527)
$L_{i,sy}$	((3, 6), (3, 55))	0.4830	0.0074	0.4984	0.4775	0.0004	0.4995
		(0.0538)	(0.0524)	(0.0141)	(0.0516)	(0.0272)	(0.0058)
	((10,16), (2, 45))	0.5006	−0.0055	0.4984	0.4762	−0.0391	0.4986
		(0.0491)	(0.0752)	(0.0151)	(0.0563)	(0.0743)	(0.0063)
	((7, 20), (14, 41))	0.4804	0.0270	0.5005	0.4852	0.0163	0.4984
		(0.0577)	(0.0697)	(0.0174)	(0.0494)	(0.0525)	(0.0089)

References

Andrieu C, Roberts GO (2009) The pseudo-marginal approach for efficient Monte Carlo computations. Ann Stat 37:697–725

Bardenet R, Doucet A, Holmes C (2014) Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach. In: Proceedings of the 31st international conference on machine learning (ICML-14), pp 405–413

Billard L (2011) Brief overview of symbolic data and analytic issues. Stat Anal Data Min 4:149–156

Billard L, Diday E (2003) From the statistics of data to the statistics of knowledge: symbolic data analysis. J Am Stat Assoc 98:470–487

Billard L, Diday E (2006) Symbolic data analysis. Wiley Series in Computational Statistics. Wiley, Chichester

Bland M (2015) Estimating mean and standard deviation from the sample size, three quartiles, minimum and maximum. Int J Stat Med Res 4:57–64

Bock HH, Diday E (eds) (2000) Analysis of symbolic data. Springer, Berlin

Brito P, Duarte Silva AP (2012) Modelling interval data with normal and skew-normal distributions. J Appl Stat 39:3–20

Cariou V, Billard L (2015) Generalization method when manipulating relational databases. In: Brito P, Venturini G (eds) Symbolic data analysis & visualisation, RNTI-E-29, pp 59–88

Dias S, Brito P (2015) Linear regression model with histogram-valued variables. Stat Anal Data Min 8:75–113

Dias S, Brito P (2017) Off the beaten track: a new linear model for interval data. Eur J Oper Res 258(3):1118–1130

Diday E (1988) The symbolic approach in clustering and related methods of data analysis: the basic choices. In: Brock HH (ed) Classification and related methods of data analysis, proceedings of IFCS87, pp 673–684

Duarte Silva AP, Brito P (2015) Discriminant analysis of interval data: an assessment of parametric and distance-based approaches. J Classif 32:516–541

Fisher R, O’Leary RA, Low-Choy S, Mengersen K, Knowlton N, Brainard RE, Caley MJ (2015) Species richness on coral reefs and the pursuit of convergent global estimates. Curr Biol 25:500–505

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) Bayesian data analysis, 3rd edn. Chapman and Hall, Boca Raton

Guha S, Hafen R, Rounds J, Xia J, Li J, Xi B, Cleveland WS (2012) Large complex data: divide and recombine (D&R) with RHIPE. Stat 1:53–67

Heitjan DF, Rubin DB (1991) Ignorability and coarse data. Ann Stat 19:2244–2253

Hozo SP, Djulbegovic B, Hozo I (2005) Estimating the mean and variance from the median, range and the size of a sample. BMC Med Res Methodol 5:13

Hron K, Brito P, Filzmoser P (2017) Exploratory data analysis for interval compositional data. Adv Data Anal Class 11:223–241

Ichino M (2011) The quantile method for symbolic principal component analysis. Stat Anal Data Min 4:184–198

Ioannidis Y (2003) The history of histograms (abridged). In: Freytag JC, Lockemann P, Abiteboul S, Carey M, Selinger P, Heuer A (eds) Proceedings of the VLDB conferences. Morgan Kaufmann, pp 19–30

Irpino A, Verde R (2015) Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein distance. Adv Data Anal Classif 9:81–106

Jordan MI, Lee JD, Yang Y (2019) Communication-efficient distributed statistical inference. J Am Stat Assoc 114:668–681

Kosmelj K, Le-Rademacher J, Billard L (2014) Symbolic covariance matrix for interval-valued variables and its application to principal component analysis: a case study. Metod Zvezki 11:1–20

Le-Rademacher J, Billard L (2011) Likelihood functions and some maximum likelihood estimators for symbolic data. J Stat Plan Inference 141:1593–1602

Le-Rademacher J, Billard L (2013) Principal component analysis for histogram-valued data. Advances in data analysis and classification, pp 1–25

Lin H, Caley MJ, Sisson SA (2022) Estimating global species richness using symbolic data meta-analysis. Ecography 2022:e05617

Lin W, González-Rivera G (2016) Interval-valued time series models: estimation based on order statistics exploring the Agriculture Marketing Service data. Comput Stat Data Anal 100:694–711

Luo D, Wan X, Liu J, Tong T (2018) Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range. Stat Methods Med Res 27:1785–1805

McLachlan GJ, Jones PN (1988) Fitting mixture models to grouped and truncated data via the EM algorithm. Biometrics 44:571–578

Mousavi H, Zaniolo C (2011) Fast and accurate computation of equi-depth histograms over data streams. In: Proceedings of the 14th international conference on extending database technology, pp 69–80

Neto EAL, Corderio GM, de Carvalho FAT (2011) Bivarite symbolic regression models for interval-valued variables. J Stat Comput Simul 81:1727–1744

Noirhomme-Fraiture M, Brito P (2011) Far beyond the classical data models: symbolic data analysis. Stat Anal Data Min 4:157–170

Quiroz M, Tran MN, Villani M, Kohn R (2018) Speeding up MCMC by delayed acceptance and data subsampling. J Comput Graph Stat 27:12–22

Quiroz M, Kohn R, Villani M, Tran MN (2019) Speeding up mcmc by efficient data subsampling. J Am Stat Assoc 114(526):831–843

Rahman P, Beranger B, Sisson S, Roughan M (2022) Likelihood-based inference for modelling packet transit from thinned flow summaries. IEEE Trans Signal Inf Process Netw 8:571–583. https://doi.org/10.1109/TSIPN.2022.3188457

Rendell LJ, Johansen AM, Lee A, Whiteley N (2020) Global consensus Monte Carlo. J Comput Graph Stat 30:1–29

Rodrigues GS, Nott DJ, Sisson SA (2016) Functional regression approximate Bayesian computation for Gaussian process density estimation. Comput Stat Data Anal 103:229–241

Rubin DB (1981) Estimation in parallel randomised experiments. J Educ Stat 6:377–401

Schweizer B (1984) Distributions are the numbers of the future. In: Proceedings of the mathematics of fuzzy systems, pp 137–149

Shi J, Luo D, Weng H, Zeng XT, Lin L, Tong T (2018) How to estimate the sample mean and standard deviation from the five number summary? arXiv:1801.01267

Sisson SA, Fan Y, Beaumont MA (eds) (2018) Handbook of approximate bayesian computation. Chapman & Hall, Boca Raton

Vardeman SB, Lee CS (2005) Likelihood-based statistical estimation from quantised data. IEEE Trans Instrum Meas 54:409–414

Vono M, Dobigeon N, Chainais P (2019) Split-and-augmented Gibbs sampler—application to large-scale inference problems. IEEE Trans Signal Process 67(6):1648–1661

Wan X, Wang W, Liu J, Tong T (2014) Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC Med Res Methodol 14:135

Whitaker T, Beranger B, Sisson SA (2020) Composite likelihood methods for histogram-valued random variables. Stat Comput 30:1459–1477

Whitaker T, Beranger B, Sisson SA (2021) Logistic regression models for aggregated data. J Comput Graph Stat 30:1049–1067

Zhang X, Beranger B, Sisson SA (2020) Constructing likelihood functions for interval-valued random variables. Scand J Stat 47(1):1–35

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.