Location and association measures for interval data based on Mallows' distance

M. Rosário Oliveira[1*], Diogo Pinheiro[2] and Lina Oliveira[3]

[1]CEMAT and Department of Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisbon, 1049-001, Portugal.
[2]Department of Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisbon, 1049-001, Portugal.
[3]CAMGSD and Department of Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisbon, 1049-001, Portugal.

*Corresponding author(s). E-mail(s): rosario.oliveira@tecnico.ulisboa.pt;
Contributing authors: diogo.pinheiro.99@tecnico.ulisboa.pt; lina.oliveira@tecnico.ulisboa.pt;

Abstract

The increasing need to analyse large volumes of data has led to the development of Symbolic Data Analysis as a promising field to tackle the data challenges of our time. New data types, such as interval-valued data, have brought fresh theoretical and methodological problems to be solved. In this paper, we derive explicit formulas for computing the Mallows' distance, also known as $\boldsymbol{L_2}$ Wasserstein distance, between two p-dimensional intervals, using information regarding the distribution of the microdata. We establish this distance as a Mahalanobis' distance between two $2p$-dimensional vectors. Our comprehensive analysis leads to the generalisation of the definitions of the expected value and covariance matrix of an interval-valued random vector. These novel results bring theoretical support and interpretability to state-of-the-art contributions. Additionally, we discuss real examples that illustrate how we can model different levels of available information on the microdata, leading to proper estimates of the measures of location and association.

Keywords: Symbolic Data Analysis; Interval Data; $\boldsymbol{L_2}$ Wasserstein Distance; Barycentre; Symbolic Covariance

1 Introduction

The explosion of data volume has motivated the appearance of new data types and the need for more complex statistical techniques to address them. Symbolic Data Analysis (SDA) is a field of Statistics that studies data with internal variation, of which histograms and intervals are two of its most important examples. At its core, it builds on statistical methods (exploratory and inferential) to learn patterns from individual observations, said microdata, based on aggregate observations, the so-called macrodata. The data aggregation may be due to sample size issues, privacy reasons, a result of the researchers' interest, or just a natural outcome of the data recording. Another source of symbolic observations is the elicitation of experts' prior knowledge about quantities of interest, as in Bayesian Statistics (see [1] for details). For a thorough review of symbolic data types and their analysis see [2–4].

SDA has been mainly approached from a sampling perspective and its techniques are mostly descriptive (see [1] and the arguments presented therein). The works [3, 5, 6] introduced measures of location, dispersion, and association between symbolic random variables, formalised as a function of the observed macrodata values and implicit assumptions about the microdata, capturing their inherent characteristics. An example of this principle is the initial proposal of Bertrand and Goupil [5] of the sample mean and sample variance of a set of interval-valued observations as the sample mean and sample variance of the respective centres.

In [7] this approach was called "*SDA two-level paradigm*". The authors proposed an alternative where the location measure was the Fréchet mean, also called barycentre, of the set of interval-valued observations. They considered the space of real bounded intervals and the L_2 Wasserstein distance, also known as the Mallows' distance, a denomination to be used in the rest of this paper, based on the assumption that the microdata spread in each observed interval according to a uniform distribution. Under this approach, the location measure is an interval, by contrast with the previous definition of this measure as a real number, and the variance is a non-negative real number, as usual. In this paper, we generalise the barycentre approach to the population framework, admitting any possible absolutely continuous distribution with finite second moment for the microdata.

The sample covariance and sample correlation matrices were also addressed in the context of symbolic principal component analysis in [8–10]. Specifically, in [10] the authors established relationships between several proposed methods of symbolic principal component analysis and available definitions of sample symbolic variance and covariance. Later, in [11] the principal components were derived as the linear combinations of the original interval-valued random variables which maximised the symbolic variance.

Other areas of Statistics have also been addressed by SDA, like clustering (see, e.g., [2, 12, 13]), discriminant analysis (see, e.g., [14–16]), regression analysis (see, e.g., [17–20]), time series (see, e.g., [21–23]), Bayesian hierarchical modelling (see, e.g., [24]), and network sciences (see, e.g., [25, 26]), amongst others.

Parametric approaches for interval-valued variables have also been considered. In [27], the authors derived maximum likelihood estimators for the mean and the variance of interval-valued and histogram-valued variables, assuming that the microdata follow uniform or symmetric triangular distributions. In the follow-up paper [28], the authors revised the initial work and derived the maximum likelihood estimators for all important covariance statistics. In [17], interval-valued variables were formulated as bivariate random vectors to introduce a symbolic regression model based on the theory of generalised linear models. The contributions in [14, 29, 30] followed a different approach. In that line of work, the centres and logarithms of the ranges were collected in a random vector with a multivariate normal or skew-normal distribution, which was used to derive methods for the analysis of variance [29], discriminant analysis [14], and outlier detection [30] of interval-valued variables. More recently, a line of research was developed using likelihood-based methods that fitted models for the microdata when only the macrodata were observed [31, 32].

In this paper, we consider the interval data model establishing the link between macrodata and microdata, introduced in [33]. The model and the needed background are presented in Section 2. In Section 3 we derive general formulations of the Mallows' distance between two p-dimensional intervals as a function of their centres, ranges, and the first two moments of the associated microdata distributions. We also show that the Mallows' distance between two p-dimensional intervals can be seen as a special case of a Mahalanobis' distance (see Section 3.2). In Section 4 these results provide support for the derivation of new contributions related to the multivariate definitions of expected value and covariance matrices based on the barycentre approach. The derived results are illustrated in Section 5 using three examples where different levels of information about the microdata are available. We discuss the choices of distributions and their parameters to model real data. Finally, in Section 6 we present the main conclusions. Appendix A and Appendix B contain the proofs of the results in Section 3 and Section 4, respectively.

The R code related to this work can be found in https://github.com/MROS13/MallowSymbCov.

2 Preliminaries

A prime example of symbolic data is interval-valued data. Consider an athlete training for the 800 metre race. On each attempt, the time taken to complete the distance is recorded. Since the athlete wants to improve his performance, he trains every week, several days a week. To see if there is any improvement, the times are compared between weeks. With conventional methods it is common to do this by comparing one or more summary statistics, such as the mean and the standard deviation, to name a few, which summarise the athlete's weekly race times. While this is practical, it does not take into account how the times are distributed. By aggregating the weekly times into an interval whose endpoints are the lowest and highest times, we are able to preserve the aggregated information. In this way, the athlete's weekly race times are represented by an interval with intrinsic variability described by the distribution of the individual times. The interval as a set of points between two real numbers is known as macrodata, and the individual points are known as the microdata, that will be characterised by a certain distribution. It should be noted that we are interested in analysing intervals as symbolic objects to which we can attach the distribution of the microdata and have more information about the data.

At the level of the macrodata, we can define interval-valued random variables together with the extension to this setting of classical notions of conventional data analysis, such as expected value, variance, and covariance. We address next the set on which the interval-valued random variables are defined.

Definition 2.1. *Let* $\mathbb{IR} = \{[a, b] \colon a, b \in \mathbb{R},\, a \leq b\}$ *be the set of all real closed bounded intervals. For a positive integer* p*, let* $\mathbb{IR}^p$ *be the cartesian product of* p *copies of* $\mathbb{IR}$*, that is,*

$$\mathbb{IR}^p = \{([a_1, b_1], \ldots, [a_p, b_p])^T : a_i, b_i \in \mathbb{R},\, a_i \leq b_i,\, i = 1, \ldots, p\}.$$

Here $\mathbb{IR}^1$ *is denoted by* $\mathbb{IR}$*.*

Given the interval $[a, b] \in \mathbb{IR}$ (the macrodata), consider its *centre* $c = (a + b)/2$ and *range* $r = b - a$. It follows that there exists a correspondence between $\mathbb{IR}$ and $\mathbb{R} \times \mathbb{R}_0^+$ through the injective mapping which sends the interval $[a, b]$ to $(c, r)^T \in \mathbb{R} \times \mathbb{R}_0^+$, where c can be any real number and r must be non-negative. Considering this, we can extend this representation of intervals to $\mathbb{IR}^p$. For a hyperrectangle $\boldsymbol{x} = ([a_1, b_1], \ldots, [a_p, b_p])^T$ in $\mathbb{IR}^p$, $p \in \mathbb{N}$, let $(\boldsymbol{c}^T, \boldsymbol{r}^T)^T$ be the corresponding vector in $\mathbb{R}^p \times \mathbb{R}^p$ of centres and ranges, that is,

$$c(\boldsymbol{x}) = \boldsymbol{c} = \left(\frac{a_1 + b_1}{2}, \ldots, \frac{a_p + b_p}{2}\right)^T \quad \text{and} \quad r(\boldsymbol{x}) = \boldsymbol{r} = (b_1 - a_1, \ldots, b_p - a_p)^T.$$

Hence, we have an injective mapping from $\mathbb{IR}^p$ to $\mathbb{R}^{2p}$, which allows for writing, by a slight abuse of notation, that

$$\boldsymbol{x} = ([a_1, b_1], \ldots, [a_p, b_p])^T = (\boldsymbol{c}^T, \boldsymbol{r}^T)^T.$$

Notice that we are identifying $\mathbb{R}^{2p}$ with $\mathbb{R}^p \times \mathbb{R}^p$. Note also that the codomain of this mapping is $\mathbb{R}^p \times (\mathbb{R}_0^+)^p$, since the ranges cannot be negative.

We are now ready to define random variables whose realisations lie in $\mathbb{IR}^p$.

Definition 2.2. *For* $p \in \mathbb{N}$ *and the (real-valued) random vectors* $\boldsymbol{A} = (A_1, \ldots, A_p)^T$ *and* $\boldsymbol{B} = (B_1, \ldots, B_p)^T$*, with* $\mathrm{P}(A_i \leq B_i) = 1$*,* $i = 1, \ldots, p$*, let* $X_i = [A_i, B_i]$ *be an* interval-valued random variable *with realisations in* $\mathbb{IR}$ *and* $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ *be an* interval-valued random vector *with realisations in* $\mathbb{IR}^p$*. Alternatively,* $\boldsymbol{X}$ *can also be represented by its random vectors of centres and ranges,*

$$c(\boldsymbol{X}) = \boldsymbol{C} = (C_1, \ldots, C_p)^T \quad \text{and} \quad r(\boldsymbol{X}) = \boldsymbol{R} = (R_1, \ldots, R_p)^T,$$

respectively, where $C_i = (A_i + B_i)/2$ *and* $R_i = B_i - A_i$*,* $i = 1, \ldots, p$*.*
Conventional random vectors are obtained as a particular case by setting $\mathrm{P}(R_i = 0) = 1$*,* $i = 1, \ldots, p$*.*

In SDA, macrodata can be seen as the manifest variable that gives information about the behaviour of the microdata, which may not be observed and, in that case, can be understood as the realisations of a latent random variable. A model that establishes a natural link between macrodata and microdata was proposed by Oliveira and co-authors in [33]. It was suggested there to unify and add interpretability to a group of definitions of sample interval-valued covariance matrices available in the literature. Population counterparts of location, scale, and association were also proposed. The model, which has proved its relevance in other areas of SDA [11, 34], is introduced in the next definition. It is important to distinguish between the cases where the random variables of the ranges take value 0 with the probabilities 0, 1, and any probability in between. The latter case is not discussed in this work.

Definition 2.3. *Let* $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ *be an interval-valued random vector with realisations in* $\mathbb{IR}^p$*, and let* $\boldsymbol{C} = (C_1, \ldots, C_p)^T$ *and* $\boldsymbol{R} = (R_1, \ldots, R_p)^T$ *be the corresponding random vectors of centres and ranges, respectively. If* $\mathrm{P}(R_i = 0) = 0$*, then the real-valued random vector* $\boldsymbol{V} = (V_1, \ldots, V_p)^T \in \mathbb{R}^p$ *describing the microdata in* $\boldsymbol{X}$ *is defined by*

$$V_i = C_i + U_i \frac{R_i}{2},\ i = 1, \ldots, p, \tag{2.1}$$

where the weights U_i *are absolutely continuous latent random variables with support* $[-1, 1]$*.*
If $\mathrm{P}(R_i = 0) = 1$*, then* $V_i = X_i = C_i$ *and set* $\mathrm{P}(U_i = 0) = 1$*.*

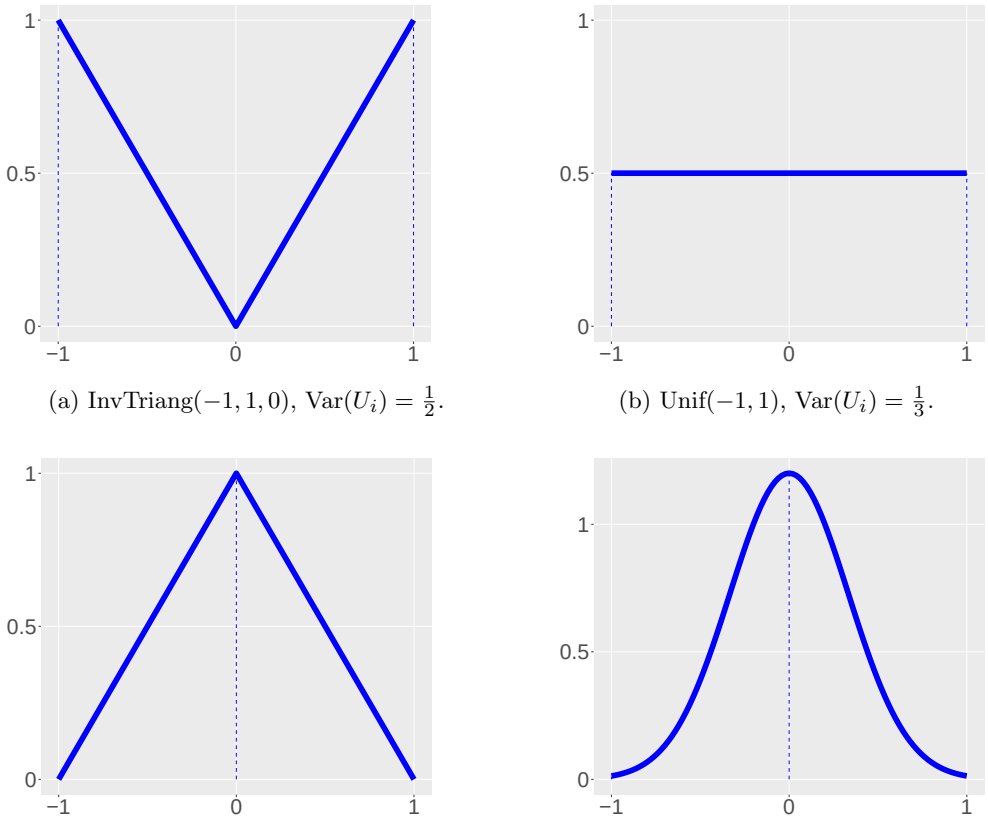
Remark 2.4. *Note that we can eliminate the contribution of the ranges in* (2.1) *by considering* $\mathrm{P}(R_i = 0) = 1$*. For coherence, we impose that in this case the weights are also equal to* 0 *with probability* 1*.*

Observe that a realisation of $\boldsymbol{V}$ is a point in the hyperrectangle related to the interval-valued random vector $\boldsymbol{X}$, characterised by its centre $\boldsymbol{C}$ and range $\boldsymbol{R}$. According to this model, the microdata for a specific hyperrectangle, say $\boldsymbol{x} = (\boldsymbol{c}^T, \boldsymbol{r}^T)^T$, are described by the random vector $\tilde{\boldsymbol{V}} = (\tilde{V}_1, \ldots, \tilde{V}_p)^T$, where $\tilde{V}_i = c_i + \tilde{U}_i r_i/2$ is the random variable V_i conditioned on realisations of C_i and R_i. Likewise, $\tilde{U}_i = U_i | (C_i = c_i, R_i = r_i)$ is the corresponding latent random variable.

This work addresses explicitly the case of non-degenerate interval-valued data, in the sense that $\mathrm{P}(R_i = 0) = 0$, $i = 1, \ldots, p$. However, all the conclusions can be extended to the case where $k \leq p$ components of $\boldsymbol{X}$ are real-valued random variables. Without loss of generality, we can assume that these real-valued random variables are the first k components of $\boldsymbol{X}$, i.e., $\mathrm{P}(R_i = 0) = 1$, $i = 1, \ldots, k$. The main results under this assumption are outlined as remarks.

Often, the assumptions about U_i are based on the knowledge of the real problem in hand and, at most, on goodness of fit measures of the methods that rely on this formulation. However, a couple of appealing assumptions could be considered to simplify the proposed model. Namely, $\boldsymbol{U} = (U_1, \ldots, U_p)^T$ has a zero mean and is independent of the random vector $(\boldsymbol{C}^T, \boldsymbol{R}^T)^T$. This latter assumption has an important contribution to the definition of the covariance matrix of an interval-valued random vector explored in this work.

In the literature, the most common distribution assumption about the microdata is that they follow a continuous uniform distribution. However, in [33] and [35] other symmetric alternatives were discussed. Figure 1 illustrates some of those distributions, organised according to the variance of U_i, from highest to lowest. In this order, we present the symmetric inverted triangular distribution $\mathrm{InvTriang}(-1, 1, 0)$ with variance equal to $1/2$, the continuous uniform distribution $\mathrm{Unif}(-1, 1)$ with variance equal to $1/3$, the symmetric triangular distribution $\mathrm{Triang}(-1, 1, 0)$ with variance equal to $1/6$, and the truncated normal distribution $\mathcal{N}(0, 1/9)|[-1, 1]$ with variance equal to $1/9 - 2\phi(3)/(6\Phi(3) - 3) \simeq 1/9$, where $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and distribution function of a standard normal distribution, respectively.



(a) $\mathrm{InvTriang}(-1, 1, 0)$, $\mathrm{Var}(U_i) = \frac{1}{2}$.

(b) $\mathrm{Unif}(-1, 1)$, $\mathrm{Var}(U_i) = \frac{1}{3}$.

(c) $\mathrm{Triang}(-1, 1, 0)$, $\mathrm{Var}(U_i) = \frac{1}{6}$.

(d) $\mathcal{N}\left(0, \frac{1}{9}\right)|[-1, 1]$, $\mathrm{Var}(U_i) \simeq \frac{1}{9}$.

Fig. 1: Examples of density functions of continuous symmetric distributions of U_i (see Definition 2.3).

3 Mallows' distance

The Mallows' distance has been widely used in SDA to compare two intervals and plays an important role in this area (see, for example, [4]). One of the reasons for this is that the Mallows' distance may be seen as a generalisation of the Euclidean distance in $\mathbb{R}^2$ and shares many of its properties. We devote the next two sections to extending and discussing relevant results already present in the literature.

3.1 Mallows' distance in $\mathbb{IR}$

We begin by defining the Mallows' distance between two univariate intervals.

Definition 3.1. *Let x_1 and x_2 be intervals in $\mathbb{IR}$. Let F_j and F_j^{-1} be, respectively, the distribution function (assumed to have finite second moment) and quantile function related with the microdata in $x_j = [a_j, b_j]$, with centre $c_j = (a_j - b_j)/2$ and range $r_j = b_j - a_j$, $j = 1, 2$. The Mallows' distance $d_M(x_1, x_2)$ between x_1 and x_2 is defined by*

$$d_M(x_1, x_2) = \left(\int_0^1 \left(F_1^{-1}(t) - F_2^{-1}(t) \right)^2 dt \right)^{1/2}. \tag{3.1}$$

Note that (3.1) is a distance between the quantile functions of the microdata within x_1 and x_2. In this sense, two intervals are at Mallows' distance 0 if and only if they have the same quantile function. In fact, using the model described in Definition 2.3, we can show that the quantile function of the microdata is a transformation of the quantile function of the latent random variable. Firstly, we prove the following auxiliary proposition.

Proposition 3.2. *Let W be a real-valued random variable, let $Z = a + b\ W$, with $a \in \mathbb{R}$, $b \in \mathbb{R}_0^+$, and let F_W and F_W^{-1} be the distribution function and quantile function of W, respectively. Then, the quantile function of Z is*

$$F_Z^{-1}(t) = a + b\ F_W^{-1}(t),\ t \in (0, 1].$$

Proof. If $b = 0$, then Proposition 3.2 holds trivially, since, for all $t \in (0, 1]$, we have $\mathrm{P}(Z = a) = 1$ and $F_Z^{-1}(t) = a$. Suppose now that $b > 0$. For $z \in \mathbb{R}$,

$$F_Z(z) = \mathrm{P}(Z \leq z) = \mathrm{P}\Big(a + b\ W \leq z\Big) = \mathrm{P}\Big(W \leq \frac{z-a}{b}\Big) = F_W\Big(\frac{z-a}{b}\Big)\ .$$

By the definition of $F_Z^{-1}(t)$, it follows that

$$F_Z^{-1}(t) = \inf\{z \in \mathbb{R} :\ t \leq F_Z(z)\} = \inf\Big\{z \in \mathbb{R} :\ t \leq F_W\Big(\frac{z-a}{b}\Big)\Big\} = \inf\Big\{a + b\ u \in \mathbb{R} :\ t \leq F_W(u)\Big\}.$$

Since $b > 0$, we obtain

$$F_Z^{-1}(t) = a + b\ \inf\{u \in \mathbb{R} :\ t \leq F_W(u)\} = a + b\ F_W^{-1}(t),$$

as required. □

Let $\tilde{V}_j = V_j | \, (C_j = c_j, R_j = r_j) = c_j + r_j\ \tilde{U}_j/2$ be the random variable describing the microdata in the interval $x_j = (c_j, r_j)^T$, where $\tilde{U}_j = U_j | \, (C_j = c_j, R_j = r_j)$, $j = 1, 2$. Since, for all j, $r_j \geq 0$, we can use Proposition 3.2 to show that

$$F_j^{-1}(t) = c_j + \frac{r_j}{2}\ F_{\tilde{U}_j}^{-1}(t),\ j = 1, 2, \tag{3.2}$$

where F_j^{-1} and $F_{\tilde{U}_j}^{-1}$ are the quantile functions of $\tilde{V}_j$ and $\tilde{U}_j$, respectively. According to (3.2), the microdata in the interval $x_j = (c_j, r_j)^T$ are identified by the centre c_j, the range r_j, and the distribution function of the latent random variable $\tilde{U}_j$. Therefore, we can introduce a more appropriate notation and state that $x_j = \Big((c_j, r_j)^T, F_{\tilde{U}_j}\Big)$, where $(c_j, r_j)^T$ refers to the macrodata and $F_{\tilde{U}_j}$ is the distribution function of the latent random variable $\tilde{U}_j$ that describes the microdata within the macrodata.

Using (3.2), we can compute the Mallows' distance between x_1 and x_2 as a function of the centres and the ranges of the two intervals and of their respective latent random variables $\tilde{U}_1$ and $\tilde{U}_2$.

Theorem 3.3. *Let x_1, x_2 be intervals such that $x_j = \Big((c_j, r_j)^T, F_{\tilde{U}_j}\Big)$, $j = 1, 2$, where $\tilde{U}_j$ is an absolutely continuous random variable with support $[-1, 1]$, assumed to have finite second moment. Then, the square of the Mallows' distance between x_1 and x_2 is*

$$\begin{aligned} d_M(x_1, x_2)^2 &= (c_1 - c_2)^2 + 2(c_1 - c_2)\left(\frac{r_1}{2}\,\mathrm{E}\,(\tilde{U}_1) - \frac{r_2}{2}\,\mathrm{E}\,(\tilde{U}_2)\right) \\ &+ \frac{r_1^2}{4}\,\mathrm{E}\,(\tilde{U}_1^2) + \frac{r_2^2}{4}\,\mathrm{E}\,(\tilde{U}_2^2) - \frac{r_1 r_2}{2}\,\mathcal{E}(\tilde{U}_1, \tilde{U}_2), \qquad (3.3) \\ &= (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2 + 2\sigma_1\sigma_2(1 - \rho_{12}), \qquad (3.4) \end{aligned}$$

where $\mathcal{E}(\tilde{U}_1,\tilde{U}_2)=\int_0^1 F^{-1}_{\tilde{U}_1}(t)F^{-1}_{\tilde{U}_2}(t)\,dt$, $\mu_j=\mathrm{E}\,(\tilde{V}_j)=c_j+r_j\,\mathrm{E}\,(\tilde{U}_j)/2$, $\sigma_j^2=r_j^2\,\mathrm{Var}(\tilde{U}_j)/4$, *and*

$$\rho_{12}=\frac{\mathcal{E}(\tilde{U}_1,\tilde{U}_2)-\mathrm{E}\,(\tilde{U}_1)\,\mathrm{E}\,(\tilde{U}_2)}{\sqrt{\mathrm{Var}(\tilde{U}_1)\mathrm{Var}(\tilde{U}_2)}}.$$

Proof. By Definition 3.1 and Proposition 3.2,

$$\begin{aligned} d_M(x_1,x_2)^2 &= \int_0^1 \left(F_1^{-1}(t)-F_2^{-1}(t)\right)^2 dt \\ &= \int_0^1 \left(c_1-c_2+\frac{r_1}{2}\,F^{-1}_{\tilde{U}_1}(t)-\frac{r_2}{2}\,F^{-1}_{\tilde{U}_2}(t)\right)^2 dt \\ &= (c_1-c_2)^2\;+\;2(c_1-c_2)\left(\frac{r_1}{2}\int_0^1 F^{-1}_{\tilde{U}_1}(t)\,dt-\frac{r_2}{2}\int_0^1 F^{-1}_{\tilde{U}_2}(t)\,dt\right) \\ &+ \int_0^1 \left(\frac{r_1}{2}\,F^{-1}_{\tilde{U}_1}(t)-\frac{r_2}{2}\,F^{-1}_{\tilde{U}_2}(t)\right)^2 dt. \end{aligned} \tag{3.5}$$

Notice that for $k=1,2$,

$$\int_0^1 \left(F^{-1}_{\tilde{U}_j}(t)\right)^k dt \;=\int_{\mathbb{R}}\left[F^{-1}_{\tilde{U}_j}\left(F_{\tilde{U}_j}(u)\right)\right]^k f_{\tilde{U}_j}(u)\,du \;=\int_{\mathbb{R}} u^k\,f_{\tilde{U}_j}(u)\,du \;=\;\mathrm{E}\,(\tilde{U}_j^k), \tag{3.6}$$

where $f_{\tilde{U}_j}$ is the probability density function of the absolutely continuous latent random variable $\tilde{U}_j$, $j=1,2$. Denoting $\mathcal{E}(\tilde{U}_1,\tilde{U}_2)=\int_0^1 F^{-1}_{\tilde{U}_1}(t)F^{-1}_{\tilde{U}_2}(t)\,dt$, and replacing (3.6) in (3.5), we obtain

$$\begin{aligned} d_M(x_1,x_2)^2 &= (c_1-c_2)^2\;+\;2(c_1-c_2)\left(\frac{r_1}{2}\,\mathrm{E}\,(\tilde{U}_1)-\frac{r_2}{2}\,\mathrm{E}\,(\tilde{U}_2)\right) \\ &+\;\frac{r_1^2}{4}\,\mathrm{E}\,(\tilde{U}_1^2)\;+\;\frac{r_2^2}{4}\,\mathrm{E}\,(\tilde{U}_2^2)\;-\;\frac{r_1r_2}{2}\,\mathcal{E}(\tilde{U}_1,\tilde{U}_2). \end{aligned} \tag{3.7}$$

By adding and subtracting $\left(r_1\,\mathrm{E}\,(\tilde{U}_1)/2-r_2\,\mathrm{E}\,(\tilde{U}_2)/2\right)^2$ to (3.7), and considering $\mu_j=\mathrm{E}(\tilde{V}_j)=c_j+r_j\,\mathrm{E}(\tilde{U}_j)/2$, and $\sigma_j^2=\mathrm{Var}(\tilde{V}_j)=r_j^2\mathrm{Var}(\tilde{U}_j)/4$, it follows that

$$\begin{aligned} d_M(x_1,x_2)^2 &= (\mu_1-\mu_2)^2+\sigma_1^2+\sigma_2^2-\frac{r_1r_2}{2}\left(\mathcal{E}(\tilde{U}_1,\tilde{U}_2)-\mathrm{E}\,(\tilde{U}_1)\,\mathrm{E}\,(\tilde{U}_2)\right) \\ &= (\mu_1-\mu_2)^2+(\sigma_1-\sigma_2)^2+2\sigma_1\sigma_2\left(1-\frac{\mathcal{E}(\tilde{U}_1,\tilde{U}_2)-\mathrm{E}\,(\tilde{U}_1)\,\mathrm{E}\,(\tilde{U}_2)}{\sqrt{\mathrm{Var}(\tilde{U}_1)\mathrm{Var}(\tilde{U}_2)}}\right) \\ &= (\mu_1-\mu_2)^2+(\sigma_1-\sigma_2)^2+2\sigma_1\sigma_2\,(1-\rho_{12})\,, \end{aligned}$$

concluding the proof. □

Remark 3.4. *Observe that although* $\mathcal{E}(\tilde{U}_1,\tilde{U}_2)$ *is not the usual* $\mathrm{E}\,(\tilde{U}_1\tilde{U}_2)$*, it is akin to an expected value. This is the case, since*

$$\mathcal{E}(\tilde{U}_1,\tilde{U}_2)\;=\;\int_0^1 F^{-1}_{\tilde{U}_1}(t)F^{-1}_{\tilde{U}_2}(t)\,dt\;=\;\mathrm{E}\left(F^{-1}_{\tilde{U}_1}(T)F^{-1}_{\tilde{U}_2}(T)\right)\;=\;\int_{\mathbb{R}} u\;F^{-1}_{\tilde{U}_2}\left(F_{\tilde{U}_1}(u)\right)\,f_{\tilde{U}_1}(u)\,du\;,$$

where T *is a real-valued random variable following a uniform distribution in* $[0,1]$*. Each value of* $\tilde{U}_1$ *is multiplied by the value of* $\tilde{U}_2$ *that shares the same value of the respective distribution function, weighted by the probability density function of* $\tilde{U}_1$*. Nevertheless, if* $\tilde{U}_1$ *and* $\tilde{U}_2$ *are identically distributed, then* $\mathcal{E}(\tilde{U}_1,\tilde{U}_2)=\mathrm{E}(\tilde{U}_1^2)=\mathrm{E}(\tilde{U}_2^2)$*.*

Remark 3.5. *In [7, Prop. 2] Irpino and Verde deduced that* $d_M(x_1,x_2)^2 = (\mu_1-\mu_2)^2+(\sigma_1-\sigma_2)^2+2\sigma_1\sigma_2(1-\rho_{12})$ *and called*

$$\rho_{12}=\frac{\int_0^1 F_{\tilde{U}_1}^{-1}(t)F_{\tilde{U}_2}^{-1}(t)dt-\mu_1\mu_2}{\sigma_1\sigma_2}=\frac{\mathcal{E}(\tilde{U}_1,\tilde{U}_2)-\mathrm{E}\,(\tilde{U}_1)\,\mathrm{E}\,(\tilde{U}_2)}{\sqrt{\mathrm{Var}(\tilde{U}_1)\mathrm{Var}(\tilde{U}_2)}},$$

the correlation coefficient between two quantile functions, since $\mathrm{E}\,(\tilde{U}_j) = \mathrm{E}\,(F_{\tilde{U}_j}^{-1}(T))$, $\mathrm{E}\,(\tilde{U}_j^2) = \mathrm{E}\left(\left(F_{\tilde{U}_j}^{-1}(T)\right)^2\right)$ *and* $\mathcal{E}(\tilde{U}_1,\tilde{U}_2)=\mathrm{E}\left(F_{\tilde{U}_1}^{-1}(T)F_{\tilde{U}_2}^{-1}(T)\right)$. *Here, the real-valued random variable* T *follows a continuous uniform distribution in* $[0,1]$. *In Theorem 3.3, we derived the same result based on the model presented in Definition 2.3.*

Remark 3.6. *Notice that the expressions (3.3) and (3.4) still hold when computing the Mallows' distance between an interval and a point. To see this, note that if* $x_2=\left((c_2,r_2)^T,F_{\tilde{U}_2}\right)$ *represents a point, which means that* $r_2=0$ *and* $F_{\tilde{U}_2}^{-1}(t)=0,\ t\in(0,1]$, *according to Definition 2.3, it follows from (3.5) that*

$$d_M(x_1,x_2)^2 = (c_1-c_2)^2 + (c_1-c_2)\,r_1\,\mathrm{E}\,(\tilde{U}_1) + \frac{r_1^2}{4}\,\mathrm{E}\,(\tilde{U}_1^2), \tag{3.8}$$

which also holds if we set $r_2=0$ *in (3.3). Additionally, (3.8) can be written as*

$$d_M(x_1,x_2)^2 = (\mu_1-c_2)^2 + \sigma_1^2,$$

which coincides with (3.4) for $r_2=0$.

The main challenge in Theorem 3.3 is to find the quantity $\mathcal{E}(\tilde{U}_1,\tilde{U}_2)$, that is to say, calculating the integral of the product of the quantile functions of $\tilde{U}_1$ and $\tilde{U}_2$. The next example illustrates how to compute the square of the Mallows' distance between two intervals, when the distribution of the latent random variables $\tilde{U}_1$ and $\tilde{U}_2$ is known.

Example 3.7. *Let* x_1, x_2 *be intervals such that* $x_j=\left((c_j,r_j)^T,F_{\tilde{U}_j}\right)$, $j=1,2$, *where* $\tilde{U}_1$ *follows a continuous uniform distribution in* $[-1,1]$ *and* $\tilde{U}_2$ *follows a symmetric triangular distribution (with mode* 0*) in* $[-1,1]$. *The square of the Mallows' distance between* x_1 *and* x_2 *can be computed using expression (3.3). We firstly note that, since both distributions are symmetric,* $\mathrm{E}(\tilde{U}_1)=\mathrm{E}(\tilde{U}_2)=0$. *Furthermore, it can easily be shown that* $\mathrm{E}(\tilde{U}_1^2)=1/3$ *and* $\mathrm{E}(\tilde{U}_2^2)=1/6$. *It only remains to compute the quantity* $\mathcal{E}(\tilde{U}_1,\tilde{U}_2)$. *Noting that the quantile function of* $\tilde{U}_1$ *is* $F_{\tilde{U}_1}^{-1}(t)=2t-1$, $t\in(0,1]$, *and the quantile function of* $\tilde{U}_2$ *is* $F_{\tilde{U}_2}^{-1}(t)=-1+\sqrt{2t}$, *if* $t\in(0,1/2]$, *and* $F_{\tilde{U}_2}^{-1}(t)=1-\sqrt{2(1-t)}$, *if* $t\in(1/2,1]$, *we have*

$$\begin{aligned}\mathcal{E}(\tilde{U}_1,\tilde{U}_2) &= \int_0^1 F_{\tilde{U}_1}^{-1}(t)F_{\tilde{U}_2}^{-1}(t)\,dt\\ &= \int_0^{\frac{1}{2}}(2t-1)\left(-1+\sqrt{2t}\right)dt + \int_{\frac{1}{2}}^1(2t-1)\left(1-\sqrt{2(1-t)}\right)dt = \frac{7}{30}.\end{aligned}$$

Hence, the square of the Mallows' distance between x_1 *and* x_2 *is*

$$d_M(x_1,x_2)^2 = (c_1-c_2)^2 + \frac{1}{12}r_1^2 + \frac{1}{24}r_2^2 - \frac{7}{60}r_1r_2.$$

When $\tilde{U}_1$ and $\tilde{U}_2$ are identically distributed, $\mathcal{E}(\tilde{U}_1,\tilde{U}_2)=\mathrm{E}(\tilde{U}_1^2)=\mathrm{E}(\tilde{U}_2^2)$ and $\rho_{12}=1$. This leads to interesting simplifications of the results stated in Theorem 3.3, as shown in the next corollary.

Corollary 3.8. *Under the conditions of Theorem 3.3, when* $\tilde{U}_1$ *and* $\tilde{U}_2$ *are identically distributed, the square of the Mallows' distance between* x_1 *and* x_2 *is*

$$\begin{aligned}d_M(x_1,x_2)^2 &= (\mu_1-\mu_2)^2+(\sigma_1-\sigma_2)^2\\ &= (c_1-c_2)^2+\frac{\mathrm{E}\,(\tilde{U}_1^2)}{4}\,(r_1-r_2)^2+\mathrm{E}\,(\tilde{U}_1)\,(c_1-c_2)\,(r_1-r_2)\,.\end{aligned} \tag{3.9}$$

Moreover, if $\tilde{U}_1$ *and* $\tilde{U}_2$ *are symmetric random variables, then* $\mathrm{E}(\tilde{U}_1) = \mathrm{E}(\tilde{U}_2) = 0$ *and*

$$d_M(x_1, x_2)^2 \;=\; (c_1 - c_2)^2 + \delta(r_1 - r_2)^2, \tag{3.10}$$

where $\delta = \mathrm{Var}(\tilde{U}_1)/4$.

Remark 3.9. *It is possible to establish that the parameter* $\delta = \mathrm{Var}(\tilde{U}_1)/4$ *takes values in the interval* $[0, 1/4]$. *Trivially, the variance is a non-negative number. To show the upper bound, let* $f_{\tilde{U}_1}$ *be the probability density function of* $\tilde{U}_1$. *Since* $\tilde{U}_1$ *is a symmetric random variable with support* $[-1, 1]$, *then*

$$\mathrm{Var}(\tilde{U}_1) \;=\; \mathrm{E}(\tilde{U}_1^2) \;=\; \int_{-1}^{1} x^2 f_{\tilde{U}_1}(x)\, dx \;\leq\; \int_{-1}^{1} f_{\tilde{U}_1}(x)\, dx \;=1 \;\Rightarrow\; \delta = \frac{\mathrm{Var}(\tilde{U}_1)}{4} \;\leq\; \frac{1}{4}.$$

To have a geometric interpretation of the Mallows' distance, assuming that $\tilde{U}_1$ and $\tilde{U}_2$ are symmetric and identically distributed random variables (see Corollary 3.8), we consider the set of intervals whose distance to the interval with macrodata $x_0 = [-3, 5]$ is one unit. Since the centre and range of x_0 are, respectively, $c_0 = 1$ and $r_0 = 8$, we have, for $\delta = \mathrm{Var}(\tilde{U}_1)/4$,

$$\mathcal{A}_\delta = \{x \in \mathbb{IR} : d_M(x, x_0) = 1\} = \left\{(c, r)^T \in \mathbb{R} \times \mathbb{R}_0^+ : (c-1)^2 + \delta(r-8)^2 = 1\right\},$$

where the distribution of the microdata within the intervals is given by $\tilde{U}_1$. The sets $\mathcal{A}_\delta$ are ellipses and are represented in Figure 2 for different distributions (listed in Figure 1) and, consequently, different values of δ. The figure shows that the lower the variance of $\tilde{U}_1$, the greater the concentration of microdata around the centres of the intervals, and the larger the area of the region whose values $x \in \mathbb{IR}$ verify $d_M(x, x_0) \leq 1$.

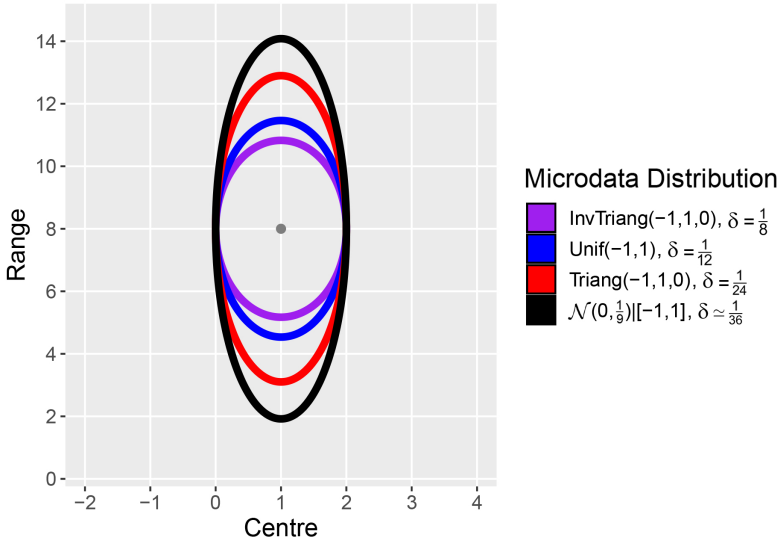


Fig. 2: Set of intervals whose Mallows' distance to $x_0 = [-3, 5]$ is one unit, according to the assumed symmetric distribution for the latent random variable $\tilde{U}_1$.

In [7, 36] the authors derived (3.4) based on the quantile functions related to the intervals x_j, $j = 1, 2$, and considered the special case when the microdata within x_j followed a continuous uniform distribution, resulting in $\delta = 1/12$. In [19] a similar result was obtained for a symmetric triangular distribution leading to $\delta = 1/24$ in (3.10).

It is also worth noticing that in the formulation (3.10) of the Mallows' distance, the weight associated with the ranges contribution, $\delta = \mathrm{Var}(\tilde{U}_1)/4$, is always smaller or equal to $1/4$, since $\tilde{U}_1$ has support in $[-1, 1]$. This emphasises the lesser role of the ranges when compared with the centres. If the variance of $\tilde{U}_1$ goes to zero, then the microdata is more and more concentrated around the centre of the interval c_1. See Figure 1 for several examples of distributions.

3.2 Mallows' distance in $\mathbb{IR}^p$

In SDA, the generalisation of the Mallows' distance to $\mathbb{IR}^p$ is defined using the Mallows' distance between each component of the two vectors.

Definition 3.10. *Let* $\boldsymbol{x}_1 = (x_{11}, \ldots, x_{1p})^T \in \mathbb{IR}^p$ *and* $\boldsymbol{x}_2 = (x_{21}, \ldots, x_{2p})^T \in \mathbb{IR}^p$. *Let* $d_M(x_{1i}, x_{2i})$, $i = 1, \ldots, p$, *be the univariate Mallows' distance between the intervals* $x_{1i}, x_{2i} \in \mathbb{IR}$. *The Mallows' distance between* $\boldsymbol{x}_1$ *and* $\boldsymbol{x}_2$ *is*

$$d_M(\boldsymbol{x}_1, \boldsymbol{x}_2) = \left(\sum_{i=1}^{p} d_M(x_{1i}, x_{2i})^2 \right)^{1/2}. \tag{3.11}$$

Since the Mallows' distance between hyperrectangles is defined using the distance between each component, we are interested in comparing the componentwise quantile functions of the microdata. As such, a hyperrectangle is uniquely identified by the vector of the centres, the vector of the ranges, and the collection of quantile functions of the microdata in each dimension. We use the notation $\boldsymbol{x}_j = \left((\boldsymbol{c}_j^T, \boldsymbol{r}_j^T)^T, F_{\tilde{U}_{j1}}, \ldots, F_{\tilde{U}_{jp}} \right)$, where each component is of the form $x_{ji} = \left((c_{ji}, r_{ji})^T, F_{\tilde{U}_{ji}} \right)$, $j = 1, 2$, $i = 1, \ldots, p$.

The following results require that, within the same dimension, the latent random variables of the hyperrectangles be identically distributed. In this scenario, we can introduce a latent random variable, $\tilde{U}_i$, $i = 1, \ldots, p$, whose distribution is the same as $\tilde{U}_{1i}$ and $\tilde{U}_{2i}$, and we have $\mathrm{E}(U_{1i}) = \mathrm{E}(U_{2i}) = \mathrm{E}(U_i)$ and $\mathrm{E}(U_{1i}^2) = \mathrm{E}(U_{2i}^2) = \mathrm{E}(U_i^2)$. We write $\boldsymbol{x}_j = \left((\boldsymbol{c}_j^T, \boldsymbol{r}_j^T)^T, F_{\tilde{U}_1}, \ldots, F_{\tilde{U}_p} \right)$, or simply $\boldsymbol{x}_j = (\boldsymbol{c}_j^T, \boldsymbol{r}_j^T)^T$, $j = 1, 2$, if no confusion arises. Given a vector $\boldsymbol{v}$, we define $\mathrm{Diag}(\boldsymbol{v})$ to be the diagonal matrix whose main diagonal is $\boldsymbol{v}$. In addition, given a matrix $\boldsymbol{A}$, we define $\mathrm{Diag}(\boldsymbol{A})$ to be the diagonal matrix with the same diagonal as $\boldsymbol{A}$.

Theorem 3.11. *For* $j = 1, 2$, *let* $\boldsymbol{x}_j = \left((\boldsymbol{c}_j^T, \boldsymbol{r}_j^T)^T, F_{\tilde{U}_1}, \ldots, F_{\tilde{U}_p} \right)$, *with* $\boldsymbol{c}_j = (c_{j1}, \ldots, c_{jp})^T \in \mathbb{R}^p$ *and* $\boldsymbol{r}_j = (r_{j1}, \ldots, r_{jp})^T \in (\mathbb{R}_0^+)^p$, *and suppose that, within the same dimension, the latent random variables are identically distributed. Then, the square of the Mallows' distance between* $\boldsymbol{x}_1$ *and* $\boldsymbol{x}_2$ *is*

$$\begin{aligned} d_M(\boldsymbol{x}_1, \boldsymbol{x}_2)^2 &= \sum_{i=1}^{p} \left((c_{1i} - c_{2i})^2 + \frac{\mathrm{E}\,(\tilde{U}_i^2)}{4} (r_{1i} - r_{2i})^2 + \mathrm{E}\,(\tilde{U}_i)(c_{1i} - c_{2i})(r_{1i} - r_{2i}) \right) \\ &= (\boldsymbol{c}_1 - \boldsymbol{c}_2)^T (\boldsymbol{c}_1 - \boldsymbol{c}_2) + (\boldsymbol{r}_1 - \boldsymbol{r}_2)^T \boldsymbol{\Delta} (\boldsymbol{r}_1 - \boldsymbol{r}_2) + (\boldsymbol{c}_1 - \boldsymbol{c}_2)^T \boldsymbol{\Psi} (\boldsymbol{r}_1 - \boldsymbol{r}_2), \end{aligned} \tag{3.12}$$

where $\boldsymbol{\Delta} = \mathrm{Diag}\,(\delta_1, \ldots, \delta_p)$ *with* $\delta_i = \mathrm{E}(\tilde{U}_i^2)/4$, *and* $\boldsymbol{\Psi} = \mathrm{Diag}\left(\mathrm{E}(\tilde{U}_1), \ldots, \mathrm{E}(\tilde{U}_p) \right)$.

Proof. The result follows from replacing the general expression of Corollary 3.8 in each component of the Mallows' distance (3.11) and using matrix notation. □

Remark 3.12. *Observe that, when the random variables* $\tilde{U}_i$ *are symmetric,* $\boldsymbol{\Psi} = \boldsymbol{0}$ *and* (3.12) *becomes*

$$d_M(\boldsymbol{x}_1, \boldsymbol{x}_2)^2 = (\boldsymbol{c}_1 - \boldsymbol{c}_2)^T (\boldsymbol{c}_1 - \boldsymbol{c}_2) + (\boldsymbol{r}_1 - \boldsymbol{r}_2)^T \boldsymbol{\Delta} (\boldsymbol{r}_1 - \boldsymbol{r}_2) = \sum_{i=1}^{p} \left((c_{1i} - c_{2i})^2 + \delta_i (r_{1i} - r_{2i})^2 \right),$$

where $\delta_i = \mathrm{E}(\tilde{U}_i^2)/4 = \mathrm{Var}(\tilde{U}_i)/4$. *In particular, if* $\delta_i = \delta$, $i = 1, \ldots, p$, *we obtain* $\boldsymbol{\Delta} = \delta \boldsymbol{I}_p$, *where* $\boldsymbol{I}_p$ *represents the* $p \times p$ *identity matrix, and*

$$d_M(\boldsymbol{x}_1, \boldsymbol{x}_2)^2 = (\boldsymbol{c}_1 - \boldsymbol{c}_2)^T (\boldsymbol{c}_1 - \boldsymbol{c}_2) + \delta (\boldsymbol{r}_1 - \boldsymbol{r}_2)^T (\boldsymbol{r}_1 - \boldsymbol{r}_2).$$

As an alternative to (3.12), Corollary 3.13 proposes the identification of the Mallows' distance d_M with a Mahalanobis' distance d_{Mah} in the space of the joint vector of the centres and ranges. This follows from the fact that within each dimension the latent random variables have the same distribution function, yielding that the only difference between the hyperrectangles is in the centres and the ranges.

Corollary 3.13. *Under the conditions of Theorem 3.11, the square of the Mallows' distance* (3.12) *between* $\boldsymbol{x}_1 = \left(\boldsymbol{y}_1, F_{\tilde{U}_1}, \ldots, F_{\tilde{U}_p} \right)$ *and* $\boldsymbol{x}_2 = \left(\boldsymbol{y}_2, F_{\tilde{U}_1}, \ldots, F_{\tilde{U}_p} \right)$, *with* $\boldsymbol{y}_1 = (\boldsymbol{c}_1^T, \boldsymbol{r}_1^T)^T$ *and* $\boldsymbol{y}_2 = (\boldsymbol{c}_2^T, \boldsymbol{r}_2^T)^T$ *representing the macrodata, can be expressed as*

$$d_M(\boldsymbol{x}_1, \boldsymbol{x}_2)^2 = d_{\mathrm{Mah}}(\boldsymbol{y}_1, \boldsymbol{y}_2; \boldsymbol{H})^2 = (\boldsymbol{y}_1 - \boldsymbol{y}_2)^T \boldsymbol{H} (\boldsymbol{y}_1 - \boldsymbol{y}_2), \text{ where } \boldsymbol{H} = \begin{pmatrix} \boldsymbol{I}_p & \frac{1}{2}\boldsymbol{\Psi} \\ \frac{1}{2}\boldsymbol{\Psi} & \boldsymbol{\Delta} \end{pmatrix}. \tag{3.13}$$

If $\boldsymbol{H}$ *is positive definite (i.e., the random variables* $\tilde{U}_i$*,* $i = 1, \ldots, p$*, are non-degenerate), then* $d_{\text{Mah}}(\boldsymbol{y}_1, \boldsymbol{y}_2; \boldsymbol{H})$ *is a Mahalanobis' distance in* $\mathbb{R}^p \times (\mathbb{R}_0^+)^p$*, whose corresponding covariance matrix is*

$$\boldsymbol{H}^{-1} = \begin{pmatrix} \boldsymbol{I}_p + \frac{1}{4}\boldsymbol{\Psi}^2\boldsymbol{Q} & -\frac{1}{2}\boldsymbol{\Psi}\boldsymbol{Q} \\ -\frac{1}{2}\boldsymbol{\Psi}\boldsymbol{Q} & \boldsymbol{Q} \end{pmatrix},$$

where $\boldsymbol{Q} = (\boldsymbol{\Delta} - \frac{1}{4}\boldsymbol{\Psi}^2)^{-1} = 4\,\text{Diag}\,(\text{Var}(\tilde{U}_1)^{-1}, \ldots, \text{Var}(\tilde{U}_p)^{-1})$*,* $\text{Var}(\tilde{U}_i) > 0$*,* $i = 1, \ldots, p$*.*

Proof. See Appendix A. □

Remark 3.14. *If any* $\tilde{U}_i$ *is a degenerate random variable, say* $\tilde{U}_1$ *for simplicity, then we can define a Mahalanobis' distance in a space of smaller dimension,* $\mathbb{R}^p \times (\mathbb{R}_0^+)^{p-1}$ *(for this case), induced by the submatrix* $\boldsymbol{H}^{\diamondsuit}$*, obtained by removing the* $(p+1)$*-th row and column of* $\boldsymbol{H}$*. Furthermore, the points defined in this space do not correspond to proper hyperrectangles, since the centres are in* $\mathbb{R}^p$ *and the ranges are in* $(\mathbb{R}_0^+)^{p-1}$*. We introduce the notation* $\boldsymbol{y}_j^{\diamondsuit}$ *to represent* $\boldsymbol{y}_j = (\boldsymbol{c}_j^T, \boldsymbol{r}_j^T)^T$ *with the* $(p+1)$*-th component removed,* $j = 1, 2$*. Hence,*

$$d_M(\boldsymbol{x}_1, \boldsymbol{x}_2)^2 = d_{\text{Mah}}(\boldsymbol{y}_1^{\diamondsuit}, \boldsymbol{y}_2^{\diamondsuit}; \boldsymbol{H}^{\diamondsuit})^2 = \left(\boldsymbol{y}_1^{\diamondsuit} - \boldsymbol{y}_2^{\diamondsuit}\right)^T \boldsymbol{H}^{\diamondsuit}(\boldsymbol{y}_1^{\diamondsuit} - \boldsymbol{y}_2^{\diamondsuit}),$$

where $d_{\text{Mah}}(\boldsymbol{y}_1^{\diamondsuit}, \boldsymbol{y}_2^{\diamondsuit}; \boldsymbol{H}^{\diamondsuit})$ *defines a Mahalanobis' distance in* $\mathbb{R}^p \times (\mathbb{R}_0^+)^{p-1}$*.*

4 Location, scale, and association between interval-valued variables

In [7], Irpino and Verde proposed an approach to derive interval and histogram-valued descriptive measures for location, scale, and association between two symbolic characteristics measured on the same set of objects. The authors rely on the (sample) Fréchet mean, also known as (sample) barycentre. Given a set of points in a metric space, the Fréchet mean is the point that minimises the weighted sum of the squares of the distance to all given points (see [37], for details). The minimum of this sum is called the Fréchet variance. The definition of Fréchet mean can be extended to the population if, instead of the weighted sum, we consider the expected value. In the case of a multivariate interval-valued random vector, $\boldsymbol{X}$, we can define the population Fréchet mean or population barycentre, using the Mallows' distance, d_M. In the rest of the paper, we identify a hyperrectangle using its centre and range since, within each component, the latent random variables are assumed to have the same distribution.

Definition 4.1. *Let* $\boldsymbol{X}$ *be an interval-valued random variable with support in* $\mathbb{IR}^p$ *characterised by the random vector of centres and ranges,* $\boldsymbol{C}$ *and* $\boldsymbol{R}$*, assumed to have finite expected values,* $\boldsymbol{\mu}_C$ *and* $\boldsymbol{\mu}_R$*, and covariance matrices,* $\boldsymbol{\Sigma}_{CC}$ *and* $\boldsymbol{\Sigma}_{RR}$*, respectively, and such that each component of* $\boldsymbol{R}$ *is equal to* 0 *with probability* 0 *or probability* 1*. Let* $\boldsymbol{\Sigma}_{CR}$ *be the covariance matrix between* $\boldsymbol{C}$ *and* $\boldsymbol{R}$*, let* $\boldsymbol{U} = (U_1, \ldots, U_p)^T$ *be a real-valued random vector, independent from* $(\boldsymbol{C}^T, \boldsymbol{R}^T)^T$*, let* $\boldsymbol{\Delta} = \text{Diag}\,(\delta_1, \ldots, \delta_p)$*, where* $\delta_i = \text{E}\,(U_i^2)/4$*,* $i = 1, \ldots, p$*, and let* $\boldsymbol{\Psi} = \text{Diag}\,(\text{E}(U_1), \ldots, \text{E}(U_p))$*. The population barycentre of* $\boldsymbol{X}$*, denoted by* $\boldsymbol{\mu}_B \in \mathbb{IR}^p$*, is the hyperrectangle* $\boldsymbol{x} = (\boldsymbol{c}^T, \boldsymbol{r}^T)^T$ *in* $\mathbb{IR}^p$ *that globally minimises the function*

$$f(\boldsymbol{c}, \boldsymbol{r}) = \text{E}\left(d_M(\boldsymbol{X}, \boldsymbol{x})^2\right) = \text{E}\left((\boldsymbol{C} - \boldsymbol{c})^T(\boldsymbol{C} - \boldsymbol{c}) + (\boldsymbol{R} - \boldsymbol{r})^T\boldsymbol{\Delta}(\boldsymbol{R} - \boldsymbol{r}) + (\boldsymbol{C} - \boldsymbol{c})^T\boldsymbol{\Psi}(\boldsymbol{R} - \boldsymbol{r})\right), \quad (4.1)$$

where the Mallows' distance is given by (3.12)*.*

The solution of the minimisation of (4.1) is given in the following theorem.

Theorem 4.2. *Under the conditions of Definition 4.1, the population barycentre of* $\boldsymbol{X}$ *is*

$$\boldsymbol{\mu}_B = (\boldsymbol{\mu}_C^T, \boldsymbol{\mu}_R^T)^T, \quad (4.2)$$

and the corresponding Fréchet variance is

$$V_F(\boldsymbol{\mu}_B) = \text{E}\left(d_M(\boldsymbol{X}, \boldsymbol{\mu}_B)^2\right) = \text{tr}(\boldsymbol{\Sigma}_{CC} + \boldsymbol{\Delta}\boldsymbol{\Sigma}_{RR} + \boldsymbol{\Sigma}_{CR}\boldsymbol{\Psi}). \quad (4.3)$$

Proof. See Appendix B.1. □

As a particular case, notice that, if we observe a sample of $\boldsymbol{X}$, say $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{IR}^p$, where $\boldsymbol{x}_j = (\boldsymbol{c}_j, \boldsymbol{r}_j)^T$, $j = 1, \ldots, n$, and consider the empirical distribution, Theorem 4.2 guarantees that the (sample) barycentre

is

$$\overline{\mathrm{x}}_B \;=\; \underset{(\boldsymbol{c},\boldsymbol{r})}{\arg\min}\; \frac{1}{n}\sum_{j=1}^{n}\left[(\boldsymbol{c}_j-\boldsymbol{c})^T(\boldsymbol{c}_j-\boldsymbol{c}) \;+\; (\boldsymbol{r}_j-\boldsymbol{r})^T\boldsymbol{\Delta}(\boldsymbol{r}_j-\boldsymbol{r}) \;+\; (\boldsymbol{c}_j-\boldsymbol{c})^T\boldsymbol{\Psi}(\boldsymbol{r}_j-\boldsymbol{r})\right], \tag{4.4}$$

leading to $\overline{\mathrm{x}}_B = (\overline{\boldsymbol{c}}_n^T, \overline{\boldsymbol{r}}_n^T)^T = \left(\sum_{j=1}^n \boldsymbol{c}_j^T/n, \sum_{j=1}^n \boldsymbol{r}_j^T/n\right)^T$.

In [7], the authors propose the definition of sample variance and sample covariance between two interval-valued variables based on the Mallows' distance that follows.

Definition 4.3. *Let $(x_{11}, x_{12})^T, \ldots, (x_{n1}, x_{n2})^T$ be a sample of size n from $(X_1, X_2)^T$, a bivariate interval-valued random vector with support in $\mathbb{IR}^2$, with sample barycentre $(\overline{x}_{B_1}, \overline{x}_{B_2})^T$. The sample covariance between X_1 and X_2 is*

$$s_{12,B} \;=\; \widehat{\mathrm{Cov}}_B(X_1,X_2) \;=\; \frac{1}{n}\sum_{j=1}^{n}\int_0^1 \left(F_{j1}^{-1}(t)-F_{B_1}^{-1}(t)\right)\left(F_{j2}^{-1}(t)-F_{B_2}^{-1}(t)\right)dt, \tag{4.5}$$

where $F_{ji}^{-1}(t)$ and $F_{B_i}^{-1}(t)$ are the quantile functions of the latent random variables describing x_{ji} and $\overline{x}_{B_i}$, respectively, for $j = 1, \ldots, n$, $i = 1, 2$. Additionally, the sample variance of X_i is $s_{ii,B} = \widehat{\mathrm{Cov}}_B(X_i, X_i)$.

In the particular case $p = 1$, $s_{11,B}$ is the minimum value of the objective function in (4.4) and corresponds to the univariate sample Fréchet variance.

The population counterpart to the symbolic sample covariance can be defined by a natural adaptation of (4.5).

Definition 4.4. *Let $X_1 = (C_1, R_1)^T$ and $X_2 = (C_2, R_2)^T$ be two interval-valued random variables with baricentres μ_{B_1} and μ_{B_2}. Let $G_i^{-1}(t) = C_i + R_i F_{U_i}^{-1}(t)/2$ be the random variable whose realisations on specific intervals are the quantile functions of the microdata within, where $F_{U_i}^{-1}$ is the quantile function of the latent random variable U_i, and let $F_{B_i}^{-1}(t)$ be the quantile function of the microdata in μ_{B_i}, $i = 1, 2$. The covariance $\mathrm{Cov}_B(X_1, X_2)$ between X_1 and X_2 is defined by*

$$\mathrm{Cov}_B(X_1,X_2) \;=\; \mathrm{E}\left(\int_0^1 \left(G_1^{-1}(t)-F_{B_1}^{-1}(t)\right)\left(G_2^{-1}(t)-F_{B_2}^{-1}(t)\right)dt\right). \tag{4.6}$$

Furthermore, the variance of X_i is $\mathrm{Cov}_B(X_i, X_i)$, $i = 1, 2$.

Taking into consideration that $\tilde{V}_{B_i} = \mu_{C_i} + \tilde{U}_{B_i}\,\mu_{R_i}/2$ (see Definition 2.3), and that U_i and $\tilde{U}_{B_i}$ are identically distributed, thus sharing the same quantile function, we can simplify the covariance matrix for the bivariate case, and then extend it for any p-dimensional interval-valued random vector in Corollary 4.6.

Corollary 4.5. *Under the conditions of Definition 4.1, the covariance between two interval-valued random variables $X_1 = (C_1, R_1)^T$ and $X_2 = (C_2, R_2)^T$ with barycentres μ_{B_1} and μ_{B_2}, respectively, is*

$$\begin{aligned}\mathrm{Cov}_B(X_1,X_2) \;&=\; \mathrm{E}\left(\int_0^1 \left(G_1^{-1}(t)-F_{B_1}^{-1}(t)\right)\left(G_2^{-1}(t)-F_{B_2}^{-1}(t)\right)dt\right)\\ &=\; \mathrm{Cov}(C_1,C_2) \;+\; \frac{\mathcal{E}(U_1,U_2)}{4}\mathrm{Cov}(R_1,R_2) \;+\; \frac{\mathrm{E}(U_2)}{2}\mathrm{Cov}(C_1,R_2)\\ &+\; \frac{\mathrm{E}(U_1)}{2}\mathrm{Cov}(C_2,R_1),\end{aligned}$$

where $\mathcal{E}(U_1, U_2) = \int_0^1 F_{U_1}^{-1}(t) F_{U_2}^{-1}(t)\, dt$. Moreover,

$$\mathrm{Var}_B(X_1) \;=\; \mathrm{Cov}_B(X_1,X_1) \;=\; \mathrm{Var}(C_1) \;+\; \frac{\mathrm{E}(U_1^2)}{4}\mathrm{Var}(R_1) \;+\; \mathrm{E}(U_1)\,\mathrm{Cov}(C_1,R_1). \tag{4.7}$$

Proof. See Appendix B.2. □

Applying Corollary 4.5 to a p-dimensional random vector results in a covariance matrix, $\mathrm{Var}_B(\boldsymbol{X}) = \boldsymbol{\Sigma}_B$, as stated in the following corollary. Here, we introduce the notation $[\boldsymbol{A}]_{ij}$ to represent the entry (i, j) of matrix $\boldsymbol{A}$.

Corollary 4.6. *Under the conditions of Definition 4.1, the covariance matrix of an interval-valued random vector* $\boldsymbol{X} \in \mathbb{IR}^p$ *is*

$$\mathrm{Var}_B(\boldsymbol{X}) \;=\; \boldsymbol{\Sigma}_B \;=\; \boldsymbol{\Sigma}_{CC} \;+\; \frac{1}{4}\boldsymbol{\mathfrak{E}}_{UU} \bullet \boldsymbol{\Sigma}_{RR} \;+\; \frac{1}{2}\boldsymbol{\Sigma}_{CR}\boldsymbol{\Psi} \;+\; \frac{1}{2}\boldsymbol{\Psi}\boldsymbol{\Sigma}_{RC}, \tag{4.8}$$

where $\boldsymbol{\Sigma}_{CC}$ *and* $\boldsymbol{\Sigma}_{RR}$ *are the respective covariance matrices of* $\boldsymbol{C}$ *and* $\boldsymbol{R}$, $\boldsymbol{\Sigma}_{CR} = \boldsymbol{\Sigma}_{RC}^T$ *is the covariance matrix between* $\boldsymbol{C}$ *and* $\boldsymbol{R}$, $\boldsymbol{\Psi} = \mathrm{Diag}\left(\mathrm{E}(U_1), \ldots, \mathrm{E}(U_p)\right)$, $\mathcal{E}(U_i, U_j) = \int_0^1 F_{U_i}^{-1}(t) F_{U_j}^{-1}(t)\, dt$, $[\boldsymbol{\mathfrak{E}}_{UU}]_{ij} = \mathcal{E}(U_i, U_j)$, $i \neq j$, $[\boldsymbol{\mathfrak{E}}_{UU}]_{ii} = \mathrm{E}(U_i^2)$, $i, j = 1, \ldots, p$, *and* $\bullet$ *denotes the Schur (or entrywise) product of matrices. The corresponding correlation matrix is* $\mathrm{Cor}_B(\boldsymbol{X}) = \boldsymbol{D}^{-1/2}\boldsymbol{\Sigma}_B\boldsymbol{D}^{-1/2}$, *where* $\boldsymbol{D} = \mathrm{Diag}\left([\boldsymbol{\Sigma}_B]_{11}, \ldots, [\boldsymbol{\Sigma}_B]_{pp}\right)$.

As before, specific assumptions on the random variables U_i, $i = 1, \ldots, p$, lead to simpler covariance matrices $\boldsymbol{\Sigma}_B$. For example, if all U_i are identically distributed to a random variable U, then $\mathcal{E}(U_i, U_j) = \mathrm{E}(U^2) = 4\delta$, and

$$\mathrm{Var}_B(\boldsymbol{X}) \;=\; \boldsymbol{\Sigma}_B \;=\; \boldsymbol{\Sigma}_{CC} \;+\; \delta\boldsymbol{\Sigma}_{RR} \;+\; \frac{\mathrm{E}(U)}{2}\left(\boldsymbol{\Sigma}_{CR} \;+\; \boldsymbol{\Sigma}_{RC}\right). \tag{4.9}$$

Moreover, if $\mathrm{E}(U) = 0$, then

$$\mathrm{Var}_B(\boldsymbol{X}) \;=\; \boldsymbol{\Sigma}_B \;=\; \boldsymbol{\Sigma}_{CC} \;+\; \delta\boldsymbol{\Sigma}_{RR}. \tag{4.10}$$

Having in mind (4.3), it can be argued that the Fréchet variance is a non-negative number, interpreted as the total variance of the covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_{CC} + \boldsymbol{\Delta}\boldsymbol{\Sigma}_{RR} + \boldsymbol{\Sigma}_{CR}\boldsymbol{\Psi}$. However, $\boldsymbol{\Omega}$ is not exactly the covariance matrix $\boldsymbol{\Sigma}_B$ based on the barycentre approach. Nevertheless, they have the same trace (that is, the same total variance), since $\mathrm{tr}(\boldsymbol{\Sigma}_{CR}\boldsymbol{\Psi}) = \mathrm{tr}(\boldsymbol{\Psi}\boldsymbol{\Sigma}_{RC})$ and $[\boldsymbol{\mathfrak{E}}_{UU}]_{ii} = \mathrm{E}(U_i^2)$, leading to $\mathrm{tr}(\boldsymbol{\mathfrak{E}}_{UU} \bullet \boldsymbol{\Sigma}_{RR}/4) = \mathrm{tr}(\boldsymbol{\Delta}\boldsymbol{\Sigma}_{RR})$. In conclusion, we have established that $V_F(\boldsymbol{\mu}_B) = \mathrm{tr}(\boldsymbol{\Sigma}_B)$.

5 Examples

In this section, we compare several estimates of the sample mean, sample covariance, and sample correlation matrix based on three distinct datasets and consider different strategies to model the latent random variables U_i, $i = 1, \ldots, p$.

The first example uses the *credit card* dataset [3, 33, 38], where the microdata are available, and the choice of the distribution of U_i was discussed in [33]. In this paper, the authors found evidence that these random variables followed a symmetric triangular distribution (mode zero). We revisit this problem by comparing the estimated correlation matrix of [33] with the one based on the barycentre approach. The second example considers the dataset *nycflights.int*, listed in [39]. This dataset contains information about all flights that departed from the three major New York airports in 2013. The data is aggregated by month and carrier. The microdata associated with two of the four interval variables does not suggest any obvious known family of distributions. Thus, non-parametric probability density estimators and associated quantile functions are considered. A third dataset related to Internet traffic redirection attacks [11] is analysed under the new proposals for location and association for interval data. In this case, only microdata measures of location are available. The distribution of U_i is chosen based on the empirical knowledge of the experts and the partial information available about the microdata.

The code related to the analysis of the datasets can be found in https://github.com/MROS13/MallowSymbCov.

5.1 Credit cards

The *credit card* dataset [3, 33, 38] refers to five interval-valued random variables measuring the monthly expenses of three credit card users on Food (x_1), Social Entertainment (x_2), Travel (x_3), Gas (x_4), and Clothes (x_5), during one year, leading to a total of $n = 36$ observations on $p = 5$ variables.

The sample barycentre (see equation (4.2)) is

$$\overline{\mathrm{x}}_B = (\,[21.52,\ 30.66],\ [8.68,\ 18.92],\ [177.46,\ 190.47],\ [20.36,\ 29.32],\ [43.37,\ 55.26]\,)^T,$$

since $\overline{\boldsymbol{c}}_n = (26.09,\ 13.80,\ 183.97,\ 24.84,\ 49.32)^T$ and $\overline{\boldsymbol{r}}_n = (9.15, 10.23, 13.01, 8.96, 11.89)^T$ are the vectors of centres' means and ranges' means, respectively.

Figure 3 displays a 5×5 matrix, where the entries below the main diagonal are the symbolic bivariate scatter plots of the interval-valued random variables. In green is represented the respective sample bivariate barycentres. The names of the variables appear in the main diagonal. Figure 3 supports the idea that the user marked in red is the one with higher expenses on Clothes (x_5) and Food (x_1). Clothes (x_5) is the variable that best separates the users' credit card monthly expenses. Additionally, the barycentre indicates that Travel (x_3) is the type of expense that users allocated the highest amount of credit card expenses, followed by Clothes (x_5). In opposition, Social Entertainment (x_2) is, on average, where the lowest amount of money is spent. The components of the barycentre's range, $\overline{\boldsymbol{r}}_n$, are fairly similar, indicating that the inner variability among the types of expenses is also similar. The symbolic bivariate scatter plot seems to suggest a moderate positive association between Food (x_1) and Clothes (x_5), a mild negative one between Gas (x_4) and Clothes (x_5), and probably a weaker negative association between Food (x_1) and Gas (x_4). These findings are confirmed by the estimated correlation values, appearing above the main diagonal of the matrix in Figure 3.

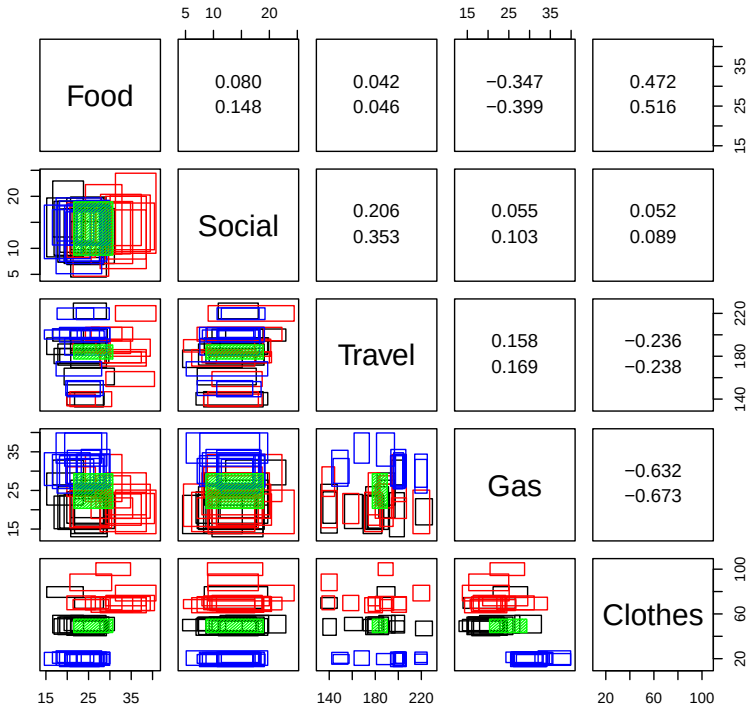


Fig. 3: Symbolic bivariate scatter plots of credit card data and respective symbolic estimated correlations, based on $\boldsymbol{\Sigma}_7 = \boldsymbol{S}_{CC} + \text{Diag}\left(\text{E}(\boldsymbol{R}\boldsymbol{R}^T)\right)/24$ (top value) and (4.10) (bottom value), assuming in both cases that $U_i \sim \text{Triang}(-1, 1, 0)$. There are three subjects with monthly expenses measured over a year, coloured differently. The bivariate barycentres are in green.

The *credit card* dataset was fully explored in [33], where eight different symbolic estimates of covariance matrices (and respective correlation matrices) were considered. Quantile-quantile plots (with 95% pointwise envelopes) of microdata values support the assumption that the U_i follow a symmetric triangular distribution, i.e., $\text{Triang}(-1, 1, m = 0)$. Under the appropriate model ($k = 7$ in [33]), it was assumed that $U_1, \ldots, U_p$ are zero mean uncorrelated random variables independent from the random vector of centres and ranges $(\boldsymbol{C}^T, \boldsymbol{R}^T)^T$. The estimated correlation matrix was presented in [33, pp. 516] and is reproduced above the main diagonal of the matrix in Figure 3, in the top value of each entrance. The respective symbolic covariance matrix is then computed as $\widehat{\boldsymbol{\Sigma}}_7 = \boldsymbol{S}_{CC} + \text{Diag}\left(\widehat{\text{E}}(\boldsymbol{R}\boldsymbol{R}^T)\right)/24$, where $\widehat{\text{E}}(\boldsymbol{R}\boldsymbol{R}^T) = \boldsymbol{S}_{RR} + \overline{\boldsymbol{r}}_n \overline{\boldsymbol{r}}_n^T$, and $\boldsymbol{S}_{CC}$ ($\boldsymbol{S}_{RR}$) is the sample covariance matrix of the centres (ranges).

In the barycentre approach, assuming that all the U_i follow a symmetric triangular distribution, we computed the estimated symbolic correlation matrix, based on $\boldsymbol{S}_B = \boldsymbol{S}_{CC} + \boldsymbol{S}_{RR}/24$ (see equation (4.10)). The values are shown above the main diagonal of the matrix of Figure 3, in the bottom value in each

entrance. As expected, according to the microdata study presented in [33], the two scenarios lead to similar estimates.

5.2 New York city flights interval dataset

This example illustrates a case where the microdata are available, but the fitting of their distribution reveals an apparent major difficulty. We suggest the use of univariate kernel density estimators (KDE) to overcome this issue, as illustrated in this example.

The dataset *nycflights.int* (see [40], for details) refers to all flights that departed from the three New York airports to destinations in the United States, Puerto Rico, and the American Virgin Islands, in 2013. Each flight is characterised by its departure delay (x_1), arrival delay (x_2), amount of time spent in the air (x_3), and distance between airports (x_4), for a total of 327 345 flights. The data was aggregated by month and carrier, leading to $n = 142$ multivariate interval-valued observations and $p = 4$ variables. In [40] authors used a robust aggregation strategy by filtering out the 5% lowest and highest values of the microdata in each interval-valued variable. Additionally, degenerate intervals (range zero) were eliminated. The histograms of the microdata per variable are shown in Figure 4. For the first two variables, the associated latent variables are modelled as a shifted Beta distribution, i.e., $U_i = 2W_i - 1$, with $W_i \sim \text{Beta}(\alpha_i, \beta_i)$, $i = 1, 2$. The parameters were estimated by the method of moments, and the estimated probability density functions are plotted in Figure 4a and Figure 4b, respectively, in blue. The cases of U_3 and U_4 illustrate the difficulty in fitting the latent distribution, as shown in Figure 4c and Figure 4d. Alternatively, a univariate kernel density estimator (KDE) is used to estimate the needed quantities. The fitted KDEs were obtained using the R package *kde1d* (see [41], for further information) and are represented in blue in the two bottom figures.

The covariance matrix based on the barycentre approach can be estimated by computing the estimates of each matrix in (4.8). The elements of the main diagonal of $\hat{\boldsymbol{\Psi}}$ are the sample estimates of the first moments of the U_i. We used the sample means, valued $\hat{\boldsymbol{\Psi}} = \text{Diag}(-0.66, -0.42, -0.21, -0.21)$. This shows a right-skewed tendency of the latent distributions. For the computation of $\hat{\mathfrak{C}}_{UU}$, we applied two methods: (i) the elements of the main diagonal, $\text{E}(U_i^2)$, were estimated as the sample second moments of the U_i; (ii) the elements outside the main diagonal are $\hat{\mathcal{E}}(U_i, U_j) = \int_0^1 \hat{F}_{U_i}^{-1}(t)\hat{F}_{U_j}^{-1}(t)\, dt$, $i \neq j$, whose integrals were computed using numerical routines from the R package *calculus* (see [42] for details). According to Proposition 3.2, $\hat{F}_{U_i}^{-1}(t) = 2\hat{F}_{W_i}^{-1}(t) - 1$, where $\hat{F}_{W_i}^{-1}(t)$ is the estimated quantile function of the fitted distribution, and $W_i \sim \text{Beta}(\hat{a}_i, \hat{b}_i)$, $i = 1, 2$. For $i = 3, 4$, $\hat{F}_{U_i}^{-1}(t)$ was computed using the function *qkde1d* (see [41] for details). This led to

$$\hat{\mathfrak{C}}_{UU} = \begin{bmatrix} 0.59 & & & \\ 0.44 & 0.35 & & \\ 0.35 & 0.32 & 0.37 & \\ 0.34 & 0.31 & 0.35 & 0.34 \end{bmatrix}.$$

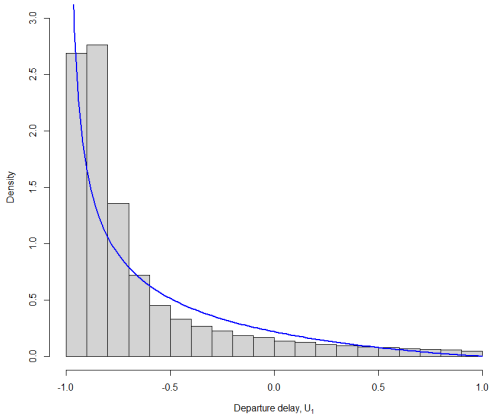
After estimating the remaining matrices in (4.8), i.e., covariance matrices of the centres and ranges, we obtained the following sample symbolic standard deviations: 10.22, 15.83, 75.25, and 574.45 for $i = 1, \ldots, 4$, respectively. Furthermore, the sample correlation matrix is

$$\widehat{\text{Cor}}_B(\boldsymbol{X}) = \begin{bmatrix} 1.00 & & & \\ 0.85 & 1.00 & & \\ -0.18 & -0.40 & 1.00 & \\ -0.17 & -0.39 & 0.99 & 1.00 \end{bmatrix}.$$

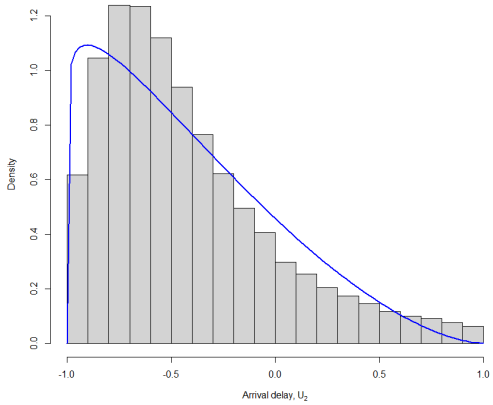
The sample correlation matrix anticipates that departure delays (x_1) and arrival delays (x_2), as well as time spent in the air (x_3) and distance between airports (x_4), are highly positively correlated (0.85 and 0.99, respectively). The remaining pairs of variables show low to moderate negative correlations. For example, in long-distance flights (or flights of longer duration), it is expected that the pilots can compensate for potential delays, resulting in lower arrival delays. This is expressed by the sample correlation of -0.39 (-0.40) between x_2 and x_4 (x_2 and x_3).

5.3 RTT dataset

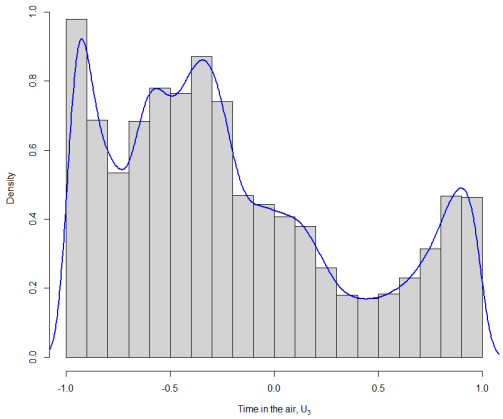
Salvador and Nogueira [43] introduced a framework to identify traffic redirection attacks, using a group of monitoring probes located across various geographic locations. These probes regularly measured the time it took for a set of 10 data packages to be sent to a target and return, the round-trip time (RTT). The aim of the study was to detect when the data packages relay through a third entity before reaching the target. This dataset was fully analysed in [44]. The intervals are built from each set of 10 data packages. In this



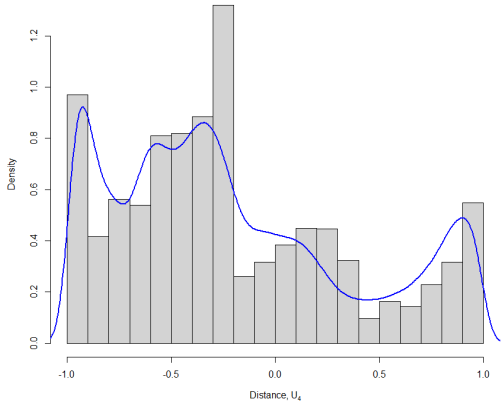
(a) Departure delay (U_1) and the pdf of $(2W_1 - 1)$, where $W_1 \sim \text{Beta}(0.44, 2.15)$ (parameters estimated by the method of moments).



(b) Arrival delay (U_2) and the pdf of $(2W_2 - 1)$, where $W_2 \sim \text{Beta}(1.08, 2.65)$ (parameters estimated by the method of moments).



(c) Time in the air (U_3) and the respective KDE of U_3.



(d) Distance (U_4) and the respective KDE of U_4.

Fig. 4: Histogram of the latent microdata related to each interval-valued variable of the New York City flights example. In the first two cases, the parameters of the beta distributions were estimated using the method of moments. In the other two cases, kernel density estimation was used.

example, we considered the target in Hong Kong, and eight monitoring probes, each corresponding to a variable ($p = 8$), located in Amsterdam (x_1), Chicago (x_2), Viña del Mar (x_3), Frankfurt (x_4), Hafnarfjordur (x_5), São Paulo (x_6), and two in Johannesburg, named Johannesburg1 (x_7) and Johannesburg2 (x_8). We only considered traffic redirected to the Madrid relay, resulting in $n = 564$ observations.

At each timestamp, h, and probe, i, only a few descriptive statistics were recorded, like the sample mean ($\bar{a}_{hi}$) and sample median ($\tilde{a}_{hi}$) together with the minimum ($a_{h,\min}$) and maximum ($a_{h,\max}$) values of the 10 RTT measures. The limited information about the microdata makes it impossible to fit any distribution to the U_i using the traditional methods. Nevertheless, Pearson's empirical "rule of thumb" allowed us to estimate the mode of each set of microdata, per timestamp and probe: $\text{mo}_{hi} = 3\tilde{a}_{hi} - 2\bar{a}_{hi}$, $h = 1, \ldots, n$, $i = 1, \ldots, p$. The sample means of the modes per probe $\hat{m}_i = \sum_{h=1}^{n} mo_{hi}/n$, $i = 1, \ldots, p$ are $(-0.14, -0.13, -0.34, -0.58, -0.69, -0.34, -0.17, -0.09)^T$, respectively. This suggests that the distributions of the latent variables are not symmetric. Additionally, the Kruskal-Wallis test concluded that the medians of the probes' modes are different. To test if each latent variable is symmetric we followed the recommendations in [45, pp. 245], and tested if the proportion of modes higher than zero is 1/2 (meaning its median is zero) using the exact test for a binomial proportion (called Clopper-Pearson test), with the Bonferroni correction. The results indicate medians are differ significantly from zero, except for the probes Chicago (x_2) and Johannesburg2 (x_8). Hence, we considered $m_2 = m_8 = 0$. For simplicity, we assumed that the latent

variables followed a triangular distribution, i.e., $U_i \sim \text{Triang}(-1, 1, m_i)$, where $m_2 = m_8 = 0$, and the other modes, m_i, were estimated by $\hat{m}_i$.

Under the assumption of a triangular distribution, $\text{E}(U_i) = m_i/3$ and $\text{Var}(U_i) = (m_i^2 + 3)/18$. As a result, $[\hat{\mathfrak{E}}_{UU}]_{ii} = \text{E}(U_i^2) = (m_i^2 + 3)/6$. Additionally, the quantile function of U_i can be written as

$$F_{U_i}^{-1}(t) = \begin{cases} -1 + \sqrt{2t(m_i+1)}, & 0 < t \leq \frac{m_i+1}{2} \\ 1 - \sqrt{2(1-t)(1-m_i)}, & \frac{m_i+1}{2} < t \leq 1 \end{cases}.$$

Without loss of generality, we can assume $m_i < m_j$, $i \neq j$, and

$$\begin{aligned} \mathcal{E}(U_i, U_j) = & \int_0^{\frac{1}{2}(m_i+1)} \left(-1 + \sqrt{2t(m_i+1)}\right)\left(-1 + \sqrt{2t(m_j+1)}\right) dt \\ & + \int_{\frac{1}{2}(m_i+1)}^{\frac{1}{2}(m_j+1)} \left(1 - \sqrt{2(1-t)(1-m_i)}\right)\left(-1 + \sqrt{2t(m_j+1)}\right) dt \\ & + \int_{\frac{1}{2}(m_j+1)}^{1} \left(1 - \sqrt{2(1-t)(1-m_i)}\right)\left(1 - \sqrt{2(1-t)(1-m_j)}\right) dt. \end{aligned}$$

For simplicity, the previous integrals were computed numerically using the routines from the R package *calculus* (see [42] for details), leading to the estimated matrix

$$\hat{\mathfrak{E}}_{UU} = \begin{bmatrix} 0.170 & & & & & & & \\ 0.167 & 0.167 & & & & & & \\ 0.168 & 0.167 & 0.186 & & & & & \\ 0.169 & 0.167 & 0.170 & 0.223 & & & & \\ 0.169 & 0.168 & 0.171 & 0.174 & 0.247 & & & \\ 0.168 & 0.167 & 0.169 & 0.170 & 0.171 & 0.186 & & \\ 0.167 & 0.167 & 0.168 & 0.169 & 0.170 & 0.168 & 0.172 & \\ 0.167 & 0.167 & 0.167 & 0.167 & 0.168 & 0.167 & 0.167 & 0.167 \end{bmatrix}.$$

Using (4.8), the previously estimated quantities resulted in the sample symbolic standard deviations: 9.44, 8.80, 56.35, 9.97, 11.50, 8.13, 8.37, and 8.34, for $i = 1, \ldots, 8$, and the sample correlation matrix

$$\widehat{\text{Cor}}_B(\boldsymbol{X}) = \begin{bmatrix} 1.00 & & & & & & & \\ 0.70 & 1.00 & & & & & & \\ 0.15 & 0.18 & 1.00 & & & & & \\ 0.72 & 0.85 & 0.22 & 1.00 & & & & \\ 0.61 & 0.72 & 0.19 & 0.75 & 1.00 & & & \\ -0.02 & -0.03 & -0.11 & -0.08 & -0.05 & 1.00 & & \\ 0.68 & 0.78 & 0.16 & 0.78 & 0.66 & -0.01 & 1.00 & \\ 0.66 & 0.77 & 0.16 & 0.79 & 0.67 & -0.05 & 0.80 & 1.00 \end{bmatrix}.$$

Notice that the probe located in São Paulo (x_6) exhibits a very small negative linear association with the other probes. This corroborates the findings in [44], where the authors suggest that this probe is not useful for the detection of Internet attacks. Additionally, note that the estimated correlations related to the probe in Viña del Mar (x_3) are much smaller in absolute value than the other probes (except São Paulo). The existence of an atypical behaviour of the probe in Viña del Mar was also detected and discussed in [44]. The remaining probes, x_1, x_2, x_4, x_5, x_7, and x_8, reveal a moderate (0.61) to high (0.85) positive correlation among themselves.

To conclude, we remark that even though the results allure to the symmetric triangular distribution, that is, the entries of $\hat{\mathfrak{E}}_{UU}$ look similar to 1/6 and the main diagonal of $\hat{\boldsymbol{\Psi}} = \text{Diag}(0.05, 0, -0.11, -0.19, -0.23, -0.11, -0.06, 0)$ is close to the zero vector, there is an added value in considering a non-symmetric triangular distribution. The Frobenius norm of the difference between $\widehat{\text{Cor}}_B(\boldsymbol{X})$ and the correlation matrix based on $\boldsymbol{S}_{CC} + \boldsymbol{S}_{RR}/24$ (under the assumption of a symmetric triangular distribution) is 0.174. The magnitude of this value accentuates the difference in choosing an assumption that uses partial information about the data as opposed to assuming the symmetric triangular distribution, which is a more simplified approach.

A common assumption amongst the SDA community is the uniform distribution, which assigns the same level of uncertainty to the microdata. The Frobenius norm of the difference between $\widehat{\mathrm{Cor}}_B(\boldsymbol{X})$ and the correlation matrix based on $\boldsymbol{S}_{CC} + \boldsymbol{S}_{RR}/12$ increases to 0.288.

6 Conclusions and discussion

In this work, we introduced the framework of interval data paired with the distribution of the microdata. Using the model proposed in [33], we started by scaling the microdata into the interval $[-1, 1]$, represented by the latent random variable $\tilde{U}_i$. This transformation has the merit of simplifying the theoretical derivations presented.

Based on it, we deduced explicit formulas for the Mallows' distance between two p-dimensional intervals, under very mild assumptions on the distribution of the microdata. In its most general form, the squared Mallows' distance can be decomposed into three terms: the squared Euclidean distance between the two vectors of the centres, the weighted squared Euclidean distance between the two vectors of the ranges, where the weights depend on the second moment of the latent random variables, and a third term that balances the contribution of the centres and ranges, weighted by the expected value of $\tilde{U}_i$. Assuming a symmetric distribution for the latent random variables eliminates this cross-term, turning the squared Mallows' distance into the sum of two squared Euclidean distances: one based on the distance between the centres and the other on the weighted distance between the ranges. In this case, the ranges' weights are quantities in $[0, 1/4]$, which accentuates the unbalanced contribution of the centres and ranges to the distance.

The general expression also allowed us to argue that the Mallows' distance between two hyperrectangles in $\mathbb{IR}^p$ with the same distribution of the latent random variables in each dimension can be expressed as a Mahalanobis' distance between two points in $\mathbb{R}^p \times (\mathbb{R}_0^+)^p$ composed by the vector of the centre and range combined. The associated covariance matrix $\boldsymbol{H}^{-1}$ is the inverse of a 2 by 2 block matrix; each of the blocks is a $p \times p$ diagonal matrix. These block matrices only depend on the first two moments of the latent random variables.

The closed form of the Mallows' distance led to generalising the definitions of the expected value and covariance matrix of an interval-valued random vector. The expected value is defined as the interval that minimises the expected value of the square of the Mallows' distance to the interval-valued random vector called population barycentre or Fréchet population mean. The minimum value of the function to be minimised is the Fréchet variance, which coincides with the trace of the deduced symbolic covariance matrix called the total variance. The deduction of the symbolic covariance matrix, based on the barycentre approach, highlights the contribution of the covariance between centres and ranges $\boldsymbol{\Sigma}_{CR}$. This is a novelty, since most of the works in SDA assume a symmetric distribution for the microdata, concealing the role of this matrix.

In practice, we may not have full information about the microdata, and even if we do, it may be difficult to fit a parametric distribution. Our examples illustrate the use of kernel density estimators to overcome this issue. Additionally, we discussed an example where only limited information about the microdata is available.

A Proof of Corollary 3.13

To prove that $\boldsymbol{H}^{-1}$ is a covariance matrix, assuming $\mathrm{Var}(\tilde{U}_i) > 0$, $i = 1, \ldots, p$, we need to show that it is a symmetric positive definite matrix. This is equivalent to showing that the $2p \times 2p$ matrix $\boldsymbol{H}$ is itself a symmetric positive definite matrix. Symmetry is easily seen from the definition of $\boldsymbol{H}$. It remains to show positive definiteness. Let $\boldsymbol{v} = (\boldsymbol{v}_1^T, \boldsymbol{v}_2^T)^T$ be a non-zero real vector, where $\boldsymbol{v}_j = (v_{j1}, \ldots, v_{jp})^T$, $j = 1, 2$, and $\boldsymbol{v} \neq \boldsymbol{0}$. We have

$$\begin{aligned} \boldsymbol{v}^T \boldsymbol{H} \boldsymbol{v} &= 2\boldsymbol{v}_1^T \boldsymbol{v}_1 + 2\boldsymbol{v}_2^T \boldsymbol{\Delta} \boldsymbol{v}_2 + 2\boldsymbol{v}_1^T \boldsymbol{\Psi} \boldsymbol{v}_2 \\ &= 2\sum_{i=1}^{p} v_{1i}^2 + \frac{1}{2}\sum_{i=1}^{p} v_{2i}^2 \,\mathrm{E}(\tilde{U}_i^2) + 2\sum_{i=1}^{p} v_{1i} v_{2i} \,\mathrm{E}(\tilde{U}_i). \end{aligned} \tag{A.1}$$

Considering $\mathrm{E}(\tilde{U}_i^2) = \mathrm{Var}(\tilde{U}_i) + \mathrm{E}(\tilde{U}_i)^2$ in (A.1), we obtain

$$\boldsymbol{v}^T \boldsymbol{H} \boldsymbol{v} = 2\sum_{i=1}^{p} \left(v_{1i} + \mathrm{E}(\tilde{U}_i) \frac{v_{2i}}{2} \right)^2 + \frac{1}{2}\sum_{i=1}^{p} v_{2i}^2 \mathrm{Var}(\tilde{U}_i). \tag{A.2}$$

We now show that $\boldsymbol{v}^T\boldsymbol{H}\boldsymbol{v} > 0$, when $\boldsymbol{v} \neq \boldsymbol{0}$. The first term in (A.2) is always non-negative, and the second is strictly positive, for $\boldsymbol{v}_2 \neq \boldsymbol{0}$ (assuming $\mathrm{Var}(\tilde{U}_i) > 0$, $i = 1, \ldots, p$). Hence, the inequality holds.
Suppose now that $\boldsymbol{v}_2 = \boldsymbol{0}$, that is, $\boldsymbol{v}^T\boldsymbol{H}\boldsymbol{v} = 2\boldsymbol{v}_1^T\boldsymbol{v}_1$. Since, by hypothesis, $\boldsymbol{v} \neq \boldsymbol{0}$, there exists at least one component of $\boldsymbol{v}_1$ different from zero, yielding $\boldsymbol{v}^T\boldsymbol{H}\boldsymbol{v} > 0$. Hence, $\boldsymbol{H}$ is positive definite.
Note that without the assumption of positive variance, one can only ascertain that $\boldsymbol{H}$ is positive semi-definite.

B Proof of results in Section 4

In this section, we prove the results formulated in Section 4.

B.1 Proof of Theorem 4.2

We firstly note that the function f defined in (4.1) is convex. This is due to the fact that the square of the Mallows' distance is a convex function and the expected value preserves convexity. Following [46], the Mallows' distance (3.13) is convex because $\boldsymbol{H}$ is a positive semi-definite matrix, as seen in Appendix A.
Now, since f is convex, any critical point is necessarily a global minimum. Hence, we find the points that make the partial derivatives of the objective function relative to $\boldsymbol{c}$ and $\boldsymbol{r}$ equal to zero. This leads to

$$\begin{cases} \dfrac{\partial}{\partial \boldsymbol{c}} f(\boldsymbol{c}, \boldsymbol{r}) = \boldsymbol{0} \\ \dfrac{\partial}{\partial \boldsymbol{r}} f(\boldsymbol{c}, \boldsymbol{r}) = \boldsymbol{0} \end{cases} \Leftrightarrow \begin{cases} -2\,\mathrm{E}\left(\boldsymbol{C} - \boldsymbol{c} + \dfrac{1}{2}\boldsymbol{\Psi}(\boldsymbol{R} - \boldsymbol{r})\right) = \boldsymbol{0} \\ -2\,\mathrm{E}\left(\boldsymbol{\Delta}(\boldsymbol{R} - \boldsymbol{r}) + \dfrac{1}{2}\boldsymbol{\Psi}(\boldsymbol{C} - \boldsymbol{c})\right) = \boldsymbol{0} \end{cases} \Leftrightarrow \begin{cases} \mathrm{E}\,(\boldsymbol{C} - \boldsymbol{c}) = -\dfrac{1}{2}\boldsymbol{\Psi}\,\mathrm{E}\,(\boldsymbol{R} - \boldsymbol{r}) \\ \left(\boldsymbol{\Delta} - \dfrac{1}{4}\boldsymbol{\Psi}^2\right)\mathrm{E}\,(\boldsymbol{R} - \boldsymbol{r}) = \boldsymbol{0} \end{cases}, \tag{B.1}$$

where $\boldsymbol{\Delta} - \boldsymbol{\Psi}^2/4 = \mathrm{Diag}\left(\mathrm{Var}(\tilde{U}_1), \ldots, \mathrm{Var}(\tilde{U}_p)\right)/4$ is a diagonal matrix that depends on the variances of the latent random variables. According to Definition 2.3, if $\mathrm{P}(R_i = 0) = 0$, $i = 1, \ldots, p$, the latent random variables are absolutely continuous, and the only solution to (B.1) is the hyperrectangle $\boldsymbol{\mu}_B = (\boldsymbol{\mu}_C^T, \boldsymbol{\mu}_R^T)^T \in \mathbb{IR}^p$.
Consider now the case where some of the components of $\boldsymbol{R} = (R_1, \ldots, R_p)^T$ are equal to 0 with probability 1. It follows from Definition 2.3 that, for these components, the corresponding latent random variables are degenerate. Since the barycentre is a realisation of the model, the ranges are zero in these components, and since (B.1) provides a specific solution for all other components, it remains true that $\boldsymbol{\mu}_B = (\boldsymbol{\mu}_C^T, \boldsymbol{\mu}_R^T)^T$.
The minimum value of the objective function (4.1), called Fréchet variance, is a non-negative real number. Thus, we can use the properties of the trace of a matrix to obtain:

$$\begin{aligned} V_F(\boldsymbol{\mu}_B) &= \mathrm{E}\left((\boldsymbol{C} - \boldsymbol{\mu}_C)^T(\boldsymbol{C} - \boldsymbol{\mu}_C) + (\boldsymbol{R} - \boldsymbol{\mu}_R)^T\boldsymbol{\Delta}(\boldsymbol{R} - \boldsymbol{\mu}_R) + (\boldsymbol{C} - \boldsymbol{\mu}_C)^T\boldsymbol{\Psi}(\boldsymbol{R} - \boldsymbol{\mu}_R)\right) \\ &= \mathrm{tr}\left(\mathrm{E}\left((\boldsymbol{C} - \boldsymbol{\mu}_C)(\boldsymbol{C} - \boldsymbol{\mu}_C)^T\right) + \boldsymbol{\Delta}\,\mathrm{E}\left((\boldsymbol{R} - \boldsymbol{\mu}_R)(\boldsymbol{R} - \boldsymbol{\mu}_R)^T\right)\right) \\ &+ \mathrm{tr}\left(\boldsymbol{\Psi}\,\mathrm{E}\left((\boldsymbol{R} - \boldsymbol{\mu}_R)(\boldsymbol{C} - \boldsymbol{\mu}_C)^T\right)\right) \\ &= \mathrm{tr}\left(\boldsymbol{\Sigma}_{CC} + \boldsymbol{\Delta}\boldsymbol{\Sigma}_{RR} + \boldsymbol{\Psi}\boldsymbol{\Sigma}_{RC}\right) = \mathrm{tr}\left(\boldsymbol{\Sigma}_{CC} + \boldsymbol{\Delta}\boldsymbol{\Sigma}_{RR} + \boldsymbol{\Sigma}_{CR}\boldsymbol{\Psi}\right), \end{aligned}$$

concluding the proof.

B.2 Proof of Corollary 4.5

According to Proposition 3.2, for $i = 1, 2$, we can write $F_{B_i}^{-1}(t) = \mu_{C_i} + \mu_{R_i} F_{\tilde{U}_{Bi}}^{-1}(t)/2$, where $\mu_{C_i} = \mathrm{E}(C_i)$ and $\mu_{R_i} = \mathrm{E}(R_i)$. Furthermore, we have $\mathrm{E}(U_i) = \int_0^1 F_{U_i}^{-1}(t)\,dt$ and $\mathcal{E}(U_1, U_2) = \int_0^1 F_{U_1}^{-1}(t) F_{U_2}^{-1}(t)\,dt$. Therefore,

$$\begin{aligned} \int_0^1 \left(G_1^{-1}(t) - F_{B_1}^{-1}(t)\right)\left(G_2^{-1}(t) - F_{B_2}^{-1}(t)\right) dt &= (C_1 - \mu_{C_1})(C_2 - \mu_{R_2}) + \\ \tfrac{1}{2}(C_1 - \mu_{C_1})(R_2 - \mu_{R_2})\,\mathrm{E}(U_2) &+ \tfrac{1}{2}(C_2 - \mu_{R_2})(R_1 - \mu_{R_1})\,\mathrm{E}(U_1) + \\ \tfrac{1}{4}(R_1 - \mu_{R_1})(R_2 - \mu_{R_2})\,\mathcal{E}(U_1, U_2), \end{aligned}$$

whose expected value is the required expression.
If U_1 and U_2 are identically distributed, then $\mathcal{E}(U_1, U_2) = \mathrm{E}(U_1^2)$, and we immediately obtain (4.7).

Datasets

The Credit Card dataset is available in Github at https://github.com/MROS13/MallowSymbCov. The New York city flights dataset is available in CRAN, at https://dx.doi.org/10.32614/CRAN.package.nycflights13. The RTT dataset will be shared on reasonable request to the corresponding author.

Funding

This work was supported by Fundação para a Ciência e Tecnologia, Portugal, through the grants UIDB/04621/2020, UIDB/04459/2020, and UIDP/04459/2020.

Acknowledgments

The authors would like to thank Dr Paulo Salvador and Dr Ana Subtil for sharing the RTT dataset used in one of the examples.

References

[1] Beranger, B., Lin, H., Sisson, S.: New models for symbolic data analysis. Adv. Data Anal. Classif., 1–41 (2022) https://doi.org/10.1007/s11634-022-00520-8

[2] Billard, L., Diday, E.: Clustering Methodology for Symbolic Data. John Wiley & Sons, Hoboken, NJ (2020). https://doi.org/10.1002/9781119010401

[3] Diday, E., Billard, L.: Symbolic Data Analysis: Conceptual Statistics and Data Mining. John Wiley & Sons, (2006). https://doi.org/10.1002/9780470090183

[4] Brito, P.: Symbolic Data Analysis: another look at the interaction of Data Mining and Statistics. WIREs Data Mining and Knowledge Discovery **4**(4), 281–295 (2014) https://doi.org/10.1002/widm.1133

[5] Bertrand, P., Goupil, F.: Descriptive statistics for symbolic data. In: Bock, H.-H., Diday, E. (eds.) Analysis of Symbolic Data. Studies in Classification, Data Analysis, and Knowledge Organization, pp. 106–124. Springer, (2000). https://doi.org/10.1007/978-3-642-57155-8_6

[6] Billard, L.: Sample covariance functions for complex quantitative data. In: Proceedings of World IASC Conference, Yokohama, Japan, pp. 157–163 (2008)

[7] Irpino, A., Verde, R.: Basic statistics for distributional symbolic variables: a new metric-based approach. Adv. Data Anal. Classif. **9**(2), 143–175 (2015) https://doi.org/10.1007/s11634-014-0176-4

[8] Le-Rademacher, J., Billard, L.: Symbolic covariance principal component analysis and visualization for interval-valued data. J. Comput. Graph. Statist. **21**(2), 413–432 (2012) https://doi.org/10.1080/10618600.2012.679895

[9] Wang, H., Guan, R., Wu, J.: CIPCA: Complete-information-based principal component analysis for interval-valued data. Neurocomputing **86**, 158–169 (2012) https://doi.org/10.1016/j.neucom.2012.01.018

[10] Oliveira, M.R., Vilela, M., Pacheco, A., Valadas, R., Salvador, P.: Extracting information from interval data using symbolic principal component analysis. Austrian Journal of Statistics **46**(3-4), 79–87 (2017) https://doi.org/10.17713/ajs.v46i3-4.673

[11] Girão Serrão, R., Oliveira, M.R., Oliveira, L.: Theoretical derivation of interval principal component analysis. Information Sciences **621**, 227–247 (2023) https://doi.org/10.1016/j.ins.2022.11.093

[12] De Carvalho, F.A.T., Lechevallier, Y.: Partitional clustering algorithms for symbolic interval data based on single adaptive distances. Pattern Recognition **42**(7), 1223–1236 (2009) https://doi.org/10.1016/j.patcog.2008.11.016

[13] Sato-Ilic, M.: Symbolic clustering with interval-valued data. Procedia Comput. Sci. **6**, 358–363 (2011) https://doi.org/10.1016/j.procs.2011.08.066

[14] Silva, A.P.D., Brito, P.: Discriminant analysis of interval data: An assessment of parametric and distance-based approaches. J. Classification **32**(3), 516–541 (2015) https://doi.org/10.1007/s00357-015-9189-8

[15] Queiroz, D.C.F., Souza, R.M.C.R., A. Cysneiros, F.J., Araújo, M.C.: Kernelized inner product-based discriminant analysis for interval data. Pattern Anal. Appl. **21**(3), 731–740 (2018) https://doi.org/10.1007/s10044-017-0601-3

[16] Dias, S., Brito, P., Amaral, P.: Discriminant analysis of distributional data via fractional programming. European J. Oper. Res. **294**(1), 206–218 (2021) https://doi.org/10.1016/j.ejor.2021.01.025

[17] A. Lima Neto, E., Cordeiro, G.M., Carvalho, F.: Bivariate symbolic regression models for interval-valued variables. J. Stat. Comput. Simul. **81**(11), 1727–1744 (2011) https://doi.org/10.1080/00949655.2010.500470

[18] Irpino, A., Verde, R.: Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein Distance. Adv. Data Anal. Classif. **9**, 81–106 (2015) https://doi.org/10.1007/s11634-015-0197-7

[19] Dias, S., Brito, P.: Off the beaten track: A new linear model for interval data. European J. Oper. Res. **258**(3), 1118–1130 (2017) https://doi.org/10.1016/j.ejor.2016.09.006

[20] Whitaker, T., Beranger, B., Sisson, S.: Logistic regression models for aggregated data. J. Comput. Graph. Statist. **30**, 1049–1067 (2021) https://doi.org/10.1080/10618600.2021.1895816

[21] Maia, A.L.S., Carvalho, F.d.A.T., Ludermir, T.B.: Forecasting models for interval-valued time series. Neurocomput. **71**(16-18), 3344–3352 (2008) https://doi.org/10.1016/j.neucom.2008.02.022

[22] Teles, P., Brito, P.: Modeling interval time series with space–time processes. Comm. Statist. Theory Methods **44**(17), 3599–3627 (2015) https://doi.org/10.1080/03610926.2013.782200

[23] Lin, W., González-Rivera, G.: Interval-valued time series models: Estimation based on order statistics exploring the Agriculture Marketing Service data. Comput. Statist. Data Anal. **100**, 694–711 (2016) https://doi.org/10.1016/j.csda.2015.07.008

[24] Lin, H., Caley, M., Sisson, S.: Estimating global species richness using symbolic data meta-analysis. Ecography **2022**, 05617 (2022) https://doi.org/10.1111/ecog.05617

[25] Alves, H., Brito, P., Campos, P.: Centrality measures in interval-weighted networks. J. Complex Netw. **10**(4) (2022) https://doi.org/10.1093/comnet/cnac031

[26] Ponti, A., Irpino, A., Candelieri, A., Bosio, A., Giordani, I., Archetti, F.: Network vulnerability analysis in wasserstein spaces. In: Simos, D.E., Rasskazova, V.A., Archetti, F., Kotsireas, I.S., Pardalos, P.M. (eds.) Learning and Intelligent Optimization, pp. 263–277. Springer, Cham (2022)

[27] Le-Rademacher, J., Billard, L.: Likelihood functions and some maximum likelihood estimators for symbolic data. J. Statist. Plann. Inference **141**(4), 1593–1602 (2011) https://doi.org/10.1016/j.jspi.2010.11.016

[28] Samadi, S.Y., Billard, L., Guo, J.-H., Xu, W.: Mle for the parameters of bivariate interval-valued models. Adv. Data Anal. Classif. (2023) https://doi.org/10.1007/s11634-023-00546-6

[29] Brito, P., Silva, A.P.D.: Modelling interval data with normal and skew-normal distributions. J. Appl. Stat. **39**(1), 3–20 (2012) https://doi.org/10.1080/02664763.2011.575125

[30] Silva, A.P.D., Filzmoser, P., Brito, P.: Outlier detection in interval data. Adv. Data Anal. Classif. **12**(3), 785–822 (2018) https://doi.org/10.1007/s11634-017-0305-y

[31] Zhang, X., Beranger, B., Sisson, S.: Constructing likelihood functions for interval-valued random variables. Scand. J. Stat. **47**, 1–35 (2020) https://doi.org/10.1111/sjos.12395

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

[33] Oliveira, M.R., Azeitona, M., Pacheco, A., Valadas, R.: Association measures for interval variables. Adv. Data Anal. Classif. **16**(3), 491–520 (2022) https://doi.org/10.1007/s11634-021-00445-8

[34] Pinheiro, D., Oliveira, M.R., Kravchenko, I., Oliveira, L.: Interval fisher's discriminant analysis and visualisation. Under preparation (2024)

[35] Cazes, P., Douzal, A., Diday, E., Schektman, Y.: Extensions de l'analyse en composantes principales à des données de type intervalle. Revue de Statistique Appliquée **XIV**, 5–24 (1997)

[36] Irpino, A., Romano, E.: Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation. In: Extraction et Gestion des Connaissances, EGC 2007, vol. E-9 (2007)

[37] Agueh, M., Carlier, G.: Barycenters in the Wasserstein space. SIAM J. Math. Anal. **43**(2), 904–924 (2011) https://doi.org/10.1137/100805741

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

[39] Silva, A.P.D., Brito, P.: MAINT.Data: Model and Analyse Interval Data. (2023). R package version 2.7.1. https://CRAN.R-project.org/package=MAINT.Data

[40] Silva, A.P.D., Brito, P., Filzmoser, P., Dias, J.G.: MAINT.Data: Modelling and Analysing Interval Data in R. The R Journal **13**(2), 336–364 (2021) https://doi.org/10.32614/RJ-2021-074

[41] Nagler, T., Vatter, T.: Kde1d: Univariate Kernel Density Estimation. (2024). R package version 1.0.7. https://CRAN.R-project.org/package=kde1d

[42] Guidotti, E.: calculus: High-dimensional numerical and symbolic calculus in R. Journal of Statistical Software **104**(5), 1–37 (2022) https://doi.org/10.18637/jss.v104.i05

[43] Salvador, P., Nogueira, A.: Customer-side detection of internet-scale traffic redirection. In: 2014 16th International Telecommunications Network Strategy and Planning Symposium (Networks), pp. 1–5 (2014). https://doi.org/10.1109/NETWKS.2014.6958532

[44] Subtil, A., Oliveira, M.R., Valadas, R., Salvador, P., Pacheco, A.: Detection of internet-wide traffic redirection attacks using machine learning techniques. IET Networks **12**(4), 179–195 (2023) https://doi.org/10.1049/ntw2.12085

[45] Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures, 5th edn. Chapman & Hall/CRC, (2011)

[46] Boyd, S.P., Vandenberghe, L.: Convex Optimization. Cambridge University Press, (2004). https://doi.org/10.1017/CBO9780511804441