Received: 19 January 2018 | Revised: 25 February 2019 | Accepted: 27 March 2019

DOI: 10.1111/sjos.12395

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Constructing likelihood functions for interval-valued random variables

X. Zhang | B. Beranger | S. A. Sisson

School of Mathematics and Statistics,
University of New South Wales,
Sydney, Australia

Correspondence
S. A. Sisson, School of Mathematics and Statistics, University of New South Wales, Sydney NSW 2052, Australia.
Email: Scott.Sisson@unsw.edu.au

Funding information
China Scholarship Council; Australian Research Council, Grant/Award Number: FT170100079; Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers of Big Data, Big Models, New Insights, Grant/Award Number: CE140100049

Abstract

There is a growing need for flexible methods to analyze interval-valued data, which can provide efficient data representations for very large data sets. However, the existing descriptive frameworks to achieve this ignore the process by which interval-valued data are typically constructed, namely, by the aggregation of real-valued data generated from some underlying process. In this paper, we develop the foundations of likelihood-based statistical inference for intervals that directly incorporates the underlying data generating procedure into the analysis. That is, it permits the direct fitting of models for the underlying real-valued data given only the interval-valued summaries. This generative approach overcomes several problems associated with existing methods, including the rarely satisfied assumption of within-interval uniformity. The new methods are illustrated by simulated and real data analyses.

KEYWORDS

aggregate data, interval-valued data, likelihood theory, symbolic data analysis

1 | INTRODUCTION

As we move inevitably toward a more data-centric society, there is a growing need for the ability to analyze data that are constructed in nonstandard forms, rather than represented as continuous points in $\mathbb{R}^p$ (Billard & Diday, 2003). The simplest and most popular of these is interval-valued data.

Interval-valued observations can arise naturally through the data recording process and essentially result as a way to characterize the measurement error or uncertainty of an observation. Examples include blood pressure, which is typically reported as an interval due to the inherent continual changes within an individual (Billard & Diday, 2006); data quantization, such as rounding or truncation, which results in observations being known to lie within some interval (McLachlan & Jones, 1988; Vardeman & Lee, 2005); and the expression of expert-elicited intervals that contain some quantity of interest (Lin, Caley, & Sisson, 2017; Fisher et al., 2015), among others.

The use of intervals as a summary representation of a collection of classical real-valued data is also rapidly gaining traction. Here, the aggregation of a large and complex data set into a smaller collection of suitably constructed intervals can enable a statistical analysis that would otherwise be computationally unviable (Billard & Diday, 2003). Where interest in the outcome of an analysis exists at the level of a group, rather than at an individual level, interval-valued data provide a convenient group-level aggregation device (Neto & de Carvalho, 2010; Noirhomme-Fraiture & Brito, 2011). Similarly, aggregation of individual observations within an interval structure allows for some preservation of privacy for the individual (Domingues, de Souza, & Cysneiros, 2010).

The earliest systematic study of interval-valued data is in numerical analysis, where Moore (1966) used intervals as a description for imprecise data. Random intervals are also special cases of random sets (Molchanov, 2005), the theory of which brings together elements of topology, convex geometry, and probability theory to develop a coherent mathematical framework for their analysis. Matheron (1975) gave the first self-contained development of statistical models for random sets, including central limit theorems and the law of large numbers, and Beresteanu and Molinari (2008) derived these limit theorems specifically for random intervals. In this framework, interval-valued random variables $[X] = [\underline{X}, \overline{X}] \subset \mathbb{R}$ are modeled as a bivariate real-valued random vector $(\underline{X}, \overline{X})$, where $\underline{X} \leq \overline{X}$, using standard inferential techniques. This approach is also used for partially identified models, where the object of economic and statistical interest is a set rather than a point (Beresteanu, Molchanov, & Molinari, 2012; Molchanov & Molinari, 2014). In probabilistic modeling, Lyashenko (1983) introduced normal random compact convex sets in Euclidean space and showed that a normal random interval is simply a Gaussian displacement of a fixed closed bounded interval. Sun and Ralescu (2015) subsequently extended this idea to normal hierarchical models for random intervals.

A more popular framework for the analysis of interval-valued data, and one which we focus on here, is symbolic data analysis (Billard & Diday, 2006). Symbols can be considered as distributions of real-valued data points in $\mathbb{R}^p$, such as intervals and histograms, or more general structures including lists. They are typically constructed as the aggregation into summary form of real-valued data within some group, and so, the symbol is interpreted as taking values as described by the summary distribution. As a result, symbols have internal variations and structures that do not exist in real-valued data, and methods for analyzing them must account for within-symbol variation in addition to between-symbol variation. In practice, the most common form of symbol is the interval or its p-dimensional extension, the p-hyper-rectangle. See Billard and Diday (2003, 2006) and Noirhomme-Fraiture and Brito (2011) for a review of recent results.

While many exploratory and descriptive data analysis techniques for symbolic data have been developed (see, e.g., Billard & Diday, 2006 for an overview), there is paucity of results for developing a robust statistical inferential framework for these data. The most significant of these (Le-Rademacher & Billard, 2011) maps the parameterization of the symbol into a real-valued random vector and then uses the standard likelihood framework to specify a suitable model. In the random interval setting, this is equivalent to the random set theory approach, which models the interval-valued random variables $[X] = [\underline{X}, \overline{X}] \subset \mathbb{R}$ by the constrained real-valued random vector $(\underline{X}, \overline{X}) \in \mathbb{R}^2$ or, more commonly, a reparameterization to the unconstrained interval center and half-range $(X_c, X_r) = ((\underline{X} + \overline{X})/2, (\overline{X} - \underline{X})/2)$, which is then more easily modeled, for example, $(X_c, \log X_r) \sim N_2(\mu, \Sigma)$. This likelihood framework has been used for analysis of variance (Brito & Duarte Silva, 2012), time-series forecasting (Arroyo, Espínola, & Maté, 2011), and interval-based regression models (Xu, 2010), among others.

While sensible, by nature, the above methods for modeling real-valued random variables only permit descriptive modeling at the level of the real-valued random vector $(\underline{X}, \overline{X})$ (or its equivalent for p-hyper-rectangles). However, this descriptive approach completely ignores the data generating procedure commonly assumed and implemented for the construction of observed intervals, namely, that the underlying real-valued data are produced from some data generating model $X_1, \dots, X_m \sim f(x_1, \dots, x_m \mid \alpha)$, and the interval is then constructed via some aggregation process, for example, $\underline{X} = \min\{X_k\}$ and $\overline{X} = \max\{X_k\}$. If interest is then in fitting the underlying data generating model $f(x_1, \dots, x_m \mid \alpha)$ for inferential or predictive purposes, while only observing interval-valued data rather than the underlying real-valued data set, or in having the interpretation of model parameters be independent of the form of the interval construction process, then the above descriptive models will be inadequate. Furthermore, the existing descriptive models for random intervals typically assume that the distribution of latent data points within the interval is uniform. Under the above data generating procedure, except in specific cases, this will almost always be untrue. This assumption is generally accepted as false in practice, but is typically ignored.

In this paper, we develop the methodological foundations of statistical models for interval-valued data that are directly constructed from an assumed underlying data generating model $f(x_1, \dots, x_m \mid \alpha)$ and a data aggregation function $\varphi(\cdot)$ that maps the space of real-valued data to the space of intervals.

To the best of our knowledge, this represents the first attempt to move beyond the restrictive descriptive models, which are prevalent in the literature, and provide an inferential framework that aligns with the generative interval construction process that is typical in practice. In addition to providing more directly interpretable parameters, it also provides a natural mechanism for departure from the uncomfortable uniformity-within-intervals assumption of descriptive models.

In Section 2, after establishing the containment distribution function, $F_{[X]}(\cdot)$, for random intervals $[X]$ based on the idea of containment functionals (Molchanov, 2005), we demonstrate the one-to-one mapping between $F_{[X]}(\cdot)$ and $f_{[X]}(\cdot)$, which is the density function of the bivariate real-valued random vector $(\underline{X}, \overline{X})$, thereby lending some support to the current best practice for modeling random intervals. All proofs are provided in the Appendix. In Section 3, these results naturally lead to the construction of likelihood functions for generative models that are directly constructed from likelihood functions for the underlying real-valued data. We demonstrate the recovery of existing results on the distribution of the order statistics of a random sample under certain conditions. We are also able to show that a limiting case of the derived generative models results in a valid descriptive model in the sense of Le-Rademacher and Billard (2011), implying that existing descriptive models, in fact, have a direct interpretation in terms of an underlying generative model.

All results are naturally extended from intervals to p-hyper-rectangles in Section 4. In Section 5, we contrast the performance of generative and descriptive models for interval-valued random variables on both simulated data and for a reanalysis of a credit card data set previously examined by Brito and Duarte Silva (2012). Here, we establish that the use of existing descriptive models to analyze interval-valued data constructed under a data generating process (which is typical in practice) can result in misinterpretable and biased parameter estimates and poorer overall fits to the observed interval-valued data than those obtained under generative models. We also examine the robustness of the generative model to model and aggregation function mis-specification. Finally, we conclude with a discussion.

2 | DISTRIBUTIONS OF RANDOM INTERVALS

We first investigate the distribution for a random (closed) interval $[X] = [\underline{X}, \overline{X}]$ defined on the space of $\mathbb{I} = \{[x, y] : -\infty < x \leq y < +\infty\}$. The current practice of constructing models for $[X]$ is by constructing models for the two real-valued random variables $\underline{X}$ and $\overline{X}$ with $\underline{X} \leq \overline{X}$ (Le-Rademacher & Billard, 2011). We term this approach the *descriptive model*.

Throughout this paper, we only consider closed intervals (hyper-rectangles). Results for other types of intervals (hyper-rectangles) can be obtained in a similar way. We denote a vector of $m < \infty$ real-valued random variables by $X_{1:m} = (X_1, \dots, X_m)'$, where $X_k \in \mathbb{R}$ for $k = 1, \dots, m$, and x_k is a realization of X_k. We can then define a data aggregation function $\varphi : \mathbb{R}^m \mapsto \mathbb{I}$ that maps a vector $X_{1:m}$ to the space of intervals $\mathbb{I}$ via $[X] = \varphi(X_{1:m})$, so that $[X]$ is a random (closed) interval. For example, a useful specification for random intervals might construct the bivariate real-valued random variable $(\underline{X}, \overline{X})$ from the minimum $(\underline{X})$ and maximum $(\overline{X})$ of the components of $X_{1:m}$.

2.1 | Descriptive models

A descriptive model treats $[X] = [\underline{X}, \overline{X}]$ as a bivariate real-valued random variable $(\underline{X}, \overline{X})$ with $\underline{X} \leq \overline{X}$. We write $f_{[X]}(\underline{x}, \overline{x} \mid \alpha) = f(\underline{x}, \overline{x} \mid \alpha)$ as the likelihood function of $(\underline{X}, \overline{X})$, where $f(\underline{x}, \overline{x} \mid \alpha)$ is a valid density function and α denotes the parameter vector of interest. Rather than construct models directly on $(\underline{X}, \overline{X})$ with the awkward constraint $\underline{X} \leq \overline{X}$, a simpler approach is to remove this constraint through reparameterization. For example, defining the interval center $X_c = \frac{\underline{X}+\overline{X}}{2}$ and half-range $X_r = \frac{\overline{X}-\underline{X}}{2}$, we obtain $f_{[X]}(\underline{x}, \overline{x} \mid \alpha) = \frac{1}{2} g(\frac{\underline{x}+\overline{x}}{2}, \frac{\overline{x}-\underline{x}}{2} \mid \alpha)$, where $g(x_c, x_r \mid \alpha)$ is a density function for X_c and X_r.

Most existing methods to model random intervals (e.g., Arroyo et al., 2011; Le-Rademacher & Billard, 2011; Brito & Duarte Silva, 2012) can be classified as descriptive models. Their interpretation is simple, and they are convenient to use. However, by construction, they are only models for interval endpoints and, as a consequence, have limitations in providing information about the distribution of the latent data points $X_{1:m}$.

In both symbolic data analysis (Billard & Diday, 2006) and theory of random sets (Molchanov, 2005), the distribution of $[X]$ can be uniquely identified by a density function for bivariate real-valued random variables, that is, $f(\underline{x}, \overline{x})$ with $\underline{x} \leq \overline{x}$.

2.2 | Containment distribution functions

In the theory of random sets, two types of functionals, the capacity functional and the containment functional, are commonly used to identify a unique distribution for random sets. For random intervals, the capacity functional and the containment functional are $T_{[X]}([x]) = P([X] \cap [x])$ and $C^{\star}_{[X]}([x]) = P([X] \subset [x])$, respectively.

In the present setting, we consider a variant of the containment functional, $C_{[X]}([x]) = P([X] \subseteq [x])$, which is more convenient for model construction. Due to its similarity to $C^{\star}_{[X]}(\cdot)$ in both functionality and interpretation, we still refer to $C_{[X]}(\cdot)$ as the containment functional throughout this paper.

Similar to $C^{\star}_{[X]}(\cdot)$, a *containment functional* of a random interval $[X]$ is a functional $C_{[X]} : \mathbb{I} \mapsto [0, 1]$ having the following properties:

(i) $C_{[X]}([\underline{x}, \overline{x}]) \to 1$, when $\underline{x} \to -\infty$ and $\overline{x} \to +\infty$;

(ii) if $[x_1] \supseteq [x_2] \supseteq \cdots \supseteq [x_n] \supseteq \cdots$ and $\cap_{n=1}^{\infty}[x_n] \in \mathbb{I}$, then

$$\lim_{n\to\infty} C_{[X]}([x_n]) = C_{[X]}\left(\cap_{n=1}^{\infty}[x_n]\right);$$

(iii) for any $[x] \subseteq [y]$, $C_{[X]}([x]) \leq C_{[X]}([y])$ and

$$C_{[X]}([y]) - C_{[X]}\left(\left[\underline{y},\overline{x}\right]\right) - C_{[X]}\left(\left[\underline{x},\overline{y}\right]\right) + C_{[X]}([x]) \geq 0.$$

However, it is more convenient to work with functions defined on the real plane; hence, we equivalently define the *containment distribution function* as $F_{[X]}(\underline{x},\overline{x}) = C_{[X]}([x])$.

Definition 1. The containment distribution function $F_{[X]} : \mathbb{R}^2 \mapsto [0,1]$ of the random interval $[X]$ has the following properties:

(i) $F_{[X]}(-\infty, +\infty) = 1$ and $F_{[X]}(\underline{x},\overline{x}) = 0$ when $\underline{x} > \overline{x}$;
(ii) $F_{[X]}(\underline{x},\overline{x})$ is left-continuous in $\underline{x}$ and right-continuous in $\overline{x}$;
(iii) $F_{[X]}(\underline{x},\overline{x})$ is nonincreasing in $\underline{x}$ and nondecreasing in $\overline{x}$;
(iv) for $\underline{y} \leq \underline{x} \leq \overline{x} \leq \overline{y}$, $F_{[X]}(\underline{y},\overline{y}) - F_{[X]}(\underline{y},\overline{x}) - F_{[X]}(\underline{x},\overline{y}) + F_{[X]}(\underline{x},\overline{x}) \geq 0$.

The containment distribution function of $[X]$ can be obtained by integration of a valid density function for random intervals.

Theorem 1. *Provided that* $f_{[X]} : \mathbb{R}^2 \mapsto \mathbb{R}$ *is the density function of a random interval* $[X]$*, the containment distribution function of* $[X]$ *can be derived as* $F_{[X]}(\underline{x},\overline{x}) = \int_{\underline{x}}^{\overline{x}}\int_{\underline{x}}^{\overline{x}} f_{[X]}(\underline{x}',\overline{x}')\,\mathrm{d}\underline{x}'\mathrm{d}\overline{x}'$.

Conversely, the density function of $[X]$ can be obtained by differentiation of a containment distribution function.

Theorem 2. *Let* $F_{[X]} : \mathbb{R}^2 \mapsto [0,1]$ *be the containment distribution function of a random interval* $[X]$*. If* $F_{[X]}(\cdot)$ *is twice differentiable, then the density function of* $[X]$ *is*

$$f_{[X]}\left(\underline{x},\overline{x}\right) = -\frac{\partial^2}{\partial\underline{x}\partial\overline{x}}F_{[X]}\left(\underline{x},\overline{x}\right). \tag{1}$$

Given the data generating process, $F_{[X]}(\underline{x},\overline{x})$ can be naturally constructed from the generative framework, where $[X] = \varphi(X_{1:m})$, by noting that the two events, $\{\varphi(X_{1:m}) \subseteq [x]\}$ and $\{[X] \subseteq [x]\}$, are equal. If φ is measurable, we may compute the probability of $\{[X] \subseteq [x]\}$ via $P(\varphi(X_{1:m}) \subseteq [x])$, given the distribution of latent data points $X_{1:m}$. Accordingly, the containment distribution function of $[X]$ can be constructed as

$$F_{[X]}\left(\underline{x},\overline{x}\right) = P(\varphi(X_{1:m}) \subseteq [x]). \tag{2}$$

Note that $[X]$ degenerates to a scalar random variable when it only contains a single point, that is, when $\underline{X} = \overline{X} = X$, and so, $P([X] \subseteq [x]) = P(X \in [x])$ identifies the distribution of a univariate real-valued random variable. In the generative framework, a univariate real-valued random variable is produced either when $m = 1$ or when $X_1 = \cdots = X_m = X$ for $m > 1$. Accordingly, this theory for random intervals is consistent with standard statistical theory. For the following sections, we assume that the data aggregation function $\varphi(\cdot)$ is always measurable.

2.3 | Density functions

We can formally establish the distribution of random intervals by constructing a measurable space of $\mathbb{I}$.

Theorem 3. *The containment distribution function* $F_{[X]}$ *determines a unique distribution of* $[X]$*, such that* $P([X] \subseteq [x]) = F_{[X]}(\underline{x},\overline{x})$ *for all* $[x] \in \mathbb{I}$.

From the above, $1 - F_{[X]}(\underline{x}, +\infty)$ and $F_{[X]}(-\infty, \overline{x})$ are the marginal cumulative distribution functions of the lower bound $\underline{X}$ and the upper bound $\overline{X}$, respectively.

The density function of $[X]$ is formally defined as the Radon–Nikodym derivative (Durrett, 2010) of a probability measure on $\mathbb{I}$ over the uniform measure as the reference measure, as described in Theorem 2.

Note that a valid density function of $[X]$ is also a density function for a bivariate real-valued random variable. Being able to express the density function $f_{[X]}(\underline{x}, \overline{x})$ of the random interval $[X]$ as the joint density of two real-valued random variables, $\underline{X}$ and $\overline{X}$, justifies those existing (descriptive) methods for modeling random intervals (e.g., Arroyo et al., 2011; Le-Rademacher and Billard, 2011; Brito & Duarte Silva, 2012—see Section 2.1) that directly specify a joint distribution for $\underline{X}, \overline{X} \mid \underline{X} \leq \overline{X}$, or some reparameterization that circumvents bounding the parameter space.

3 | GENERATIVE MODELS

One approach for constructing models for $[X]$ is by constructing models for the two real-valued random variables $\underline{X}$ and $\overline{X}$ with $\underline{X} \leq \overline{X}$, that is, descriptive models. While it can describe the structure and variation between intervals, it is unable to model the distribution of latent data points within an interval, as it is simply a model for the interval endpoints. This approach is almost universal in the symbolic data analysis literature. As an alternative, we develop the *generative model*, which is constructed directly at the level of the latent data points $X_{1:m}$ through the data aggregation function $\varphi(\cdot)$. In the following, we use $F_{[X]}(\cdot)$ and $f_{[X]}(\cdot)$ for interval-valued random variables and $F(\cdot)$ and $f(\cdot)$ for real-valued random variables.

A generative model of the random interval may be constructed bottom-up from the distribution of latent data points $X_{1:m}$ and the data aggregation function $\varphi(\cdot)$, based on (2). Here, the random interval $[X]$ is constructed from $X_{1:m}$ and $\varphi(\cdot)$ via $[X] = \varphi(X_{1:m})$. If $f(x_{1:m} \mid \alpha)$ is the likelihood function of the m data points, then, from (2), we may form the containment distribution function of $[X]$ as

$$F_{[X]}\left(\underline{x}, \overline{x} \mid \alpha, m\right) = \int_A f(x_{1:m} \mid \alpha) \mathrm{d}x_{1:m}, \tag{3}$$

where $A = \{\varphi(x_{1:m}) \subseteq [x]\}$ denotes the collection of $x_{1:m}$, for which the corresponding interval is a subset of or equal to $[x]$. If $\varphi(\cdot)$ is continuous, the containment distribution function (3) is twice differentiable, and so, from (1), its contribution to the likelihood function would be

$$f_{[X]}\left(\underline{x}, \overline{x} \mid \alpha, m\right) = -\frac{\partial^2}{\partial \underline{x} \partial \overline{x}} \int_A f(x_{1:m} \mid \alpha) \mathrm{d}x_{1:m}. \tag{4}$$

Note that containment distribution functions (3) and density functions (4) of generative models contain a parameter m specifying the number of latent data points within $[X]$.

When m is large, the evaluation of (4) can be challenging as it involves high-dimensional integration. This integration can be simplified in the case where $X_{1:m}$ is a sequence of independent and identically distributed (i.i.d.) random variables with $X_k \sim f(x \mid \theta)$ for $k = 1, \ldots, m$. We denote the likelihood function of $[X]$ with the i.i.d. latent data points by

$$f^{\star}_{[X]}\left(\underline{x}, \overline{x} \mid \theta, m\right) = -\frac{\partial^2}{\partial \underline{x} \partial \overline{x}} \int_A \prod_{k=1}^{m} f(x_k \mid \theta) \mathrm{d}x_{1:m} \tag{5}$$

and term it the *i.i.d. generative model.*

In practice, the data aggregation function $\varphi(\cdot)$ will typically depend on the order statistics of the latent data points, so that $\varphi_{l,u}(x_{1:m}) = [x_{(l)}, x_{(u)}]$, where $x_{(l)}$ and $x_{(u)}$ are, respectively, the lth (lower) and uth (upper) order statistics of $x_{1:m}$. The region for integration in (3) and (4) then becomes $A = \{x_{1:m} : \underline{x} \leq x_{(l)} \leq x_{(u)} \leq \overline{x}\}$—the collection of $x_{1:m}$ for which the lth-order statistic is no less than $\underline{x}$ and the uth-order statistic is no greater than $\overline{x}$. In this case and for i.i.d. random variables $X_k \sim f(x \mid \theta)$ for $k = 1, \dots, m$, the likelihood function (5) becomes

$$f^{\star}_{[X]}\left(\underline{x}, \overline{x} \mid \theta, m, l, u\right) = \frac{m!}{(l-1)!(u-l-1)!(m-u)!}\left[F\left(\underline{x} \mid \theta\right)\right]^{l-1} \\ \times \left[F\left(\overline{x} \mid \theta\right) - F\left(\underline{x} \mid \theta\right)\right]^{u-l-1}\left[1 - F\left(\overline{x} \mid \theta\right)\right]^{m-u} f\left(\underline{x} \mid \theta\right) f\left(\overline{x} \mid \theta\right), \tag{6}$$

where $F(x \mid \theta) = \int_{-\infty}^{x} f(z \mid \theta)\,\mathrm{d}z$ is the cumulative distribution function of X_k. That is, (5) reduces to (6), which is the joint likelihood function of the lth- and uth-order statistics of m i.i.d. samples. Consequently, if $l/(m+1) \to \underline{p}$ and $u/(m+1) \to \overline{p}$ as $m \to \infty$, the distribution of $[X]$ converges to a point mass at $[Q(\underline{p};\theta), Q(\overline{p};\theta)]$, where $Q(\cdot;\theta)$ is the quantile function of $f(x \mid \theta)$.

Further simplification is possible when $[X]$ is constructed from the minimum and maximum values of $X_{1:m}$ (so that $l = 1$ and $u = m$). Here, $A = \{x_{1:m} : \underline{x} \leq x_k \leq \overline{x}, k = 1, \dots, m\}$ is a hyper-rectangle in $\mathbb{R}^m$ with identical length in each dimension, and the likelihood function (6) becomes

$$f^{\star\star}_{[X]}\left(\underline{x}, \overline{x} \mid \theta, m\right) = m(m-1)\left[F\left(\overline{x} \mid \theta\right) - F\left(\underline{x} \mid \theta\right)\right]^{m-2} f\left(\underline{x} \mid \theta\right) f\left(\overline{x} \mid \theta\right). \tag{7}$$

In this case, if the support of $f(x \mid \theta)$ is bounded on $[a, b]$, then as $m \to \infty$, the distribution of $[X]$ converges to a point mass at $[a, b]$. However, if $f(x \mid \theta)$ has unbounded support, the distribution of $[X]$ will diverge to $(-\infty, +\infty)$.

From the above, we may conclude that, for i.i.d. generative models, when m is large, all interval-valued observations will be similar. As in practice, we may expect significant variation in interval-valued observations, even for a large m value; this indicates that the usefulness of an i.i.d. model may be restricted to specific settings.

3.1 | Hierarchical generative models

Evaluating the likelihood function (4) of the generative model for general latent distributions $f(x_{1:m} \mid \alpha)$ of latent data points is challenging, except in simplified settings. Here, we consider a special class of the generative model for which the latent data points $X_{1:m}$ are exchangeable. This exchangeability leads to a hierarchical generative model, which can capture both inter- and intra-interval structure and variability.

Suppose that $X_{1:m}$ are exchangeable, that is, their joint distribution is invariant to any permutation of $X_{1:m}$. From de Finetti's theorem (Aldous, 1985), the distribution of $X_{1:m}$ may be represented as a mixture, that is,

$$P(X_{1:m} \in A) = \int P_{\star}^{(m)}(X_{1:m} \in A)\, \mu_{P_\star}(\mathrm{d}P_\star), \tag{8}$$

where $\mu_{P_\star}$ is the distribution on the space of all probability measures of $\mathbb{R}$, and $P_{\star}^{(m)} = \prod_m P_\star$ is the product measure on $\mathbb{R}^m$. In other words, all X_k for $k = 1, \dots, m$ are i.i.d. from $P_\star$ with $P_\star \sim \mu_{P_\star}$. By recalling from (3) and (4) that $A = \{\varphi(x_{1:m}) \subseteq [x]\}$, then the mixture component $P_{\star}^{(m)}(X_{1:m} \in A)$ is equal to $P_{\star}^{(m)}([X] \subseteq [x])$, which is the containment distribution function for an i.i.d. generative model of $[X]$, with $X_k \sim P_\star$ for $k = 1, \dots, m$ and the same data aggregation function $\varphi(\cdot)$. This means that $P([X] \subseteq [x])$, which is equal to $P(X_{1:m} \in A)$, may be represented as the mixture of $P_\star^{(m)}([X] \subseteq [x])$ with $P_\star \sim \mu_{P_\star}$, that is, as a mixture of i.i.d. generative models.

In the following, we consider the case when $P_\star$ belongs to some parametric family, so that $\mathrm{d}P_\star = f(x\,|\,\theta)\,\mathrm{d}x$. From (8), the joint density function of $X_{1:m}$ is then given by the mixture representation $\int \prod_{k=1}^{m} f(x_k \mid \theta)\pi(\theta)\mathrm{d}\theta$, where the mixing distribution $\pi(\theta)$ may be nonparametric or parametric $\pi(\theta\,|\,\alpha)$ with parameter α. The resulting containment distribution function of $[X]$ is then the mixture of $F_{[X]}(\underline{x}, \overline{x} \mid \theta, m)$ given in (3), with $f(x_{1:m} \mid \theta) = \prod_{k=1}^{m} f(x_k \mid \theta)$, with respect to (w.r.t.) $\pi(\theta\,|\,\alpha)$. If $\varphi(\cdot)$ is continuous, we obtain the likelihood function of such a generative model as

$$f_{[X]}\left(\underline{x}, \overline{x} \mid \alpha, m\right) = \int f_{[X]}^{\star}\left(\underline{x}, \overline{x} \mid \theta, m\right) \pi(\theta \mid \alpha)\mathrm{d}\theta, \tag{9}$$

where $f_{[X]}^{\star}(\underline{x}, \overline{x} \mid \theta, m)$ is the likelihood function of i.i.d. generative model (5).

In practice, the latent data points $X_{1:m}$ may not be exchangeable. However, the data aggregation function $\varphi(\cdot)$ may be symmetric. Let Γ be the set of all permutations of the indices from 1 to m and X_γ be the latent data points $X_{1:m}$ permuted according to $\gamma \in \Gamma$ with density function $f(x_\gamma)$. As $\varphi(\cdot)$ is symmetric, $\varphi(x_\gamma) = \varphi(x_{1:m})$, and thus, $[X_\gamma] = \varphi(X_\gamma)$ has the same containment distribution function as $[X]$. As a result, for the exchangeable random variables defined as $Y_{1:m} \sim \frac{1}{m!}\sum_{\gamma\in\Gamma} f(X_\gamma)$, $[Y] = \varphi(Y_{1:m})$ has the same containment distribution function as $[X]$.

The existence of such $Y_{1:m}$ implies that when the latent data points $X_{1:m}$ are aggregated into intervals $[X]$ by symmetric data aggregation methods, information on the order-related dependence structure will vanish. As a result, it is unnecessary to model the distribution of $X_{1:m}$ with a more complex dependence structure than exchangeability—modeling the exchangeable $Y_{1:m}$ will be sufficient.

Accordingly, for random intervals $[X_1], \dots, [X_n]$, the generative model (9) can be directly interpreted as the hierarchical model

$$\begin{aligned} [X_i] &= \varphi(X_{i,1:m}), \\ X_{i,k} &\sim f(x \mid \theta_i),\ k = 1, \dots, m_i, \\ \theta_i &\sim \pi(\theta \mid \alpha), \end{aligned}$$

with known m_i for $i = 1, \dots, n$. Thus, we term them *hierarchical generative models*. The contribution to the integrated likelihood (9) for the first two terms is given by $f_{[X]}^{\star}(\underline{x}_i, \overline{x}_i \mid \theta_i, m_i)$—the likelihood function of the i.i.d. generative model (5) for the interval-valued observation $[x_i]$, with the density function of each (conditionally) i.i.d. latent data point $X_{i,1:m_i}$ given by $f(x_{i,k}\,|\,\theta_i)$ and where $\pi(\theta\,|\,\alpha)$ is the mixing distribution for θ_i given the parameter α. Given such interpretation, $f(x_{i,k}\,|\,\theta_i)$ (or θ_i) is the *local* density function (or parameter) for $[X_i]$, whereas $\pi(\theta\,|\,\alpha)$ (or α) is the *global* density function (or parameter) among all intervals. Therefore, the intra-interval structure is described by the local density function and m, whereas the inter-interval variability is modeled by the global density function.

As a result, inference on this model permits direct analysis of the underlying distribution of data points $X_{1:m}$ within each interval $[X_i]$ and its model parameter θ_i—an advantageous property of the generative model over the descriptive model. For example, if the global density $\pi(\theta\,|\,\alpha)$ works as the prior distribution, in the Bayesian framework, for the local parameter θ_i, $\pi(\theta_i \mid \alpha, [x_i]) \propto f_{[X]}^{\star}(\underline{x}_i, \overline{x}_i \mid \theta_i, m_i)\pi(\theta_i \mid \alpha)$ is the posterior distribution of the parameter of the local density $f(x\,|\,\theta_i)$ underlying $[x_i]$. Similarly, the posterior predictive distribution of latent data points underlying $[x_i]$ is directly available as $\pi(x \mid \alpha, [x_i]) \propto \int f(x \mid \theta_i)\pi(\theta_i \mid \alpha, [x_i])\mathrm{d}\theta_i$.

3.2 | Asymptotic properties

Although they are constructed quite distinctly, it is possible to directly relate the descriptive and generative models under specific circumstances. In particular, for standard (descriptive) symbolic analysis techniques, when there is no prior knowledge on the distribution of data within an interval, this distribution is commonly assumed to be uniform $U(a,b)$ with $a \leq b$ (e.g., Le-Rademacher & Billard, 2011). Let $I(\underline{x},\overline{x} : a \leq \underline{x} \leq \overline{x} \leq b)$ be an indicator function of $\underline{x}$ and $\overline{x}$, which is equal to 1 when $a \leq \underline{x} \leq \overline{x} \leq b$ and 0 elsewhere. Defining $f(x \mid \theta)$ so that $X_k \sim U(a,b)$ for $k = 1, \ldots, m$ and constructing $[X] = \varphi_{1,m}(X_{1:m})$ from the minimum and maximum values of these latent data points, then the density function of $[X]$ given by (7) becomes

$$f^{\star\star}_{[X]}\left(\underline{x},\overline{x} \mid a,b,m\right) = m(m-1)\left(\overline{x}-\underline{x}\right)^{m-2}(b-a)^{-m} I\left(\underline{x},\overline{x} : a \leq \underline{x} \leq \overline{x} \leq b\right),$$

which converges to a point mass at $[a,b]$ as $m \to \infty$ (Section 3). Then, by substituting $f^{\star\star}_{[X]}(\underline{x},\overline{x} \mid a,b,m)$ into (9), the hierarchical generative model becomes

$$f_{[X]}\left(\underline{x},\overline{x} \mid m\right) = \iint_{\{a \leq \underline{x}, b \geq \overline{x}\}} m(m-1)\frac{\left(\overline{x}-\underline{x}\right)^{m-2}}{(b-a)^{m}}\pi(a,b)\,\mathrm{d}a\mathrm{d}b, \tag{10}$$

where $\pi(a,b)$ describes the inter-interval parameter variability. When m is large, the following theorem states that this hierarchical generative model converges to $\pi(\underline{x},\overline{x})$, which is a valid descriptive model.

Theorem 4. *Suppose that $[X] = \varphi_{1,m}(X_{1:m})$ with $X_k \sim U(a,b)$ for $k = 1, \ldots, m$ and the global density function $\pi(a,b)$ is bounded, continuous, and equal to 0 when $a > b$. Then, as $m \to \infty$, the density function of $[X]$ (10) converges to $\pi(\underline{x},\overline{x})$ pointwise, that is,*

$$\lim_{m\to\infty} f_{[X]}\left(\underline{x},\overline{x} \mid m\right) = \pi\left(\underline{x},\overline{x}\right).$$

This result is interesting in that it reveals that descriptive models for $[X] \sim f_{[X]}(\underline{x},\overline{x} \mid \theta)$ described in Section 2.1 (e.g., Arroyo et al., 2011; Le-Rademacher & Billard, 2011; Brito & Duarte Silva, 2012) actually possess an underlying and implicit generative structure. Specifically, the sampling process of the descriptive model $[X] \sim f_{[X]}(\underline{x},\overline{x}) = \pi(\underline{x},\overline{x})$ can be expressed via the generative process

$$\begin{aligned}
[X] &= \lim_{m\to\infty} \varphi_{1,m}\left(X_{1:m}\right), \\
X_1, X_2 \ldots &\sim U\left(\underline{X}_{\star}, \overline{X}_{\star}\right), \\
\left(\underline{X}_{\star}, \overline{X}_{\star}\right) &\sim \pi\left(\underline{x},\overline{x}\right).
\end{aligned}$$

That is, to obtain a sample realization of $[X]$, values of lower- and upper-bound parameters, $(\underline{X}_{\star}, \overline{X}_{\star})$, of local uniform distribution are first drawn from the descriptive model $\pi(\underline{x},\overline{x})$, which, in this case, is exactly equivalent to the global density for the associated underlying hierarchical generative model. As the resulting infinite collection of latent data points $X_k \sim U(\underline{X}_{\star}, \overline{X}_{\star})$ fully identifies the local density and $\min\{X_k\} = \underline{X}_{\star}$, $\max\{X_k\} = \overline{X}_{\star}$ are sufficient statistics for uniform distributions, the generated interval $[X]$ is then determined as $[X] = [\underline{X}_{\star}, \overline{X}_{\star}]$ with $(\underline{X}_{\star}, \overline{X}_{\star}) \sim \pi(\underline{x}_{\star}, \overline{x}_{\star})$. As a result, there is no loss of information from the data aggregation procedure, and the variation of $[X]$ is completely due to the variation permitted in the distribution of local parameters, which is the global distribution. In this manner, the descriptive model is a special case of and directly interpretable as a particular generative model.

This idea can be extended to a more general class of hierarchical generative models in which the local distribution is only governed by location (μ) and scale ($\tau > 0$) parameters, so that $X_k \sim f(x \mid \mu, \tau)$ for $k = 1, \dots, m$. Suppose $\underline{x}$ and $\overline{x}$ are the lth- and uth-order statistics, respectively. The associated values of μ and τ are available by solving

$$\begin{cases} Q(l/(m+1); \mu, \tau) = \underline{x} \\ Q(u/(m+1); \mu, \tau) = \overline{x}, \end{cases} \tag{11}$$

where $Q(\cdot; \mu, \tau)$ is the quantile function of $f(x \mid \mu, \tau)$. If a unique solution exists for (11), then $f(x \mid \mu, \tau)$ is an *interval-identifiable* local distribution.

We previously discussed that, under the order statistic–based data aggregation function, the i.i.d. generative model (6) will converge to a point mass as $m \to \infty$. Similar to Theorem 4, those hierarchical generative models (9) with interval-identifiable local density functions $f(x \mid \mu, \tau)$ will also converge to descriptive models.

Theorem 5. *Suppose that* $[X] = \varphi_{l,u}(X_{1:m})$ *with* $X_k \sim f(x \mid \mu, \tau)$ *for* $k = 1, \dots, m$, *where the local density function* $f(x \mid \mu, \tau)$ *is interval identifiable with location parameter* μ *and scale parameter* $\tau > 0$. *Further suppose that* $l/(m+1) \to \underline{p} > 0$ *and* $u/(m+1) \to \overline{p} < 1$ *as* $m \to \infty$ *and that*

(i) *the global density function* $\pi(\mu, \tau)$ *is twice differentiable,*
(ii) $f(x \mid \mu, \tau)$ *is positive and continuous in neighborhoods of* $Q(\underline{p}; \mu, \tau)$ *and* $Q(\overline{p}; \mu, \tau)$, *and*
(iii) $\iint |f^{\star}_{[X]}(\underline{x}, \overline{x} \mid \mu, \tau, m, l, u) \mid \pi(\mu, \tau) \mathrm{d}\mu \mathrm{d}\tau < \infty$ *for any* $0 < l < u < m$.

Then, as $m \to \infty$, *the density function of* $[X]$ *for the hierarchical generative model (9) converges pointwise to*

$$\pi_{\star}\left(\underline{x}, \overline{x}\right) = \pi\left(\mu\left(\underline{x}, \overline{x}; \underline{p}, \overline{p}\right), \tau\left(\underline{x}, \overline{x}; \underline{p}, \overline{p}\right)\right) \times \left| J\left(\mu\left(\underline{x}, \overline{x}; \underline{p}, \overline{p}\right), \tau\left(\underline{x}, \overline{x}; \underline{p}, \overline{p}\right); \underline{p}, \overline{p}\right)\right|^{-1},$$

where $\mu(\underline{x}, \overline{x}; \underline{p}, \overline{p})$ *and* $\tau(\underline{x}, \overline{x}; \underline{p}, \overline{p})$ *are the solution of (11) and*

$$J\left(\mu, \tau; \underline{p}, \overline{p}\right) = \begin{pmatrix} \frac{\partial}{\partial \mu} Q\left(\underline{p} \mid \mu, \tau\right) & \frac{\partial}{\partial \tau} Q\left(\underline{p} \mid \mu, \tau\right) \\ \frac{\partial}{\partial \mu} Q\left(\overline{p} \mid \mu, \tau\right) & \frac{\partial}{\partial \tau} Q\left(\overline{p} \mid \mu, \tau\right) \end{pmatrix}.$$

In the specific case where $f(x \mid a, b)$ is a $U[a, b]$ local density function, with quantile function $Q(p \mid a, b) = (1 - p)a + pb$, the hierarchical generative model (9) converges to the distribution of

$$\left[\left(1 - \underline{p}\right) \underline{X}_{\star} + \underline{p} \overline{X}_{\star}, \left(1 - \overline{p}\right) \underline{X}_{\star} + \overline{p} \overline{X}_{\star}\right],$$

where $(\underline{X}_{\star}, \overline{X}_{\star}) \sim \pi(\underline{x}, \overline{x})$.

4 | MULTIVARIATE MODELS FOR HYPER-RECTANGLES

The p-dimensional equivalent of the univariate interval-valued random variable $[X]$ is the random p-hyper-rectangle, which corresponds to a p-tuple of random intervals. Specifically, we denote $[\boldsymbol{x}] = ([x_1], \dots, [x_p]) \in \mathbb{I}^p$ as a hyper-rectangle in the space of p-hyper-rectangles and $\boldsymbol{x} = (x_1, \dots, x_p) \in \mathbb{R}^p$ as one p-dimensional latent data point. It is straightforward to extend the previous theory on containment distribution functions and likelihood functions for random intervals (Sections 2 and 3) to random hyper-rectangles.

4.1 | Containment distribution functions

Similar to Section 2.1, descriptive models for random p-hyper-rectangles may be constructed through direct specification of the $2p$-dimensional density function $f_{[\boldsymbol{X}]}(\underline{x}_1, \overline{x}_1, \dots, \underline{x}_p, \overline{x}_p)$. These models are easily constructed and simple to use but have the same limitations as the descriptive models for random intervals discussed in Section 2.1.

The containment distribution function of $[\boldsymbol{X}]$, denoted $F_{[\boldsymbol{X}]} : \mathbb{R}^{2p} \mapsto [0, 1]$, is a function on the real hyperplane, having similar properties to those described in Definition 1 (not stated here for brevity). The following theorems show the connection between the containment distribution function and the density function for $[\boldsymbol{X}]$.

Theorem 6. *Provided that $f_{[\boldsymbol{X}]} : \mathbb{R}^{2p} \mapsto \mathbb{R}$ is the density function of a random p-hyper-rectangle $[\boldsymbol{X}]$, the containment distribution function can be derived as follows:*

$$F_{[\boldsymbol{X}]}\left(\underline{x}_1, \overline{x}_1, \dots, \underline{x}_p, \overline{x}_p\right) = \int_{\underline{x}_p}^{\overline{x}_p} \cdots \int_{\underline{x}_1}^{\overline{x}_1} f_{[\boldsymbol{X}]}\left(\underline{x}'_1, \overline{x}'_1, \dots, \underline{x}'_p, \overline{x}'_p\right) \mathrm{d}\underline{x}'_1 \mathrm{d}\overline{x}'_1 \dots d\underline{x}'_p \mathrm{d}\overline{x}'_p.$$

Theorem 7. *Let $F_{[\boldsymbol{X}]} : \mathbb{R}^{2\mathrm{p}} \mapsto [0{,}1]$ be the containment distribution function of a random hyper-rectangle $[\boldsymbol{X}]$. If $F_{[\boldsymbol{X}]}$ is $2p$-times differentiable, then the density function of $[\boldsymbol{X}]$ is*

$$f_{[\boldsymbol{X}]}\left(\underline{x}_1, \overline{x}_1, \dots, \underline{x}_p, \overline{x}_p\right) = (-1)^p \frac{\partial^{2p}}{\partial \underline{x}_1 \partial \overline{x}_1 \dots \partial \underline{x}_p \partial \overline{x}_p} F_{[\boldsymbol{X}]}\left(\underline{x}_1, \overline{x}_1, \dots, \underline{x}_p, \overline{x}_p\right). \tag{12}$$

4.2 | Generative models

Containment distribution functions and likelihood functions of generative models may be formulated using the same ideas as in (3) and (4). However, due to the necessity of calculating $2p$th-order mixed derivatives in (12), although intuitive, the structure of the resulting likelihood functions would be highly complex, even for i.i.d. generative models of random rectangles. The full form of the likelihood function for an i.i.d. generative model in the bivariate case $[\boldsymbol{X}] = [X_1] \times [X_2]$ is presented in the Appendix A.5.

The complex form of the likelihood function of an i.i.d. generative model accordingly induces a similarly complex hierarchical generative model. One option to produce more tractable models is to impose a conditional independence structure within each p-dimensional latent data point, so that $\boldsymbol{x}_k \sim f(\boldsymbol{x} \mid \theta_{1:p}) = \prod_{j=1}^{p} f(x_j \mid \theta_j)$. Consequently, each random interval marginal distribution of the p-hyper-rectangle is conditionally independent of others, that is,

$$f^{\star}_{[\boldsymbol{X}]}\left(\underline{x}_1, \overline{x}_1, \dots, \underline{x}_p, \overline{x}_p \mid \theta_{1:p}\right) = \prod_{j=1}^{p} f^{\star}_{[X_j]}\left(\underline{x}_j, \overline{x}_j \mid \theta_j\right),$$

where $f^{\star}_{[X_j]}(\underline{x}_j, \overline{x}_j \mid \theta_j)$ is the likelihood function of the i.i.d. generative model (5) for $[X_j]$. Although this choice will result in clear modeling consequences, the resulting likelihood function for the hierarchical generative model

$$f_{[\boldsymbol{X}]}\left(\underline{x}_1, \overline{x}_1, \dots, \underline{x}_p, \overline{x}_p \mid m, \alpha\right) = \int \prod_{j=1}^{p} f^{\star}_{[X_j]}\left(\underline{x}_j, \overline{x}_j \mid \theta_j\right) \pi(\theta_{1:p} \mid \alpha) \mathrm{d}\theta_{1:p} \tag{13}$$

will only then require p second-order mixed derivatives.

In this scenario, dependencies between the random interval marginal distributions of $[\boldsymbol{X}]$, such as temporal or spatial dependencies, are controlled only by the dependence among local parameters $\theta_{1:p}$ as introduced by the global distribution $\pi(\theta_{1:p} \mid \alpha)$. As a result, beyond any a priori information on the joint distribution of the p-dimensional latent data points underlying construction of the random interval $[\boldsymbol{X}]$ being incorporated within $\pi(\theta_{1:p} \mid \alpha)$, it will be impossible to identify any further dependence based on the observed p-hyper-rectangles. If this is inadequate for a given analysis, the full multivariate likelihood will need to be derived (see, e.g., the Appendix A.5).

5 | APPLICATIONS

We illustrate our new models by first comparing the performance of the generative models to the existing descriptive models for simulated univariate (random interval) data. We then provide a generative model reanalysis of a real data set of 5000 credit card customers, as previously analyzed by Brito and Duarte Silva (2012) using a descriptive model. The size of this data set does not merit the use of symbolic data methods for its analysis; however, it does serve as a useful illustration of the benefits of generative models. We conclude with an examination of the robustness of the generative method to model mis-specification.

5.1 | Simulated data analysis

In order to provide a direct comparison between descriptive and generative models, we construct our observed random intervals under the generative model as $[x_i] = [\underline{x}_i, \overline{x}_i]$, where $\underline{x}_i$ and $\overline{x}_i$ are, respectively, the observed minimum and maximum values of $x_{i1}, \dots, x_{im_i}$ under the mixture model

$$\begin{aligned} x_{i1}, \dots, x_{im_i} \quad &\sim \quad U(c_i - \mathrm{e}^{\tau_i}, c_i + \mathrm{e}^{\tau_i}), \\ c_i \sim N\left(\mu_c, \sigma_c^2\right) \quad &\text{and} \quad \tau_i \sim N\left(\mu_\tau, \sigma_\tau^2\right), \end{aligned} \tag{14}$$

for $i = 1, \dots, n$. From Theorem 4, this hierarchical model is asymptotically equivalent (as $m_i \to \infty$ for each i) to a descriptive model with $[x_i^\star] = [c_i^\star - \mathrm{e}^{\tau_i^\star}, c_i^\star + \mathrm{e}^{\tau_i^\star}]$, where $(c_i^\star, \tau_i^\star)$ follows the same joint distribution as (c_i, τ_i). While, in practice, random intervals will generally be constructed from different numbers of random samples, $x_{i1}, \dots, x_{im_i}$ (e.g., see Section 5.2), here, we specify $m_i = m$ for all $i = 1, \dots, n$. In this analysis, we will compare the maximum likelihood estimators (MLEs) of parameters for both generative and descriptive models obtained using data simulated from each model.

For each random interval $[x_i]$ under the mixture model, the two-dimensional integration (9), with $\theta = (c_i, \tau_i)$, can be reduced to a one-dimensional integration by first integrating out c_i and then reparameterizing to $z_i = m(\tau_i - \log \frac{1}{2}(\overline{x}_i - \underline{x}_i))$. This leads to the likelihood function of a single interval observation $[x_i]$ given by

$$\begin{aligned} &\int_0^\infty \left(\overline{x}_i - \underline{x}_i\right)^{-2}(m-1)\mathrm{e}^{-z_i}\phi\left(m^{-1}z_i + \log\frac{\overline{x}_i - \underline{x}_i}{2}; \mu_\tau, \sigma_\tau^2\right) \\ &\times \left\{\Phi\left(\underline{x}_i + \frac{\overline{x}_i - \underline{x}_i}{2}\mathrm{e}^{m^{-1}z_i}; \mu_c, \sigma_c^2\right) - \Phi\left(\overline{x}_i - \frac{\overline{x}_i - \underline{x}_i}{2}\mathrm{e}^{m^{-1}z_i}; \mu_c, \sigma_c^2\right)\right\} \mathrm{d}z_i, \end{aligned} \tag{15}$$

where ϕ and Φ respectively denote the Gaussian density and distribution function. This form may be quickly and accurately approximated by Gauss–Laguerre quadrature methods (e.g., Evans & Swartz, 2000). The form of the integrand in (15) for varying m and the resulting negative log-likelihood function are shown in Figure 1 for $\underline{x}_i = -1, \overline{x}_i = 1$, $\mu_c = \mu_\tau = 0$, and $\sigma_c^2 = \sigma_\tau^2 = 1$.

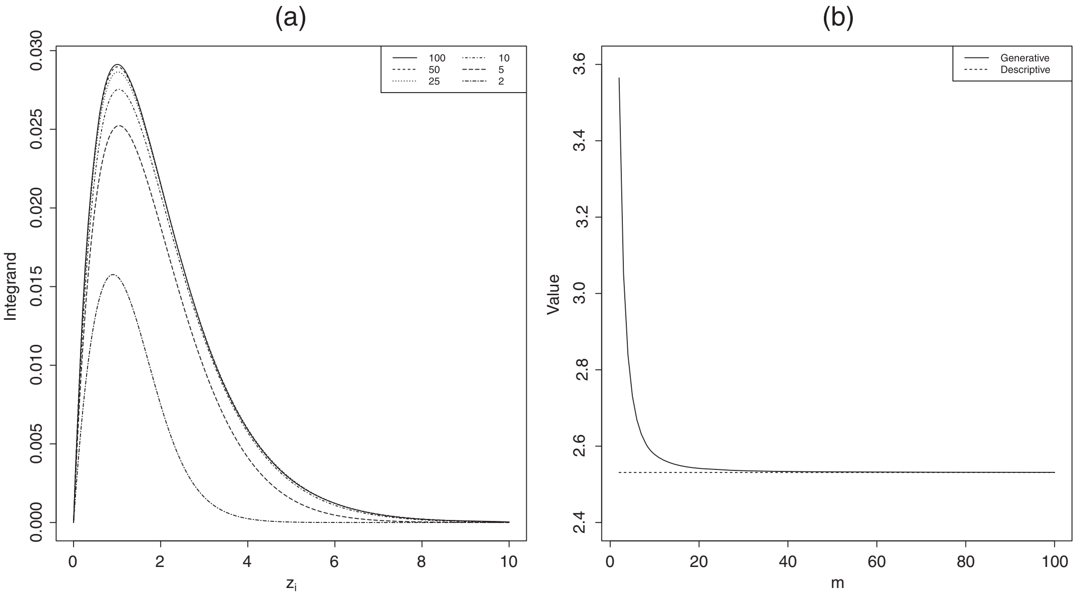


FIGURE 1 (a) Forms of the integrand in (15), with $m = 2, 5, 10, 25, 50, 100$, as a function of z_i. (b) Negative log-likelihood function as a function of m, when $\underline{x}_i = -1, \overline{x}_i = 1, \mu_c = \mu_\tau = 0$, and $\sigma_c^2 = \sigma_\tau^2 = 1$

These plots illustrate the convergence of the generative model to the descriptive model as m gets large (Theorem 4), with only very minor differences observed for $m > 30$, and suggest (panel (a)) that quadrature integration methods will be accurate with around 20 nodes.

We simulate 1000 replicate data sets, each comprising $n = 100$ intervals, from the descriptive model with $c_i^\star, \tau_i^\star \sim N(0, 1)$ for $i = 1, \dots, n$ (i.e., $\mu_c = \mu_\tau = 0$ and $\sigma_c^2 = \sigma_\tau^2 = 1$). MLEs of the model parameters $(\mu_c, \mu_\tau, \sigma_c^2, \sigma_\tau^2)$ are obtained from fitting both descriptive and generative models, with the latter assuming a specified number of latent variables, m. Note that, in practice, the number of latent variables, m, will typically be known (and finite). The first column of Figure 2 illustrates the differences between the resulting descriptive and generative model parameter MLEs (e.g., $\hat{\mu}_c^{(D)} - \hat{\mu}_c^{(G)}$, where the superscripts indicate parameters of the descriptive (D) and generative (G) models), with the solid line indicating the mean and the dotted lines indicating the central 95% interval, computed over the 1000 replicates.

First, we notice that the difference between the estimates is large for small m values and becomes gradually smaller as m increases. This is not surprising as, in this model specification, the generative model approaches the descriptive mode as $m \to \infty$. However, as both models are identically centered, the mean difference between the location parameter estimates $\hat{\mu}_c^{(D)}$ and $\hat{\mu}_c^{(G)}$ is zero, regardless of the number of latent variables.

An obvious area of difference is that the point estimates of the interval half-range (modeled by μ_τ) are much smaller for the (correct) descriptive model than for the generative model. This occurs as the expected range of $x_{i1}, \dots, x_{im}$ under a generative model is lower for small m values than it is for large m values. As a result, the generative model will determine that μ_τ should be sufficiently larger for small m values than it would be for large m values, given the same observed $[\underline{x}_i, \overline{x}_i]$. That is, if the data are truly generated from the descriptive model, parameters estimated from the generative model are effectively biased for any finite m and overestimate the true model parameters, with the magnitude of the bias determined by the assumed value of m. Of course, this bias can be reduced by setting m to be large in this case.

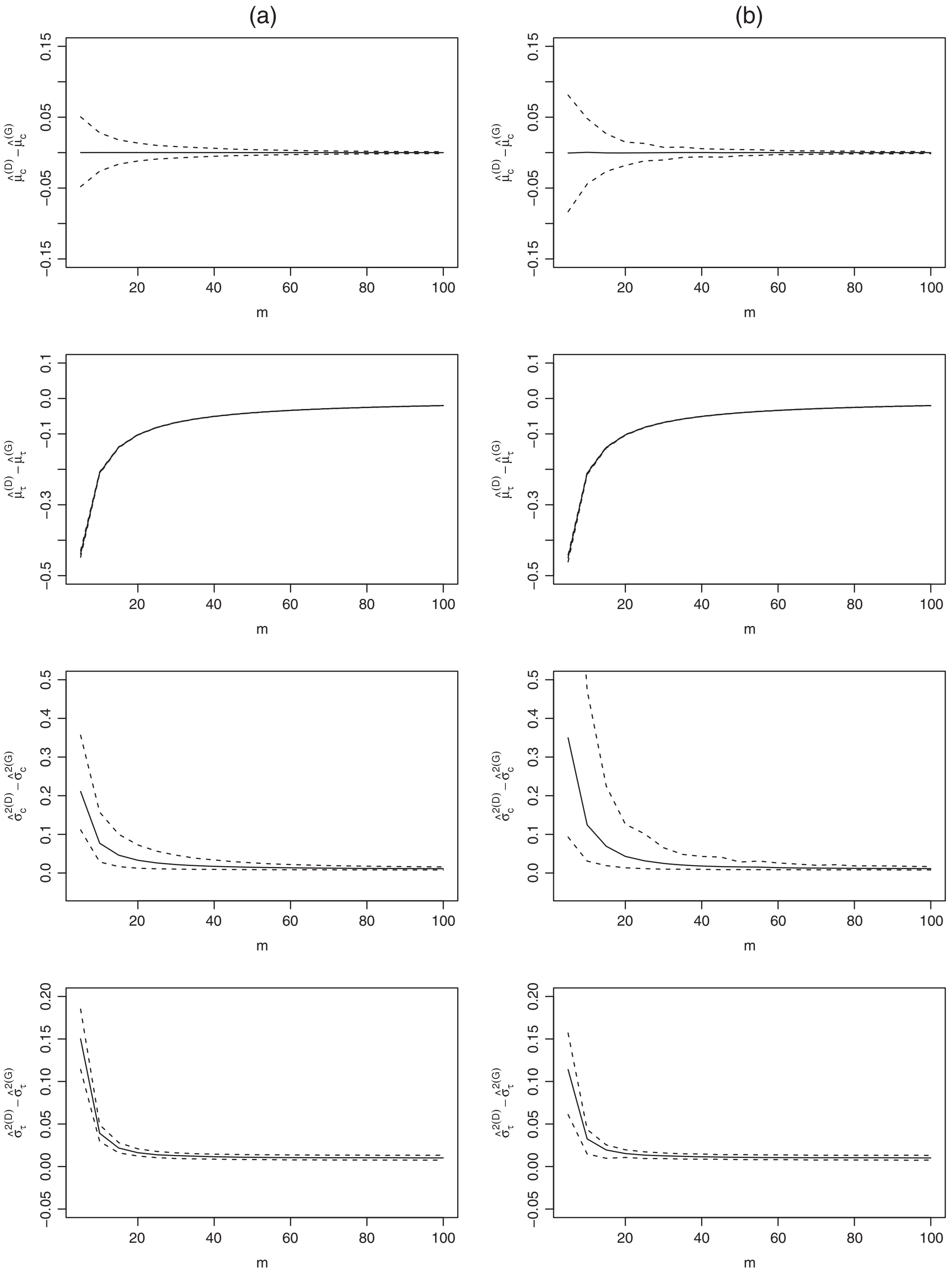


FIGURE 2 Differences between the maximum likelihood estimators (MLEs) of the descriptive model and the generative hierarchical model, based on data generated from each model (left column: descriptive model data; right column: generative model data), as a function of $m = 5, \ldots, 100$, the number of latent data points $x_{i1}, \ldots, x_{im}$ in the generative model. Lines indicate the MLE means (solid lines) and 2.5% and 97.5% quantiles (dashed lines) based on 1000 replicate data sets

The second area of difference is that the estimated variability of the point estimates of interval location and scale ($\hat{\sigma}_c^2$ and $\hat{\sigma}_\tau^2$) is higher under the descriptive model than that under the generative model. This occurs as the generative model assumes that the variability of, for example, $\frac{\underline{x}_i+\bar{x}_i}{2}$ comprises both the variability of the latent data $x_{i1}, \dots, x_m$ within interval i, in addition to the variability of interval locations c_i between intervals. Under the descriptive model, this first source of variability is zero, and therefore, $\hat{\sigma}_c^{2(D)}$ will always be greater than $\hat{\sigma}_c^{2(G)}$ for finite m values. Similar reasoning explains why $\hat{\sigma}_\tau^{2(D)}$ is always greater than $\hat{\sigma}_\tau^{2(G)}$.

The second column of Figure 2 shows the same output as the first column, but based on data simulated from the generative model with the same parameter settings as before and for varying (true) numbers of latent data points $m = 5, \dots, 100$. The results are similar to before, except critically with the interpretation that the generative model with fixed m is now correct. This means that, for example, if intervals are constructed using the generative process (which is the most likely scenario in practice) but are then analyzed with a descriptive model, the point estimates of the interval range (μ_τ) can be substantially underestimated by assuming $m \to \infty$ under the descriptive model, when, in fact, m is small and finite. Similarly, the estimates of σ_c^2 and σ_τ^2 will always be overestimated when assuming an incorrect descriptive model. These scenarios will obviously be problematic for data analysis in practice.

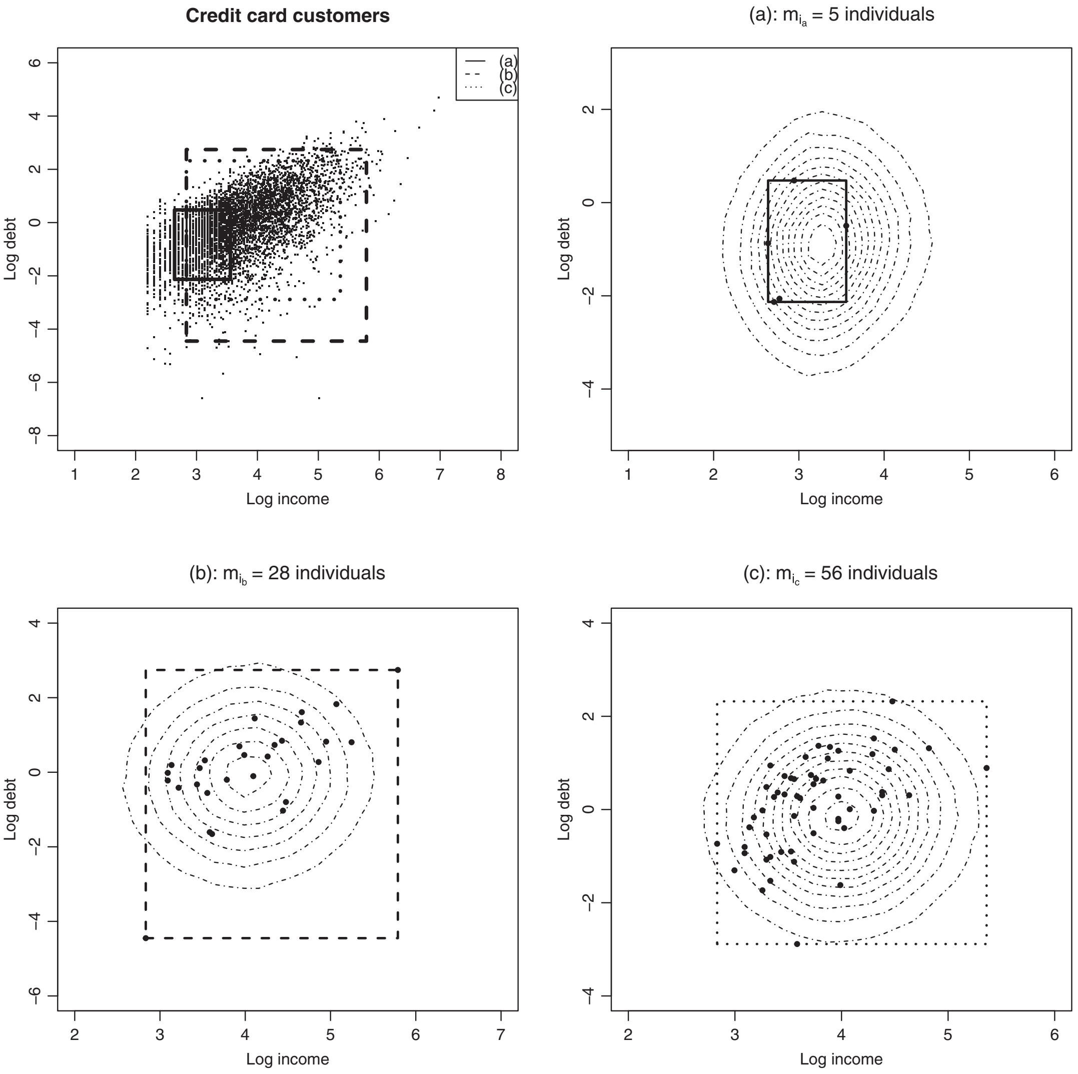
The takeaway message of this analysis is that it is important to fit the model (descriptive or generative) that matches the interval (or p-hyper-rectangle) construction process. Failure to do so can result in misinterpretation of model parameters, resulting in severe biases in parameter estimates, which can then detrimentally impact an analysis. In practice, intervals tend to be constructed from underlying classical data (e.g., see Section 5.2), using a known process and where m is also known. This implies that the generative model is a more natural construction than the descriptive model and with parameters that more directly relate to the observed data.

While this analysis has assumed uniformity of the generative process (14) in order that the descriptive model is obtained as $m \to \infty$ and, hence, that the parameter estimates between the two models can be directly compared, the same principles of interpretation and bias occur regardless of the generative model. The parameters are simply less directly comparable with each other.

5.2 | Analysis of credit card data

The data (available in the SPSS package *customer.dbase*) comprise log income and log credit card debt in thousands of U.S. dollars of 5000 credit card customers. In a previous analysis using descriptive models by Brito and Duarte Silva (2012), these data were aggregated into random bivariate rectangles by stratifying individuals according to gender, age category (18–24, 25–34, 35–49, 50–64, 65+ years old), level of education (did not complete high school, high school degree, some college, college degree, undergraduate degree+), and designation of primary credit card (none, gold, platinum, other). This leads to 192 nonempty groups, each producing a random rectangle $[x_{i1}] \times [x_{i2}]$ constructed by the intervals bounded by the minimum and maximum observed values on log income and log credit card debt.

The data are illustrated in Figure 3, along with the underlying data and constructed random rectangles for three of the 192 groups, containing (a) $m_{i_a} = 5$, (b) $m_{i_b} = 28$, and (c) $m_{i_c} = 56$ individuals. The number of individuals in all groups varies greatly (from 5 to 56), and it is noticeable that the distribution of individuals within each group comes from a nonuniform distribution. As a result, the usual uniformity assumption of descriptive models for random rectangles is clearly inappropriate. The generative model is more suited to dealing with these heterogeneous rectangle-valued data containing complex intra-rectangle structures.



Credit card customers

(a): $m_{i_a} = 5$ individuals

(b): $m_{i_b} = 28$ individuals

(c): $m_{i_c} = 56$ individuals

FIGURE 3 Log income and log credit card debt in thousands of U.S. dollars for 5000 customers. Panels illustrate rectangle-valued observations constructed from three groups of customers comprising (a) $m_{i_a} = 5$, (b) $m_{i_b} = 28$, and (c) $m_{i_c} = 56$ individuals. The contours in the last three panels indicate the predictive distributions of individuals for each group conditional on the corresponding rectangle-valued observations, based on the generative model

Given the clear nonuniformity within each group i, we assume that the underlying data are Gaussian with group-specific means and covariances. That is,

$$(x_{i1}, x_{i2}) \sim N_2(\mu_i, \Sigma_i)$$

for $i = 1, \dots, n = 192$, where $\mu_i = (\mu_{i1}, \mu_{i2})$ and $\Sigma_i = \text{diag}(\sigma_{i1}^2, \sigma_{i2}^2)$. Note that we choose to model log income and log credit card debt as uncorrelated, despite there being some visual evidence of positive correlation in the data underlying each random rectangle. It is worth briefly explaining this decision in detail. For a small number of latent data points m_i, it is possible for a single

TABLE 1 Maximum likelihood estimates (MLEs) and 95% asymptotic confidence intervals (CIs) for the parameters of the generative and descriptive models for the credit card data set

		θ_1	λ_1^2	θ_2	λ_2^2	ρ_μ	η_1	ϵ_1^2	η_2	ϵ_2^2
Generative	MLE	3.76	0.13	−0.36	0.21	0.90	−1.20	0.48	0.41	0.09
	95% CI	3.70	0.10	−0.44	0.13	0.83	−1.31	0.35	0.34	0.04
		3.82	0.17	−0.26	0.29	0.98	−1.09	0.61	0.48	0.13
Descriptive	MLE	3.79	0.17	−0.42	0.52	0.57	0.02	0.20	0.82	0.09
	95% CI	3.74	0.13	−0.53	0.41	0.47	−0.04	0.16	0.78	0.07
		3.85	0.20	−0.32	0.62	0.67	0.08	0.24	0.87	0.11

point to determine both upper (or lower) ranges of the random rectangle, and the probability of this occurring increases as the correlation of the underlying data increases. Hence, in principle, there is some information about the correlation structure of the underlying data available through the associated random rectangle. However, for groups with larger m_i values, the upper and lower ranges of the random rectangles are more likely to be determined by four individual data points, in which case it is not then possible to discern the underlying correlation structure. Although we have several groups with small numbers of latent data points (e.g., $m_{i_a} = 5$), in principle allowing their correlation to be estimated, note that the same random rectangles will arise whether the latent data are positively or negatively correlated. That is, the correlation parameter is nonidentifiable from the observed rectangle data. As such, we proceed without attempting to estimate this parameter, despite information on the magnitude of the correlation being available in principle for some groups.

We model the group-specific (local) parameters as

$$\begin{aligned}(\mu_{i1}, \mu_{i2}) &\sim N_2\left(\theta_1, \theta_2, \lambda_1^2, \lambda_2^2, \rho_\mu\right), \\ \log \sigma_{ij}^2 &\sim N\left(\eta_j, \epsilon_j^2\right),\end{aligned} \tag{16}$$

for $j = 1, 2$ and $i = 1, \dots, 192$. The integration in the generative model (13) is achieved using Gauss–Hermite quadratures with 20^4 nodes to integrate over the four parameters.

Maximum likelihood estimates and 95% confidence intervals for each model parameter are illustrated in Table 1 for both generative and descriptive models. Similar to the results for the simulated examples, the point estimates of location (θ_1 and θ_2) are broadly insensitive to the choice of model; however, the estimated values for many of other parameters differ between the two models. Most importantly, the estimated values of ρ_μ are considerably larger for the generative model ($\hat{\rho}_\mu = 0.9040$) compared to those for the descriptive model ($\hat{\rho}_\mu = 0.5695$). While both of these indicate a positive relationship between income and credit card debt, which is evident in the underlying data in Figure 3, there is a clear difference in the strength of that relationship. The descriptive model results in a weaker estimated value in the correlation because it does not take the noisy data–generating process into account. While we suspect that the generative model may be the more accurate of the two given the data generating procedure used to construct the random rectangles, in terms of drawing inferential conclusions about the underlying data, it is critical that we are certain in this regard.

For the generative models, the distribution of the local parameters $(\mu_{ij}, \sigma_{ij}^2)$ for each rectangle-valued observation can be computed by empirical Bayesian methods (previously, these parameters were integrated out for the optimization in Table 1). The prior distribution for the local parameters is the global distribution (16) with its parameter values given by the estimates in Table 1, and the likelihood function is the local density function of one observed rectangle.

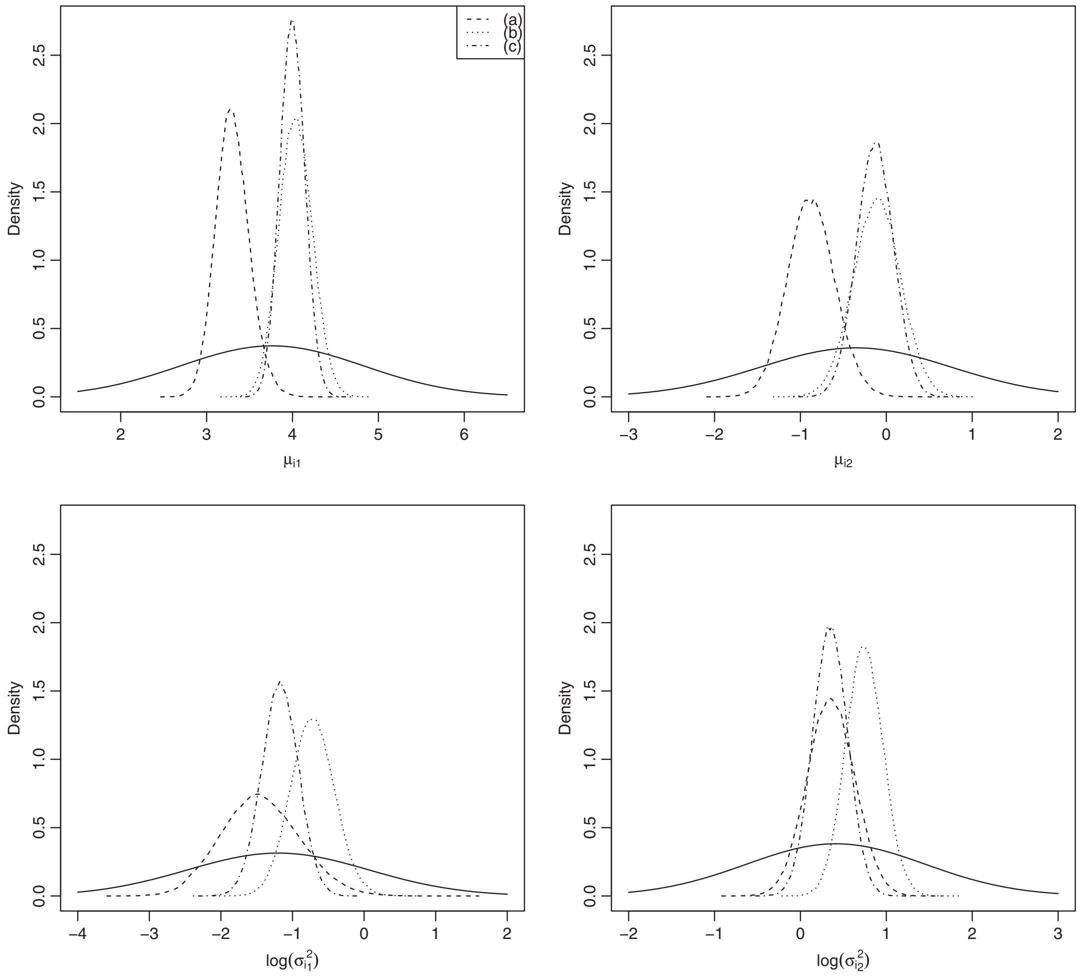


FIGURE 4 Estimated marginal posterior distributions of the local parameters μ_{i1}, μ_{i2}, σ_{i1}^2, and σ_{i2}^2 associated with the three groups (a)–(c) shown in Figure 3. Solid lines correspond to the prior distributions for local parameters

The resulting marginal posterior distributions for the parameters of the observed rectangles (a)–(c) (Figure 3) are shown in Figure 4. Compared to the prior (solid line), the parameters are well informed, even for rectangle (a) with $m_{i_a} = 5$ observations, with the level of precision increasing with the number of individuals within each rectangle.

Goodness of fit for both descriptive and generative models can be evaluated through model predictive distributions of random rectangles, in addition to predictive distributions for individual data points for the generative model. In the latter case, based on the posterior distributions of the local parameters in Figure 4, the predictive distributions of individual data points within the random intervals (a)–(c), conditional on observing the associated random interval, are shown in Figure 3. While the predictive distributions are marginally independent due to model specification, their coverage describes the observed data well. For group (a), the predictive distribution covers a wider region than the observed rectangle, as this rectangle is constructed from only five individuals. As the number of individuals increases in groups (b) and (c), the predictive regions more closely represent the region of the observed rectangle, indicating that the generative model

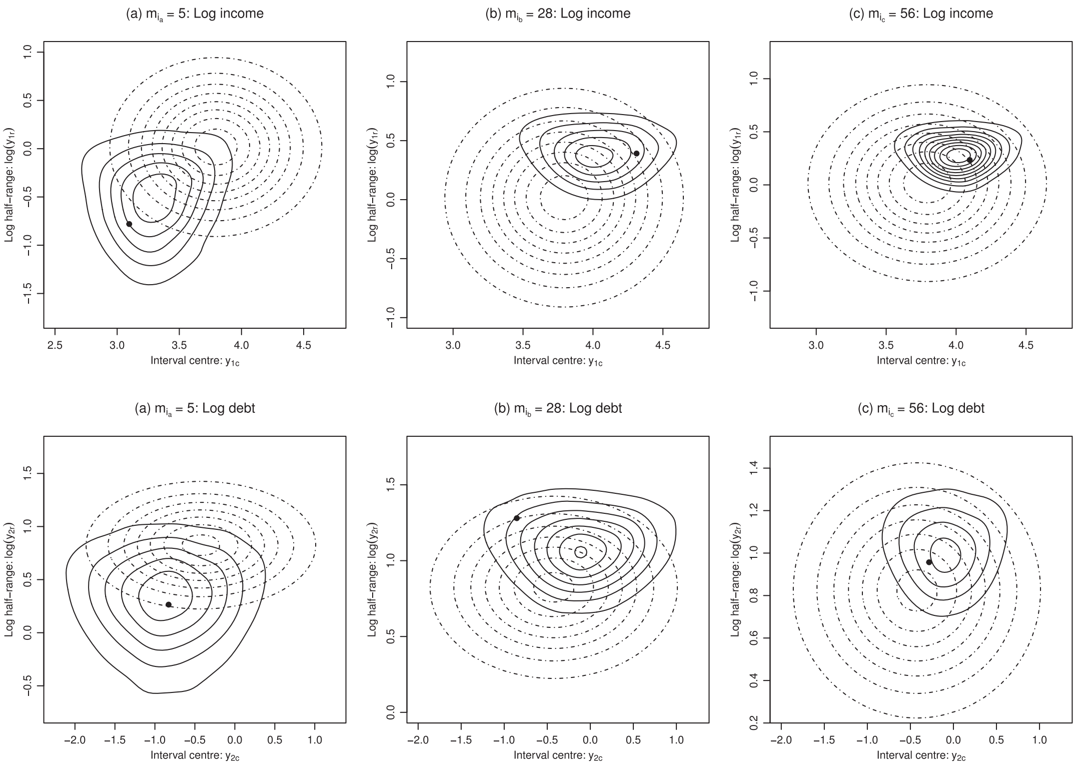


FIGURE 5 Posterior predictive distribution of a random rectangle $[y_1] \times [y_2]$ for each of the groups (a)–(c) (left column to right) in Figure 3. Columns illustrate the marginal random intervals of log income ($[y_1]$, top row) and log debt ($[y_2]$, bottom row), with each interval $[y_j] = [y_{j1}, y_{j2}]$ expressed in interval center and half-range form $(y_{jc}, y_{jr}) = ((y_{j1} + y_{j2})/2, (y_{j2} - y_{j1})/2)$ for $j = 1, 2$. Solid and dashed lines indicate predictive distributions of generative and descriptive models, respectively. The dot indicates the observed interval $[x_{i1}] \times [x_{i2}]$ used for model fitting

has the ability to correctly account for the different numbers of individuals used to construct each rectangle. The predictive distribution for group (b) individuals also indicates some robustness to the two outliers that completely define the observed rectangle. This occurs as the model correctly accounts for the fact that rectangle (b) is constructed from half the number of observations used to construct the rectangle of group (c), although both rectangles are roughly of the same size.

The predictive distributions of random rectangles for groups (a)–(c) are illustrated in Figure 5 for both descriptive (dashed lines) and generative (solid lines) models. Shown are the bivariate predictive distributions of interval center and log half-range, for both log income (top row) and log debt (bottom row). The dot indicates the observed interval. Under the generative model, these distributions are obtained directly from the predictive distributions for individuals (Figure 3).

In all cases, the predictive distributions of the generative model more accurately and more precisely identify the location of the observed data. This is particularly the case in group (a) in which the descriptive model is clearly indicating a lack of model fit. The predicted interval for log debt in group (b) is not fully centered on the observed interval, as the model attempts to account for the unlikely (under the model) construction of the observed interval by outliers (Figure 3).

However, the observed data are still well predicted under the generative model. The overall fit to the observed data is better under the generative model than the descriptive model, indicating that it more accurately describes the complexities of the observed data.

5.3 | Robustness to model mis-specification

Until now, we have focused on the setting where both the underlying model $f(x \mid \theta)$ and the data aggregation function $\varphi(\cdot)$ are known. When the true $f(x \mid \theta)$ is not known, this is the standard setting of statistical model mis-specification. There are two possible mis-specification scenarios in which $\varphi(\cdot)$ may not be known. Firstly, $\varphi(\cdot)$ may have been misreported, so that, for example, different quantiles were used to construct intervals from data than were modeled in $\varphi(\cdot)$. Second, $\varphi(\cdot)$ may simply be unknown, so that the task is to analyze data having quantiles $\underline{X}$ and $\overline{X}$ but where such quantiles are unknown. In this second scenario, at best, the generative likelihood could be integrated over all possible $\varphi(\cdot)$ with respect to some prior measure. It is possible that with informative prior information, this could yield *some* viable inference, but this would likely be circumstantial and not ideal.

The following analysis aims to examine the effect of mis-specifying the fitted model and $\varphi(\cdot)$. We consider data $x_{1:m}$, with $m = 1000$, generated independently from either normal or uniform distribution, both with mean $\mu = 0$ and standard deviation $\sigma = 2$. To evaluate the effect of outliers, we create additional data sets that replace 5% of each original data set by observations drawn from the (normal or uniform) generating distribution with $\mu = 0$ and $\sigma = 5$. For each data set, observed intervals are constructed through the aggregation function $\varphi_i := \varphi_{i,m-i+1}(x_{1:m}) = [x_{(i)}, x_{(m-i+1)}]$, with $i = 1$ and $i = 250$ corresponding to constructing intervals based on sample minimum/maximum and the first/third quartiles. For each of these interval data sets, we fit both normal and uniform models and assess the impact of knowing the aggregation function $\varphi(\cdot)$ by supposing that the observed intervals are obtained from φ_i with $i = 1, 50, 100, \ldots, 450$.

Figure 6 shows boxplots of 500 replicate maximum likelihood estimates of μ (top row) and $\log(\sigma)$ (middle row), when the true underlying data distribution is normal, as a function of the aggregation function φ_i used to fit the model. The true interval aggregation function is φ_1 ($i = 1$; left two columns) and φ_{250} ($i = 250$; right two columns), and use of this is indicated by the shaded boxplots. In each panel, the horizontal line denotes the true parameter value, and the rightmost boxplot shows the impact of using the true aggregation function with the outlier data sets.

The mean (μ; top row) is consistently well estimated, regardless of the model being fitted or the aggregation function. This is not surprising, as changing φ_i affects the scale of the intervals and not their location. However, for $\log \sigma$ (middle row), when the model being fitted is correct (columns 1 and 3), using generative model aggregation functions that use narrower (wider) quantiles than actually used to construct the empirical interval leads to larger (smaller) estimates of σ. This observation also holds when fitting the uniform model, although the picture is distorted due to model mis-specification (fitting a uniform model to normal data). That is, when the model is correctly specified under the true data aggregation process, the maximum likelihood estimates are accurate.

A goodness-of-fit check between predicted and observed intervals would not reveal problems in any of the above analyses: Both models are in the location-and-scale family, and so, each can describe all observed interval data sets well. However, differences can easily be seen by comparing to the original underlying data. The bottom row of Figure 6 denotes quantile–quantile

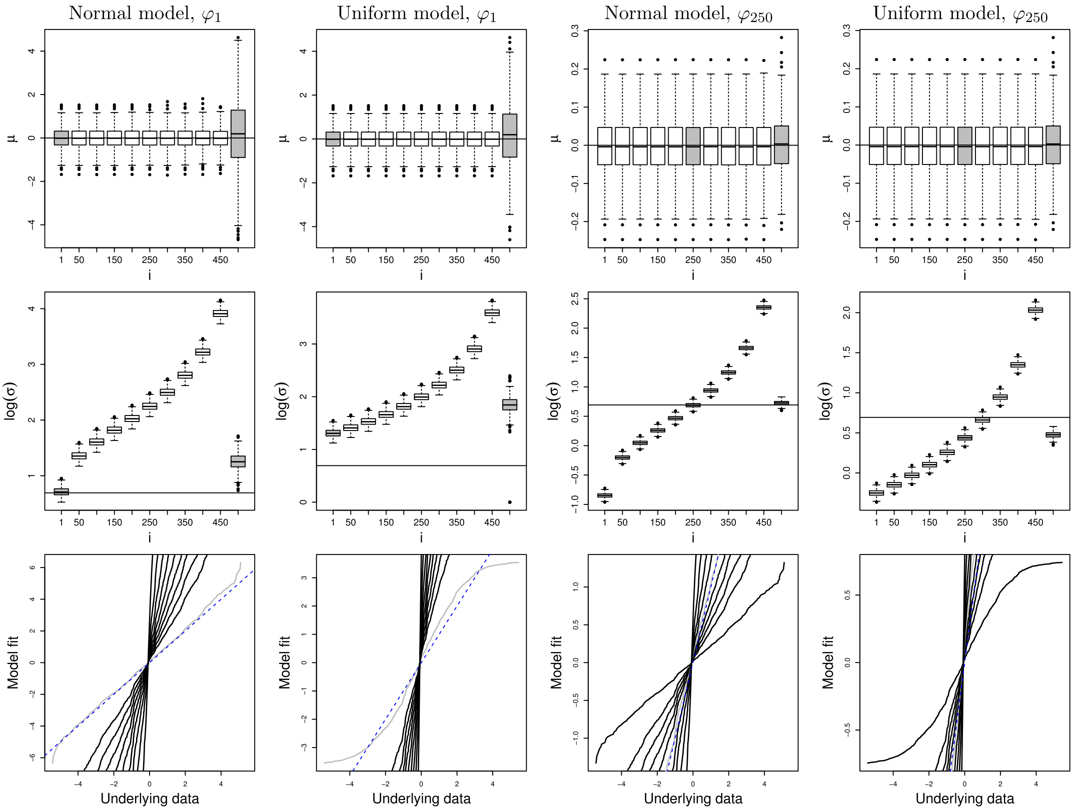


FIGURE 6 Boxplots of 500 replicate maximum likelihood estimates of μ and $\log \sigma$ under an $N(0, 2^2)$ true data generating process with $m = 1000$ and assuming data aggregation function ϕ_i, $i = 1, 50, 100, \ldots, 450$. The true aggregation functions are ϕ_1 (left two columns) and ϕ_{250} (right two columns). The models fitted are the normal (columns 1 and 3) and uniform (columns 2 and 4) distributions. In each panel, the rightmost boxplot indicates the outcome using the data set with 5% outliers. The bottom row shows quantile–quantile curves of the fitted model (y-axis) versus the empirical underlying data quantiles (x-axis). Gray curves indicate use of the correct $\varphi(\cdot)$ function. The dashed line denotes $y = x$ [Colour figure can be viewed at wileyonlinelibrary.com]

(Q–Q) plots of the fitted model (y-axis) against the original sample $x_{1:m}$ (in practice, this would be constructed from a subsample of the data when dealing with very large data sets). In all cases, only when the model and aggregation function are correct does the Q–Q plot align on the $y = x$ axis. A deviation from this indicates that either the model or $\varphi(\cdot)$, or both, is incorrect. As the data aggregation function will typically be known, this would usually suggest that it is the fitted model that needs further requirement. However, when the data aggregation function is mis-specified, then it may be difficult to identify a fitted model that, in combination with the mis-specified $\varphi(\cdot)$, will fit the data well. Failure to improve on a model's goodness of fit when modifying the model could therefore indicate that the data aggregation function is mis-specified.

In the presence of outliers in the original data set (rightmost boxplots in each panel), as might be expected, constructing intervals that are robust to these (e.g., using the first/third quartiles) produce more sensible results than less robust intervals (e.g., using min/max). Qualitatively, similar conclusions to the above can be drawn when the true data generating process is uniform rather than normal (see Figure A2 in the Appendix).

6 | DISCUSSION

Current techniques for modeling random intervals (p-hyper-rectangles) are based on constructing models directly at the level of the interval-valued data (e.g., Arroyo et al., 2011; Le-Rademacher & Billard, 2011; Brito & Duarte Silva, 2012). These approaches are additionally based on the assumption that the unobserved individual data points from which the interval is constructed are uniformly distributed within the interval. As we have demonstrated in Section 5, using these descriptive methods when the data are constructed from underlying individual data points, which is typical in practical applications, can result in misleading and biased parameter estimates and, therefore, unreliable inferences.

In this paper, we have established the distribution theory for interval-valued random variables that are constructed bottom-up from distributions of latent real-valued data and aggregation functions used to construct the random intervals. These generative models explicitly permit the fitting of standard statistical models for latent data points through likelihood-based techniques, while accounting for the manner in which the observed interval-valued data are constructed. This approach directly accounts for the nonuniformity of latent individual data points within intervals and provides a natural way to handle the differing number of latent data points m_i within each random interval, which is, again, typical in practice. The method as presented is fully parametric, although extending these ideas to the nonparametric framework would be of some interest (e.g., Jeon, Ahn, & Park, 2015).

By deriving a descriptive model as the limiting case of a generative model (i.e., as $m_i \to \infty$ for each i), we have demonstrated that these descriptive models have an explicit underlying generative model interpretation. In turn, this indicates why inferences from descriptive models may be potentially misleading in practice.

In order to evaluate the integrated generative likelihood function (13) for the unimodal distributions considered in Section 5, we have used Gaussian quadrature methods. This technique will be less useful when integrating over more than six parameters (Evans & Swartz, 1995) or when there are strong dependencies between local parameters. In these cases, approximate MLEs can be obtained using, for example, Monte Carlo maximum likelihood estimation (Geyer & Thompson, 1992) or Monte Carlo expectation–maximization techniques (Wei & Tanner, 1990) or, in the Bayesian framework, Gibbs sampling (Geman & Geman, 1984) or pseudo-marginal and other likelihood-free Monte Carlo methods (Andrieu & Roberts, 2009; Sisson, Fan, & Beaumont, 2018).

In order to construct the likelihood function (13) for p-hyper-rectangles, we assumed independence among all margins in local distributions to avoid the $2p$th-order mixed differentiation of $F_{[X]}(\underline{x}_1, \overline{x}_1, \ldots, \underline{x}_p, \overline{x}_p)$. Although this differentiation may be achieved using symbolic computation software, the resulting likelihood functions are complex even when $p = 2$ (see Appendix A.5), and the alternative of numerical differentiation would be highly computational. However, this independence assumption does not hold if there is a priori information on the dependence structure within each latent data point $\boldsymbol{x}$. As pointed out by Billard and Diday (2006), this is often the case because the structure of symbolic data might determine inherent dependencies, such as logical, taxonomic, and hierarchical dependencies, but not statistical dependencies. In the generative model, those dependencies as well as statistical dependencies can be addressed simultaneously through the local distribution function $f(\boldsymbol{x} \mid \boldsymbol{\theta})$. However, without the marginal independence assumption, inference for these models can be challenging.

While our examples have primarily focused on minimum-and-maximum–based data aggregation functions $\varphi(x_{1:m})$, there is clear interest in parameter estimation and inference for more robust order-based functions $\varphi_{l,u}(x_{1:m})$, as the resulting intervals will be less sensitive to outliers, as demonstrated in Section 5.3. The procedures for constructing the associated likelihood functions are analogous to those presented here, and Theorem 5 provides their limiting descriptive model counterpart. An additional practical question for inference using order-based aggregation functions is which order-based statistics to use. As this choice will impact the efficiency of the resulting inference, it is an open question to understand what method of random interval construction would be optimal for any given analysis (e.g., Beranger, Lin, & Sisson, 2018).

Finally, we have derived an approximation $\hat{L}$ of the likelihood function of the underlying data, $L(x_{1:m} \mid \theta)$, based on constructing random intervals or p-hyper-rectangles through the data aggregation function $\varphi(\cdot)$, so that $\hat{L}(\varphi(x_{1:m}) \mid \theta) \approx L(x_{1:m} \mid \theta)$. Clearly, there can be some information loss when moving from $x_{1:m}$ to $\varphi(x_{1:m})$. Understanding the quality of this approximation is important both for quantifying inferential accuracy and for guiding the design of the aggregation function (where possible) to increase the performance of an analysis. This is the focus of current research.

ACKNOWLEDGEMENTS

X. Zhang was supported by the China Scholarship Council. B. Beranger and S. A. Sisson were supported by the Australian Research Council through the Discovery Project scheme (FT170100079) and the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers of Big Data, Big Models, New Insights (CE140100049).

ORCID

X. Zhang https://orcid.org/0000-0001-9894-2634
B. Beranger https://orcid.org/0000-0002-7944-3925
S. A. Sisson https://orcid.org/0000-0001-8943-067X

REFERENCES

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII — 1983* (pp. 1–198), Berlin, Germany: Springer.

Andrieu, C. & Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, *37*, 697–725.

Arroyo, J., Espínola, R., & Maté, C. (2011). Different approaches to forecast interval time series: A comparison in finance. *Computational Economics 37*, 169–191.

Beranger, B., Lin, H., & Sisson, S. A. (2018). New models for symbolic data analysis. https://arxiv.org/abs/1809.03659

Beresteanu, A., Molchanov, I., & Molinari, F. (2012). Partial identification using random set theory. *Journal of Econometrics*, *166*, 17–32.

Beresteanu, A., & Molinari, F. (2008). Asymptotic properties for a class of partially identified models. *Econometrica*, *76*, 763–814.

Billard, L., & Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic data analysis. *J Amer Statist Assoc 98*, 470–487.

Billard, L., & Diday, E. (2006). *Symbolic data analysis: Conceptual statistics and data mining*. Chichester, England: John Wiley & Sons.

Brito, P. & Duarte Silva, A. P. (2012). Modelling interval data with Normal and Skew-Normal distributions. *Journal of Applied Statistics*, *39*, 3–20.

Domingues, M. A. O., de Souza, R. M. C. R., & Cysneiros, F. J. A. (2010). A robust method for linear regression of symbolic interval data. *Pattern Recognition Letters 31*, 1991–1996.

Durrett, R. (2010). *Probability: Theory and examples*. New York, NY: Cambridge University Press.

Evans, M., & Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science*, *10*, 254–272.

Evans, M., & Swartz, T. (2000). *Approximating integrals via Monte Carlo and deterministic methods*. New York, NY: Oxford University Press.

Fisher, R., O'Leary, R. A., Low-Choy, S., Mengersen, K., Knowlton, N., Brainard, R. E., & Caley, M. J. (2015). Species richness on coral reefs and the pursuit of convergent global estimates. *Current Biology 25*, 500–505.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Geyer, C. J., & Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *54*, 657–683.

Jeon, Y., Ahn, J., & Park, C. (2015). A nonparametric kernel approach to inverval-valued data analysis. *Technometrics*, *57*, 566–575.

Le-Rademacher, J., & Billard, L. (2011). Likelihood functions and some maximum likelihood estimators for symbolic data. *Journal of Statistical Planning and Inference*, *141*, 1593–1602.

Lin, H., Caley, M. J., & Sisson, S. A. (2017). Estimating global species richness using symbolic data meta-analysis. https://arxiv.org/abs/1711.03202

Lyashenko, N. N. (1983). Statistics of random compacts in Euclidean space. *Journal of Soviet Mathematics*, *21*, 76–92.

Matheron, G. (1975). *Random sets and integral geometry*. New York, NY: Wiley.

McLachlan, G. J., & Jones, P. N. (1988). Fitting mixture models to grouped and truncated data via the EM algorithm. *Biometrics*, *44*, 571–578.

Molchanov, I. (2005). *Theory of random sets*. London, England: Springer.

Molchanov, I., & Molinari, F. (2014). Applications of random set theory in Econometrics. *Annual Review of Economics*, *6*, 229–251.

Moore, R. E. (1966). *Interval analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Munkres, J. R. (2000). *Topology*. London, UK: Pearson.

Neto, E. A. L., & de Carvalho, F. A. T. (2010). Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics and Data Analysis*, *54*, 333–347.

Noirhomme-Fraiture, M., & Brito, P. (2011). Far beyond the classical data models: Symbolic data analysis. *Statistical Analysis and Data Mining*, *4*, 157–170.

Reiss, R.-D. (1989). *Approximate distributions of order statistics: With applications to nonparamteric statistics*. New York, NY: Springer.

Sisson, S. A., Fan, Y., & Beaumont, M. A. (2018). *Handbook of approximate Bayesian computation*. New York, NY: Chapman and Hall/CRC Press.

Sun, Y., & Ralescu, D. (2015). A normal hierarchical model and minimum contrast estimation for random intervals. *Annals of the Institute of Statistical Mathematics*, *67*, 313–333.

Vardeman, S. B., & Lee, C.-S. (2005). Likelihood-based statistical estimation from quantized data. *IEEE Transactions on Instrumentation and Measurement*, *54*, 409–414.

Wei, G. C. G., & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, *85*, 699–704.

Xu, W. (2010). Symbolic data analysis: Interval-valued data regression (Doctoral dissertation). University of Georgia, Athens, GA.

How to cite this article: Zhang X, Beranger B, Sisson SA. Constructing likelihood functions for interval-valued random variables. *Scand J Statist*. 2020;47:1–35. https://doi.org/10.1111/sjos.12395

APPENDIX

A.1 | Constructing a measurable space

We denote Ω as a sample space equipped with a σ-algebra $\mathscr{F}$ and a probability measure $P(\cdot)$. In order to construct a measurable space of $\mathbb{I}$, we identify those subsets of $\mathbb{I}$, which are equivalent to particular subsets of $\mathbb{R}^m$. A subset of interest is $\{[x'] \subseteq [x]\} = \{[x'] : [x'] \subseteq [x]\}$, which corresponds to the collection of all intervals that are a subinterval of or equal to $[x]$. This subset is the image of the event $\{[X] \subseteq [x]\} = \{\omega \in \Omega : [X](\omega) \subseteq [x]\}$ on $\mathbb{I}$. The subset $\{[X] \subseteq [x]\}$ may also be written as $\{\varphi(X_{1:m}) \subseteq [x]\} = \{\omega \in \Omega : \varphi(X_{1:m}(\omega)) \subseteq [x]\}$, of which the image on $\mathbb{R}^m$ is $\{\varphi(x'_{1:m}) \subseteq [x]\} = \{x'_{1:m} : \varphi(x'_{1:m}) \subseteq [x]\}$, that is, the subset of $\mathbb{R}^m$ containing those $x'_{1:m}$ that can generate an interval that is a subinterval of or equal to $[x]$. The two subsets $\{[x'] \subseteq [x]\}$ and $\{\varphi(x'_{1:m}) \subseteq [x]\}$ are equivalent as their pre-images on Ω are identical. As a result, given a probability measure on $\mathbb{R}^m$, the probability of $\{\varphi(x'_{1:m}) \subseteq [x]\}$ and, hence, of $\{[x'] \subseteq [x]\}$ can be calculated if only if it is measurable. This implies that in a measurable space of $\mathbb{I}$, $\{[x'] \subseteq [x]\}$ should be measurable.

We construct the metric topology on $\mathbb{I}$, denoted by $\mathscr{T}_{\mathbb{I}}$, induced by the Hausdorff metric, which specifies the distance between elements $[a]$ and $[b]$ as

$$d_H([a],[b]) = \max\left\{ \left|\underline{a}-\underline{b}\right|, \left|\overline{a}-\overline{b}\right| \right\},$$

where $|\cdot|$ denotes an absolute value. If we consider the mapping $h([x]) = (\underline{x}, \overline{x})$ from $\mathbb{I}$ to $\mathbb{R}^2$, then we have $d_2((\underline{a},\overline{a}),(\underline{b},\overline{b})) = d_H([a],[b])$ for any $[a],[b] \in \mathbb{I}$, where $d_2(\cdot)$ is the square metric on $\mathbb{R}^2$. That is, h is a distance-preserving map, or isometry, and hence, $(\mathbb{I}, \mathscr{T}_{\mathbb{I}})$ is isometrically embedded into the metric topological space on $\mathbb{R}^2$ induced by $d_2(\cdot)$, which is also known as the standard topology. The standard topology on $\mathbb{R}^2$ is generated by the open rectangles (Munkres, 2000). This implies that $\mathscr{T}_{\mathbb{I}}$ inherits properties of the standard topology on $\mathbb{R}^2$, such as completeness, local compactness, and separability (see Section A.2 for details).

Let $\mathcal{F} = \{\{[x'] \subseteq [x]\} : [x] \in \mathbb{I}\}$ be the collection of subsets of interest. We can now construct a measurable space involving $\mathcal{F}$ from the topology $\mathscr{T}_{\mathbb{I}}$. Let $\mathscr{B}_{\mathbb{I}}$ be the smallest σ-algebra containing all open subsets $\mathscr{B}_{\mathbb{I}} = \sigma(\mathscr{T}_{\mathbb{I}})$, that is, the Borel σ-algebra on $\mathbb{I}$. The topology $\mathscr{T}_{\mathbb{I}}$ is the collection of all open subsets of $\mathbb{I}$, and the Borel σ-algebra is the smallest σ-algebra containing all open

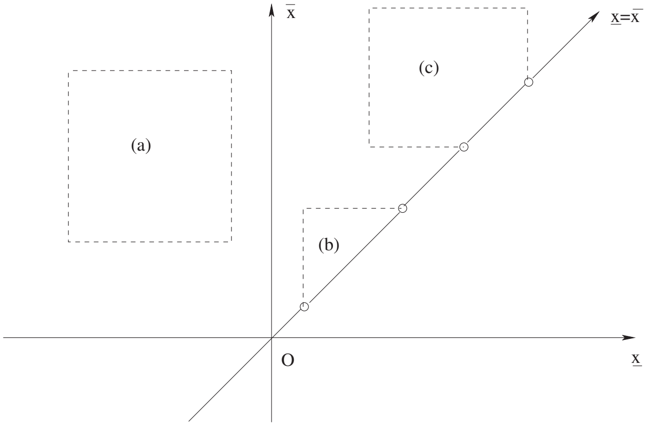


FIGURE A1 $B([a],[b])$ and $W\{[a,b]\}$ are (a) and (b), respectively. (a), (b), and (c) constitute the basis of $\mathscr{T}_{\mathbb{I}}$

subsets (Munkres, 2000). This Borel σ-algebra contains $\mathcal{F}$, as all elements of $\mathcal{F}$ are closures of some elements of $\mathscr{T}_{\mathbb{I}}$ (Section A.2). The following lemma provides a stronger result that $\mathcal{F}$ is sufficient to construct $\mathscr{B}_{\mathbb{I}}$.

Lemma 1. *The Borel σ-algebra on $\mathbb{I}$ is the smallest σ-algebra generated by $\mathcal{F}$, that is, $\mathscr{B}_{\mathbb{I}} = \sigma(\mathcal{F})$.*

This property indicates that $\mathscr{B}_{\mathbb{I}}$ is rich enough to ensure that all elements in $\mathcal{F}$ are measurable. It also suggests that if we only define a proper nonnegative function on $\mathcal{F}$, we can extend it to a measure on $(\mathbb{I}, \mathscr{B}_{\mathbb{I}})$. In particular, if the induced measure is a probability measure, it would then be the distribution function of $[X]$.

Based on the isometry $h([x]) = (\underline{x}, \overline{x})$ between $\mathbb{I}$ and $\mathbb{R}^2$, we now construct a measure on $(\mathbb{I}, \mathscr{B}_{\mathbb{I}})$, representing the uniform measure on $\mathbb{I}$, which gives equal weight to all intervals. Let the Borel σ-algebra on $\mathbb{R}^2$ be $\mathscr{B}_{\mathbb{R}^2}$ and $\mu\colon \mathscr{B}_{\mathbb{R}^2} \mapsto [0, +\infty)$ be the Lebesgue measure on $(\mathbb{R}^2, \mathscr{B}_{\mathbb{R}^2})$. Due to the isometry $h([x]) = (\underline{x}, \overline{x})$, we then have that $\mu_{\mathbb{I}} = \mu \circ h$ is the uniform measure on $(\mathbb{I}, \mathscr{B}_{\mathbb{I}})$. Consequently, the uniform measure of every Borel subset of $\mathbb{I}$ can be calculated via $\mu(\cdot)$ and $h(\cdot)$. Specifically, for every element of $\mathcal{F}$, we have

$$\mu_{\mathbb{I}}(\{[x'] \subseteq [x]\}) = \mu(h(\{[x'] \subseteq [x]\})) = \frac{1}{2}\left(\overline{x} - \underline{x}\right)^2,$$

as $h(\{[x'] \subseteq [x]\}) = \{(\underline{x}', \overline{x}') : \underline{x} \leq \underline{x}' \leq \overline{x}' \leq \overline{x}\}$ is the region of an isosceles right triangle on the real plane. From Lemma 1, the uniform measure of all Borel subsets $E \in \mathscr{B}_{\mathbb{I}}$ is also available.

Lemma 2. *Define the infinitesimal neighborhood of* $[x]$ *as*

$$\mathrm{d}[x] = \left\{[x'] \in \mathbb{I} \mid \underline{x} - \mathrm{d}\underline{x} < \underline{x}' \leq \underline{x} \leq \overline{x} \leq \overline{x}' < \overline{x} + \mathrm{d}\overline{x}\right\},$$

where $\mathrm{d}\underline{x}, \mathrm{d}\overline{x} > 0$. *Its uniform measure is* $\mu_{\mathbb{I}}(\mathrm{d}\,[x]) = \mathrm{d}\underline{x} \times \mathrm{d}\overline{x}$.

From the above, we note that $\mu_{\mathbb{I}}(\cdot)$ is a nonatomic measure, that is, $\mu_{\mathbb{I}}(\{[x]\}) = 0$, where $\{[x]\}$ is a set containing a single interval $[x]$. Furthermore, there is a convenient way to compute the value of $\mu_{\mathbb{I}}(\cdot)$ for any Borel subset via the Lebesgue integration on $\mathbb{R}^2$. That is, for any subset $E \in \mathscr{B}_{\mathbb{I}}$, we have

$$\mu_{\mathbb{I}}(E) = \int_E \mu_{\mathbb{I}}\,(\mathrm{d}[x]) = \iint_{h(E)} \mathrm{d}\underline{x}\mathrm{d}\overline{x}.$$

Accordingly, through such isometry, the measurable space of intervals $(\mathbb{I}, \mathscr{B}_{\mathbb{I}})$ inherits the convenient structure and properties of the real plane. These results permit the construction of distribution and density functions of random intervals.

A.2 | Topology

The basis of the standard topology on $\mathbb{R}^2$ is the collection of all open rectangles. Its subspace topology induced by $\{(x, y) : x \leq y\}$, as shown in Figure A1, has the basis of which element is the remaining part of an open rectangle on the top-left half-plane. Therefore, the collection of their counterparts on $\mathbb{I}$ via the isometry, $h([x]) = (\underline{x}, \overline{x})$, is the basis of $\mathscr{T}_{\mathbb{I}}$.

The open subset of $\mathbb{I}$ corresponding to rectangle (a) in Figure A1 is

$$B([a], [b]) = \left\{[x] : \underline{b} < \underline{x} < \underline{a} \leq \overline{a} < \overline{x} < \overline{b}\right\}.$$

This is the collection of all intervals for which the lower bounds are bounded between $\underline{a}$ and $\underline{b}$, whereas the upper bounds are bounded between $\overline{a}$ and $\overline{b}$. The open subset of $\mathbb{I}$ corresponding to triangle (b) is

$$W([c]) = \left\{[x] \,:\, \underline{c} < \underline{x} \leq \overline{x} < \overline{c}\right\}.$$

This is the collection of all intervals for which the lower bounds are greater than $\underline{c}$, whereas the upper bounds are smaller than $\overline{c}$.

Lemma 3. *Suppose that $\mathcal{E}$ is the collection of all $B([a],[b])$ and $W([c])$. Then, $\mathcal{E}$ is a basis for $\mathscr{T}_{\mathbb{I}}$.*

Lemma 4. $B([a],[b]) = W([b])\backslash[\{[x] \subseteq [\underline{a}, \overline{b}]\} \cup \{[x] \subseteq [\underline{b}, \overline{a}]\}]$.

Lemma 5. *$\mathscr{T}_{\mathbb{I}}$ is the smallest topology containing all $W([c])$ and $\{[x] \subseteq [c]\}^{\mathrm{c}}$.*

A.3 | Hypercubes

Similarly, through the property of isometry, $h_p([\boldsymbol{x}]) = (\underline{x}_1, \overline{x}_1, \ldots, \underline{x}_p, \overline{x}_p)$, it can be shown that a basis of the topology $\mathscr{T}_{\mathbb{I}^p}$ is the collection of the following two classes of subsets:

$$B_p([\boldsymbol{a}],[\boldsymbol{b}]) = \left\{[\boldsymbol{x}] \,:\, \underline{b}_j < \underline{x}_j < \underline{a}_j \leq \overline{a}_j < \overline{x}_j < \overline{b}_j,\, j = 1, \ldots, p\right\},$$
$$W_p([\boldsymbol{c}]) = \left\{[\boldsymbol{x}] \,:\, \underline{c}_j < \underline{x}_j \leq \overline{x}_j < \overline{c}_j,\, j = 1, \ldots, p\right\}.$$

The next lemma shows an analogous result of Lemma 4.

Lemma 6. $B_p([\boldsymbol{a}],[\boldsymbol{b}]) = W_p([\boldsymbol{b}])\backslash \cup_{j=1}^{p} \left[\{[\boldsymbol{x}] \subseteq [\boldsymbol{a}_{j1}]\} \cup \{[\boldsymbol{x}] \subseteq [\boldsymbol{a}_{j2}]\}\right]$, *where*

$$[\boldsymbol{a}_{j1}] = \left([a_1], \ldots, \left[\underline{a}_j, \overline{b}_j\right], \ldots, [a_p]\right),$$
$$\left[\boldsymbol{a}_{j2}\right] = \left([a_1], \ldots, \left[\underline{b}_j, \overline{a}_j\right], \ldots, [a_p]\right).$$

Similar to the proof of Lemma 4, based on the above lemma, the hypercube's version of Lemma 1 can be proved in a similar way.

A.4 | Proofs

A.4.1 | Proof of Lemma 1

As an isometric embedding to the standard topology of the real plane, the topology $\mathscr{T}_{\mathbb{I}}$ is separable, and thus, it has a countable basis. We define rational intervals $[q] \in \mathbb{I}$, where $\underline{q}, \overline{q}$ are rational numbers. Then, the collection of all rational intervals, $\mathbb{I}_{\mathrm{Q}}$, is dense in $\mathbb{I}$.

We first show that $\mathcal{E}_{\mathrm{Q}}$ is a countable basis of $\mathscr{T}_{\mathbb{I}}$. Let $\mathcal{E}_{\mathrm{Q}}$ be the collection of all $B([q_1],[q_2])$ and $W([q])$. As rational numbers are countable, $\mathcal{E}_{\mathrm{Q}}$ is countable. It can be shown that $\mathcal{E}_{\mathrm{Q}}$ is a basis of a topology and that its generated topology is $\mathscr{T}_{\mathbb{I}}$, similarly as in Lemma 3. As a result, $\mathcal{E}_{\mathrm{Q}}$ is a countable basis of $\mathscr{T}_{\mathbb{I}}$.

Then, we show that $\sigma(\mathcal{F}) = \sigma(\mathcal{E}_{\mathrm{Q}})$. For any $\{[x'] \subseteq [x]\} \in \mathcal{F}$, $\{[x'] \subseteq [x]\} = W([x])^{\mathrm{c}}$ and $W([x]) \in \mathscr{T}_{\mathbb{I}}$ can be generated by set operations over countable elements from $\mathcal{E}_{\mathrm{Q}}$, as $\mathcal{E}_{\mathrm{Q}}$ is a countable basis of $\mathscr{T}_{\mathbb{I}}$. Hence, $\sigma(\mathcal{F}) \subseteq \sigma(\mathcal{E}_{\mathrm{Q}})$. On the other hand, for any $W([q]) \in \mathcal{E}_{\mathrm{Q}}$, we have $W([q]) = \cup_{n=k}^{\infty}\{[q'] \subseteq [\underline{q} + 1/n, \overline{q} - 1/n]\}$, where $\overline{q} - \underline{q} \geq 2/k$, and for any $B([q_1],[q_2]) \in \mathcal{E}_{\mathrm{Q}}$, we have $B([q_1],[q_2]) = W([q_2]) \backslash [\{[q] \subseteq [\underline{q_1}, \overline{q_2}]\} \cup \{[q] \subseteq [\underline{q_2}, \overline{q_1}]\}]$ (Lemma 4). Thus, $\sigma(\mathcal{E}_{\mathrm{Q}}) \subseteq \sigma(\mathcal{F})$.

That is, $\sigma(\mathcal{F}) = \sigma(\mathcal{E}_{\mathrm{Q}}) = \sigma(\mathscr{T}_{\mathbb{I}})$.

A.4.2 | Proof of Lemma 2

We let

$$B^{\star}([a],[b]) = \left\{ [x] \,:\, \underline{b} < \underline{x} \le \underline{a} \le \bar{a} \le \bar{x} < \bar{b} \right\}. \tag{A1}$$

In a way analogous to Lemma 4, we have

$$B^{\star}([a],[b]) = W([b]) \backslash \left[W\left(\left[\underline{a}, \bar{b} \right] \right) \cup W\left(\left[\underline{b}, \bar{a} \right] \right) \right]. \tag{A2}$$

By the continuity of the measure,

$$\begin{aligned} \mu_{\mathbb{I}}(W([x])) &= \mu_{\mathbb{I}} \left(\overset{\infty}{\cup} \left\{ [x]' \subseteq \left[\underline{x} + 1/n, \bar{x} - 1/n \right] \right\} \right) \\ &= \lim_{n \to \infty} \mu_{\mathbb{I}} \left(\left\{ [x]' \subseteq \left[\underline{x} + 1/n, \bar{x} - 1/n \right] \right\} \right) \\ &= \lim_{n \to \infty} \frac{1}{2} \left(\bar{x} - \underline{x} - 2/n \right)^2 = \frac{1}{2} \left(\bar{x} - \underline{x} \right)^2. \end{aligned}$$

Note that $W([\underline{a}, \bar{b}]) \cap W([\underline{b}, \bar{a}]) = W([a])$. We then have

$$\begin{aligned} \mu_{\mathbb{I}}(B^{\star}([a],[b])) &= \mu_{\mathbb{I}}(W([b])) - \mu_{\mathbb{I}} \left(W\left(\left[\underline{a}, \bar{b} \right] \right) \right) - \mu_{\mathbb{I}} \left(W\left(\left[\underline{b}, \bar{a} \right] \right) \right) + \mu_{\mathbb{I}}(W([a])) \\ &= \left(\underline{a} - \underline{b} \right) \left(\bar{b} - \bar{a} \right). \end{aligned}$$

Therefore, $\mu_{\mathbb{I}}(\mathrm{d}[x]) = \mu_{\mathbb{I}} \left(B^{\star}([x], [\underline{x} - \mathrm{d}\underline{x}, \bar{x} + \mathrm{d}\bar{x}]) \right) = \mathrm{d}\underline{x} \times \mathrm{d}\bar{x}$.

A.4.3 | Proof of Lemma 3

We first show that $\mathcal{E}$ is a basis for a topology. Note that, for any $[x] \in \mathbb{I}$, there exists at least one $E \in \mathcal{E}$ s.t. $[x] \in E$. Then, we show in the following that, for any $E_1, E_2 \in \mathcal{E}$, if $[x] \in E_1 \cap E_2$, then there exists $E_3 \in \mathcal{E}$ s.t. $[x] \in E_3$ and $E_3 \subset E_1 \cap E_2$. Note that $\vee$ and $\wedge$ take the maximum and the minimum of two operands, respectively.

(i) Consider $[x] \in B([a],[b]) \cap B([a'],[b']) \neq \emptyset$. Then, $\underline{b} \vee \underline{b}' < \underline{a} \wedge \underline{a}'$ and $\bar{a} \vee \bar{a}' < \bar{b} \wedge \bar{b}'$. From $[x] \in B([a],[b])$, we have that $\underline{b} < \underline{x} < \underline{a} \le \bar{a} < \bar{x} < \bar{b}$. From $[x] \in B([a'],[b'])$, we have that $\underline{b}' < \underline{x} < \underline{a}' \le \bar{a}' < \bar{x} < \bar{b}'$. Therefore, $\underline{b} \vee \underline{b}' < \underline{x} < \underline{a} \wedge \underline{a}'$ and $\bar{a} \vee \bar{a}' < \bar{x} < \bar{b} \wedge \bar{b}'$. There exists $[a''], [b''] \in \mathbb{I}$ s.t. $\underline{b} \vee \underline{b}' < \underline{b}'' < \underline{x} < \underline{a}'' < \underline{a} \wedge \underline{a}'$ and $\bar{a} \vee \bar{a}' < \bar{a}'' < \bar{x} < \bar{b}'' < \bar{b} \wedge \bar{b}'$. That is, $[x] \in B([a''],[b''])$ and $B([a''],[b'']) \subset B([a],[b]) \cap B([a'],[b'])$.

(ii) Consider $[x] \in W([c_1]) \cap W([c_2]) \neq \emptyset$. Then, $\underline{c}_1 \vee \underline{c}_2 < \bar{c}_1 \wedge \bar{c}_2$. From $[x] \in W([c_1])$, we have that $\underline{c}_1 < \underline{x} \le \bar{x} < \bar{c}_1$. From $[x] \in W([c_2])$, we have that $\underline{c}_2 < \underline{x} \le \bar{x} < \bar{c}_2$. Therefore, $\underline{c}_1 \vee \underline{c}_2 < \underline{x} \le \bar{x} < \bar{c}_1 \wedge \bar{c}_2$. There exists $[c] \in \mathbb{I}$ s.t. $\underline{c}_1 \vee \underline{c}_2 < \underline{c} < \underline{x} \le \bar{x} < \bar{c} < \bar{c}_1 \wedge \bar{c}_2$. That is, $[x] \in W([c])$ and $W([c]) \subset W([c_1]) \cap W([c_2])$.

(iii) Consider $[x] \in B([a],[b]) \cap W([c]) \neq \emptyset$. Then, $\underline{c} < \underline{a}$ and $\bar{c} > \bar{a}$. From $[x] \in B([a],[b])$, we have that $\underline{b} < \underline{x} < \underline{a} \le \bar{a} < \bar{x} < \bar{b}$. From $[x] \in W([c])$, we have that $\underline{c} < \underline{x} \le \bar{x} < \bar{c}$. Therefore, $\underline{c} \vee \underline{b} < \underline{x} < \underline{a} \le \bar{a} < \bar{x} < \bar{c} \wedge \bar{b}$. There exists $[a'], [b'] \in \mathbb{I}$ s.t. $\underline{c} \vee \underline{b} < \underline{b}' < \underline{x} < \underline{a}' < \underline{a}$ and $\bar{a} < \bar{a}' < \bar{x} < \bar{b}' < \bar{c} \wedge \bar{b}$. That is, $[x] \in B([a'],[b'])$ and $B([a'],[b']) \subset B([a],[b]) \cap W([c])$.

That is, $\mathcal{E}$ is a basis for a topology. Next, we show that $\mathcal{E}$ is a basis for $\mathscr{T}_{\mathbb{I}}$. Figure A1 shows that the basis of $\mathscr{T}_{\mathbb{I}}$ consists of three types of subsets. As $B([a],[b])$ is an (a)-type subset and $W\{[c]\}$ is a (b)-type subset, the topology generated by $\mathcal{E}$ is coarser than $\mathscr{T}_{\mathbb{I}}$. On the other hand, for any $[x]$ in a (c)-type subset, we can find at least one (a)-type subset or (b)-type subset that contains that

$[x]$ and subsets of that (c)-type subset. Therefore, the topology generated by $\mathcal{E}$ is finer than $\mathscr{T}_{\mathbb{I}}$. In conclusion, the topology generated by $\mathcal{E}$ is $\mathscr{T}_{\mathbb{I}}$.

A.4.4 | Proof of Lemma 4

For any $[x] \in B([a],[b])$, that is, $\underline{b} < \underline{x} < \underline{a} \leq \overline{a} < \overline{x} < \overline{b}$, we have $[x] \in W([b])$. Moreover, $[x] \nsubseteq [\underline{a},\overline{b}]$ and $[x] \nsubseteq [\underline{b},\overline{a}]$, that is, $[x] \notin \{[x] \subseteq [\underline{a},\overline{b}]\} \cup \{[x] \subseteq [\underline{b},\overline{a}]\}$. Therefore, $B([a],[b]) \subseteq W([b]) \setminus [\{[x] \subseteq [\underline{a},\overline{b}]\} \cup \{[x] \subseteq [\underline{b},\overline{a}]\}]$.

On the other hand, for any $[x] \in W([b]) \setminus [\{[x] \subseteq [\underline{a},\overline{b}]\} \cup \{[x] \subseteq [\underline{b},\overline{a}]\}]$, we have $[x] \in W([b])$, that is, $\underline{b} < \underline{x} \leq \overline{x} < \overline{b}$. Moreover, $[x] \nsubseteq [\underline{a},\overline{b}]$ and $[x] \nsubseteq [\underline{b},\overline{a}]$, that is, $\underline{x} < \underline{a}$ and $\overline{x} > \overline{a}$. Hence, $\underline{b} < \underline{x} < \underline{a}$ and $\overline{a} < \overline{x} < \overline{b}$, that is, $[x] \in B([a],[b])$. Therefore, $W([b]) \setminus [\{[x] \subseteq [\underline{a},\overline{b}]\} \cup \{[x] \subseteq [\underline{b},\overline{a}]\}] \subseteq B([a],[b])$.

In conclusion, $B([a],[b]) = W([b]) \setminus [\{[x] \subseteq [\underline{a},\overline{b}]\} \cup \{[x] \subseteq [\underline{b},\overline{a}]\}]$.

A.4.5 | Proof of Lemma 5

$\{[x] \subseteq [c]\}$ is a closed subset, as it is the closure of $W\{[c]\}$. Accordingly, its complement $\{[x] \subseteq [c]\}^{\mathrm{c}}$ is open, and thus, $\{[x] \subseteq [c]\}^{\mathrm{c}} \in \mathscr{T}_{\mathbb{I}}$. From Lemma 4, $B([a],[b]) = W([b]) \cap \{[x] \subseteq [\underline{a},\overline{b}]\}^{\mathrm{c}} \cap \{[x] \subseteq [\underline{b},\overline{a}]\}^{\mathrm{c}}$. Hence, every element in $\mathcal{E}$ can be generated by set operations over finite elements of $W\{[c]\}$ and $\{[x] \subseteq [c]\}^{\mathrm{c}}$. As $\mathcal{E}$ is a basis of $\mathscr{T}_{\mathbb{I}}$, every element in $\mathscr{T}_{\mathbb{I}}$ can be generated by set operations over finite elements of $W\{[c]\}$ and $\{[x] \subseteq [c]\}^{\mathrm{c}}$. Therefore, $\mathscr{T}_{\mathbb{I}}$ is the smallest topology containing $W([c])$ and $\{[x] \subseteq [c]\}^{\mathrm{c}}$.

A.4.6 | Proof of Theorem 1

For any function $f_{[X]}(\underline{x},\overline{x})$ satisfying the conditions in the theorem, we can construct its containment distribution function $F_{[X]}(\underline{x},\overline{x})$ as

$$F_{[X]}\left(\underline{x},\overline{x}\right) = \int_{\underline{x}}^{\overline{x}} \int_{\underline{x}}^{b} f_{[X]}(a,b)\,\mathrm{d}a\mathrm{d}b \text{ or } F_{[X]}\left(\underline{x},\overline{x}\right) = \int_{\underline{x}}^{\overline{x}} \int_{a}^{\overline{x}} f_{[X]}(a,b)\,\mathrm{d}b\mathrm{d}a.$$

It is easy to check that $F_{[X]}(\underline{x},\overline{x})$ satisfies the conditions in Definition 1.

A.4.7 | Proof of Theorem 2

Let $C_{[X]}([x]) = F_{[X]}(\underline{x},\overline{x})$ be the containment functional. From Theorem 3 and its proof, it determines a unique probability measure $P_{[X]}\colon \mathscr{B}_{\mathbb{I}} \mapsto [0,1]$ on the space of intervals subject to $P_{[X]}([x]) = F_{[X]}(\underline{x},\overline{x})$. As $\mathrm{d}[x] = B_{\star}([x],[\underline{x}-\mathrm{d}\underline{x},\overline{x}+\mathrm{d}\overline{x}])$, from (A1) and (A2), we have

$$B_{\star}\left([x],\left[\underline{x}-\mathrm{d}\underline{x},\overline{x}+\mathrm{d}\overline{x}\right]\right) = W\left(\left[\underline{x}-\mathrm{d}\underline{x},\overline{x}+\mathrm{d}\overline{x}\right]\right) \setminus \left[W\left(\left[\underline{x},\overline{x}+\mathrm{d}\overline{x}\right]\right) \cup W\left(\left[\underline{x}-\mathrm{d}\underline{x},\overline{x}\right]\right)\right].$$

Therefore, we obtain

$$\begin{aligned} P_{[X]}(\mathrm{d}[x]) = {} & P_{[X]}\left(W\left(\left[\underline{x}-\mathrm{d}\underline{x},\overline{x}+\mathrm{d}\overline{x}\right]\right)\right) - P_{[X]}\left(W\left(\left[\underline{x},\overline{x}+\mathrm{d}\overline{x}\right]\right)\right) \\ & - P_{[X]}\left(W\left(\left[\underline{x}-\mathrm{d}\underline{x},\overline{x}\right]\right)\right) + P_{[X]}\left(W\left(\left[\underline{x},\overline{x}\right]\right)\right). \end{aligned}$$

By the continuity of the measure and $W([x]) = \cup_{n=k}^{\infty}\left\{[x]' \subseteq \left[\underline{x}+\frac{1}{n},\overline{x}-\frac{1}{n}\right]\right\}$,

$$P_{[X]}(W([x])) = \lim_{\underline{x}'\to\underline{x}+}\lim_{\overline{x}'\to\overline{x}-} P_{[X]}([X] \subseteq [x]') = \lim_{\underline{x}'\to\underline{x}+}\lim_{\overline{x}'\to\overline{x}-} F_{[X]}\left(\underline{x}',\overline{x}'\right).$$

As $F_{[X]}$ is twice differentiable (thus continuous), $P_{[X]}(W([x])) = F_{[X]}(\underline{x}, \overline{x})$. Therefore, we have

$$P_{[X]}(\mathrm{d}[x]) = F_{[X]}\left(\underline{x} - \mathrm{d}\underline{x}, \overline{x} + \mathrm{d}\overline{x}\right) - F_{[X]}\left(\underline{x}, \overline{x} + \mathrm{d}\overline{x}\right) - F_{[X]}\left(\underline{x} - \mathrm{d}\underline{x}, \overline{x}\right) + F_{[X]}\left(\underline{x}, \overline{x}\right).$$

Substituting second-order Taylor expansions for the first three terms in the above equation, we obtain

$$P_{[X]}(\mathrm{d}[x]) = -\frac{\partial^2}{\partial \underline{x} \partial \overline{x}} F_{[X]}\left(\underline{x}, \overline{x}\right) \mathrm{d}\underline{x}\mathrm{d}\overline{x} + o\left(\mathrm{d}\underline{x}\mathrm{d}\overline{x}\right).$$

Note that $\mu_{\mathbb{I}}(\mathrm{d}[x]) = \mathrm{d}\underline{x}\mathrm{d}\overline{x}$ (Theorem 2), and so,

$$P_{[X]}(\mathrm{d}[x]) = -\frac{\partial^2}{\partial \underline{x} \partial \overline{x}} F_{[X]}\left(\underline{x}, \overline{x}\right) \mu_{\mathbb{I}}(\mathrm{d}[x]) + o(\mu_{\mathbb{I}}(\mathrm{d}[x])).$$

In addition, $P_{[X]}(\mathrm{d}[x]) = 0$ when $\mu_{\mathbb{I}}(\mathrm{d}[x]) = 0$, that is, $P_{[X]}(\cdot)$ is absolute continuous w.r.t. $\mu_{\mathbb{I}}(\cdot)$. Therefore, the Radon–Nikodym derivative exists, and

$$\frac{P_{[X]}(\mathrm{d}[x])}{\mu_{\mathbb{I}}(\mathrm{d}[x])} = -\frac{\partial^2}{\partial \underline{x} \partial \overline{x}} F_{[X]}\left(\underline{x}, \overline{x}\right).$$

A.4.8 | Proof of Theorem 3

As $\mathscr{B}_{\mathbb{I}} = \sigma(\mathcal{F})$ (Lemma 1), any $E \in \mathscr{B}_{\mathbb{I}}$ can be generated by set operations over, at most, countable elements from $\mathcal{F}$. Hence, its probability measure $P([X] \in E)$ will be available if $P([X] \subseteq [x])$ is known for any $[x]$. Therefore, the uniqueness has been proved.

Next, we prove the existence of a probability measure $P_{[X]}: \mathscr{B}_{\mathbb{I}} \mapsto [0,1]$ on the space of intervals subject to $P_{[X]}(\{[x'] \subseteq [x]\}) = C_{[X]}([x])$. Let $\mathcal{G}$ be the collection of all $B'([x],[y]) = \{[x'] : \underline{y} \le \underline{x}' < \underline{x} \le \overline{x} < \overline{x}' \le \overline{y}\}$. Similar to Lemma 4, we have $B'([x],[y]) = \{[x]' \subseteq [y]\} \backslash [\{[x]' \subseteq [\underline{x}, \overline{y}]\} \cup \{[x]' \subseteq [\underline{y}, \overline{x}]\}]$. Then, define $\mathcal{H} = \mathcal{F} \cup \mathcal{G} \cup \{\emptyset, \mathbb{I}\}$ and extend $C_{[X]}(\cdot)$ to a function $P_C(\cdot)$ on $\mathcal{H}$ s.t. $P_C(\emptyset) = 0$, $P_C(\mathbb{I}) = 1$, $P_C(\{[x'] \subseteq [x]\}) = C_{[X]}([x])$, and

$$P_C(B'([x],[y])) = C_{[X]}([y]) - C_{[X]}\left([\underline{x}, \overline{y}]\right) - C_{[X]}\left(\left[\underline{y}, \overline{x}\right]\right) + C_{[X]}([x]) \ge 0,$$

by condition (*iii*) of the definition of $C_{[X]}(\cdot)$ in Section 2.2. That is, $P_C(\cdot)$ is nonnegative.

In addition, as $\mathbb{I}$ is locally compact, for any $A \subset \mathbb{I}$ and $\delta > 0$, there exists $E_1, \ldots, E_N \in \mathcal{H}$ with all $\mu_{\mathbb{I}}(E_i) \le \delta$, such that $A \subset \cup_{i=1}^N E_i$. Therefore, we can use Carathéodory construction (Durrett, 2010) to define a metric outer measure. Let $P^\star_{[X]}(A) = \lim_{\delta \to 0} P_\delta(A)$, where

$$P_\delta(A) = \inf\left\{\sum_{i=1}^{\infty} P_C(E_i) : E_i \in \mathcal{H}, \mathrm{diam}(E_i) \le \delta, \cup_{i=1}^{\infty} E_i \supseteq A\right\},$$

where $\mathrm{diam}(E_i)$ is the diameter of E_i. Hence, $P^\star_{[X]}(\cdot)$ is a metric outer measure, and thus, the Borel subsets on $\mathbb{I}$ are measurable w.r.t. $P^\star_{[X]}(\cdot)$. That is, there exists a probability measure $P_{[X]}: \mathscr{B}_{\mathbb{I}} \mapsto [0,1]$, such that $P_{[X]}(E) = P^\star_{[X]}(E)$ for any $E \in \mathscr{B}_{\mathbb{I}}$.

Finally, we can check that $P_{[X]}(\{[x'] \subseteq [x]\}) = C_{[X]}([x])$. For any $n = 1, 2, \ldots,$ there exists $\delta_n \to 0$ as $n \to \infty$ s.t. $P_{\delta_n}(\{[x'] \subseteq [x]\}) \le C_{[X]}([\underline{x} - \frac{1}{n}, \overline{x} + \frac{1}{n}])$. Moreover, $P_\delta(\{[x'] \subseteq [x]\}) \ge C_{[X]}([x])$ by definition for any $\delta > 0$. Therefore, we have

$$C_{[X]}([x]) \le \lim_{n\to\infty} P_{\delta_n}(\{[x'] \subseteq [x]\}) \le \lim_{n\to\infty} C_{[X]}\left(\left[\underline{x} - 1/n, \overline{x} + 1/n\right]\right).$$

By condition (*ii*) of the definition of $C_{[X]}(\cdot)$ in Section 2.2, we have

$$\lim_{n\to\infty} C_{[X]}\left(\left[\underline{x} - 1/n, \overline{x} + 1/n\right]\right) = C_{[X]}\left(\cap_{n=1}^{\infty}\left[\underline{x} - 1/n, \overline{x} + 1/n\right]\right) = C_{[X]}([x]).$$

Therefore,

$$P_{[X]}(\{[x'] \subseteq [x]\}) = \lim_{n\to\infty} P_{\delta_n}(\{[x'] \subseteq [x]\}) = C_{[X]}([x]).$$

As a result, given a random interval $[X] : \Omega \mapsto \mathbb{I}$, we obtain a probability measure $P : \sigma([X]) \mapsto [0,1]$ s.t. $P([X] \subseteq [x]) = P_{[X]}(\{[x'] \subseteq [x]\}) = C_{[X]}([x])$.

A.4.9 | Proof of Theorem 4

Let $c = \frac{a+b}{2} \in (-\infty, +\infty)$ and $r = \frac{b-a}{2} \geq 0$. We can rewrite

$$f_{[X]}\left(\underline{x}, \overline{x} \mid m\right) = \iint_{\{a\leq \underline{x}, b\geq \overline{x}\}} m(m-1)\frac{\left(\overline{x}-\underline{x}\right)^{m-2}}{(b-a)^m}\pi(a,b)\mathrm{d}a\mathrm{d}b$$

as

$$f_{[X]}(\underline{x}, \overline{x} \mid m) = 2^{-m}m(m-1)\left(\overline{x}-\underline{x}\right)^{m-2}\iint_A r^{-m}g(c,r)\mathrm{d}c\mathrm{d}r,$$

where $g(c,r) = 2\pi(c-r, c+r)$ is the density function of (c,r) and $A = \left\{(c,r) : \overline{x} - r \leq c \leq \underline{x} + r, r \geq \frac{\overline{x}-\underline{x}}{2}\right\}$. As $\pi(\cdot)$ is bounded continuously, $\int_{-\infty}^{+\infty} g(c,r)\mathrm{d}c < \infty$. Let $g(r) = \int_{-\infty}^{+\infty} g(c,r)\mathrm{d}c$, $B_0 = \{r : g(r) = 0\}$, and $B_1 = \{r : g(r) \neq 0\}$. When $g(r) \neq 0$, we have $g(c \mid r) = \frac{g(c,r)}{g(r)}$. The above integration can be decomposed into the following two cases. In the case that $g(r) \neq 0$, we replace $g(c,r)$ with $g(r)g(c \mid r)$ and integrate out c, that is,

$$\iint_{A\cap B_1} r^{-m}g(c,r)\mathrm{d}c\mathrm{d}r = \int_{\frac{\overline{x}-\underline{x}}{2}}^{\infty} r^{-m}g(r)\left\{\int_{\overline{x}-r}^{\underline{x}+r} g(c \mid r)\mathrm{d}c\right\}\mathrm{d}r.$$

In the case that $g(r) = 0$, we have $g(c,r) = 0$ and $\iint_{A\cap B_0} r^{-m}g(c,r)\mathrm{d}c\mathrm{d}r = 0$.

Then, writing $z = (m-1)(\log r - \log \frac{\overline{x}-\underline{x}}{2})$, we have

$$f_{[X]}\left(\underline{x}, \overline{x} \mid m\right) = \frac{1}{2}m\left(\overline{x}-\underline{x}\right)^{-1}$$
$$\times \int_0^{\infty} \mathrm{e}^{-z}g\left(\frac{\overline{x}-\underline{x}}{2}\mathrm{e}^{(m-1)^{-1}z}\right)\left\{\int_{\overline{x}-\frac{\overline{x}-\underline{x}}{2}\mathrm{e}^{(m-1)^{-1}z}}^{\underline{x}+\frac{\overline{x}-\underline{x}}{2}\mathrm{e}^{(m-1)^{-1}z}} g\left(c \mid \frac{\overline{x}-\underline{x}}{2}\mathrm{e}^{(m-1)^{-1}z}\right)\mathrm{d}c\right\}\mathrm{d}z.$$

As $\pi(\cdot)$ is bounded continuously, $g(c,r) = 2\pi(c-r, c+r)$ is bounded continuously. Due to the mean value theorem, the above term can be simplified as

$$f_{[X]}\left(\underline{x}, \overline{x} \mid m\right) = \frac{1}{2}\int_0^{\infty} m\left\{\mathrm{e}^{(m-1)^{-1}z} - 1\right\}\mathrm{e}^{-z}g\left(\xi, \frac{\overline{x}-\underline{x}}{2}\mathrm{e}^{(m-1)^{-1}z}\right)\mathrm{d}z,$$

where $\underline{x} + \frac{\overline{x}-\underline{x}}{2}\mathrm{e}^{(m-1)^{-1}z} \leq \xi \leq \overline{x} - \frac{\overline{x}-\underline{x}}{2}\mathrm{e}^{(m-1)^{-1}z}$. Let $M(\xi) = \sup_{z\geq 0} g(\xi, \frac{\overline{x}-\underline{x}}{2}\mathrm{e}^{(m-1)^{-1}z})$. $M(\xi)$ is bounded as $g(c,r)$ is bounded. When $m \geq 3$, we have

$$f_{[X]}\left(\underline{x}, \overline{x} \mid m\right) \leq \frac{M(\xi)}{2}\int_0^{\infty} m\left\{\mathrm{e}^{(m-1)^{-1}z} - 1\right\}\mathrm{e}^{-z}\mathrm{d}z = \frac{m}{2(m-2)}M(\xi) \leq \frac{3}{2}M(\xi).$$

Therefore, $f_{[X]}(\underline{x}, \overline{x} \mid m)$ is bounded when $m \to \infty$, and thus,

$$\lim_{m\to\infty} f_{[X]}\left(\underline{x}, \overline{x} \mid m\right) = \frac{1}{2}\int_0^{\infty} \lim_{m\to\infty} m\left\{\mathrm{e}^{(m-1)^{-1}z} - 1\right\}\mathrm{e}^{-z}g\left(\xi, \frac{\overline{x}-\underline{x}}{2}\mathrm{e}^{(m-1)^{-1}z}\right)\mathrm{d}z$$
$$= \frac{1}{2}g\left(\frac{\underline{x}+\overline{x}}{2}, \frac{\overline{x}-\underline{x}}{2}\right) = \pi\left(\underline{x}, \overline{x}\right).$$

A.4.10 | Proof of Theorem 5

Let $f_{\mu,\tau} = f(\cdot \mid \mu,\tau)$, and denote $F_{\mu,\tau} = F(\cdot \mid \mu,\tau)$ and $Q_{\mu,\tau} = Q(\cdot;\mu,\tau)$ as its cumulative distribution function and quantile function, respectively. As $f_{\mu,\tau}$ is positive and continuous in the neighborhoods of $Q_{\mu,\tau}(\underline{p})$ and $Q_{\mu,\tau}(\overline{p})$ with $\underline{p} > 0$ and $\overline{p} < 1$, the joint density function of

$$\begin{cases} (m+1)^{\frac{1}{2}} f_{\mu,\tau}\left(Q_{\mu,\tau}\left(\underline{p}\right)\right)\left(\underline{X} - Q_{\mu,\tau}\left(\underline{p}\right)\right) \\ (m+1)^{\frac{1}{2}} f_{\mu,\tau}\left(Q_{\mu,\tau}\left(\overline{p}\right)\right)\left(\overline{X} - Q_{\mu,\tau}\left(\overline{p}\right)\right) \end{cases}$$

converges pointwise to a bivariate normal density function, with zero mean and covariance matrix

$$\Sigma = \begin{pmatrix} p(1-\underline{p}) & \underline{p}(1-\overline{p}) \\ \underline{p}(1-\overline{p}) & \overline{p}(1-\overline{p}) \end{pmatrix}$$

when $m \to \infty$ (Reiss, 1989). Thus, when m is large, the density function of the i.i.d. generative model

$$\begin{aligned} f^{\star}_{[X]}\left(\underline{x},\overline{x} \mid \theta, m, l, u\right) = {} & \frac{m!}{(l-1)!(u-l-1)!(m-u)!}\left[F\left(\underline{x} \mid \theta\right)\right]^{l-1} \\ & \times \left[F\left(\overline{x} \mid \theta\right) - F\left(\underline{x} \mid \theta\right)\right]^{u-l-1}\left[1 - F\left(\overline{x} \mid \theta\right)\right]^{m-u} f\left(\underline{x} \mid \theta\right) f\left(\overline{x} \mid \theta\right) \end{aligned}$$

is asymptotically equivalent to

$$\frac{m+1}{2\pi|\Sigma|^{\frac{1}{2}}} f_{\mu,\tau}\left(Q_{\mu,\tau}\left(\underline{p}\right)\right) f_{\mu,\tau}\left(Q_{\mu,\tau}\left(\overline{p}\right)\right) \exp\left\{-(m+1)T\left(\underline{x},\overline{x};\mu,\tau\right)\right\},$$

where

$$\begin{aligned} T\left(\underline{x},\overline{x};\mu,\tau\right) &= \frac{1}{2}\left(\underline{t}\left(\underline{x};\mu,\tau\right), \overline{t}\left(\overline{x};\mu,\tau\right)\right)\Sigma^{-1}\left(\underline{t}\left(\underline{x};\mu,\tau\right), \overline{t}\left(\overline{x};\mu,\tau\right)\right)^{\mathsf{T}}, \\ \underline{t}\left(\underline{x};\mu,\tau\right) &= f_{\mu,\tau}\left(Q_{\mu,\tau}\left(\underline{p}\right)\right)\left(\underline{x} - Q_{\mu,\tau}\left(\underline{p}\right)\right), \\ \overline{t}\left(\overline{x};\mu,\tau\right) &= f_{\mu,\tau}\left(Q_{\mu,\tau}\left(\overline{p}\right)\right)\left(\overline{x} - Q_{\mu,\tau}\left(\overline{p}\right)\right). \end{aligned}$$

That is, the density function of the hierarchical generative model

$$f_{[X]}\left(\underline{x},\overline{x} \mid \alpha, m\right) = \int f^{\star}_{[X]}\left(\underline{x},\overline{x} \mid \theta, m\right)\pi\left(\theta \mid \alpha\right)\mathrm{d}\theta$$

is asymptotically equivalent to

$$\frac{m+1}{2\pi|\Sigma|^{\frac{1}{2}}} \times H\left(\underline{x},\overline{x};\underline{p},\overline{p},m\right), \tag{A3}$$

where

$$H\left(\underline{x},\overline{x};\underline{p},\overline{p},m\right) = \iint f_{\mu,\tau}\left(Q_{\mu,\tau}\left(\underline{p}\right)\right) f_{\mu,\tau}\left(Q_{\mu,\tau}\left(\overline{p}\right)\right)\pi(\mu,\tau)\exp\left\{-(m+1)T\left(\underline{x},\overline{x};\mu,\tau\right)\right\}\mathrm{d}\mu\mathrm{d}\tau.$$

Note that Σ is positive definite, and so, $T(\underline{x},\overline{x};\mu,\tau) \geq 0$. In addition, $T(\underline{x},\overline{x};\underline{p},\overline{p},m)$ reaches its minimum 0, when $Q_{\mu,\tau}(\underline{p}) = \underline{x}$ and $Q_{\mu,\tau}(\overline{p}) = \overline{x}$. As $f_{\mu,\tau}$ is interval identifiable, the system of equations, $Q_{\mu,\tau}(\underline{p}) = \underline{x}$ and $Q_{\mu,\tau}(\overline{p}) = \overline{x}$, has a unique solution, and thus, $T(\underline{x},\overline{x};\underline{p},\overline{p},m)$ is unimodal.

As $\mu_{\star} = \mu(\underline{x},\overline{x};\underline{p},\overline{p})$ and $\tau_{\star} = \tau(\underline{x},\overline{x};\underline{p},\overline{p})$ are the solutions of $Q_{\mu,\tau}(\underline{p}) = \underline{x}$ and $Q_{\mu,\tau}(\overline{p}) = \overline{x}$, given conditions ($i$) and ($iii$) in the theorem, a Laplace approximation can be applied to $H(\underline{x},\overline{x};\underline{p},\overline{p},m)$ at the point $(\mu_{\star},\tau_{\star})$, giving

$$H\left(\underline{x},\overline{x};\underline{p},\overline{p},m\right) \approx 2\pi(m+1)^{-1}\left|\nabla^2 T\left(\underline{x},\overline{x};\mu_{\star},\tau_{\star}\right)\right|^{-\frac{1}{2}} f_{\mu_{\star},\tau_{\star}}\left(\underline{x}\right) f_{\mu_{\star},\tau_{\star}}\left(\overline{x}\right)\pi(\mu_{\star},\tau_{\star}). \tag{A4}$$

We let $T = T(\underline{x}, \bar{x}; \mu, \tau)$, $\underline{t} = \underline{t}(\underline{x}; \mu, \tau)$, $\bar{t} = \bar{t}(\bar{x}; \mu, \tau)$, and $\Sigma^{-1} = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}$; hence, we have $T = \frac{1}{2}(a_{11}\underline{t}^2 + 2a_{12}\underline{t}\bar{t} + a_{22}\bar{t}^2)$. The first-order partial derivatives of T are

$$\begin{aligned} \frac{\partial T}{\partial \mu} &= a_{11}\underline{t}\frac{\partial \underline{t}}{\partial \mu} + a_{12}\bar{t}\frac{\partial \underline{t}}{\partial \mu} + a_{12}\underline{t}\frac{\partial \bar{t}}{\partial \mu} + a_{22}\bar{t}\frac{\partial \bar{t}}{\partial \mu}, \\ \frac{\partial T}{\partial \tau} &= a_{11}\underline{t}\frac{\partial \underline{t}}{\partial \tau} + a_{12}\bar{t}\frac{\partial \underline{t}}{\partial \tau} + a_{12}\underline{t}\frac{\partial \bar{t}}{\partial \tau} + a_{22}\bar{t}\frac{\partial \bar{t}}{\partial \tau}. \end{aligned}$$

Let $T^\star$, $\underline{t}^\star$, and $\bar{t}^\star$ denote the corresponding functions and their derivatives taking values at $(\mu_\star, \tau_\star)$. As $\underline{t}^\star = \bar{t}^\star = 0$, the second-order partial derivatives at $(\mu_\star, \tau_\star)$ are

$$\begin{aligned} \frac{\partial^2 T^\star}{\partial \mu^2} &= a_{11}\left(\frac{\partial \underline{t}^\star}{\partial \mu}\right)^2 + 2a_{12}\frac{\partial \bar{t}^\star}{\partial \mu}\frac{\partial \underline{t}^\star}{\partial \mu} + a_{22}\left(\frac{\partial \bar{t}^\star}{\partial \mu}\right)^2, \\ \frac{\partial^2 T^\star}{\partial \tau^2} &= a_{11}\left(\frac{\partial \underline{t}^\star}{\partial \tau}\right)^2 + 2a_{12}\frac{\partial \bar{t}^\star}{\partial \tau}\frac{\partial \underline{t}^\star}{\partial \tau} + a_{22}\left(\frac{\partial \bar{t}^\star}{\partial \tau}\right)^2, \\ \frac{\partial^2 T^\star}{\partial \mu \partial r} &= a_{11}\frac{\partial \underline{t}^\star}{\partial \mu}\frac{\partial \underline{t}^\star}{\partial \tau} + a_{12}\frac{\partial \underline{t}^\star}{\partial \mu}\frac{\partial \bar{t}^\star}{\partial \tau} + a_{12}\frac{\partial \underline{t}^\star}{\partial \tau}\frac{\partial \bar{t}^\star}{\partial \mu} + a_{22}\frac{\partial \bar{t}^\star}{\partial \mu}\frac{\partial \bar{t}^\star}{\partial \tau}. \end{aligned}$$

Therefore, $\nabla^2 T$ at $(\mu_\star, \tau_\star)$ is

$$\nabla^2 T^\star = \begin{pmatrix} \frac{\partial \underline{t}^\star}{\partial \mu} & \frac{\partial \underline{t}^\star}{\partial \tau} \\ \frac{\partial \bar{t}^\star}{\partial \mu} & \frac{\partial \bar{t}^\star}{\partial \tau} \end{pmatrix}^{\intercal} \Sigma^{-1} \begin{pmatrix} \frac{\partial \underline{t}^\star}{\partial \mu} & \frac{\partial \underline{t}^\star}{\partial \tau} \\ \frac{\partial \bar{t}^\star}{\partial \mu} & \frac{\partial \bar{t}^\star}{\partial \tau} \end{pmatrix},$$

and its determinant is $|\nabla^2 g| = |\Sigma|^{-1}\left(\frac{\partial \underline{t}^\star}{\partial \mu}\frac{\partial \bar{t}^\star}{\partial \tau} - \frac{\partial \underline{t}^\star}{\partial \tau}\frac{\partial \bar{t}^\star}{\partial \mu}\right)^2$.

The derivatives of $\underline{t}$ and $\bar{t}$ at $(\mu^\star, \tau^\star)$ are

$$\begin{aligned} \frac{\partial \underline{t}^\star}{\partial \mu} &= -f_{\mu^\star,\tau^\star}\left(\underline{x}\right) \times \tfrac{\partial}{\partial \mu} Q_{\mu^\star,\tau^\star}\left(\underline{p}\right), \\ \frac{\partial \underline{t}^\star}{\partial \tau} &= -f_{\mu^\star,\tau^\star}\left(\underline{x}\right) \times \tfrac{\partial}{\partial \tau} Q_{\mu^\star,\tau^\star}\left(\underline{p}\right), \\ \frac{\partial \bar{t}^\star}{\partial \mu} &= -f_{\mu^\star,\tau^\star}(\bar{x}) \times \tfrac{\partial}{\partial \mu} Q_{\mu^\star,\tau^\star}\left(\bar{p}\right), \\ \frac{\partial \bar{t}^\star}{\partial \tau} &= -f_{\mu^\star,\tau^\star}\left(\bar{x}\right) \times \tfrac{\partial}{\partial \tau} Q_{\mu^\star,\tau^\star}\left(\bar{p}\right), \end{aligned}$$

and thus,

$$\left|\nabla^2 T^\star\right| = |\Sigma|^{-1} f_{\mu^\star,\tau^\star}\left(\underline{x}\right)^2 f_{\mu^\star,\tau^\star}\left(\bar{x}\right)^2 \left|J\left(\mu^\star, \tau^\star; \underline{p}, \bar{p}\right)\right|^2, \tag{A5}$$

where

$$J\left(\mu^\star, \tau^\star; \underline{p}, \bar{p}\right) = \begin{pmatrix} \frac{\partial}{\partial \mu} Q_{\mu^\star,\tau^\star}\left(\underline{p}\right) & \frac{\partial}{\partial \tau} Q_{\mu^\star,\tau^\star}\left(\underline{p}\right) \\ \frac{\partial}{\partial \mu} Q_{\mu^\star,\tau^\star}\left(\bar{p}\right) & \frac{\partial}{\partial \tau} Q_{\mu^\star,\tau^\star}\left(\bar{p}\right) \end{pmatrix}.$$

From (A4) and (A5), we obtain that the density function of the hierarchical generative model (A3) converges pointwise to $\pi(\mu^\star, \tau^\star)|J(\mu^\star, \tau^\star; \underline{p}, \bar{p})|^{-1}$.

A.4.11 | Proof of Theorems 6 and 7

Similar to the proof of Theorems 1 and 2. Use Lemma 6 and Taylor expansions.

A.5 | Likelihood function of a two-dimensional i.i.d. generative model

Let $[\boldsymbol{X}] = [X_1] \times [X_2]$ be the random rectangle generated from m i.i.d. bivariate latent data points from $f(x_1, x_2 \mid \boldsymbol{\theta})$, with the data aggregation function taking the minimum and maximum values at each margin. Let $F(x_1, x_2 \mid \boldsymbol{\theta})$ be the distribution function of $f(x_1, x_2 \mid \boldsymbol{\theta})$. The distribution function of $[X_1] \times [X_2]$ is

$$F_{[\boldsymbol{X}]}\left(\underline{x}_1, \bar{x}_1, \underline{x}_2, \bar{x}_2 \mid \boldsymbol{\theta}\right) = \left[F\left(\bar{x}_1, \bar{x}_2 \mid \boldsymbol{\theta}\right) - F\left(\underline{x}_1, \bar{x}_2 \mid \boldsymbol{\theta}\right) - F\left(\bar{x}_1, \underline{x}_2 \mid \boldsymbol{\theta}\right) + F\left(\underline{x}_1, \underline{x}_2 \mid \boldsymbol{\theta}\right)\right]^m.$$

This is the probability that all m latent data points fall within the rectangle $[x_1] \times [x_2]$. From Theorem 7, the likelihood function is the fourth-order mixed derivative shown as follows:

$$\begin{aligned}
f_{[\boldsymbol{X}]}&\left(\underline{x}_1, \bar{x}_1, \underline{x}_2, \bar{x}_2 \mid \boldsymbol{\theta}\right) = m(m-1)(m-2)(m-3)\\
&\times \left\{F\left(\bar{x}_1, \bar{x}_2 \mid \boldsymbol{\theta}\right) - F\left(\underline{x}_1, \bar{x}_2 \mid \boldsymbol{\theta}\right) - F\left(\bar{x}_1, \underline{x}_2 \mid \boldsymbol{\theta}\right) + F\left(\underline{x}_1, \underline{x}_2 \mid \boldsymbol{\theta}\right)\right\}^{m-4}\\
&\times \int_{\underline{x}_1}^{\bar{x}_1} f\left(y_1, \underline{x}_2 \mid \boldsymbol{\theta}\right) \mathrm{d}y_1 \int_{\underline{x}_1}^{\bar{x}_1} f\left(y_2, \bar{x}_2 \mid \boldsymbol{\theta}\right) \mathrm{d}y_2 \int_{\underline{x}_2}^{\bar{x}_2} f\left(\underline{x}_1, y_3 \mid \boldsymbol{\theta}\right) \mathrm{d}y_3 \int_{\underline{x}_2}^{\bar{x}_2} f\left(\bar{x}_1, y_4 \mid \boldsymbol{\theta}\right) \mathrm{d}y_4\\
&+ m(m-1)(m-2)\left\{F\left(\bar{x}_1, \bar{x}_2 \mid \boldsymbol{\theta}\right) - F\left(\underline{x}_1, \bar{x}_2 \mid \boldsymbol{\theta}\right) - F\left(\bar{x}_1, \underline{x}_2 \mid \boldsymbol{\theta}\right) + F\left(\underline{x}_1, \underline{x}_2 \mid \boldsymbol{\theta}\right)\right\}^{m-3}\\
&\times \Bigg\{ f\left(\underline{x}_1, \underline{x}_2 \mid \boldsymbol{\theta}\right) \int_{\underline{x}_1}^{\bar{x}_1} f\left(y_2, \bar{x}_2 \mid \boldsymbol{\theta}\right) \mathrm{d}y_2 \int_{\underline{x}_2}^{\bar{x}_2} f\left(\bar{x}_1, y_4 \mid \boldsymbol{\theta}\right) \mathrm{d}y_4\\
&\qquad + f\left(\underline{x}_1, \bar{x}_2 \mid \boldsymbol{\theta}\right) \int_{\underline{x}_1}^{\bar{x}_1} f\left(y_1, \underline{x}_2 \mid \boldsymbol{\theta}\right) \mathrm{d}y_1 \int_{\underline{x}_2}^{\bar{x}_2} f\left(\bar{x}_1, y_4 \mid \boldsymbol{\theta}\right) \mathrm{d}y_4\\
&\qquad + f\left(\bar{x}_1, \underline{x}_2 \mid \boldsymbol{\theta}\right) \int_{\underline{x}_1}^{\bar{x}_1} f\left(y_2, \bar{x}_2 \mid \boldsymbol{\theta}\right) \mathrm{d}y_2 \int_{\underline{x}_2}^{\bar{x}_2} f\left(\underline{x}_1, y_3 \mid \boldsymbol{\theta}\right) \mathrm{d}y_3\\
&\qquad + f\left(\bar{x}_1, \bar{x}_2 \mid \boldsymbol{\theta}\right) \int_{\underline{x}_1}^{\bar{x}_1} f\left(y_1, \underline{x}_2 \mid \boldsymbol{\theta}\right) \mathrm{d}y_1 \int_{\underline{x}_2}^{\bar{x}_2} f\left(\underline{x}_1, y_3 \mid \boldsymbol{\theta}\right) \mathrm{d}y_3 \Bigg\}\\
&+ m(m-1)\left\{F\left(\bar{x}_1, \bar{x}_2 \mid \boldsymbol{\theta}\right) - F\left(\underline{x}_1, \bar{x}_2 \mid \boldsymbol{\theta}\right) - F\left(\bar{x}_1, \underline{x}_2 \mid \boldsymbol{\theta}\right) + F\left(\underline{x}_1, \underline{x}_2 \mid \boldsymbol{\theta}\right)\right\}^{m-2}\\
&\times \left\{f\left(\underline{x}_1, \underline{x}_2 \mid \boldsymbol{\theta}\right) f\left(\bar{x}_1, \bar{x}_2 \mid \boldsymbol{\theta}\right) + f\left(\underline{x}_1, \bar{x}_2 \mid \boldsymbol{\theta}\right) f\left(\bar{x}_1, \underline{x}_2 \mid \boldsymbol{\theta}\right)\right\}.
\end{aligned}$$

Although it is rather complex, in fact, it has a similar intuitive interpretation to (7). The first term denotes the case that $m - 4$ points fall within $[x_1] \times [x_2]$, whereas the remaining four points are $(y_1, \underline{x}_2)$, $(y_2, \bar{x}_2)$, $(\underline{x}_1, y_3)$, and $(\bar{x}_1, y_4)$, where $\underline{x}_1 \leq y_1, y_2 \leq \bar{x}_1$ and $\underline{x}_2 \leq y_3, y_4 \leq \bar{x}_2$. The second term represents the case that $m - 3$ points fall within $[x_1] \times [x_2]$, whereas the remaining three points determine the boundary of the rectangle. The last terms are the case where the boundary is formed by only two points.

A.6 | Additional plots from simulation study

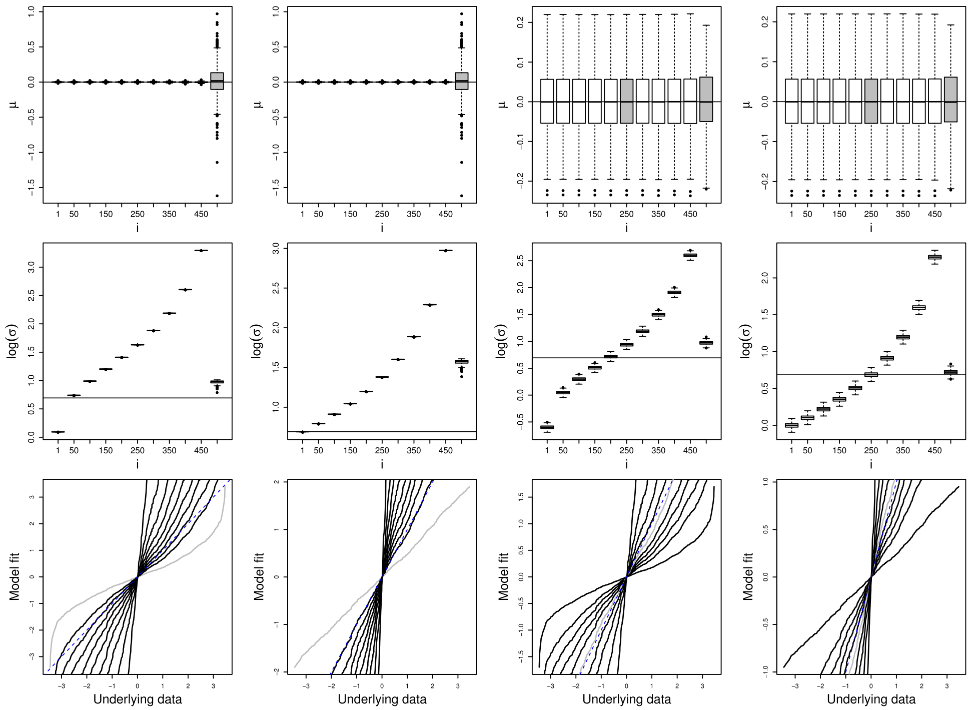


FIGURE A2 As for Figure 6 in the main text, except that the true data generating process is uniform. Boxplots of 500 replicate maximum likelihood estimates of μ and $\log \sigma$ under a uniform distribution with mean $\mu = 0$ and standard deviation $\sigma = 2$ as the true data generating process with $m = 1000$ and assuming data aggregation function ϕ_i, $i = 1, 50, 100, \dots, 450$. The true aggregation functions are ϕ_1 (left two columns) and ϕ_{250} (right two columns). The models fitted are the normal (columns 1 and 3) and uniform (columns 2 and 4) distributions. In each panel, the rightmost boxplot indicates the outcome using the data set with 5% outliers. The bottom row shows quantile–quantile curves of the fitted model (y-axis) versus the empirical underlying data quantiles (x-axis). Gray curves indicate use of the correct $\varphi(\cdot)$ function. The dashed line denotes $y = x$ [Colour figure can be viewed at wileyonlinelibrary.com]