Ann Inst Stat Math (2015) 67:313–333
DOI 10.1007/s10463-014-0453-1

A normal hierarchical model and minimum contrast estimation for random intervals

Yan Sun · **Dan Ralescu**

Received: 8 April 2013 / Revised: 31 December 2013 / Published online: 16 March 2014

Abstract Many statistical data are imprecise due to factors such as measurement errors, computation errors, and lack of information. In such cases, data are better represented by intervals rather than by single numbers. Existing methods for analyzing interval-valued data include regressions in the metric space of intervals and symbolic data analysis, the latter being proposed in a more general setting. However, there has been a lack of literature on the parametric modeling and distribution-based inferences for interval-valued data. In an attempt to fill this gap, we extend the concept of normality for random sets by Lyashenko and propose a Normal hierarchical model for random intervals. In addition, we develop a minimum contrast estimator (MCE) for the model parameters, which is both consistent and asymptotically normal. Simulation studies support our theoretical findings and show very promising results. Finally, we successfully apply our model and MCE to a real data set.

Keywords Random intervals · Uncertainty · Normality · Choquet functional · Minimum contrast estimator · Strong consistency · Asymptotic normality

Electronic supplementary material The online version of this article (doi:10.1007/s10463-014-0453-1) contains supplementary material, which is available to authorized users.

Dan Ralescu was supported by a Taft Research Grant.

Y. Sun
Department of Mathematics & Statistics, Utah State University, 3900 Old Main Hill, Logan, UT 84322-3900, USA
e-mail: yan.sun@usu.edu

D. Ralescu (✉)
Department of Mathematical Sciences, University of Cincinnati, 4199 French Hall, 2815 Commons Way, Cincinnati, OH 45221-0025, USA
e-mail: dan.ralescu@uc.edu

1 Introduction

In classical statistics, it is often assumed that the outcome of an experiment is precise and the uncertainty of observations is solely due to randomness. Under this assumption, numerical data are represented as collections of real numbers. In recent years, however, there has been increased interest in situations when exact outcomes of the experiment are very difficult or impossible to obtain, or to measure. The imprecise nature of the data thus collected is caused by various factors such as measurement errors, computational errors, loss or lack of information. Under such circumstances and, in general, any other circumstances such as grouping and censoring, when observations cannot be pinned down to single numbers, data are better represented by intervals. Practical examples include interval-valued stock prices, oil prices, temperature data, medical records, and mechanical measurements among many others.

In the statistical literature, random intervals are most often studied in the framework of random sets, for which the probability-based theory has been developed since the publication of the seminal book (Matheron 1975). Studies on the corresponding statistical methods to analyze set-valued data, while still at the early stage, have shown promising advances. See Stoyan (1998) for a comprehensive review. Specifically, to analyze interval-valued data, the earliest attempt probably dates back to 1990, when Diamond (1990) published his paper on the least-squares fitting of compact set-valued data and considered interval-valued input and output as a special case. Due to the embedding theorems started by Brunn and Minkowski and later refined by Rådström (1952) and Hörmander (1954), $\mathcal{K}(\mathbb{R}^n)$, the space of all nonempty compact convex subsets of $\mathbb{R}^n$, is embedded into the Banach space of support functions. Diamond (1990) defined an L_2 metric in this Banach space of support functions and found the regression coefficients by minimizing the L_2 metric of the sum of residuals. This idea was further studied in Gil et al. (2002), where the L_2 metric was replaced by a generalized metric on the space of nonempty compact intervals, called "W-distance", proposed earlier by Körner and Näther (1998). Separately, Billard and Diday (2003) introduced the central tendency and dispersion measures and developed the symbolic interval data analysis based on those [see also Carvalho et al. (2004)]. However, none of the existing literature considered distributions of the random intervals and the corresponding statistical methods.

It is well known that normality plays an important role in classical statistics. But the normal distribution for random sets remained undefined for a long time, until the 1980s when the concept of normality was first introduced for compact convex random sets in the Euclidean space by Lyashenko (1983). This concept is especially useful in deriving limit theorems for random sets. See, Puri et al. (1986) and Norberg (1984), among others. Since a compact convex set in $\mathbb{R}$ is a closed bounded interval, by the definition of Lyashenko (1983), a normal random interval is simply a Gaussian displacement of a fixed closed bounded interval. From the point of view of statistics, this is not enough to fully capture the randomness of a general random interval.

In this paper, we extend the definition of normality given by Lyashenko (1983) and propose a Normal hierarchical model for random intervals. With one more degree of freedom on "shape", our model conveniently captures the entire randomness of random intervals via a few parameters. It is a natural extension from Lyashenko (1983) yet a highly practical model accommodating a large class of random intervals. In particular, when the length of the random interval reduces to zero, it becomes the usual normal random variable. Therefore, it can also be viewed as an extension of the classical normal distribution that accounts for the extra uncertainty added to the randomness. In addition, there are two interesting properties regarding our Normal hierarchical model: (1) conditioning on the first hierarchy, it is exactly the normal random interval defined by Lyashenko (1983), which could be a very useful property in view of the limit theorems; (2) with certain choices of the distributions, a linear combination of our Normal hierarchical random intervals follows the same Normal hierarchical distribution. An immediate consequence of the second property is the possibility of a factor model for multi-dimensional random intervals, as the "factor" will have the same distribution as the original intervals.

For random set models, it is important, in the stage of parameter estimation, to take into account the geometric characteristics of the observations. For example, Tanaka et al. (2008) proposed an approximate maximum likelihood estimation for parameters in the Neyman–Scott point processes based on the point pattern of the observation window. For another model, Heinrich (1993) discussed several distance functions (called "contrast functions") between the parametric and the empirical contact distribution functions that are used towards parameter estimation for Boolean models. Bearing this in mind, to estimate the parameters of our Normal hierarchical model, we propose a minimum contrast estimator (MCE) based on the hitting function (capacity functional) that characterizes the distribution of a random interval by the hit-and-miss events of test sets. See Matheron (1975). In particular, we construct a contrast function based on the integral of a discrepancy function between the empirical and the parametric distribution measures. Theoretically, we show that under certain conditions our MCE satisfies a strong consistency and asymptotic normality. The simulation study is consistent with our theorems. We apply our model to analyze a daily temperature range data and, in this context, we have derived interesting and promising results.

The use of an integral measure of probability discrepancy here is not new. For example, the integral probability metrics (IPMs), widely used as tools for statistical inferences, have been defined as the supremum of the absolute differences between expectations with respect to two probability measures. See, e.g., Zolotarev (1983), Müller (1997), and Sriperumbudur et al. (2012), for references. Especially, the empirical estimation of IPMs proposed by Sriperumbudur et al. (2012) drastically reduces the computational burden, thereby emphasizing the practical use of the IPMs. This idea is potentially applicable to our MCE and we expect similar reduction in computational intensity as for IPMs.

The rest of the paper is organized as follows. Section 2 formally defines our Normal hierarchical model and discusses its statistical properties. Section 3 introduces a minimum contrast estimator for the model parameters and presents its asymptotic properties. A simulation study is reported in Sect. 4, and a real data application is demonstrated in Sect. 5. We give concluding remarks in Sect. 6. Proofs of the theorems are presented in Sect. 7. Useful lemmas and other proofs are provided in the supplementary material.

2 The normal hierarchical model

2.1 Definition

Let $(\Omega, \mathcal{L}, P)$ be a probability space. Denote by $\mathcal{K}$ the collection of all non-empty compact subsets of $\mathbb{R}^d$. A random compact set is a Borel measurable function $A : \Omega \to \mathcal{K}$, $\mathcal{K}$ being equipped with the Borel σ-algebra induced by the Hausdorff metric. If $A(\omega)$ is convex for almost all ω, then A is called a random compact convex set. [See Molchanov (2005), p. 21, p. 102.] Denote by $\mathcal{K}_{\mathcal{C}}$ the collection of all compact convex subsets of $\mathbb{R}^d$. By Theorem 1 of Lyashenko (1983), a compact convex random set A in the Euclidean space $\mathbb{R}^d$ is Gaussian if and only if A can be represented as the Minkowski sum of a fixed compact convex set M and a d-dimensional normal random vector ϵ, i.e.,

$$A = M + \{\epsilon\}\,. \tag{1}$$

As pointed out in Lyashenko (1983), Gaussian random sets are especially useful in view of the limit theorems discussed earlier in Lyashenko (1979). That is, if the conditions in those theorems are satisfied and the limit exists, then it is Gaussian in the sense of (1). Puri et al. (1986) extended these results to separable Banach spaces.

In the following, we will restrict ourselves to compact convex random sets in $\mathbb{R}^1$, that is, bounded closed random intervals. They will be called random intervals for ease of presentation.

According to (1), a random interval A is Gaussian if and only if A is representable in the form

$$A = I + \{\epsilon\}\,, \tag{2}$$

where I is a fixed bounded closed interval and ϵ is a normal random variable. Obviously, such a random interval is simply a Gaussian displacement of a fixed interval, so it is not enough to fully capture the randomness of a general random interval. In order to model the randomness of both the location and the "shape" (length), we propose the following Normal hierarchical model for random intervals:

$$A = I + \{\epsilon\}\,, \tag{3}$$

$$I = \eta I_0, \tag{4}$$

where η is another random variable and $I_0 = [a_0, b_0]$ is a fixed interval in $\mathbb{R}$. Here, the product ηI_0 is in the sense of scalar multiplication of a real number and a set. Let $\lambda(\cdot)$ denote the Lebesgue measure of $\mathbb{R}^1$. Then,

$$\lambda(A) = \lambda(\epsilon + \eta I_0) = \lambda(\eta I_0) = |\eta|\,\lambda(I_0). \tag{5}$$

That is, η is the variable that models the length of A. In particular, if $\eta \to 0$, then A reduces to a normal random variable.

Obviously, ϵ and η are "location" and "shape" variables. We assume that $\eta > 0$. Then, the Normal hierarchical random interval is explicitly expressible as

$$A = [\epsilon + a_0\eta, \epsilon + b_0\eta]\,.$$

The parameter b_0 is indeed unnecessary, as the difference $b_0 - a_0$ can be absorbed by η. As a result,

$$A = [\epsilon + a_0\eta, \epsilon + (a_0 + 1)\,\eta]\,. \tag{6}$$

Compared to the "naive" model $A = [\epsilon - \frac{1}{2}\eta, \epsilon + \frac{1}{2}\eta]$, for which ϵ is precisely the center of the interval, (6) has an extra parameter a_0. Notice that the center of A is $\epsilon + \left(a_0 + \frac{1}{2}\right)\eta$, so a_0 controls the difference between ϵ and the center, and therefore is interpreted as modeling the uncertainty that the Normal random variable ϵ is not necessarily the center.

Remark 1 There are some existing works in the literature to model the randomness of intervals. For example, a random interval can be viewed as the "crisp" version of the LR-fuzzy random variable, which is often used to model the randomness of imprecise intervals such as [approximately 2, approximately 5]. See Körner (1997) for detailed descriptions. However, as far as the authors are aware, models with distribution assumptions for interval-valued data have not been studied yet. Our Normal hierarchical random interval is the first statistical approach that extends the concept of normality while modeling the full randomness of an interval.

An interesting property of the Normal hierarchical random interval is that its linear combination is still a Normal hierarchical random interval. This is seen by simply observing that

$$\sum_{i=1}^{n} a_i A_i = \sum_{i=1}^{n} a_i\,(\epsilon_i + \eta_i I_0) = \sum_{i=1}^{n} a_i\epsilon_i + I_0\left(\sum_{i=1}^{n} a_i\eta_i\right), \tag{7}$$

for arbitrary constants $a_i, i = 1, \ldots, n$, where "+" denotes the Minkowski addition. This is very useful in developing a factor model for the analysis of multiple random intervals. Especially, if we assume $\eta_i \sim N(\mu_i, \sigma_i^2), i = 1, \ldots, n$, then the "factor" $\sum_{i=1}^{n} a_i A_i$ has exactly the same distribution as the original random intervals. We will elaborate more on this issue in Sect. 4.

Without loss of generality, we can assume in the model (3–4) that $E\epsilon = 0$. We will make this assumption throughout the rest of the paper.

2.2 Model properties

According to the Choquet theorem (Molchanov 2005, p.10), the distribution of a random closed set (and random compact convex set as a special case) A is completely characterized by the hitting function T defined as

$$T(K) = P(K \cap A \neq \emptyset), \quad \forall K \in \mathcal{K}_{\mathcal{C}}. \tag{8}$$

Writing $I_0 = [a_0, b_0]$ with $a_0 \leq b_0$, the Normal hierarchical random interval in (3–4) has the following hitting function: for $K = [a, b]$:

$$\begin{aligned}
&T_A([a,b])\\
&= P([a,b]\cap A \neq \emptyset)\\
&= P([a,b]\cap A \neq \emptyset, \eta \geq 0) + P([a,b]\cap A \neq \emptyset, \eta < 0)\\
&= P(a-\eta b_0 \leq \epsilon \leq b-\eta a_0, \eta \geq 0) + P(a-\eta a_0 \leq \epsilon \leq b-\eta b_0, \eta < 0).
\end{aligned}$$

The expectation of a compact convex random set A is defined by the Aumann integral [see Aumann (1965), Artstein and Vitale (1975)] as

$$EA = \{E\xi : \xi \in A \text{ almost surely}\}.$$

In particular, the Aumann expectation of a random interval A is given by

$$EA = [EA_l, EA_u], \tag{9}$$

where A_l and A_u are the interval ends. Therefore, the Aumann expectation of the Normal hierarchical random interval A is

$$\begin{aligned}
EA &= E(\epsilon + \eta I_0) = E\epsilon + E(\eta I_0) = E(\eta I_0)\\
&= E\left\{[a_0\eta, b_0\eta]I_{(\eta\geq 0)} + [b_0\eta, a_0\eta]I_{(\eta<0)}\right\}\\
&= E\left[a_0\eta I_{(\eta\geq 0)} + b_0\eta I_{(\eta<0)}, b_0\eta I_{(\eta\geq 0)} + a_0\eta I_{(\eta<0)}\right]\\
&= \left[a_0E\eta_+ + b_0E\eta_-, b_0E\eta_+ + a_0E\eta_-\right],
\end{aligned}$$

where

$$\begin{aligned}
\eta_+ &= \eta I_{(\eta\geq 0)},\\
\eta_- &= \eta I_{(\eta<0)}.
\end{aligned}$$

Notice that η_+ can be interpreted as the positive part of η, but η_- is not the negative part of η, as $\eta_- < 0$ when $\eta < 0$.

The variance of a compact convex random set A in $\mathbb{R}^d$ is defined via its support function. In the special case when $d = 1$, it is shown by straightforward calculations that

$$Var(A) = \frac{1}{2}Var(A_l) + \frac{1}{2}Var(A_u), \tag{10}$$

or equivalently,

$$Var(A) = Var(A_c) + Var(A_r), \tag{11}$$

where A_c and A_r denote the center and radius of a random interval A. See Körner (1995). Again, as we pointed out in Remark 1, a random interval can be viewed as a special case of the LR-fuzzy random variable. Therefore, formulae (10) and (11) coincide with the variance of the LR-fuzzy random variable, when letting the left and right spread both equal to 0, i.e., $l = r = 0$. See Körner (1997). For the Normal hierarchical random interval A,

$$\begin{aligned}
&Var(A_l)\\
&= Var\,(\epsilon + a_0\eta_+ + b_0\eta_-)\\
&= E\,(\epsilon + a_0\eta_+ + b_0\eta_-)^2 - \left[E\,(\epsilon + a_0\eta_+ + b_0\eta_-)\right]^2\\
&= E\epsilon^2 + a_0^2 Var(\eta_+) + b_0^2 Var(\eta_-)\\
&\quad +2\,(a_0 E\epsilon\eta_+ + b_0 E\epsilon\eta_- - a_0 b_0 E\eta_+ E\eta_-)\,,
\end{aligned}$$

and, analogously,

$$\begin{aligned}
&Var(A_u)\\
&= E\epsilon^2 + b_0^2 Var(\eta_+) + a_0^2 Var(\eta_-)\\
&\quad +2\,(b_0 E\epsilon\eta_+ + a_0 E\epsilon\eta_- - a_0 b_0 E\eta_+ E\eta_-)\,.
\end{aligned}$$

The variance of A is then found to be

$$\begin{aligned}
Var(A) &= \frac{1}{2}Var(A_l) + \frac{1}{2}Var(A_u)\\
&= E\epsilon^2 + \frac{1}{2}\left(a_0^2 + b_0^2\right)\left[Var(\eta_+) + Var(\eta_-)\right]\\
&\quad +(a_0 + b_0)E\epsilon\eta - 2a_0 b_0 E\eta_+\eta_-.
\end{aligned}$$

Remark 2 Assuming $\eta > 0$, we have

$$\begin{aligned}
Var(A) &= E\epsilon^2 + \frac{1}{2}(a_0^2 + b_0^2)Var(\eta) + (a_0 + b_0)E\epsilon\eta\\
&= Var(\epsilon) + \frac{1}{2}(a_0^2 + b_0^2)Var(\eta) + (a_0 + b_0)Cov(\epsilon, \eta),
\end{aligned}$$

with $E\epsilon = 0$. This formula certainly includes the special case of the "naive" model $A = [\epsilon - \frac{1}{2}\eta, \epsilon + \frac{1}{2}\eta]$, by letting $a_0 = -\frac{1}{2}$ and $b_0 = \frac{1}{2}$. It is more general because it also accounts for the covariance between "location" and "length" in calculating the total variance of the random interval, while the "naive" model simply has $Var\,(A) = Var\,(\epsilon) + Var\,(\eta)$.

3 The minimum contrast estimation

3.1 Definitions

We study MCE of the parameters of the Normal hierarchical random interval (3–4) as well as its asymptotic properties. Since $d = 1$, from now on we let $\mathcal{K}$ be the space of all non-empty compact subsets in $\mathbb{R}$ restrictively, and let $\mathcal{F}$ be the Borel σ-algebra on $\mathcal{K}$ induced by the Hausdorff metric. Let $\mathcal{K}_\mathcal{C}$ denote the space of all non-empty compact convex subsets, i.e., bounded closed intervals, in $\mathbb{R}$. As mentioned in the previous section, a random interval X is a Borel measurable function from a probability space $(\Omega, \mathcal{L}, P)$ to $(\mathcal{K}, \mathcal{F})$ such that $X \in \mathcal{K}_\mathcal{C}$ almost surely.

Throughout this section, we assume observing a sample of i.i.d. random intervals $X(n) = \{X_1, X_2, \ldots, X_n\}$. Let $\boldsymbol{\theta}$ denote a $p \times 1$ vector containing all the parameters in the model, which takes on a value from a parameter space $\Theta \subset \mathbb{R}^p$. Here, p is the number of parameters. Let $\boldsymbol{\theta}_0$ denote the true value of the parameter vector. Denote by $T_{\boldsymbol{\theta}}([a, b])$ the hitting function of X_i with parameter $\boldsymbol{\theta}$.

In order to introduce the MCE, we will need some extra notations. Let $\mathbf{X}$ be a basic set and $\mathcal{A}$ be a σ-field over it. Let $\mathcal{B}$ denote a family of probability measures on $(\mathbf{X},\mathcal{A})$ and τ be a mapping from $\mathcal{B}$ to some topological space T. $\tau(P)$ denotes the parameter value pertaining to P, $\forall P \in \mathcal{B}$. The classical definition of MCE given in Pfanzagl (1969) is quoted below.

Definition 1 (Pfanzagl 1969) A family of $\mathcal{A}$-measurable functions $f_t : \mathbf{X} \to \mathbb{R}$, $t \in T$ is a family of contrast functions if

$$E_P[f_t] < \infty, \quad \forall t \in T, \forall P \in \mathcal{B} \tag{12}$$

and

$$E_P\left[f_{\tau(P)}\right] < E_P[f_t], \quad \forall t \in T, \forall P \in \mathcal{B}, t \neq \tau(P). \tag{13}$$

In other words, a contrast function is a measurable function of the random variable(s) whose expected value reaches its minimum under the probability measure that generates the random variable(s). From the view of probability, with the true parameters, a contrast function tends to have a smaller value than with other parameters.

Adopting notation from Pfanzagl (1969), we let $\mathcal{B}$ denote a family of probability measures on $(\mathcal{K}_{\mathcal{C}}, \mathcal{F})$ and τ be a mapping from $\mathcal{B}$ to some topological space T. Similarly, $\tau(P)$ denotes the parameter value pertaining to P, $\forall P \in \mathcal{B}$. In a similar fashion to the contrast function in Heinrich (1993) for Boolean models, we give our definition of contrast function for random intervals in the following. And then the MCE is defined as the minimizer of the contrast function.

Definition 2 A family of $\mathcal{F}^n$-measurable functions $M(X(n); \boldsymbol{\theta})$: $\mathcal{K}_{\mathcal{C}}^n \to [-\infty, +\infty]$, $n \in \mathbb{N}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is a family of contrast functions for $\mathcal{B}$, if there exists a function $N(\cdot, \cdot)$: $\boldsymbol{\Theta} \times \boldsymbol{\Theta} \to \mathbb{R}$ such that

$$P_{\boldsymbol{\theta}}\left(\left\{\omega : \lim_{n\to\infty} M(X(n); \boldsymbol{\zeta}) = N(\boldsymbol{\theta}, \boldsymbol{\zeta})\right\}\right) = 1, \quad \forall\, \boldsymbol{\theta}, \boldsymbol{\zeta} \in \boldsymbol{\Theta}, \tag{14}$$

and

$$N(\boldsymbol{\theta}, \boldsymbol{\theta}) < N(\boldsymbol{\theta}, \boldsymbol{\zeta})\ \forall\, \boldsymbol{\theta}, \quad \boldsymbol{\zeta} \in \boldsymbol{\Theta}, \ \boldsymbol{\theta} \neq \boldsymbol{\zeta}. \tag{15}$$

Definition 3 A $\mathcal{F}^n$-measurable function $\hat{\boldsymbol{\theta}}_n$: $\mathcal{K}_{\mathcal{C}}^n \to \tau(\mathcal{B})$, which depends on $X(n)$ only, is called a minimum contrast estimator (MCE) if

$$M(X(n); \hat{\boldsymbol{\theta}}_n) = \inf\left\{M(X(n); \boldsymbol{\theta}) : \boldsymbol{\theta} \in \tau(\mathcal{B})\right\}. \tag{16}$$

3.2 Theoretical results

We make the following assumptions to present the theoretical results in this section.

Assumption 1 Θ is compact, and $\boldsymbol{\theta}_0$ is an interior point of Θ.

Assumption 2 The model is identifiable.

Assumption 3 $T_{\boldsymbol{\theta}}([\cdot, \cdot])$ is continuous with respect to $\boldsymbol{\theta}$.

Assumption 4 $\frac{\partial T_{\boldsymbol{\theta}_0}}{\partial \theta_i}([\cdot, \cdot]), i = 1, \ldots, p$, exist and are finite on a bounded region $S^0 \subset \mathbb{R}^2$.

Assumption 5 $\frac{\partial T_{\boldsymbol{\theta}}}{\partial \theta_j}([\cdot, \cdot])$, $\frac{\partial^2 T_{\boldsymbol{\theta}}}{\partial \theta_j \partial \theta_k}([\cdot, \cdot])$, and $\frac{\partial^3 T_{\boldsymbol{\theta}}}{\partial \theta_j \partial \theta_k \partial \theta_l}([\cdot, \cdot])$, $i, j, k = 1, \ldots, p$, exist and are finite on S^0 for $\boldsymbol{\theta} \in \Theta$.

Assumptions 4 and 5 are essential to establish the asymptotic normality for the MCE $\hat{\boldsymbol{\theta}}_n$. They are rather mild and can be met by a large class of capacity functionals. For example, if S^0 is closed, then each $T_{\boldsymbol{\theta}_0}$ with continuous up to third-order partial derivatives satisfies both assumptions, as a continuous function on a compact region is always bounded. The following theorem gives sufficient conditions under which the minimum contrast estimator $\hat{\boldsymbol{\theta}}_n$ defined above is strongly consistent.

Theorem 1 *Let $M(X(n); \boldsymbol{\theta})$ be a contrast function as in Definition 2 and let $\hat{\boldsymbol{\theta}}_n$ be the corresponding MCE. Under the hypothesis of Assumption 1 and in addition if $M(X(n); \boldsymbol{\theta})$ is equicontinuous w.r.t. $\boldsymbol{\theta}$ for all $X(n)$, $n = 1, 2, \ldots$, then,*

$$\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0 \;\; a.s., \;\; as\; n \to \infty.$$

Let $[a, b] \in \mathcal{K}_{\mathcal{C}}$. Define an empirical estimator $\hat{T}([a, b]; X(n))$ for $T([a, b])$ as

$$\hat{T}([a, b]; X(n)) = \frac{\#\{X_i : [a, b] \cap X_i \neq \emptyset, i = 1, \ldots, n\}}{n}. \tag{17}$$

Extending the contrast function defined in Heinrich (1993) (for parameters in the Boolean model), we construct a family of functions:

$$H(X(n); \boldsymbol{\theta}) = \iint_S \left[T_{\boldsymbol{\theta}}([a, b]) - \hat{T}([a, b]; X(n))\right]^2 W(a, b) \mathrm{d}a \mathrm{d}b, \tag{18}$$

for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $S \subset S^0 \subset \mathbb{R}^2$, and $W(a, b)$ is a weight function on $[a, b]$ satisfying $0 < W(a, b) < C$, $\forall [a, b] \in \mathcal{K}_{\mathcal{C}}$.

We show in the next Proposition that $H(X(n); \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}$ defined in (18) is a family of contrast functions for $\boldsymbol{\theta}$. This, together with Theorem 1, immediately yields the strong consistency of the associated MCE. This result is summarized in Corollary 1.

Proposition 1 *Suppose that Assumptions 2 and 3 are satisfied. Then $H(X(n); \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, as defined in (18), is a family of contrast functions with limiting function*

$$N(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \iint_S \left[T_{\boldsymbol{\theta}}([a, b]) - T_{\boldsymbol{\zeta}}([a, b])\right]^2 W(a, b) \mathrm{d}a \mathrm{d}b. \tag{19}$$

In addition, $H(X(n); \boldsymbol{\theta})$ is equicontinuous w.r.t. $\boldsymbol{\theta}$.

Corollary 1 *Suppose that Assumptions 1, 2, and 3 are satisfied. Let $H(X(n);\boldsymbol{\theta})$ be defined as in (18), and*

$$\boldsymbol{\theta}_n^H = \arg\min_{\boldsymbol{\theta}\in\Theta} H\left(X(n);\boldsymbol{\theta}\right). \tag{20}$$

Then

$$\boldsymbol{\theta}_n^H \to \boldsymbol{\theta}_0,\ a.s.,$$

as $n \to \infty$.

Next, we show the asymptotic normality for $\boldsymbol{\theta}_n^H$. As a preparation, we first prove the following proposition. The central limit theorem for $\boldsymbol{\theta}_n^H$ is then presented afterwards.

Proposition 2 *Assume the conditions of Lemma 1 (in the supplementary material). Define*

$$\frac{\partial H}{\partial \boldsymbol{\theta}}\left(X(n);\boldsymbol{\theta}\right) := \left[\frac{\partial H}{\partial \theta_1}\left(X(n);\boldsymbol{\theta}\right),\ldots,\frac{\partial H}{\partial \theta_p}\left(X(n);\boldsymbol{\theta}\right)\right]^T,$$

as the $p\times 1$ gradient vector of $H\left(X(n);\boldsymbol{\theta}\right)$ w.r.t. $\boldsymbol{\theta}$. Then,

$$\sqrt{n}\left[\frac{\partial H}{\partial \boldsymbol{\theta}}\left(X(n);\boldsymbol{\theta}_0\right)\right] \xrightarrow{\mathcal{D}} N\left(0,\Xi\right),$$

where Ξ is the $p\times p$ symmetric matrix with the (i,j)th component

$$\begin{aligned}\Xi(i,j) = 4\iiiint_{S\times S} &\left\{P\left(X_1\cap[a,b]\neq\emptyset, X_1\cap[c,d]\neq\emptyset\right)\right.\\ &\left.-T_{\boldsymbol{\theta}_0}\left([a,b]\right)T_{\boldsymbol{\theta}_0}\left([c,d]\right)\right\}\\ &\frac{\partial T_{\boldsymbol{\theta}_0}}{\partial \theta_i}\left([a,b]\right)\frac{\partial T_{\boldsymbol{\theta}_0}}{\partial \theta_j}\left([c,d]\right)W(a,b)W(c,d)\mathrm{d}a\mathrm{d}b\mathrm{d}c\mathrm{d}d.\end{aligned} \tag{21}$$

Theorem 2 *Let $H(X(n);\boldsymbol{\theta})$ be defined in (18) and $\boldsymbol{\theta}_n^H$ be defined in (20). Assume the conditions of Corollary 1. If additionally Assumption 5 is satisfied, then*

$$\sqrt{n}\left(\boldsymbol{\theta}_n^H-\boldsymbol{\theta}_0\right) \xrightarrow{\mathcal{D}} N\left(0, C(T_{\boldsymbol{\theta}_0})^{-1}\,\Xi C(T_{\boldsymbol{\theta}_0})^{-1}\right), \tag{22}$$

where $C(T_{\boldsymbol{\theta}_0}) = 2\iint_S \left(\frac{\partial T_{\boldsymbol{\theta}_0}}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial T_{\boldsymbol{\theta}_0}}{\partial\boldsymbol{\theta}}\right)^T([a,b])W(a,b)\mathrm{d}a\mathrm{d}b$, and Ξ is defined in (21).

4 Simulation

We carry out a small simulation to investigate the performance of the MCE introduced in Definition 3. Assume, in the Normal hierarchical model (3–4), that

$$\begin{bmatrix}\epsilon\\ \eta\end{bmatrix} \sim \mathrm{BVN}\left(\begin{bmatrix}0\\ \mu\end{bmatrix}, \Sigma = \begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}\right), \tag{23}$$

and

$$b_0 = a_0 + 1. \tag{24}$$

The bivariate normal distribution conveniently takes care of the variances and covariance of the location variable ϵ and the shape variable η. The removal of the freedom of b_0 is for model identifiability purposes; it is seen that the hitting function T_A is defined via ηa_0 and ηb_0 only. For the simulation, we assign the following parameter values:

$$a_0 = 1, \quad \mu = 20, \quad \Sigma = \begin{bmatrix} 10 & 1 \\ 1 & 10 \end{bmatrix}. \tag{25}$$

4.1 Hitting function

Under the bivariate normal distribution assumption, the hitting function of our Normal hierarchical model is found to be

$$\begin{aligned}
&T_{\boldsymbol{\theta}}([a, b]) \\
&= P(a - \eta b_0 \le \epsilon \le b - \eta a_0, \eta \ge 0) + P(a - \eta a_0 \le \epsilon \le b - \eta b_0, \eta < 0) \\
&= P(\epsilon \le b - \eta a_0, \eta \ge 0) - P(\epsilon < a - \eta b_0, \eta \ge 0) \\
&\quad + P(\epsilon \le b - \eta b_0, \eta < 0) - P(\epsilon < a - \eta a_0, \eta < 0) \\
&= P\left(\begin{bmatrix} 1 & a_0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} \epsilon \\ \eta \end{bmatrix} \le \begin{bmatrix} b \\ 0 \end{bmatrix}\right) - P\left(\begin{bmatrix} 1 & b_0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} \epsilon \\ \eta \end{bmatrix} \le \begin{bmatrix} a \\ 0 \end{bmatrix}\right) \\
&\quad + P\left(\begin{bmatrix} 1 & b_0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \epsilon \\ \eta \end{bmatrix} \le \begin{bmatrix} b \\ 0 \end{bmatrix}\right) - P\left(\begin{bmatrix} 1 & a_0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \epsilon \\ \eta \end{bmatrix} \le \begin{bmatrix} a \\ 0 \end{bmatrix}\right) \\
&= \Phi\left(\begin{bmatrix} b \\ 0 \end{bmatrix}; D_1 \begin{bmatrix} 0 \\ \mu \end{bmatrix}, D_1 \Sigma D_1^{'}\right) - \Phi\left(\begin{bmatrix} a \\ 0 \end{bmatrix}; D_2 \begin{bmatrix} 0 \\ \mu \end{bmatrix}, D_2 \Sigma D_2^{'}\right) \\
&\quad + \Phi\left(\begin{bmatrix} b \\ 0 \end{bmatrix}; D_3 \begin{bmatrix} 0 \\ \mu \end{bmatrix}, D_3 \Sigma D_3^{'}\right) - \Phi\left(\begin{bmatrix} a \\ 0 \end{bmatrix}; D_4 \begin{bmatrix} 0 \\ \mu \end{bmatrix}, D_4 \Sigma D_4^{'}\right),
\end{aligned} \tag{26}$$

where $\Phi(\mathbf{x}; \boldsymbol{\mu}, \Omega)$ is the bivariate normal cdf with mean $\boldsymbol{\mu}$ and covariance Ω, and

$$D_1 = \begin{bmatrix} 1 & a_0 \\ 0 & -1 \end{bmatrix}, \quad D_2 = \begin{bmatrix} 1 & b_0 \\ 0 & -1 \end{bmatrix}, \quad D_3 = \begin{bmatrix} 1 & b_0 \\ 0 & 1 \end{bmatrix}, \quad D_4 = \begin{bmatrix} 1 & a_0 \\ 0 & 1 \end{bmatrix}.$$

After linear transformation of variables, the terms in formula (26) is calculated via the standard bivariate normal cdf. By absolute continuity, $T_{\boldsymbol{\theta}}([a, b])$ in this case is continuous and also infinitely continuously differentiable. Therefore, all the assumptions are satisfied and the corresponding MCE achieves the strong consistency and asymptotic normality.

According to the assigned parameter values given in (25), $P(\eta < 0) < 10^{-10}$. Therefore, the hitting function is well approximated by

$$
\begin{aligned}
& T_{\boldsymbol{\theta}}([a, b]) \\
& \approx P(a - \eta b_0 \leq \epsilon \leq b - \eta a_0, \eta \geq 0) \\
& \approx P(a - \eta b_0 \leq \epsilon \leq b - \eta a_0) \\
& = P\left(\begin{bmatrix} 1 & a_0 \\ -1 & -a_0 - 1 \end{bmatrix} \begin{bmatrix} \epsilon \\ \eta \end{bmatrix} \leq \begin{bmatrix} b \\ -a \end{bmatrix}\right) \\
& = \Phi\left(\begin{bmatrix} b \\ -a \end{bmatrix}; D \begin{bmatrix} 0 \\ \mu \end{bmatrix}, D \Sigma D^{'}\right),
\end{aligned}
$$

where

$$
D = \begin{bmatrix} 1 & a_0 \\ -1 & -a_0 - 1 \end{bmatrix}.
$$

We use this approximate hitting function to simplify the computation in our simulation study.

4.2 Parameter initialization

The model parameters can be estimated by the method of moments. In most cases it is reasonable to assume $\eta^{-} \approx 0$, and consequently, $\eta \approx |\eta|$. So the moment estimates for μ and a_0 are approximately given by

$$
\tilde{\mu} \leftarrow \bar{X_u} - \bar{X_l}, \tag{27}
$$

$$
\tilde{a_0} \leftarrow \bar{X_l} / \tilde{\mu}, \tag{28}
$$

where $\bar{X_u}$ and $\bar{X_l}$ denote the sample means of A_u and A_l, respectively. Denoting by A_c the center of the random interval A, we further notice that $A_c = \epsilon + \frac{1}{2}(a_0 + b_0)\eta = \epsilon + \left(a_0 + \frac{1}{2}\right)\eta$. By the same approximation, we have $\epsilon \approx A_c - \left(a_0 + \frac{1}{2}\right)|\eta|$. Define a random variable

$$
A_\delta = A_c - \left(a_0 + \frac{1}{2}\right)|\eta|.
$$

Then, the moment estimate for Σ is approximately given by the sample variance–covariance matrix of A_δ and $A_u - A_l$, i.e.,

$$
\tilde{\Sigma} \leftarrow \Sigma_s\left(A_\delta, A_u - A_l\right). \tag{29}
$$

4.3 Performance of MCE

Our simulation experiment is designed as follows: we first simulate an i.i.d. random sample of size n from model (3–4) with the assigned parameter values, then find the initial parameter values by (27–29) based on the simulated sample, and lastly the initial values are updated to the MCE using the function *fminsearch.m* in Matlab 2011a. The process is repeated 10 times independently for each n, and we let $n = 100, 200, 300, 400, 500$, successively, to study the consistency and efficiency of the MCE's.

Figure 1 shows one random sample of 100 observations generated from the model. We show the average biases and standard errors of the estimates as functions of the sample size in Fig. 2. Here, the average bias and standard error of the estimates of Σ are the L_2 norms of the average bias and standard error matrices, respectively. As expected from Corollary 1 and Theorem 2, both the bias and the standard error reduce to 0 as sample size grows to infinity. The numerical results are summarized in Table 1.

Finally, we point out that the choice of the region of integration S is important. A larger S usually leads to more accurate estimates and could also result in more computational complexity. We do not investigate this issue in this paper. However, based on our simulation experience, an S that covers most of the points $(a, b) \in \mathbb{R}^2$, such that $[a, b]$ hits some of the observed intervals, is a good choice as a rule of thumb. In our simulation, $E(A) \approx [20, 40]$, by ignoring the small probability $P(\eta < 0)$. Therefore, we choose $S = \{(x - y, x + y) : 20 \leq x \leq 40, 0 \leq y \leq 10\}$, and the estimates are satisfactory.

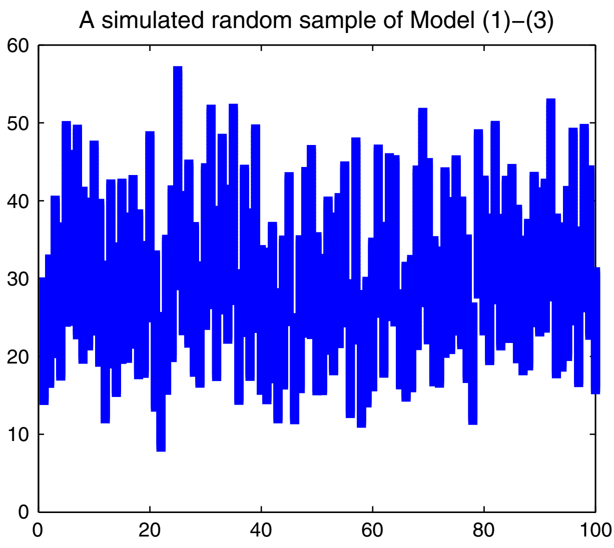

Fig. 1 Plot of a simulated sample from model (3–4) with $n = 100$

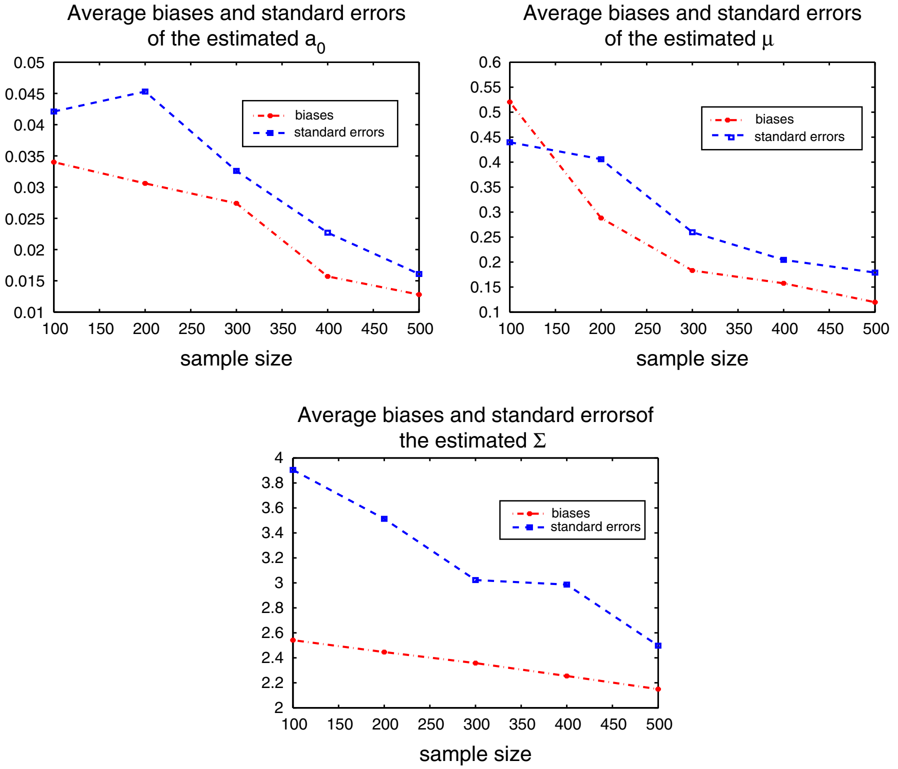

Fig. 2 Average bias and standard error of the MCE's for a_0 (*top left*), μ (*top right*), and Σ (*bottom*), as a function of the sample size n

Table 1 Average biases and standard errors of the MCE's of the model parameters in the simulation study

n	a_0=1		μ=20		Σ	
	bias	ste	bias	ste	bias	ste
100	0.0683	0.1289	1.1648	1.7784	4.1166	5.7951
200	0.0387	0.0457	0.4569	0.5924	3.8581	4.0558
300	0.0274	0.0326	0.1831	0.2598	3.0317	3.9042
400	0.0157	0.0227	0.1575	0.2044	2.8210	3.5128
500	0.0128	0.0161	0.1197	0.1790	2.1494	2.4973

5 A real data application

In this section, we apply our Normal hierarchical model and minimum contrast estimator to analyze the daily temperature range data. We consider two data sets containing 10 years of daily minimum and maximum temperatures in January, in Granite Falls, Minnesota (latitude 44.81241, longitude 95.51389) from 1901 to 1910, and from 2001 to 2010, respectively. Each data set, therefore, is constituted of 310 observations of the form: (minimum temperature, maximum temperature). We obtained these data from

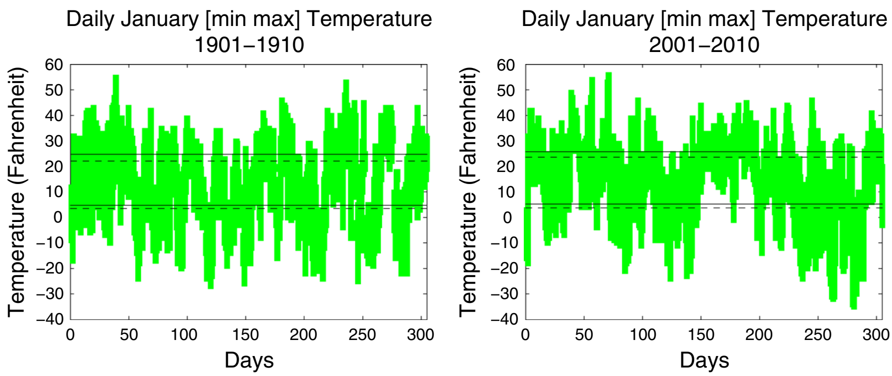

Fig. 3 Plots of daily January temperature range 1901–1910 (*left*) and 2001–2010 (*right*). On each plot, the model fitted mean is the interval between the *two horizontal lines*, and the moment estimate of mean is the interval between the *two dashed horizontal lines*

the National Weather Service, and all observations are in Fahrenheit. The plot of the data is shown in Fig. 3. The obvious correlations of the data play no roles here.

Same as in the simulation, we assume a bivariate normal distribution for (ϵ, η) and $I_0 = [a_0, a_0 + 1]$ has length 1. The initial parameter values are computed according to (27–29), and the weight function $W \equiv 1$. The minimum contrast estimates for the model parameters are

- Data set 1 (1901–1910):

$$\hat{a}_{0,1} = 0.2495, \quad \hat{\mu}_1 = 19.8573, \quad \hat{\Sigma}_1 = \begin{bmatrix} 207.1454 & -44.8547 \\ -44.8547 & 102.5263 \end{bmatrix},$$

- Data set 2 (2001–2010):

$$\hat{a}_{0,2} = 0.2614, \quad \hat{\mu}_2 = 20.4722, \quad \hat{\Sigma}_2 = \begin{bmatrix} 318.9283 & -84.0892 \\ -84.0892 & 68.4783 \end{bmatrix}.$$

Recall that the center and the length of the Normal hierarchical random interval are $\epsilon + (a_0 + \frac{1}{2})\eta$ and $|\eta|(\approx \eta$ for the two considered data sets), respectively. Therefore, they are assumed to follow Normal distributions with means $(a_0 + \frac{1}{2})\mu$ and μ, and variances $\sigma_1^2 + \left(a_0 + \frac{1}{2}\right)^2 \sigma_2^2 + (2a_0 + 1)\,\sigma_{12}^2$ and σ_2^2, respectively. To assess the goodness-of-fit, we compare the fitted Normal distributions with the corresponding empirical distributions for both the center and the length of the two data sets. The results are shown in Fig. 4. For the interval length of data 2 (2001–2010), the fitted Normal distribution is slightly more deviated from the empirical distribution, due to the skewness and heavy tail of the data. All the other three plots show very good fittings of our model to the data.

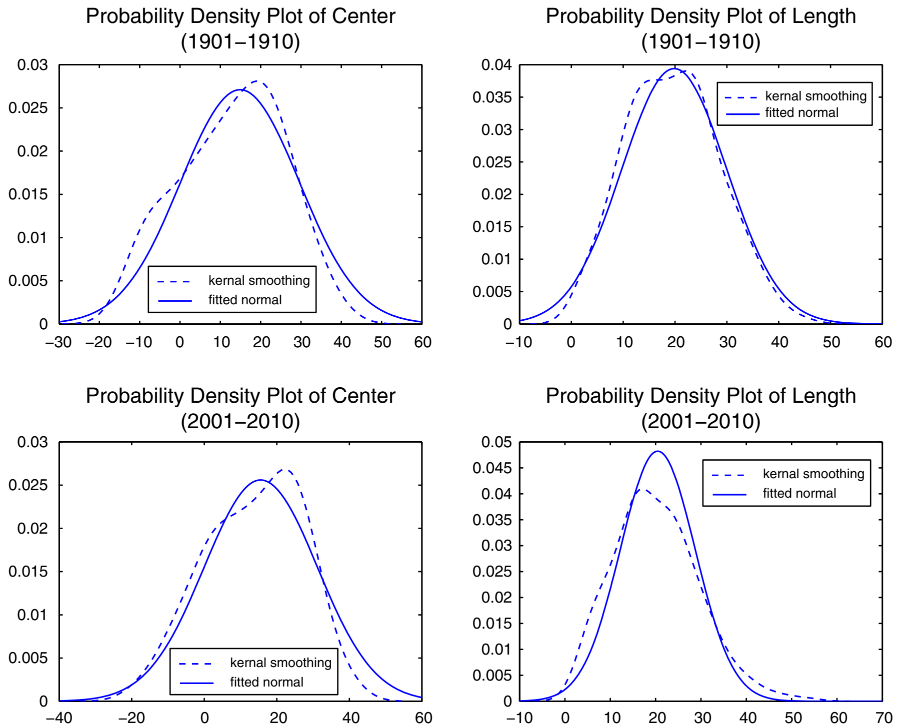

Fig. 4 Plots of the kernel smoothing density and the fitted Normal probability density for the centers and the lengths of the two data sets

Denote by A_1 and A_2 respectively the random intervals from which the two data sets are drawn. The model fitted mean and variance for A_1 and A_2 are found to be

$$\hat{\mathrm{E}}(A_1) = [4.8590, 24.9071], \quad \widehat{\mathrm{Var}}(A_1) = 221.2313;$$
$$\hat{\mathrm{E}}(A_2) = [5.3335, 25.8416], \quad \widehat{\mathrm{Var}}(A_2) = 247.3275.$$

Both mean and variance of the recent data are larger than those of the data 100 years ago. The two model fitted means are also shown on the data plots in blue as the intervals between the solid horizontal lines in Fig. 3. In addition, the correlation coefficient of (ϵ, η) is -0.3078 for data set 1 and -0.5690 for data set 2, suggesting a negative correlation between the location and the length for the January temperature range data in general. That is, colder days tend to have larger temperature ranges, and this relationship is stronger in the more recent data.

Finally, we point out that some of the parameters can be easily estimated by simple traditional methods. For example, by averaging the two interval ends respectively, we get the moment estimates for the two means:

$$\hat{\mathrm{E}}_M(A_1) = [3.5323, 22.1968],$$
$$\hat{\mathrm{E}}_M(A_2) = [3.8323, 23.6903].$$

They are shown in Fig. 3 as the intervals between the dashed horizontal lines, in comparison with our model fitted means. Further, the sample correlations between the interval centers and lengths are computed as -0.1502 and -0.3148 for data sets 1 and 2, respectively. These estimates can be viewed as a preliminary analysis. Our model and the MCE of the parameters refine it and provide a more systematic understanding of the data, by examining their geometric structure in the framework of random sets.

6 Conclusion

In this paper, we introduced a new model of random sets (specifically for random intervals). In many practical situations, data are not completely known, or are only known with some margins of error, and it is a very important issue to consider a model which extends normality for ordinary (numerical) data. Our hierarchical normal model extends normality for point-valued random variables, and is quite flexible in the sense that it is well suited for both theoretical investigations and for simulations and real data analysis. To these goals, we have defined a minimum contrast estimator for the model parameters and proved its consistency and asymptotic normality. We carry out simulation experiments, and finally, apply our model to a real data set (daily temperature range data obtained from the National Weather Service). Our approach is suitable for extensions to models in higher dimensions, e.g., a factor model for multiple random intervals, or more general random sets, including possible extensions to spherical random sets.

7 Proofs

7.1 Proof of Theorem 1

Assume by contradiction that $\hat{\boldsymbol{\theta}}_n$ does not converge to $\boldsymbol{\theta}_0$ almost surely. Then, there exists an $\epsilon > 0$ such that

$$P\left(\left\{\omega : \limsup_{n\to\infty} \left\|\hat{\boldsymbol{\theta}}_n(\omega) - \boldsymbol{\theta}_0\right\| \geq \epsilon\right\}\right) > 0.$$

Let $F := \left\{\omega : \limsup_{n\to\infty} \left\|\hat{\boldsymbol{\theta}}_n(\omega) - \boldsymbol{\theta}_0\right\| \geq \epsilon\right\}$ and $\Lambda := \Theta \cap \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \geq \epsilon\}$. By the compactness of Λ, for every $\omega \in F$, there exists a convergent subsequence $\left\{\hat{\boldsymbol{\theta}}_{n_i}(\omega)\right\}$ of $\left\{\hat{\boldsymbol{\theta}}_n(\omega)\right\}$ such that

$$\hat{\boldsymbol{\theta}}_{n_i}(\omega) \to \tilde{\boldsymbol{\theta}}(\omega) \in \Lambda,$$

as $i \to \infty$. Since $\boldsymbol{\theta}_0$ is the true underlying parameter vector that generates $X(n)$, from Definition 2, $M(X(n); \boldsymbol{\theta}_0)$ converges to $N(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$ almost surely, and any subsequence converges too. So we have

$$\lim_{i\to\infty} M(X(n_i); \boldsymbol{\theta}_0) = N(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0).$$

On the other hand, almost surely,

$$
\begin{aligned}
&\lim_{i\to\infty} M(X(n_i);\boldsymbol{\theta}_0) &(30)\\
&= \liminf_{i\to\infty} M(X(n_i);\boldsymbol{\theta}_0)\\
&\geq \liminf_{i\to\infty} M(X(n_i);\hat{\boldsymbol{\theta}}_{n_i})\\
&= \liminf_{i\to\infty} \left\{M(X(n_i);\hat{\boldsymbol{\theta}}_{n_i}) - M(X(n_i);\tilde{\boldsymbol{\theta}}) + M(X(n_i);\tilde{\boldsymbol{\theta}})\right\}\\
&\geq \liminf_{i\to\infty} \left\{M(X(n_i);\hat{\boldsymbol{\theta}}_{n_i}) - M(X(n_i);\tilde{\boldsymbol{\theta}})\right\} + \liminf_{i\to\infty} \left\{M(X(n_i);\tilde{\boldsymbol{\theta}})\right\}\\
&= \liminf_{i\to\infty} \left\{M(X(n_i);\tilde{\boldsymbol{\theta}})\right\}\\
&= \lim_{i\to\infty} \left\{M(X(n_i);\tilde{\boldsymbol{\theta}})\right\}\\
&= N(\tilde{\boldsymbol{\theta}};\boldsymbol{\theta}_0). &(31)
\end{aligned}
$$

Equation (31) follows from the equicontinuity of $M(X(n);\boldsymbol{\theta})$.

Therefore,

$$
P\left(\left\{\omega : N(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0) \geq N(\tilde{\boldsymbol{\theta}}(\omega),\boldsymbol{\theta}_0)\right\}\right) > 0, \tag{32}
$$

where $\tilde{\boldsymbol{\theta}}(\omega) \in \Lambda$ and consequently $\tilde{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_0$. But from the assumptions, $N(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0) < N(\tilde{\boldsymbol{\theta}}(\omega),\boldsymbol{\theta}_0)$, $\forall\omega$. This contradicts (32). Hence, the desired result follows.

7.2 Proof of Theorem 2

From Taylor's Theorem, we have

$$
\begin{aligned}
0 = \frac{\partial H}{\partial\theta_i}\left(X(n);\boldsymbol{\theta}_n^H\right) &= \frac{\partial H}{\partial\theta_i}(X(n);\boldsymbol{\theta}_0) + \sum_{j=1}^{p}\left(\theta_{n,j}^H - \theta_{0,j}\right)\frac{\partial^2 H}{\partial\theta_j\partial\theta_i}(X(n);\boldsymbol{\theta}_0)\\
&\quad + \frac{1}{2}\left[\sum_{j=1}^{p}\left(\theta_{n,j}^H - \theta_{0,j}\right)\frac{\partial}{\partial\theta_j}\right]^2 \frac{\partial H}{\partial\theta_i}(X(n);\boldsymbol{\epsilon}_n)\\
&= \frac{\partial H}{\partial\theta_i}(X(n);\boldsymbol{\theta}_0) + \sum_{j=1}^{p}\left(\theta_{n,j}^H - \theta_{0,j}\right)\\
&\quad \times\left[\frac{\partial^2 H}{\partial\theta_j\partial\theta_i}(X(n);\boldsymbol{\theta}_0) + \frac{1}{2}\sum_{l=1}^{p}\left(\theta_{n,l}^H - \theta_{0,l}\right)\frac{\partial^3 H}{\partial\theta_l\partial\theta_j\partial\theta_i}(X(n);\boldsymbol{\epsilon}_n)\right],
\end{aligned}
$$

for $i = 1, \ldots, p$, where $\boldsymbol{\epsilon}_n$ lies between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_n^H$. Writing the above equations in matrix form, we get

$$\begin{aligned}
&\frac{\partial H}{\partial \boldsymbol{\theta}}\left(X(n); \boldsymbol{\theta}_0\right) \\
&\quad + \left[\frac{\partial^2 H}{\partial \boldsymbol{\theta}^2}\left(X(n); \boldsymbol{\theta}_0\right) + \frac{1}{2}\sum_{j=1}^{p}\left(\theta_{n,j}^H - \theta_{0,j}\right)\left(\frac{\partial}{\partial \theta_j}\left(\frac{\partial^2 H}{\partial \boldsymbol{\theta}}\right)\left(X(n); \boldsymbol{\epsilon}_n\right)\right)\right] \\
&\quad \left(\boldsymbol{\theta}_n^H - \boldsymbol{\theta}_0\right) = 0.
\end{aligned} \tag{33}$$

Observe, by taking derivatives under the integral sign, that $\forall i, j$,

$$\begin{aligned}
&\frac{\partial^2 H}{\partial \theta_j \partial \theta_i}\left(X(n); \boldsymbol{\theta}_0\right) \\
&= \frac{\partial^2 H}{\partial \theta_j \partial \theta_i}\iint_S \left[T_{\boldsymbol{\theta}}([a,b]) - \hat{T}([a,b]; X(n))\right]^2 W(a,b)\mathrm{d}a\mathrm{d}b, \\
&= \frac{\partial}{\partial \theta_j} 2\iint_S \left[T_{\boldsymbol{\theta}}([a,b]) - \hat{T}([a,b]; X(n))\right]\frac{\partial T_{\boldsymbol{\theta}_0}}{\partial \theta_i}([a,b])W(a,b)\mathrm{d}a\mathrm{d}b, \\
&= 2\iint_S \left[T_{\boldsymbol{\theta}}([a,b]) - \hat{T}([a,b]; X(n))\right]\frac{\partial^2 T_{\boldsymbol{\theta}_0}}{\partial \theta_j \partial \theta_i}([a,b])W(a,b)\mathrm{d}a\mathrm{d}b \\
&\quad +2\iint_S \left(\frac{\partial T_{\boldsymbol{\theta}_0}}{\partial \theta_j}\frac{\partial T_{\boldsymbol{\theta}_0}}{\partial \theta_i}\right)([a,b])W(a,b)\mathrm{d}a\mathrm{d}b \\
&:= I + II.
\end{aligned}$$

The first term is

$$\begin{aligned}
I &= 2\iint_S \left(T_{\boldsymbol{\theta}_0}([a,b]) - \frac{1}{n}\sum_{k=1}^{n} Y_k(a,b)\right)\frac{\partial^2 T_{\boldsymbol{\theta}_0}}{\partial \theta_j \partial \theta_i}([a,b])W(a,b)\mathrm{d}a\mathrm{d}b \\
&= \frac{2}{n}\sum_{k=1}^{n}\iint_S \left[T_{\boldsymbol{\theta}_0}([a,b]) - Y_k(a,b)\right]\frac{\partial^2 T_{\boldsymbol{\theta}_0}}{\partial \theta_j \partial \theta_i}([a,b])W(a,b)\mathrm{d}a\mathrm{d}b \\
&= o_P(1),
\end{aligned}$$

according to the strong law of large numbers for i.i.d. random variables. Therefore,

$$\frac{\partial^2 H}{\partial \theta_j \partial \theta_i}\left(X(n); \boldsymbol{\theta}_0\right) = o_P(1) + 2\iint_S \left(\frac{\partial T_{\boldsymbol{\theta}_0}}{\partial \theta_j}\frac{\partial T_{\boldsymbol{\theta}_0}}{\partial \theta_i}\right)([a,b])W(a,b)\mathrm{d}a\mathrm{d}b, \quad \forall i, j.$$

In matrix form,

$$\frac{\partial^2 H}{\partial \boldsymbol{\theta}^2}\left(X(n); \boldsymbol{\theta}_0\right) = o_P(1) + 2\iint_S \left(\frac{\partial T_{\boldsymbol{\theta}_0}}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial T_{\boldsymbol{\theta}_0}}{\partial \boldsymbol{\theta}}\right)^T([a,b])W(a,b)\mathrm{d}a\mathrm{d}b. \tag{34}$$

Observe again that $\forall j, k, l$,

$$\begin{aligned}
&\left|\frac{\partial^3 H(X(n);\boldsymbol{\epsilon}_n)}{\partial\theta_j\partial\theta_k\partial\theta_l}\right| \\
&\le 2\iint_S \left|\left[T_{\boldsymbol{\epsilon}_n}([a,b]) - \hat{T}([a,b];X(n))\right]\frac{\partial^3 T_{\boldsymbol{\epsilon}_n}}{\partial\theta_j\partial\theta_k\partial\theta_l}([a,b])W(a,b)\mathrm{d}a\mathrm{d}b\right| \\
&\quad +2\left|\iint_S \left[\left(\frac{\partial T_{\boldsymbol{\epsilon}_n}}{\partial\theta_j}\frac{\partial^2 T_{\boldsymbol{\epsilon}_n}}{\partial\theta_k\partial\theta_l}\right)+\left(\frac{\partial^2 T_{\boldsymbol{\epsilon}_n}}{\partial\theta_j\partial\theta_k}\frac{\partial T_{\boldsymbol{\epsilon}_n}}{\partial\theta_l}\right)+\left(\frac{\partial^2 T_{\boldsymbol{\epsilon}_n}}{\partial\theta_j\partial\theta_l}\frac{\partial T_{\boldsymbol{\epsilon}_n}}{\partial\theta_k}\right)\right]\right. \\
&\quad \left.\times([a,b])W(a,b)\mathrm{d}a\mathrm{d}b\right| \le 4\iint_S\left|\frac{\partial^3 T_{\boldsymbol{\epsilon}_n}}{\partial\theta_j\partial\theta_k\partial\theta_l}([a,b])W(a,b)\mathrm{d}a\mathrm{d}b\right| \\
&\quad +2\left|\iint_S \left[\left(\frac{\partial T_{\boldsymbol{\epsilon}_n}}{\partial\theta_j}\frac{\partial^2 T_{\boldsymbol{\epsilon}_n}}{\partial\theta_k\partial\theta_l}\right)+\left(\frac{\partial^2 T_{\boldsymbol{\epsilon}_n}}{\partial\theta_j\partial\theta_k}\frac{\partial T_{\boldsymbol{\epsilon}_n}}{\partial\theta_l}\right)+\left(\frac{\partial^2 T_{\boldsymbol{\epsilon}_n}}{\partial\theta_j\partial\theta_l}\frac{\partial T_{\boldsymbol{\epsilon}_n}}{\partial\theta_k}\right)\right]\right. \\
&\quad \left.\times([a,b])W(a,b)\mathrm{d}a\mathrm{d}b\right| := C_1(\boldsymbol{\epsilon}_n) \le C_2,
\end{aligned}$$

$\forall\boldsymbol{\epsilon}_n \in \Theta$, by the compactness of Θ. This, together with the strong consistency of $\boldsymbol{\theta}_n^H$, gives

$$\begin{aligned}
&\frac{1}{2}\sum_{j=1}^{p}\left(\theta_{n,j}^H - \theta_{0,j}\right)\left(\frac{\partial}{\partial\theta_j}\left(\frac{\partial^2 H}{\partial\theta_k\partial\theta_l}\right)(X(n);\boldsymbol{\epsilon}_n)\right) \\
&= \frac{1}{2}\sum_{j=1}^{p} o_P(1)\frac{\partial^3 H(X(n);\boldsymbol{\epsilon}_n)}{\partial\theta_j\partial\theta_k\partial\theta_l} \\
&= o_P(1),
\end{aligned}$$

$\forall k, l$. Equivalently, in matrix form,

$$\frac{1}{2}\sum_{j=1}^{p}\left(\theta_{n,j}^H - \theta_{0,j}\right)\left(\frac{\partial}{\partial\theta_j}\left(\frac{\partial^2 H}{\partial\boldsymbol{\theta}}\right)(X(n);\boldsymbol{\epsilon}_n)\right) = o_P(1). \tag{35}$$

By the multivariate Slutsky's theorem, Proposition 2, together with Eqs. (33), (34), and (35), yields the desired result.

Acknowledgments The authors thank the editor and the two anonymous referees for their careful reading of the manuscript and valuable comments that improved the old version of the paper.

References

Artstein, Z., Vitale, R. A. (1975). A strong law of large numbers for random compact sets. *Annals of Probability*, *5*, 879–882.

Aumann, R. J. (1965). Integrals and set-valued functions. *Journal of Mathematical Analysis and Applications*, *12*, 1–12.

Billard, L., Diday, E. (2003). From the statistics of data to the statistics of knowledge: symbolic data analysis. *Journal of the American Statistical Association, 98*, 470–487, review article.
Carvalho, F. A. T., Neto, E. A. L., Tenorio, C. P. (2004). A New Method to Fit a Linear Regression Model for Interval-Valued Data. *Lecture Notes in Computer Science*, *3238*, 295–306.
Diamond, P. (1990). Least squares fitting of compact set-valued data. *Journal of Mathematical Analysis and Applications*, *147*, 531–544.
Gil, M. A., Lubiano, M. A., Montenegro, M., Lopez, M. T. (2002). Least squares fitting of an affine function and strength of association for interval-valued data. *Metrika*, *56*, 97–111.
Heinrich, L. (1993). Asymptotic properties of minimum contrast estimators for parameters of Boolean models. *Metrika*, *40*, 69–74.
Hörmander, L. (1954). Sur la fonction d'appui des ensembles convexes dans un espace localement convexe. *Arkiv för Matematik*, *3*, 181–186.
Körner, R. (1995). A Variance of Compact Convex Random Sets. Working paper, Institut für Stochastik, Bernhard-von-Cotta-Str. 2 09599 Freiberg.
Körner, R. (1997). On the variance of fuzzy random variables. *Fuzzy Sets and Systems*, *92*, 83–93.
Körner, R., Näther, W. (1998). Linear regression with random fuzzy variables: extended classical estimates, best linear estimates, least squares estimates. *Information Sciences*, *109*, 95–118.
Lyashenko, N. N. (1979). On limit theorems for sums of independent compact random subsets in the Euclidean space. *Translated from Zapiski Nauchnykh Seminarov Leningradskogo Otdeleniya Matematicheskogo Instituta*, *85*, 113–128.
Lyashenko, N. N. (1983). Statistics of random compacts in Euclidean space. *Translated from Zapiski Nauchnykh Seminarov Leningradskogo Otdeleniya Matematicheskogo Instituta*, *98*, 115–139.
Matheron, G. (1975). *Random sets and integral geometry*. New York: Wiley.
Molchanov, I. (2005). *Theory of random sets*. London: Springer.
Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, *29*, 429–443.
Norberg, T. (1984). Convergence and existence of random set distributions. *The Annals of Probability*, *12*(3), 726–732.
Pfanzagl, J. (1969). On the measurability and consistency of minimum contrast estimates. *Metrika*, *14*, 249–272.
Puri, M. L., Ralescu, D. A., Ralescu, S. S. (1986). Gaussian random sets in Banach space. *Theory of Probablity and its Applications*, *31*, 526–529.
Rådström, H. (1952). An embedding theorem for spaces of convex sets. *Proceedings of the American Mathematical Society*, *3*, 165–169.
Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., Lanckriet, G. R. G. (2012). On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, *6*, 1550–1599.
Stoyan, D. (1998). Random sets: models and statistics. *International Statistical Review*, *66*(1), 1–27.
Tanaka, U., Ogata, Y., Stoyan, D. (2008). Parameter estimation and model Selection for Neyman–Scott point processes. *Biometrical Journal*, *50*, 43–57.
Zolotarev, V. M. (1983). Probability metrics. *Theory of Probability and its Applications*, *28*, 278–302.