Adv. Stat. Clim. Meteorol. Oceanogr., 11, 73–87, 2025
https://doi.org/10.5194/ascmo-11-73-2025

ASCMO
Open Access

# On inference of boxplot symbolic data: applications in climatology

**Abdolnasser Sadeghkhani**[1] **and Ali Sadeghkhani**[2]

[1]Department of Mathematics and Statistics,
North Carolina Agricultural and Technical State University, Greensboro, NC, USA
[2]Department of Mathematics and Statistics, University of Windsor, Windsor, ON, Canada

**Correspondence:** Abdolnasser Sadeghkhani (asadeghkhani@ncat.edu)

Received: 20 August 2024 – Revised: 22 January 2025 – Accepted: 24 January 2025 – Published: 17 March 2025

**Abstract.** This paper presents a pioneering study on the inference of boxplot-valued data using both Bayesian and frequentist approaches within a multivariate framework. This approach leverages complex yet intuitive representations to make large datasets more manageable and enhance their interpretability, which is invaluable in the age of big data. Boxplot-valued data are particularly important due to their ability to capture the inherent variability and distributional characteristics of complex datasets.

In our study, we propose novel methodologies for parameter estimation and density estimation for boxplot-valued data and apply these techniques to climatological data. Specifically, we utilize data from the Berkeley Earth Surface Temperature Study, which aggregates 1.6 billion temperature reports from 16 pre-existing archives affiliated with the Lawrence Berkeley National Laboratory. Our methods are validated through extensive simulations comparing the efficiency and accuracy of Bayesian and frequentist estimators.

We demonstrate the practical applicability of our approach by analyzing summer average temperatures across various European countries. The proposed techniques provide robust tools for analyzing complex data structures, offering valuable insights into climatic trends and variations. Our study highlights the advantages and limitations of each inferential method, offering guidance for future research and applications in the field of climatology.

## 1 Introduction

Data analysts are now encountering increasingly large datasets whose analysis, spanning decades or even centuries, presents a complex challenge that demands more in-depth attention. There are many instances where complex and vast information, due to its intrinsic structure, cannot be adequately represented as single-valued data. As a result, symbolic data analysis (SDA) has emerged as a crucial approach, integrating elements of data science, multivariate analysis, pattern recognition, data mining, and artificial intelligence to analyze such data properly without losing information. Initially conceptualized by Diday (1988) and later formalized by Billard and Diday (2003), SDA represents a significant advancement in managing, reducing, and interpreting large datasets.

The initial steps of summarizing data to facilitate a clustering process applied to a large database appeared in the work of Diday and Noirhomme-Fraiture (2008), which, by introducing the cluster as a category, highlighted the importance of data summarization using SDA. To manage data effectively, SDA works without losing information, a common issue in traditional analyses, and encompasses various data formats, including intervals, sets, lists, histograms, trees, boxplots, and other distributional representations. By employing these complex yet intuitive representations, SDA not only makes vast amounts of data more manageable but also enhances their comprehensibility, proving to be an invaluable tool in the age of big data.

The importance of boxplot-valued data in real-world applications is particularly evident in climate science, where data often span long periods and involve significant uncertainty. For example, long-term datasets such as historical temperature records or remote sensing data generate vast volumes of observations that are often summarized as medians, quartiles, and extreme values for computational efficiency. These boxplot-valued summaries enable researchers to analyze trends in temperature variability, extreme events, or regional shifts in central tendencies, all of which are critical for understanding climate dynamics and making policy decisions.

By expanding data and converting their form from single-valued to symbolic (e.g., multivalued data such as list data, model-valued data such as histograms, or interval-valued data), the need for further theoretical development of SDA in each of its types has emerged. Diday (1995), Émilion (1997), and Diday (1988) provided the foundational mathematical background for various forms of symbolic data.

Among symbolic data types, interval-valued data have received significant attention due to their simplicity, versatility, and prevalence in various applied sciences compared to other symbolic data types. These data types, which range from univariate to multivariate, address the needs of real-world applications. In recent years, both frequentist and Bayesian approaches to interval-valued data estimators have been explored (e.g., Xu and Qin, 2022; Samadi et al., 2024; Sadeghkhani and Sadeghkhani, 2024). Additionally, various statistical tools for investigating relationships within interval-valued data have been developed. For instance, principal component analysis (PCA) using vertices to represent intervals was proposed by Douzal-Chouakria et al. (2011), a regression method that avoids the center and range approach was introduced by Billard and Diday (2000) and later by Neto and De Carvalho (2008) and Neto and De Carvalho (2010), and interval-valued time series methods were implemented by Xiong et al. (2015).

Beyond interval-valued data, boxplot-valued data provide additional distributional insights by including medians, quartiles, and extremes, making them well suited to summarizing massive datasets. In climate science, examples include datasets where raw observations (e.g., hourly or daily temperatures) are summarized into annual or monthly boxplots for computational feasibility. Historical datasets from predigital eras often come as summarized statistics (e.g., medians and extremes), making boxplot-valued data analysis indispensable. Such datasets allow for the study of trends in variability and centrality, which are critical for analyzing long-term climate change impacts.

The relevance of boxplot-valued data is also evident in studies of regional variability, where data collection is aggregated to minimize measurement noise. For example, climate model datasets often summarize temperature and precipitation data across spatial grids, which could naturally be represented and analyzed as boxplot-valued data. This approach reduces computational complexity while preserving the essential distributional features of the data.

Through this work, we aim to develop a methodological framework for analyzing boxplot-valued data and providing tools that are computationally efficient and applicable to a wide range of fields, including but not limited to climate science. By addressing the unique challenges posed by such data, we contribute to the broader goal of managing and interpreting massive datasets in the applied sciences.

While interval-valued data capture the range of variability, boxplot-valued data provide a richer summary by including additional distributional characteristics such as quartiles and the median. Boxplots are widely used for comparing distributions across different groups, detecting outliers, and understanding the spread and symmetry of the data. The concept of boxplots was first proposed by Tukey (1977) and further developed by Benjamini (1988). However, the mathematical foundation for statistical inference using boxplot-valued data has not been assessed thoroughly.

Though boxplots are valuable for comparing distributions across different groups, detecting outliers, and understanding the spread and symmetry of the data, their mathematical foundation has not been assessed thoroughly. They are widely used in various fields due to their simplicity and effectiveness in summarizing complex datasets. The concept of boxplots was first proposed by Tukey (1977) and was developed by Benjamini (1988). The importance of data visualization was further emphasized by Chambers (2018), who explored graphical methods for data analysis. Wickham and Wickham (2016) focused on modern approaches to creating boxplots using the `ggplot2` package in R, highlighting practical aspects of their implementation.

To compare the role of boxplot-valued data, Arroyo et al. (2006) contrasted these types of data with interval-valued data and histograms. Boxplot variables serve as an intermediate point between the simplicity of interval variables and the detailed information provided by histogram variables. While interval variables do not convey information about the central area of an empirical distribution, boxplot variables do so using three quartiles. In contrast, histogram variables offer detailed insights into the empirical distribution, though their structure is more complex, requiring a set of consecutive intervals with associated weights. Despite their simpler structure, boxplot variables effectively capture the shape of the distribution.

To the best of our knowledge, there are no specific mathematical considerations for boxplot-valued data parameter estimation in the context of symbolic data, whether from a frequentist or Bayesian perspective. However, Reyes et al. (2024) proposed a parameterized regression method for boxplot-valued data by applying a Box–Cox transformation. Additionally, Reyes et al. (2022) focused on forecasting time series of these types of data.

Let $X_i$ be a boxplot-valued observation summarized by its five-number summary statistics: the minimum ($a_i$), first quartile ($q_{1i}$), median ($m_i$), third quartile ($q_{3i}$), and maximum ($b_i$). This structure effectively captures the distributional characteristics of the data for each observation. To facilitate more precise parameter estimation, we decompose $X_i$ into two components, i.e., $Q_i = (q_{1i}, m_i, q_{3i})^\top$, which focuses on the central tendency and the shape of the distribution, and $R_i = b_i - a_i$, which represents the range, thus capturing the overall spread of the data. By separating these components, we are able to model the central distribution and variability independently, thereby enhancing the accuracy of our parameter estimates.

While Le-Rademacher and Billard (2011) focused on univariate interval-valued data, modeling the midpoints in a frequentist framework, we extend their methodology to handle boxplot-valued data. They assumed that the midpoints of intervals are normally distributed and derived maximum likelihood estimators for the mean and variance. By assuming a uniform distribution within each interval, we leverage the properties outlined by Le-Rademacher and Billard to justify the normality of $Q_i = [q_{1i}, m_i, q_{3i}]^\top$, where, for instance, $q_{1i}$ is the midpoint between $a_i$ and $m_i = q_{2i}$.

However, it is important to note that this extension from univariate intervals to multivariate quartiles relies on specific assumptions. These include the assumption that the quartiles are derived from sufficiently large datasets where the central limit theorem (CLT) ensures the approximation of normality. Additionally, we assume that the data within the intervals are uniformly distributed. While this assumption is commonly employed in interval-valued data analysis, deviations from uniformity or smaller datasets could impact the validity of the multivariate normal model. In such cases, a larger $Q_i$ dataset would be required to approximate normality using the CLT.

In our approach, we model $Q_i$ as following a trivariate normal distribution $\mathcal{N}_3(\mu, \Sigma)$, where $\mu = (\mu_{q_1}, \mu_m, \mu_{q_3})^\top$ is the mean vector and $\Sigma$ is the covariance matrix. This extension allows us to capture more detailed distributional characteristics inherent in boxplot data, offering a balance between theoretical rigor and practical applicability.

Therefore, we have

$$f(Q_i;\, \mu, \Sigma) = \frac{1}{(2\pi)^{3/2}|\Sigma|^{1/2}} \cdot \exp\left(-\frac{1}{2}(Q_i - \mu)^\top \Sigma^{-1}(Q_i - \mu)\right) \tag{1}$$

and

$$f(R_i;\, R_{\min}, R_{\max}) = \frac{1}{R_{\max} - R_{\min}} \quad \text{for } R_{\min} \le R_i \le R_{\max}, \tag{2}$$

where we assume that $Q_i$ (the quantiles) and $R_i$ (the range) are conditionally independent under the assumption that the internal distribution (quantiles) of the interval is not directly driven by the overall width of the interval.

It is worth noting that, in our model, the range $R_i = b_i - a_i$ is assumed to follow a uniform distribution over the interval $[R_{\min}, R_{\max}]$. This assumption is commonly employed in the context of interval-valued data analysis (e.g., Neto and De Carvalho, 2008; Zhao et al., 2023). The uniform distribution is chosen due to its simplicity and practical effectiveness, particularly when there is no prior information suggesting a different underlying distribution. The bounds $R_{\min}$ and $R_{\max}$ are determined empirically based on the observed data, ensuring that they capture the full range of variability in our dataset.

Note that, by defining the scaling parameter $\lambda$, the bounds $a_i$ and $b_i$ are hence related to the quantiles $q_{1i}$ and $q_{3i}$ as

$$a_i = q_{1i} - \lambda\left(1 - \frac{\mathrm{IQR}_i}{R_i}\right) R_i, \tag{3}$$

$$b_i = q_{3i} + (1-\lambda)\left(1 - \frac{\mathrm{IQR}_i}{R_i}\right) R_i, \tag{4}$$

where $\mathrm{IQR}_i = q_{3i} - q_{1i}$, which is known as the interquartile range.

The rest of the paper is organized as follows. In Sect. 2, we elaborate on the maximum likelihood (ML) method for estimating the unknown parameters of a $p$-variate boxplot-valued random variable where $p \ge 1$. This section also includes a simulation to assess this method as well as the asymptotic distribution of the ML estimators. Section 3 deals with Bayesian estimation of the parameters and evaluates the proposed methods. Section 4 investigates the methods of density estimation in boxplot-valued data, ranging from plugin types to posterior predictive density estimators. In Sect. 5, we illustrate the practical utility and effectiveness of the proposed techniques in analyzing and interpreting complex boxplot-valued environmental data. Finally, we conclude with a discussion in Sect. 6.

## 2 Maximum likelihood estimators

The likelihood function of the unknown parameters $\mu, \Sigma, R_{\min}, R_{\max}$, and $\lambda$ based on a random sample of size $n$, $X_i = [a_i, q_{1i}, m_i, q_{3i}, b_i]$ for $i = 1, \ldots, n$, assuming that Eqs. (1) and (2) are conditionally independent despite the inherent relationship between $R_i$ and $Q_i$, is given by

$$\mathcal{L}(\mu, \Sigma, R_{\min}, R_{\max}, \lambda) = \prod_{i=1}^{n} f(Q_i;\, \mu, \Sigma) \cdot f(R_i;\, R_{\min}, R_{\max}). \tag{5}$$

It can easily be seen that the ML estimators of $\mu$ are the sample means of $q_{1i}$, $m_i$, and $q_{3i}$, i.e.,

$$\hat{\mu}_{q_1} = \frac{1}{n}\sum_{i=1}^{n} q_{1i}, \quad \hat{\mu}_m = \frac{1}{n}\sum_{i=1}^{n} m_i, \quad \hat{\mu}_{q_3} = \frac{1}{n}\sum_{i=1}^{n} q_{3i}. \tag{6}$$

The ML estimator of $\Sigma$ is the sample covariance matrix

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(Q_i - \hat{\mu})(Q_i - \hat{\mu})^\top,$$

while the ML estimators for $R_{\min}$ and $R_{\max}$ are

$$\hat{R}_{\min} = \min_{i=1}^{n} R_i, \quad \hat{R}_{\max} = \max_{i=1}^{n} R_i.$$

The scaling parameter $\lambda$ is estimated by minimizing an objective function. That is,

$$\hat{\lambda} = \arg\min_{\lambda} \sum_{i=1}^{n} \left( R_i - \left( q_{3i} + (1-\lambda) \left( 1 - \frac{\mathrm{IQR}_i}{R_i} \right) R_i \right) \right)^2 .$$

Our assumption of conditional independence between $R_i$ and $Q_i$ is made to simplify the estimation process, recognizing that $R_i$ and $Q_i$ are functionally related via the quartiles and the scaling parameter $\lambda$. However, by assuming that $\lambda$ is constant across observations, we reduce the direct influence of $Q_i$ on $R_i$, making this approximation reasonable in practice.

The primary reason we applied an objective function for minimizing $\lambda$ was to directly assess how well the model captures the relationship between the observed ranges $R_i$ and the estimated upper quantiles $q_{3i}$. Furthermore, while maximum likelihood estimation (MLE) is a robust method for parameter estimation, it relies on certain assumptions about the distributional properties of the data. Minimizing the objective function allows for greater flexibility in modeling this relationship without imposing strict distributional assumptions. Future work could explore models that explicitly incorporate this dependency, potentially enhancing the accuracy of the parameter estimates.

In the next subsection, we generalize the univariate $X$ into the $p$-variate $\mathbf{X}_i = [X_{i1}, X_{i2}, \ldots, X_{ip}]$ with $i = 1, \ldots, n$ for $p > 1$, and we focus on estimating the multivariate boxplot data.

### 2.1 Multivariate ML estimators

Consider a $p$-variate boxplot-valued random variable $\mathbf{X}_i = [X_{i1}, X_{i2}, \ldots, X_{ip}]$, where each $X_{ij} = [a_{ij}, q_{1ij}, m_{ij}, q_{3ij}, b_{ij}]$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Moreover, we assume that

$$\mathbf{Q}_i = (Q_{i1}, \ldots, Q_{ip})^\top \sim \mathcal{N}_{3p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

such that

$$Q_{ij} = (q_{1ij}, m_{ij}, q_{3ij}) \sim \mathcal{N}_3(\mu_j, \Sigma_j)$$
for $i = 1, \ldots, n,\ j = 1, \ldots, p,$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$ is the mean vector with $\mu_j = (\mu_{q_{1j}}, \mu_{mj}, \mu_{q_{3j}})^\top$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} & \cdots & \Sigma_{1,p} \\ \Sigma_{2,1} & \Sigma_{2,2} & \cdots & \Sigma_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{p,1} & \Sigma_{p,2} & \cdots & \Sigma_{p,p} \end{pmatrix}$$

is the covariance matrix, with $\Sigma_{k,l}$ representing the covariance between $Q_k$ and $Q_l$ for $k, l = 1, \ldots, p$. Moreover, the range $R_{ij} = b_{ij} - a_{ij}$ is uniformly distributed over $[R_{\min,j}, R_{\max,j}]$.

In a similar fashion to the univariate case with $p = 1$, to estimate the parameters in a multivariate setting, we maximize the likelihood function

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{R}_{\min}, \boldsymbol{R}_{\max}, \boldsymbol{\lambda}) = \prod_{i=1}^{n} f(\mathbf{Q}_i;\ \boldsymbol{\mu}, \boldsymbol{\Sigma}) \cdot \prod_{j=1}^{p} f(R_{ij};\ R_{\min,j}, R_{\max,j}), \tag{7}$$

where $f(\mathbf{Q}_i;\ \boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents the multivariate normal density function and $f(R_{ij};\ R_{\min,j}, R_{\max,j})$ is the density function of the uniform distribution for each component $R_{ij}$, with $\boldsymbol{R}_{\min}$ and $\boldsymbol{R}_{\max}$ denoting the vectors of the minimum and maximum values for each component $j$ of the range.

This results in the ML estimator of the parameters as presented in the following theorem.

**Theorem 1.** *Consider a $p$-variate boxplot-valued random variable $\mathbf{X}_i = [X_{ii}, X_{i2}, \ldots, X_{ip}]$, where $X_{ij} = [a_{ij}, q_{1ij}, m_{ij}, q_{3ij}, b_{ij}]$ and $\mu_j = (\mu_{q_{1j}}, \mu_{mj}, \mu_{q_{3j}})^\top$. Suppose that*

$$\mathbf{Q}_i = (Q_{i1}, \ldots, Q_{ip})^\top \sim \mathcal{N}_{3p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

*with $Q_{ij} = (q_{1ij}, m_{ij}, q_{3ij})^\top$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. The ML estimators of the mean vector $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_p)^\top$, where $\hat{\mu}_j = (\hat{\mu}_{q_{1j}}, \hat{\mu}_{mj}, \hat{\mu}_{q_{3j}})^\top$, are given by*

$$\hat{\mu}_{q_{1j}} = \frac{1}{n} \sum_{i=1}^{n} q_{1ij}, \tag{8}$$

$$\hat{\mu}_{mj} = \frac{1}{n} \sum_{i=1}^{n} m_{ij}, \tag{9}$$

$$\hat{\mu}_{q_{3j}} = \frac{1}{n} \sum_{i=1}^{n} q_{3ij}. \tag{10}$$

*The ML estimator of the covariance matrix $\boldsymbol{\Sigma}$ is*

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{Q}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{Q}_i - \hat{\boldsymbol{\mu}})^\top. \tag{11}$$

*The ML estimators of the range parameters are*

$$\hat{R}_{\min,j} = \min\nolimits_{i=1}^{n} R_{ij}, \tag{12}$$

$$\hat{R}_{\max,j} = \max_{i=1}^{n} R_{ij}, \tag{13}$$

*and the ML estimator of the scaling parameter $\lambda_j$ is*

$$\hat{\lambda}_j = \arg\min_{\lambda_j} \sum_{i=1}^{n} \left( R_{ij} - \left( q_{3ij} + (1-\lambda_j) \cdot \left( 1 - \frac{q_{3ij} - q_{1ij}}{R_{ij}} \right) \cdot R_{ij} \right) \right)^2 . \tag{14}$$

*Proof.* The proof involves straightforward maximization of the likelihood function. For brevity, it is omitted here.

□

The following subsection details the simulation setup for the proposed MLE method for symbolic boxplot-valued data.

## 2.2 Simulation study of ML estimators

To demonstrate the estimation procedure, we simulate data for $p = 3$ variables with a sample size of $n = 100$ and the following true parameters:

$$\mu_1 = (5, 10, 15)^\top, \quad \mu_2 = (4, 8, 12)^\top,$$
$$\mu_3 = (6, 11, 16)^\top, \tag{15}$$

$$\Sigma_{1,1} = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}, \quad \Sigma_{1,2} = \begin{bmatrix} 1.5 & 0.8 & 0.8 \\ 0.8 & 1.5 & 0.8 \\ 0.8 & 0.8 & 1.5 \end{bmatrix},$$
$$\Sigma_{1,3} = \begin{bmatrix} 2.5 & 1.2 & 1.2 \\ 1.2 & 2.5 & 1.2 \\ 1.2 & 1.2 & 2.5 \end{bmatrix},$$
$$\Sigma_{2,2} = \begin{bmatrix} 1.2 & 0.5 & 0.5 \\ 0.5 & 1.2 & 0.5 \\ 0.5 & 0.5 & 1.2 \end{bmatrix}, \quad \Sigma_{2,3} = \begin{bmatrix} 1.8 & 0.9 & 0.9 \\ 0.9 & 1.8 & 0.9 \\ 0.9 & 0.9 & 1.8 \end{bmatrix},$$
$$\Sigma_{3,3} = \begin{bmatrix} 2 & 0.9 & 0.9 \\ 0.9 & 2 & 0.9 \\ 0.9 & 0.9 & 2 \end{bmatrix}, \tag{16}$$

$$R_{\min,j} = 5, \quad R_{\max,j} = 15, \quad j = 1, 2, 3. \tag{17}$$

The true values for the scaling parameters are $\lambda_1 = 0.48$, $\lambda_2 = 0.49$, and $\lambda_3 = 0.51$.

In the simulation, we generate the components $Q_i$ and $R_i$ separately to simplify the estimation. Specifically, $Q_i = (q_{1i}, m_i, q_{3i})$ is generated from a trivariate normal distribution $\mathcal{N}_3(\mu, \Sigma)$ as outlined in Eq. (1), and $R_i = b_i - a_i$ is assumed to follow a uniform distribution over $[R_{\min}, R_{\max}]$, as given in Eq. (2). This approach enables independent handling of each component's characteristics.

The mean squared error (MSE) is computed for the mean vector $\boldsymbol{\mu}$ and the covariance matrices $\boldsymbol{\Sigma}_{j,k}$ for $j, k = 1, 2, 3$, as follows:

$$\mathrm{MSE}(\hat{\mu}_j) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_{ji} - \mu_j)^2, \quad j = 1, 2, 3,$$

and

$$\mathrm{MSE}(\hat{\Sigma}_{j,k}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{3}\sum_{b=1}^{3}(\hat{\Sigma}_{j,k}^{(i)}[a,b] - \Sigma_{j,k}[a,b])^2, \tag{18}$$

where $\hat{\Sigma}_{j,k}^{(i)}[a,b]$ and $\Sigma_{j,k}[a,b]$ represent the $(a,b)$th elements of the estimated and true covariance matrices, respectively, in the $i$th simulation.

**Table 1.** ML estimates and MSEs.

| Parameters | Estimates (MSEs) |
|---|---|
| $\mu_1$ | $(5.21, 9.86, 14.88)^\top$ (0.025) |
| $\mu_2$ | $(3.87, 7.89, 11.99)^\top$ (0.0096) |
| $\mu_3$ | $(5.66, 10.79, 15.65)^\top$ (0.0929) |
| $\Sigma_{1,1}$ | $\begin{bmatrix} 2.25 & 1.44 & 1.67 \\ 1.44 & 2.25 & 1.56 \\ 1.67 & 1.56 & 2.89 \end{bmatrix}$ (0.1302) |
| $\Sigma_{1,2}$ | $\begin{bmatrix} 1.78 & 1.01 & 1.13 \\ 1.01 & 1.52 & 0.71 \\ 1.13 & 0.71 & 1.78 \end{bmatrix}$ (0.0541) |
| $\Sigma_{1,3}$ | $\begin{bmatrix} 3.18 & 1.70 & 1.35 \\ 1.70 & 2.28 & 1.04 \\ 1.35 & 1.04 & 1.90 \end{bmatrix}$ (0.1608) |
| $\Sigma_{2,2}$ | $\begin{bmatrix} 1.09 & 0.57 & 0.47 \\ 0.57 & 1.48 & 0.66 \\ 0.47 & 0.66 & 1.28 \end{bmatrix}$ (0.0178) |
| $\Sigma_{2,3}$ | $\begin{bmatrix} 1.89 & 0.89 & 0.87 \\ 0.89 & 1.95 & 0.96 \\ 0.87 & 0.96 & 1.55 \end{bmatrix}$ (0.0112) |
| $\Sigma_{3,3}$ | $\begin{bmatrix} 1.96 & 0.88 & 0.87 \\ 0.88 & 2.05 & 0.96 \\ 0.87 & 0.96 & 1.78 \end{bmatrix}$ (0.0174) |
| $R_{\min,1}, R_{\max,1}$ | 5.0389, 14.7589 (0.0015, 0.0580) |
| $R_{\min,2}, R_{\max,2}$ | 5.0510, 14.888 (0.001, 0.0670) |
| $R_{\min,3}, R_{\max,3}$ | 4.998, 15.008 (0.0015, 0.0580) |
| $\lambda_1, \lambda_2, \lambda_3$ | 0.48, 0.49, 0.51 (NA) |

NA – no variation across simulations.

Table 1 presents the MLE parameters along with their MSEs. Note that the MSEs for $\lambda_j$ $(j = 1, 2, 3)$ are reported as "NA", indicating no variation in these parameters across simulations.

## 2.3 Asymptotic distribution of ML estimators for boxplot-valued data

In this subsection, we derive the asymptotic distribution of the ML estimators for the parameters of a boxplot-valued random variable.

**Theorem 2.** *Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\lambda}})$ denote the ML estimators for the parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ of boxplot-valued data. Under regularity conditions (see, e.g., Ferguson, 2017), the asymptotic distribution of the ML estimators is given as*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}_{\frac{9p^2+11p}{2}}(0, \mathcal{I}(\boldsymbol{\theta})^{-1}),$$

*where $\mathcal{I}(\boldsymbol{\theta})$ is the Fisher information matrix and is given by*

$$\mathcal{I}(\boldsymbol{\theta}) = \begin{bmatrix} \mathcal{I}_{\boldsymbol{\mu}} & 0 & 0 \\ 0 & \mathcal{I}_{\boldsymbol{\Sigma}} & 0 \\ 0 & 0 & \mathcal{I}_{\boldsymbol{\lambda}} \end{bmatrix}$$

*with*

$$\mathcal{I}_{\boldsymbol{\mu}} = n\,\mathbf{I}_{3p} \otimes \boldsymbol{\Sigma}^{-1},$$
$$\mathcal{I}_{\boldsymbol{\Sigma}} = \frac{n}{2}(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}),$$
$$\mathcal{I}_{\boldsymbol{\lambda}} = n\,\mathrm{diag}\left(\frac{1}{\lambda_1^2}, \frac{1}{\lambda_2^2}, \ldots, \frac{1}{\lambda_p^2}\right),$$

*where* $\mathbf{I}_{3p}$ *is the* $3p \times 3p$ *identity matrix and* $\otimes$ *denotes the Kronecker product.*

*Proof.* The Fisher information matrix is the expected value of the outer product of the score function $\nabla\ell(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, and it is given as

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}\left[\nabla\ell(\boldsymbol{\theta})\nabla\ell(\boldsymbol{\theta})^\top\right].$$

For the multivariate normal distribution $\mathbf{Q}_i \sim \mathcal{N}_{3p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the score function for $\boldsymbol{\mu}$ is

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}} = \boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}(\mathbf{Q}_i - \boldsymbol{\mu}),$$

and hence

$$\mathcal{I}_{\boldsymbol{\mu}} = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \boldsymbol{\mu}}\right)\left(\frac{\partial \ell}{\partial \boldsymbol{\mu}}\right)^\top\right] = n\mathbf{I}_{3p} \otimes \boldsymbol{\Sigma}^{-1}.$$

Similarly, for $\boldsymbol{\Sigma}$, we have the score function

$$\frac{\partial \ell}{\partial \boldsymbol{\Sigma}} = \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}\right),$$

and therefore we have

$$\mathcal{I}_{\boldsymbol{\Sigma}} = \frac{n}{2}\mathbb{E}\big[\boldsymbol{\Sigma}^{-1}(\mathbf{Q}_i - \boldsymbol{\mu})(\mathbf{Q}_i - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}(\mathbf{Q}_i - \boldsymbol{\mu})(\mathbf{Q}_i - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}\big].$$

Since $\mathbb{E}\left[(\mathbf{Q}_i - \boldsymbol{\mu})(\mathbf{Q}_i - \boldsymbol{\mu})^\top\right] = \boldsymbol{\Sigma}$, we can write

$$\mathcal{I}_{\boldsymbol{\Sigma}} = \frac{n}{2}(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}).$$

For the uniform distribution $R_{ij} \sim \mathcal{U}(R_{\min,j}, R_{\max,j})$, the Fisher information for the parameter $\lambda_j$ (related to the range) is derived as

$$\mathcal{I}_{\boldsymbol{\lambda}} = n\,\mathrm{diag}\left(\frac{1}{\lambda_1^2}, \frac{1}{\lambda_2^2}, \ldots, \frac{1}{\lambda_p^2}\right).$$

This expression reflects the dependence of the Fisher information on the individual $\lambda_j$ values.

Using the central limit theorem, the score function evaluated at the true parameter values $\boldsymbol{\theta}$ is asymptotically normally distributed as

$$\sqrt{n}\nabla\ell(\boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}_{\frac{9p^2+11p}{2}}(0, \mathcal{I}(\boldsymbol{\theta})).$$

The covariance of the score function is related to the Fisher information matrix by the information matrix equality

$$\mathrm{Cov}(\nabla\ell(\boldsymbol{\theta})) = \mathcal{I}(\boldsymbol{\theta}).$$

The dimension of the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ is given by

$$\dim(\boldsymbol{\theta}) = 3p + \frac{3p(3p+1)}{2} + p = \frac{9p^2 + 11p}{2}.$$

This completes the proof.

□

## 3 Bayesian estimation for boxplot-valued data

Sadeghkhani and Sadeghkhani (2024) studied the Bayesian inference of symbolic interval-valued data by introducing noninformative priors, including the Jeffreys prior, into the multivariate setting. In this section, we use informative priors for boxplot-valued data, which allows us to update our prior knowledge based on the observed data. We derive the posterior distributions and find the Bayes estimators under the squared error loss (SEL) criterion.

### 3.1 Likelihood and priors

Consider the model where $\boldsymbol{Q}_1, \boldsymbol{Q}_2, \ldots, \boldsymbol{Q}_n$ are independent and identically distributed (iid) from $\mathcal{N}_{3p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with additional parameters $R_{\min j}$, $R_{\max j}$, and $\lambda_j$ for $j = 1, 2, \ldots, p$ included in the likelihood function $\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \{R_{\min j}\}, \{R_{\max j}\}, \{\lambda_j\} \mid \boldsymbol{Q})$.

We assume the following priors for the parameters:

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma} \sim \mathcal{N}_{3p}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\kappa), \tag{19}$$
$$\boldsymbol{\Sigma} \sim \mathcal{W}_{3p}^{-1}(\nu_0, \Psi_0), \tag{20}$$

where $\boldsymbol{\mu}_0$ is the prior mean, $\kappa$ is the scaling factor, $\nu_0$ is the degrees of freedom, and $\Psi_0$ is the scale matrix of the inverse Wishart distribution. Additionally, we have

$$R_{\min j} \sim \mathcal{U}(a_j, b_j), \quad R_{\max j} \sim \mathcal{U}(c_j, d_j),$$
$$\lambda_j \sim \mathcal{B}(\alpha_j, \beta_j), \tag{21}$$

where $a_j, b_j, c_j$, and $d_j$ are the bounds of the uniform distributions and $\alpha_j$ and $\beta_j$ are the parameters of the Beta distribution for each $j$.

Next, we find the posterior distributions of the parameters in the following lemma.

**Lemma 1** (posterior distributions). *Under the given priors in Eqs. (19), (20), and (21) and the observed data* $\boldsymbol{Q}_1, \boldsymbol{Q}_2, \ldots, \boldsymbol{Q}_n$*, we have the following.*

i. *The posterior distribution of* $\boldsymbol{\Sigma}$ *is given by*

$$\boldsymbol{\Sigma} \mid \boldsymbol{Q} \sim \mathcal{W}^{-1}(\nu_0 + n, \boldsymbol{\Psi}_0 + \boldsymbol{S}_n), \tag{22}$$

*where*

$$\boldsymbol{S}_n = \boldsymbol{S} + \frac{n\kappa}{n+\kappa}(\overline{\boldsymbol{Q}} - \boldsymbol{\mu}_0)(\overline{\boldsymbol{Q}} - \boldsymbol{\mu}_0)^\top$$

*and*

$$\boldsymbol{S} = \sum_{i=1}^{n}(\boldsymbol{Q}_i - \overline{\boldsymbol{Q}})(\boldsymbol{Q}_i - \overline{\boldsymbol{Q}})^\top, \quad \overline{\boldsymbol{Q}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Q}_i.$$

ii. *The posterior distribution of $\boldsymbol{\mu}$ is*

$$\boldsymbol{\mu} \mid \boldsymbol{Q} \sim \mathcal{T}_{3p}\left(\boldsymbol{\mu}_n, \frac{\Psi_0 + \boldsymbol{S}_n}{\kappa(n+\kappa)}\left(1+\frac{1}{n}\right), \nu_0 + n - p + 1\right) \tag{23}$$

*with*

$$\boldsymbol{\mu}_n = \frac{n\overline{\boldsymbol{Q}} + \kappa\boldsymbol{\mu}_0}{n+\kappa},$$

*where $\mathcal{T}_p(\boldsymbol{m}, \boldsymbol{A}, \nu)$ represents a multivariate Student's t distribution with mean vector $\boldsymbol{m}$, variance matrix $\boldsymbol{A}$, and $\nu$ degrees of freedom.*

iii. *The posterior distribution of $R_{\min j}$ is given by*

$$R_{\min j} \mid \boldsymbol{Q} \sim \mathcal{U}(\max(a_j, R_{\min j,\mathrm{obs}}), b_j),$$

*where $R_{\min j,\mathrm{obs}}$ is the minimum observed $R_{\min j}$.*

iv. *The posterior distribution of $R_{\max j}$ is given by*

$$R_{\max j} \mid \boldsymbol{Q} \sim \mathcal{U}(c_j, \min(d_j, R_{\max j,\mathrm{obs}})),$$

*where $R_{\max j,\mathrm{obs}}$ is the maximum observed $R_{\max j}$.*

v. *The posterior distribution of $\lambda_j$ is given by*

$$\lambda_j \mid \boldsymbol{Q} \sim \mathcal{B}(\alpha_j + n, \beta_j + n).$$

*Proof.* According to Bayes' rule, the posterior distributions are derived as follows:

i. The posterior of $\boldsymbol{\Sigma}$ is given by

$$\begin{aligned}\pi(\boldsymbol{\Sigma} \mid \boldsymbol{Q}) &\propto \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \{R_{\min j}\}, \{R_{\max j}\}, \{\lambda_j\} \mid \boldsymbol{Q})\pi(\boldsymbol{\Sigma})\\ &\propto |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{Q}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{Q}_i - \boldsymbol{\mu})\right)\\ &\quad \times |\boldsymbol{\Sigma}|^{-(\nu_0+3p+1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\Psi_0\boldsymbol{\Sigma}^{-1})\right)\\ &\propto |\boldsymbol{\Sigma}|^{-(\nu_0+n+3p+1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}\left((\Psi_0 + \boldsymbol{S}_n)\boldsymbol{\Sigma}^{-1}\right)\right),\end{aligned}$$

which is the kernel of an inverse Wishart distribution:

$$\boldsymbol{\Sigma} \mid \boldsymbol{Q} \sim \mathcal{W}^{-1}(\nu_0 + n, \boldsymbol{\Psi}_0 + \boldsymbol{S}_n).$$

ii. The posterior of $\boldsymbol{\mu}$ is given by

$$\begin{aligned}\pi(\boldsymbol{\mu} \mid \boldsymbol{Q}, \boldsymbol{\Sigma}) &\propto \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{Q})\pi(\boldsymbol{\mu} \mid \boldsymbol{\Sigma})\\ &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{Q}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{Q}_i - \boldsymbol{\mu})\right)\\ &\quad \times \exp\left(-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \left(\frac{\kappa}{\boldsymbol{\Sigma}}\right)^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right)\\ &\propto \exp\left(-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_n)^\top \left(\frac{1}{n+\kappa}\boldsymbol{\Sigma} + \frac{\kappa}{\boldsymbol{\Sigma}}\right)^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_n)\right),\end{aligned}$$

which results in a multivariate normal distribution for $\boldsymbol{\mu}$ given $\boldsymbol{\Sigma}$, i.e.,

$$\boldsymbol{\mu} \mid \boldsymbol{Q}, \boldsymbol{\Sigma} \sim \mathcal{N}_{3p}\left(\boldsymbol{\mu}_n, \frac{\boldsymbol{\Sigma}}{n+\kappa}\right).$$

By integrating out $\boldsymbol{\Sigma}$, we have

$$\boldsymbol{\mu} \mid \boldsymbol{Q} \sim \mathcal{T}_{3p}\left(\boldsymbol{\mu}_n, \frac{\Psi_0 + \boldsymbol{S}_n}{\kappa(n+\kappa)}\left(1+\frac{1}{n}\right), \nu_0 + n - p + 1\right).$$

iii. For $R_{\min j}$, combining the likelihood with the uniform prior results in the posterior distribution

$$R_{\min j} \mid \boldsymbol{Q} \sim \mathcal{U}(\max(a_j, R_{\min j,\mathrm{obs}}), b_j).$$

iv. For $R_{\max j}$, combining the likelihood with the uniform prior results in the posterior distribution

$$R_{\max j} \mid \boldsymbol{Q} \sim \mathcal{U}(c_j, \min(d_j, R_{\max j,\mathrm{obs}})).$$

v. For $\lambda_j$, combining the likelihood with the Beta prior results in the posterior distribution

$$\lambda_j \mid \boldsymbol{Q} \sim \mathcal{B}(\alpha_j + n, \beta_j + n).$$

□

Next, Theorem 3 provides the Bayes estimator of unknown parameters of a boxplot-valued random variable.

**Theorem 3** (Bayes estimators under SEL). *Under the assumptions of Lemma 1, the Bayes estimators under SEL, which are the means of the marginal posteriors, are obtained as follows:*

i. *The Bayes estimator of $\boldsymbol{\mu}$ is*

$$\hat{\boldsymbol{\mu}}_{\mathrm{Bayes}} = \frac{n\overline{\boldsymbol{Q}} + \kappa\boldsymbol{\mu}_0}{n+\kappa}. \tag{24}$$

ii. *The Bayes estimator of $\boldsymbol{\Sigma}$ is*

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{Bayes}} = \frac{\Psi_0 + \boldsymbol{S}_n}{\nu_0 + n - 3p - 1},$$

*where*

$$\boldsymbol{S}_n = \boldsymbol{S} + \frac{n\kappa}{n+\kappa}(\overline{\boldsymbol{Q}} - \boldsymbol{\mu}_0)(\overline{\boldsymbol{Q}} - \boldsymbol{\mu}_0)^\top$$

*and*

$$\boldsymbol{S} = \sum_{i=1}^{n}(\boldsymbol{Q}_i - \overline{\boldsymbol{Q}})(\boldsymbol{Q}_i - \overline{\boldsymbol{Q}})^\top, \quad \overline{\boldsymbol{Q}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Q}_i.$$

iii. *The Bayes estimators of* $R_{\min j}$ *and* $R_{\max j}$ *are*

$$\hat{R}_{\min j,\text{Bayes}} = \frac{\max(a_j, R_{\min j,\text{obs}}) + b_j}{2},$$

$$\hat{R}_{\max j,\text{Bayes}} = \frac{c_j + \min(d_j, R_{\max j,\text{obs}})}{2},$$

*where* $R_{\min j,\text{obs}}$ *and* $R_{\max j,\text{obs}}$ *are the minimum observed* $R_{\min j}$ *and maximum observed* $R_{\max j}$, *respectively.*

iv. *The Bayes estimator of* $\lambda_j$ *is*

$$\hat{\lambda}_{\text{Bayes}} = \frac{\alpha_j + n}{\alpha_j + \beta_j + 2n}.$$

*Proof.* The Bayes estimators are the means of the posterior distributions given in Lemma 1, and this completes the proof. □

### 3.2 Simulation study in the Bayesian setup

Similar to Sect. 2.2, we conduct a simulation study to demonstrate the Bayesian estimation procedure. The simulation is carried out with the same setup and true values as described there, i.e., for $p = 3$ variables with a sample size of $n = 100$ and the true parameters specified in Eqs. (15)–(17).

Here, we use noninformative priors to ensure that the data predominantly influence the parameter estimates. In this simulation, a normal prior distribution with mean vector $\mathbf{0}$ and covariance matrix $10^2 \mathbf{I}_p$ is used for $\mu$. An inverse Wishart prior distribution is employed for $\mathbf{\Sigma}$ with a scale matrix $\mathbf{S}_0 = \mathbf{I}_p$ and degrees of freedom $\nu = p+1$. For each $\lambda_j$, $j = 1, 2, 3$ a uniform $\mathcal{U}(0, 1)$ ($\mathcal{B}(\alpha_j = 1, \beta_j = 1)$) is used, which covers a broad range of values to avoid imposing any strong constraints.

The results of the Bayesian estimation, including the MSEs for each parameter, are summarized in Table 2.

In order to compare the efficiency of the proposed ML and Bayesian estimators for different sample sizes $n$, we use the same simulation settings as those in Sects. 2.2 and 3.2. Figure 1 illustrates the relative efficiency (RE), defined as the MSE of the ML estimator over the MSE of the Bayesian estimator for the parameters $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$. It can be seen that, for smaller sample sizes $n$, the efficiency of the Bayesian estimators is generally better (RE > 1). As $n$ increases, thanks to the consistency of the ML estimators, the ML estimators tend to outperform the Bayesian estimators (RE < 1).

**Table 2.** Bayesian estimates and MSEs.

| Parameters | Estimates (MSEs) |
|---|---|
| $\mu_1$ | $(4.94, 10.01, 15.12)^\top$ (0.0060) |
| $\mu_2$ | $(3.83, 7.99, 12.05)^\top$ (0.0105) |
| $\mu_3$ | $(6.15, 11.05, 16.00)^\top$ (0.0087) |
| $\Sigma_{1,1}$ | $\begin{bmatrix} 1.97 & 0.85 & 0.75 \\ 0.85 & 1.74 & 0.74 \\ 0.75 & 0.74 & 2.09 \end{bmatrix}$ (0.0421) |
| $\Sigma_{1,2}$ | $\begin{bmatrix} 1.70 & 0.96 & 0.82 \\ 0.96 & 1.30 & 0.70 \\ 0.82 & 0.70 & 1.34 \end{bmatrix}$ (0.0196) |
| $\Sigma_{1,3}$ | $\begin{bmatrix} 3.06 & 1.46 & 1.03 \\ 1.46 & 2.70 & 0.95 \\ 1.03 & 0.95 & 2.09 \end{bmatrix}$ (0.0933) |
| $\Sigma_{2,2}$ | $\begin{bmatrix} 1.04 & 0.31 & 0.28 \\ 0.31 & 1.43 & 0.36 \\ 0.28 & 0.36 & 1.05 \end{bmatrix}$ (0.0347) |
| $\Sigma_{2,3}$ | $\begin{bmatrix} 1.51 & 0.55 & 0.84 \\ 0.55 & 1.49 & 0.71 \\ 0.84 & 0.71 & 1.79 \end{bmatrix}$ (0.0561) |
| $\Sigma_{3,3}$ | $\begin{bmatrix} 1.98 & 0.89 & 0.88 \\ 0.89 & 2.02 & 0.96 \\ 0.87 & 0.96 & 1.88 \end{bmatrix}$ (0.0112) |
| $R_{\min,1}, R_{\max,1}$ | 4.889, 14.880 (0.003, 0.0192) |
| $R_{\min,2}, R_{\max,2}$ | 5.0270, 14.970 (0.012, 0.077) |
| $R_{\min,3}, R_{\max,3}$ | 5.002, 15.167 (0.0015, 0.0471) |
| $\lambda_1, \lambda_2, \lambda_3$ | 0.48, 0.48, 0.50 (NA) |

## 4 Posterior predictive density estimator

Given the posterior distributions of the parameters $\boldsymbol{\mu}$, $\mathbf{\Sigma}$, $\lambda_j$, $R_{\min j}$, and $R_{\max j}$ derived in Lemma 1, we can obtain the posterior predictive density for a new observation $\mathbf{Q}^*$. Theorem 4 finds the posterior predictive density estimator for a future or new random variable $\mathbf{Q}^*$.

**Theorem 4** (posterior predictive density estimator). *The posterior predictive density for a new random variable* $\mathbf{Q}^*$ *given the observable* $\mathbf{Q}_1, \ldots, \mathbf{Q}_n$ *is as follows:*

i. *When* $\lambda_j$ *is known, the posterior predictive density estimator is given by*

$$\mathbf{Q}^* \mid \mathbf{Q}_1, \ldots, \mathbf{Q}_n \sim \mathcal{T}_{3p}\left(\frac{\kappa\boldsymbol{\mu}_0 + n\overline{\mathbf{Q}}}{\kappa + n}, \frac{(\kappa + n + 1)(\Psi_0 + \mathbf{S}_n/\lambda_j)}{(\kappa + n)(\nu_0 + n - 3p + 1)}, \nu_0 + n - 3p + 1\right). \tag{25}$$

ii. *When* $\lambda_j$ *is unknown, the posterior predictive density is a mixture of multivariate Student's t distributions as given by*

$$p(\mathbf{Q}^* \mid \mathbf{Q}_1, \ldots, \mathbf{Q}_n) = \int \mathcal{T}_{3p}\left(\frac{\kappa\boldsymbol{\mu}_0 + n\overline{\mathbf{Q}}}{\kappa + n}, \frac{(\kappa + n + 1)(\Psi_0 + \frac{1}{\lambda_j}\mathbf{S}_n)}{(\kappa + n)(\nu_0 + n - 3p + 1)}, \nu_0 + n - 3p + 1\right) \times p(\lambda_j \mid \mathbf{Q}_1, \ldots, \mathbf{Q}_n)\,\mathrm{d}\lambda_j. \tag{26}$$

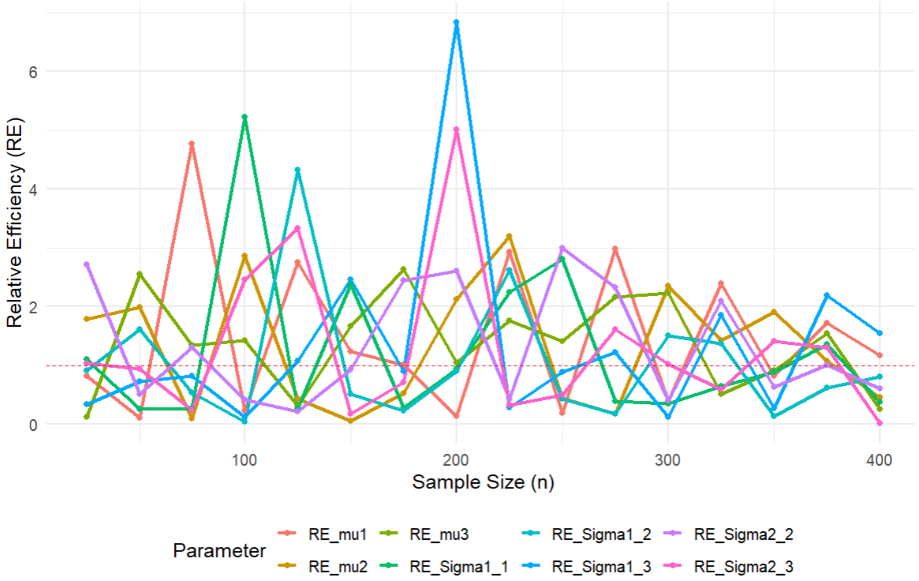

**Figure 1.** Relative efficiency of Bayesian and ML estimators in the simulation study (Sects. 2.2 and 3.2).

*Proof.* To derive the posterior predictive density, we need to integrate the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\lambda_j$ from the joint posterior distribution:

i. When $\lambda_j$ is known, from Lemma 1, the posterior distributions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given by

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \mathbf{Q} \sim \mathcal{N}_{3p}\left(\frac{\kappa\boldsymbol{\mu}_0 + n\overline{\mathbf{Q}}}{\kappa+n}, \frac{\boldsymbol{\Sigma}}{\kappa+n}\right),$$

$$\boldsymbol{\Sigma} \mid \mathbf{Q} \sim \mathcal{W}_{3p}^{-1}(\nu_0 + n, \Psi_0 + \boldsymbol{S}_n),$$

and therefore the posterior predictive density is given by

$$p(\mathbf{Q}^* \mid \mathbf{Q}_1, \ldots, \mathbf{Q}_n) = \iint p(\mathbf{Q}^* \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\, p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{Q}_1, \ldots, \mathbf{Q}_n)\, \mathrm{d}\boldsymbol{\mu}\, \mathrm{d}\boldsymbol{\Sigma}.$$

Since $p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{Q}_1, \ldots, \mathbf{Q}_n)$ follows a normal inverse Wishart distribution, the resulting predictive density is a multivariate Student's $t$ distribution:

$$\mathbf{Q}^* \mid \mathbf{Q}_1, \ldots, \mathbf{Q}_n \sim \mathcal{T}_{3p}\left(\frac{\kappa\boldsymbol{\mu}_0 + n\overline{\mathbf{Q}}}{\kappa+n}, \frac{(\kappa+n+1)(\Psi_0 + \boldsymbol{S}_n)}{(\kappa+n)(\nu_0+n-3p+1)}, \nu_0 + n - 3p + 1\right).$$

ii. When $\lambda_j$ is unknown, the posterior distributions of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\lambda_j$ are

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \lambda_j, \mathbf{Q} \sim \mathcal{N}_{3p}\left(\frac{\kappa\boldsymbol{\mu}_0 + n\overline{\mathbf{Q}}}{\kappa+n}, \frac{\boldsymbol{\Sigma}}{\lambda_j(\kappa+n)}\right),$$

$$\boldsymbol{\Sigma} \mid \lambda_j, \mathbf{Q} \sim \mathcal{W}_{3p}^{-1}\left(\nu_0 + n, \Psi_0 + \frac{1}{\lambda_j}\boldsymbol{S}_n\right),$$

$$\lambda_j \mid \mathbf{Q} \sim p(\lambda_j \mid \mathbf{Q}_1, \ldots, \mathbf{Q}_n).$$

Then, the posterior predictive density is given by

$$p(\mathbf{Q}^* \mid \mathbf{Q}_1, \ldots, \mathbf{Q}_n) = \iiint p(\mathbf{Q}^* \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \lambda_j) p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \lambda_j \mid \mathbf{Q}_1, \ldots, \mathbf{Q}_n) \cdot \mathrm{d}\boldsymbol{\mu}\, \mathrm{d}\boldsymbol{\Sigma}\, \mathrm{d}\lambda_j,$$

and this completes the proof.

□

**Remark 1.** *In addition to the posterior predictive density derived in Theorem 4, two other types of density estimators can be considered: plugin density estimators. We assume that $\boldsymbol{\lambda}$ is known or has been estimated using either Bayesian or ML methods.*

i. ***ML plugin density estimator:*** *by plugging in the ML estimators for $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\lambda}$ from Theorem 2, the density estimator is given by*

$$p(\mathbf{Q}^* \mid \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\lambda}) = \mathcal{N}_{3p}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}).$$

ii. ***Bayesian plugin density estimator:*** *by plugging in the Bayesian estimators for $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\lambda}$ from Theorem 3, the density estimator is*

$$p(\mathbf{Q}^* \mid \hat{\boldsymbol{\mu}}_{\text{Bayes}}, \hat{\boldsymbol{\Sigma}}_{\text{Bayes}}, \hat{\lambda}_{\text{Bayes}}) = \mathcal{N}_{3p}(\hat{\boldsymbol{\mu}}_{\text{Bayes}}, \hat{\boldsymbol{\Sigma}}_{\text{Bayes}}).$$

*Note that, in both cases, λ influences the range but not the normal distribution part.*

### 4.1 Comparison of density estimators using the expected Kullback–Leibler (KL) Loss

The KL loss function measures the difference between two probability distributions. For a true density $p(\mathbf{Q})$ and an estimated density $q(\mathbf{Q})$, the KL loss is defined as

$$D_{\mathrm{KL}}(p\|q) = \int p(\mathbf{Q}) \log\left(\frac{p(\mathbf{Q})}{q(\mathbf{Q})}\right) \mathrm{d}\mathbf{Q}. \tag{27}$$

The expected KL loss, also known as the KL risk function, for a density estimator $q$ is the expectation of KL loss over the distribution of the data and is given as

$$R_{\mathrm{KL}}(q) = \mathbb{E}\left[D_{\mathrm{KL}}(p\|q)\right]. \tag{28}$$

To evaluate the performance of different predictive density estimators, we compare the KL risk performance of three methods in estimating the future density, i.e., Bayesian predictive, Bayesian plugin, and ML plugin estimators. The next lemma helps to find the KL risk functions.

**Lemma 2.** *The KL loss between two multivariate normal distributions $\mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ is given by*

$$\begin{aligned}&\mathrm{KL}(\mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \,\|\, \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)) \\ &\quad = \frac{1}{2}\Bigg[\mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \\ &\quad - p + \log\frac{\det(\boldsymbol{\Sigma}_2)}{\det(\boldsymbol{\Sigma}_1)}\Bigg].\end{aligned}$$

*The KL divergence between a multivariate normal distribution and a multivariate Student's t distribution is given by*

$$\begin{aligned}&\mathrm{KL}(\mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \,\|\, \mathcal{T}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \nu)) \\ &\quad = \frac{1}{2}\Bigg[\mathrm{tr}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \\ &\quad - p + \log\frac{\det(\boldsymbol{\Sigma}_2)}{\det(\boldsymbol{\Sigma}_1)}\Bigg] - \frac{p}{2}\left(\frac{\nu - p + 1}{\nu}\right),\end{aligned}$$

*where* $\mathrm{tr}(\cdot)$ *denotes the trace of a given matrix.*

*Proof.* The KL loss function between two multivariate normal densities is straightforward and therefore omitted. Given the density function of the multivariate normal distribution as

$$p(\mathbf{Q}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_1|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{Q} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\mathbf{Q} - \boldsymbol{\mu}_1)\right)$$

and the density function of the multivariate Student's $t$ distribution as

$$\begin{aligned}q(\mathbf{Q}) = &\frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{p/2}|\boldsymbol{\Sigma}_2|^{1/2}} \\ &\cdot \left(1 + \frac{1}{\nu}(\mathbf{Q} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\mathbf{Q} - \boldsymbol{\mu}_2)\right)^{-\frac{\nu+p}{2}},\end{aligned}$$

substituting these into the KL loss function in Eq. (27) and simplifying them gives the result.

□

**Table 3.** KL risk comparison of different density estimators with different sample sizes.

| $n$ | Predictive density | Bayesian plugin density | ML plugin density |
|---|---|---|---|
| 50 | 0.3301 | 0.3010 | 0.1077 |
| 75 | 0.3138 | 0.2942 | 0.0682 |
| 100 | 0.3083 | 0.2935 | 0.0494 |
| 125 | 0.3060 | 0.2941 | 0.0391 |
| 150 | 0.3011 | 0.2912 | 0.0318 |
| 200 | 0.2981 | 0.2907 | 0.0238 |
| 250 | 0.2967 | 0.2907 | 0.0188 |
| 300 | 0.2952 | 0.2902 | 0.0151 |
| 350 | 0.2935 | 0.2893 | 0.0132 |

#### 4.1.1 KL risk function comparison simulation

We conducted a simulation study to evaluate the KL risk for different sample sizes using three methods: Bayesian predictive, Bayesian plugin, and ML plugin estimators. The true parameters and hyperparameters used in the simulation are as follows. The true parameters are defined as

$$\boldsymbol{\mu} = \begin{bmatrix}1\\2\\3\end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix}1 & 0.5 & 0.3\\0.5 & 2 & 0.4\\0.3 & 0.4 & 3\end{bmatrix},$$

and the prior parameters are defined as

$$\boldsymbol{\mu}_0 = \begin{bmatrix}0\\0\end{bmatrix}, \quad \kappa_0 = 0.0001, \quad \nu_0 = 3, \quad \boldsymbol{\Psi}_0 = \begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}.$$

As seen from Table 3 and Fig. 2, the KL risks for the Bayesian predictive and plugin density estimators are quite close. Both risk functions are fairly small, and the difference between them decreases as the sample size $n$ increases. Moreover, the KL risk for the ML plugin density estimator is smaller than that for the other two estimators, and it decreases as $n$ increases. This suggests that the ML plugin density estimator performs better when estimating the new density. However, it is important to note that both Bayesian estimators were based on noninformative priors by choosing a very small value for $\kappa_0$.

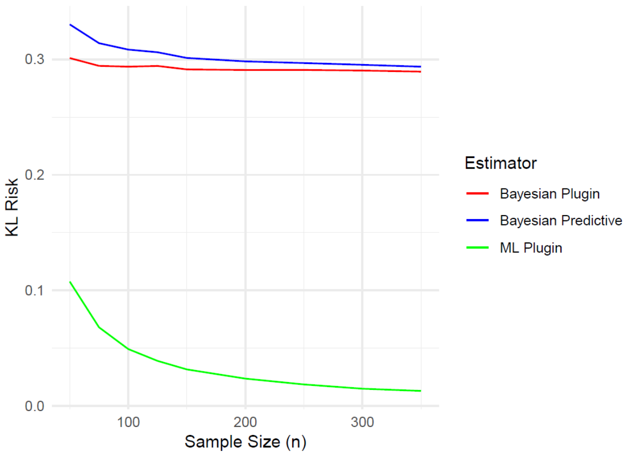

**Figure 2.** KL risk function comparison of different density estimators for different sample sizes.

To visualize the comparison of KL divergence for different sample sizes, we present a plot in Fig. 2.

## 5 Applications in climatology

In this section, we apply the methodologies discussed – specifically the point estimation methods (both ML and Bayesian) and density estimations – to real climatological data. This approach demonstrates the practical utility and effectiveness of these statistical techniques in analyzing and interpreting complex boxplot-valued environmental data. The data are sourced from the Berkeley Earth Surface Temperature Study, which aggregates 1.6 billion temperature reports from 16 pre-existing archives affiliated with the Lawrence Berkeley National Laboratory (data can be found in the Berkeley Earth Surface Temperature Study at https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data, last access: 19 July 2024).

The section is divided into two subsections, each focusing on a distinct aspect of climatological analysis. First, we examine the monthly average temperatures in the United States from January 2000 to September 2013. In the latter subsection, we analyze summer average temperatures in European countries, with data going back to the 18th century. Both datasets are vast, containing extensive data. Aggregating these data into boxplots allows for computational efficiency by representing large amounts of information in a concise form.

### 5.1 Analysis of monthly average temperatures in the United States

#### 5.1.1 Point estimators

The dataset used for this analysis includes monthly average temperatures from various cities across the United States, spanning from January 2000 to September 2013. These extensive data provide a comprehensive view of temperature trends and variations over a significant period. Here, each $X_i$ represents the monthly average temperature summaries across multiple cities for month $i$ within this period, capturing the variability of average temperatures in the form of symbolic data.

For each month, the first quartile ($q_1$), median, and third quartile ($q_3$) of the average temperatures across cities were calculated. Thus, each $Q_i = (q_{1i}, m_i, q_{3i})^\top$ represents a vector of three summary statistics for month $i$, where the dimension of the data is defined as $p = 1$, corresponding to the single temperature variable. These quartiles are then used to estimate the mean vector and covariance matrix using both ML and Bayesian methods.

The total number of monthly observations, $n = 165$, spans from January 2000 to September 2013, resulting in 165 vectors of symbolic data, each representing monthly summary statistics of average temperatures. The hyperparameters for the Bayesian estimation are as follows: $\kappa = 0.001$, $\mu_0 = (0,0,0)^\top$, $\nu_0 = 3$, $\Psi_0 = I_{3\times 3}$, $a = 0.3$, $b = 0.4$, $c = 1.8$, and $d = 2.0$.

Table 4 summarizes the results from both the ML and Bayesian estimations.

#### 5.1.2 Density estimators

In this section, we explore three methods for estimating predictive densities: posterior predictive density, the ML method, and the Bayesian plugin method. These methods provide different approaches to predicting future monthly average temperatures in the United States.

The ML and Bayesian plugin density estimators, using the ML and Bayesian estimators for $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\lambda$ from Table 4, can be obtained easily and are denoted by $p(\mathbf{Q}^* \mid \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\lambda})$ and $p(\mathbf{Q}^* \mid \hat{\boldsymbol{\mu}}_{\text{Bayes}}, \hat{\boldsymbol{\Sigma}}_{\text{Bayes}}, \hat{\lambda}_{\text{Bayes}})$, respectively. On the other hand, the posterior predictive density for a new random variable $\mathbf{Q}^*$ given the observations $\mathbf{Q}_1, \ldots, \mathbf{Q}_n$ when $\lambda$ is unknown is a mixture of multivariate Student's $t$ distributions, as given in Eq. (26). If $\lambda$ has already been estimated using ML or Bayesian methods, the posterior predictive density simplifies to a Student's $t$ distribution as shown in Eq. (25).

### 5.2 Analysis of summer average temperatures in European countries

In this subsection, we extend our analysis to examine summer average temperatures across various European countries, utilizing historical climate data that provide compre-

**Table 4.** Comparison of ML and Bayesian estimates for monthly average temperatures in the United States.

| Parameter | ML estimate | Bayesian estimate |
|---|---|---|
| $\mu$ | $(14.61265, 15.10984, 15.62950)^\top$ | $(14.61256, 15.10975, 15.62941)^\top$ |
| $\Sigma$ | $\begin{bmatrix} 57.99463 & 57.24220 & 55.76430 \\ 57.24220 & 56.55180 & 55.12211 \\ 55.76430 & 55.12211 & 53.77253 \end{bmatrix}$ | $\begin{bmatrix} 58.00203 & 57.24355 & 55.76570 \\ 57.24355 & 56.55929 & 55.12355 \\ 55.76570 & 55.12355 & 53.78012 \end{bmatrix}$ |
| $R_{\text{min}}$ | 0.3483502 | 0.3741751 |
| $R_{\text{max}}$ | 1.849506 | 1.824753 |
| $\lambda$ | 0.47 | 0.53 |

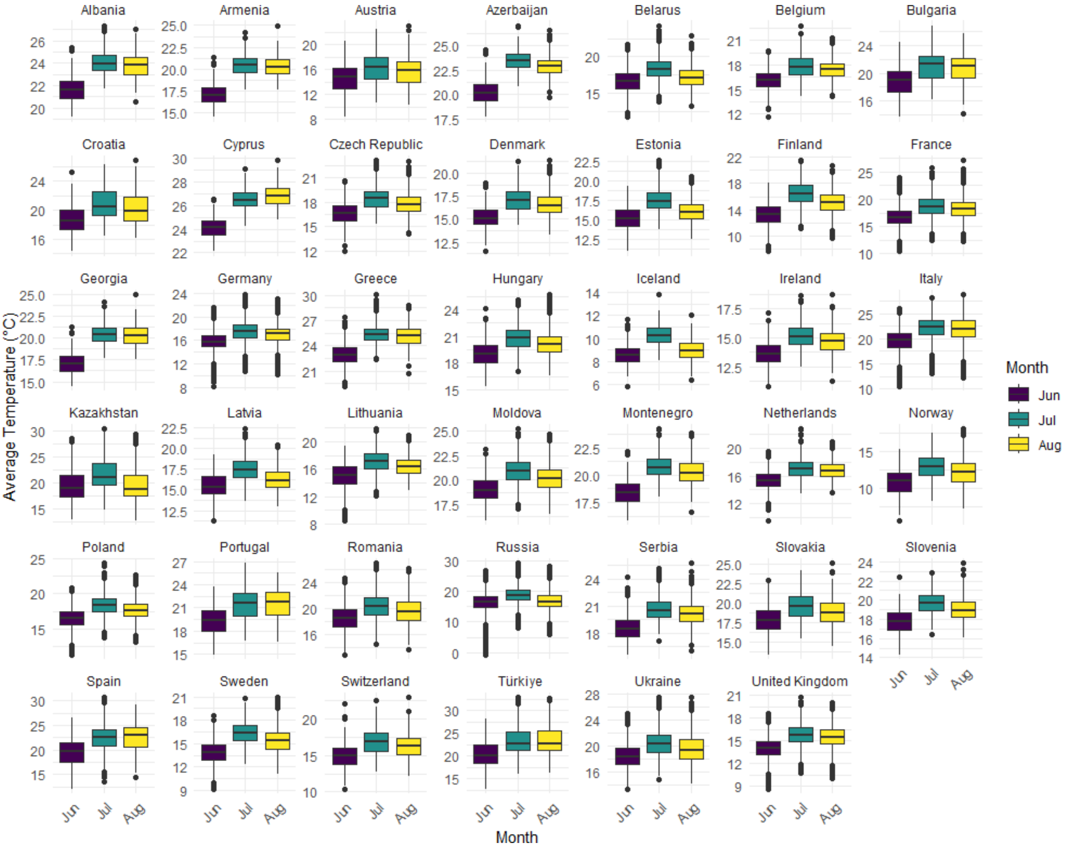

**Figure 3.** Boxplots of summer average temperatures in European countries.

hensive temperature records dating back to the 18th century. The dataset includes average monthly temperature data measured from numerous weather stations situated in major cities across Europe, allowing for an in-depth analysis of long-term climate trends and variations.

The analysis focuses on temperature records from 41 European countries, covering the three summer months June, July, and August. This extensive temporal and spatial coverage enables us to explore regional differences and similarities in temperature trends, contributing to a broader understanding of global climatic patterns.

The dataset was filtered to include only European countries, with the average temperatures for the summer months extracted for further analysis. Boxplots were then generated to visualize the distribution of average temperatures across countries during the summer season. Each boxplot represents the distribution of average temperatures for a specific country across the three summer months.

Figure 3 presents these boxplots, providing a clear illustration of the temperature variations and trends across Europe. This comparative study highlights the regional climatic pat-

**Table 5.** ML and Bayesian estimates for summer average temperatures in European countries.

| Parameters | ML estimates | Bayesian estimates |
|---|---|---|
| $\mu_1$(Jun) | $(15.9954, 17.0785, 18.2147)^\top$ | $(15.995, 17.0781, 18.2142)^\top$ |
| $\mu_2$(Jul) | $(18.1736, 19.2431, 20.3959)^\top$ | $(18.1732, 19.2426, 20.3954)^\top$ |
| $\mu_3$(Aug) | $(17.5342, 18.6084, 19.7467)^\top$ | $(17.5337, 18.6079, 19.7462)^\top$ |
| $\Sigma_{1,1}$(Jun/Jun) | $\begin{bmatrix} 10.9042 & 7.5238 & 7.3382 \\ 7.5238 & 8.0299 & 8.1470 \\ 7.3382 & 8.1470 & 12.3818 \end{bmatrix}$ | $\begin{bmatrix} 8.5485 & 8.5060 & 8.5735 \\ 8.506 & 8.6527 & 8.8184 \\ 8.5735 & 8.8184 & 9.2067 \end{bmatrix}$ |
| $\Sigma_{1,2}$(Jun/Jul) | $\begin{bmatrix} 10.2312 & 8.3206 & 7.3775 \\ 8.3901 & 10.2309 & 9.4302 \\ 7.6828 & 9.6218 & 13.0352 \end{bmatrix}$ | $\begin{bmatrix} 7.7133 & 9.8158 & 9.2316 \\ 7.0172 & 9.0069 & 8.7039 \\ 9.7522 & 8.1082 & 10.2076 \end{bmatrix}$ |
| $\Sigma_{1,3}$(Jun/Aug) | $\begin{bmatrix} 10.6484 & 8.8207 & 8.0046 \\ 9.0129 & 10.9896 & 10.3168 \\ 8.2967 & 10.3747 & 13.8746 \end{bmatrix}$ | $\begin{bmatrix} 9.7982 & 7.9931 & 7.4551 \\ 8.8901 & 9.4564 & 10.2118 \\ 8.1022 & 9.0435 & 11.2543 \end{bmatrix}$ |
| $\Sigma_{2,2}$(Jul/Jul) | $\begin{bmatrix} 10.9766 & 9.0864 & 8.2593 \\ 9.0864 & 10.9044 & 10.2101 \\ 8.2593 & 10.2101 & 13.7735 \end{bmatrix}$ | $\begin{bmatrix} 9.8133 & 9.7178 & 9.7322 \\ 9.7178 & 9.8069 & 9.9089 \\ 9.7322 & 9.9089 & 10.2876 \end{bmatrix}$ |
| $\Sigma_{2,3}$(Jul/Aug) | $\begin{bmatrix} 12.0938 & 10.5636 & 9.8549 \\ 10.5636 & 12.844 & 12.2766 \\ 9.8549 & 12.276 & 12.8815 \end{bmatrix}$ | $\begin{bmatrix} 11.6531 & 10.1128 & 9.7282 \\ 9.7776 & 10.8769 & 11.0019 \\ 9.4322 & 10.9889 & 11.7871 \end{bmatrix}$ |
| $\Sigma_{3,3}$(Aug/Aug) | $\begin{bmatrix} 11.324 & 11.4021 & 11.881 \\ 11.5691 & 11.9810 & 11.5987 \\ 11.543 & 11.9021 & 12.1243 \end{bmatrix}$ | $\begin{bmatrix} 11.2529 & 11.3064 & 11.3472 \\ 11.3064 & 11.5691 & 11.7098 \\ 11.3472 & 11.7098 & 12.1139 \end{bmatrix}$ |
| $R_{\min}$, $R_{\max}$ (Jun) | 1.5865,8.883 | 1.047, 8.545 |
| $R_{\min}$, $R_{\max}$ (Jul) | 1.4015, 9.380 | 1.035, 9.682 |
| $R_{\min}$, $R_{\max}$ (Aug) | 1.4015, 9.380 | 1.411, 8.998 |
| λ(Jun, Jul, Aug) | 0.47, 0.48, 0.49 | 0.5, 0.5, 0.49 |

terns, offering valuable insights into the broader context of global climate change.

To further understand the differences between the ML and Bayesian approaches in estimating summer average temperatures in European countries, we present a comparative analysis of the estimators for the mean vectors and covariance matrices across the three summer months: June, July, and August. Table 5 summarizes the results.

The table above presents a side-by-side comparison of the ML and Bayesian estimators for the mean vectors and covariance matrices of summer average temperatures in European countries. The results show a close alignment between the ML and Bayesian estimates for the mean vectors across the three summer months. However, the covariance matrices exhibit some differences, particularly in the off-diagonal elements, which suggest variations in how each method captures the relationships between temperature readings across different months.

For instance, the Bayesian estimators tend to produce more consistent estimates across the covariance matrices, potentially reflecting the smoothing effect of prior information. The parameters $R_{\min}$ and $R_{\max}$ for each month also indicate that the range of temperature variations is slightly narrower in the Bayesian framework, further highlighting the influence of the priors.

## 6 Discussion

In this article, we have introduced a novel method for estimating the parameters of boxplot-valued random variables from both Bayesian and frequentist perspectives. Working with boxplot-valued random variables enables us to efficiently summarize large datasets, as boxplots concisely represent extensive amounts of information. Our method offers a significant computational advantage, as we derived closed-form solutions for both ML and Bayesian estimators. These closed-form expressions allow for direct computation of parameter estimates without the need for iterative optimization procedures or Markov chain Monte Carlo (MCMC)-type methods, which are commonly required in traditional Bayesian approaches. By avoiding these computationally intensive steps, our method reduces the overall complexity and computational load, making it particularly suitable for large datasets.

Additionally, we have explored the density estimation of future random boxplots, providing a comprehensive tool for investigating and forecasting these distributions. Our approach was evaluated through simulations and demonstrated its accuracy and computational efficiency when compared with existing methods. Furthermore, we applied the proposed techniques to real environmental data to illustrate their practical utility and effectiveness in real-world scenarios.

Boxplot-valued data analysis is particularly useful in contexts where data are collected in large volumes and need to be summarized effectively without losing significant information. By adopting both Bayesian and frequentist methods, we provide a comprehensive framework that leverages the strengths of each approach. The Bayesian methods offer a probabilistic interpretation and incorporate prior information, which is valuable when data are scarce or expensive to obtain. On the other hand, the frequentist methods provide consistency and robustness when large amounts of data are available.

Furthermore, our study on the density estimation of future random boxplots enhances predictive modeling capabilities, which is crucial for applications such as environmental monitoring and climate forecasting. By accurately estimating the distribution of future data, decision-makers can make more informed predictions and plans.

Through extensive simulations, we validated the performance of our proposed methods, showcasing their accuracy and reliability. The application to real environmental data further underscores the practical relevance and adaptability of our techniques. Moreover, although the multivariate normal model theoretically allows for the possibility of overlapping quartiles, our empirical studies – including both simulations and real-world applications – consistently resulted in ordered quartiles ($q_{1i} \leq m_i \leq q_{3i}$). This empirical consistency indicates that, within the context of our data and analysis, the model effectively preserves the natural ordering of quartiles, thereby mitigating the theoretical limitation in practical scenarios. Overall, this work contributes to the growing field of symbolic data analysis by offering efficient and effective tools for handling boxplot-valued data, thereby broadening the scope and applicability of statistical methodologies in various domains.

**Code availability.** The R code used for the analysis in this study is not publicly available but can be shared upon reasonable request. Due to institutional policies and ongoing related research, we have opted not to deposit the code in a public repository at this time. However, researchers interested in accessing the code may contact the corresponding author at asadeghkhani@ncat.edu.

**Data availability.** The data used in this study are publicly available and can be accessed from the Kaggle repository at https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data (Berkeley Earth, 2017). The dataset was obtained from this source and used in accordance with its terms of use. Researchers interested in the dataset can directly access it via the provided link.

**Author contributions.** AbS and AlS contributed equally to the development of the statistical methodology, study design, and data analysis. AbS prepared the initial manuscript draft, and both authors participated in its review and revision.

**Competing interests.** The contact author has declared that neither of the authors has any competing interests.

**Disclaimer.** Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

**Acknowledgements.** The authors sincerely thank the editors and the two anonymous reviewers for their constructive feedback, which significantly improved the quality of this paper.

**Review statement.** This paper was edited by Reetam Majumder and reviewed by two anonymous referees.

## References

Arroyo, J., Maté, C., and Roque, A. M.-S.: Hierarchical clustering for boxplot variables, in: Data Science and Classification, edited by: Batagelj, V., Bock, H.-H., Ferligoj, A., and Žiberna, A., Springer, 59–66, https://doi.org/10.1007/3-540-34416-0_7, 2006.

Benjamini, Y.: Opening the box of a boxplot, Am. Stat., 42, 257–262, 1988.

Berkeley Earth: Climate Change: Earth Surface Temperature Data, Kaggle [data set], https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data (last access: 19 July 2024), 2017.

Billard, L. and Diday, E.: Regression analysis for interval-valued data, in: Data analysis, classification, and related methods, edited by: Kiers, H. A. L., Rasson, J. P., Groenen, P. J. F., and Schader, M., Springer, 369–374, https://doi.org/10.1007/978-3-642-59789-3_58, 2000.

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

Chambers, J. M.: Graphical methods for data analysis, Chapman and Hall/CRC, https://doi.org/10.1201/9781351072304, 2018.

Diday, E.: The symbolic approach in clustering and related methods of data analysis: the basic choices, in: Classification and Related Methods of Data Analysis, Proceedings of the First Conference of the International Federation of Classification Societies (IFCS-87: Technical University of Aachen, 673–684, North Holland, NII Article ID 10011477669, 1988.

Diday, E.: Probabilist, possibilist and belief objects for knowledge analysis, Ann. Oper. Res., 55, 225–276, 1995.

Diday, E. and Noirhomme-Fraiture, M.: Symbolic data analysis and the SODAS software, John Wiley & Sons, ISBN 978-0470018835, 2008.

Douzal-Chouakria, A., Billard, L., and Diday, E.: Principal component analysis for interval-valued observations, Stat. Anal. Data Min., 4, 229–246, 2011.

Émilion, R.: Différentiation des capacités et des intégrales de Choquet, C.R. Acad. Sci. I-Math., 324, 389–392, 1997.

Ferguson, T. S.: A course in large sample theory, Routledge, ISBN 0412043718 2017.

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

Neto, E. d. A. L. and De Carvalho, F. D. A.: Centre and range method for fitting a linear regression model to symbolic interval data, Comput. Stat. Data An., 52, 1500–1515, 2008.

Neto, E. d. A. L. and De Carvalho, F. D. A.: Constrained linear regression models for symbolic interval-valued variables, Comput. Stat. Data An., 54, 333–347, 2010.

Reyes, D. M., de Souza, R. M., and de Oliveira, A. L.: A three-stage approach for modeling multiple time series applied to symbolic quartile data, Expert Syst. Appl., 187, 115884, https://doi.org/10.1016/j.eswa.2021.115884, 2022.

Reyes, D. M., Souza, L. C., de Souza, R. M., and de Oliveira, A. L.: Parametrized linear regression for boxplot-multivalued data applied to the Brazilian Electric Sector, Inform. Sci., 652, 119758, https://doi.org/10.1016/j.ins.2023.119758, 2024.

Sadeghkhani, A. and Sadeghkhani, A.: Multivariate Interval-Valued Models in Frequentist and Bayesian Schemes, arXiv [preprint], https://doi.org/10.48550/arXiv.2405.06635, 2024.

Samadi, S. Y., Billard, L., Guo, J. H., and Xu, W.: MLE for the parameters of bivariate interval-valued model, Adv. Data Anal. Classif., 18, 827–850, https://doi.org/10.1007/s11634-023-00546-6, 2024.

Tukey, J.: Exploratory data analysis, Springer, https://doi.org/10.1007/978-3-031-20719-8_2, 1977.

Wickham, H. and Wickham, H.: Getting Started with ggplot2, ggplot2: Elegant graphics for data analysis, Springer, 11–31, https://doi.org/10.1007/978-3-319-24277-4_2, 2016.

Xiong, T., Li, C., Bao, Y., Hu, Z., and Zhang, L.: A combination method for interval forecasting of agricultural commodity futures prices, Knowl.-Based Syst., 77, 92–102, 2015.

Xu, M. and Qin, Z.: A bivariate Bayesian method for interval-valued regression models, Knowl.-Based Syst., 235, 107396, https://doi.org/10.1016/j.knosys.2021.107396, 2022.

Zhao, Q., Wang, H., and Wang, S.: Robust regression for interval-valued data based on midpoints and log-ranges, Adv. Data Anal. Classi., 17, 583–621, 2023.