WORKING PAPERS

N° 1585

October 2024

"ICS for complex data with application to outlier detection for density data objects"

Camille Mondon, Huong Thi Trinh, Anne Ruiz-Gazen and Christine Thomas-Agnan

Toulouse School of Economics

ICS for complex data with application to outlier detection for density data objects

Camille Mondon[a,*], Huong Thi Trinh[b], Anne Ruiz-Gazen[a], Christine Thomas-Agnan[a]

[a] *Toulouse School of Economics, Toulouse Capitole University, France*
[b] *Faculty of Mathematical Economics, Thuongmai University, Hanoi, Vietnam*

Abstract

ICS (Invariant coordinate selection) is a method aimed at dimension reduction as a preliminary step for clustering and outlier detection. It can be applied on multivariate or functional data. This work introduces a coordinate-free definition of ICS and extends the ICS method to distributional data. Indeed the inherent constraints of density functions imply a necessary adaptation of functional ICS. Our first achievement is a coordinate-free version of ICS within the framework of Hilbert spaces, assuming that the data lies almost surely in a finite dimensional subspace. Using the Bayes space framework tailored for density functions, we express the centred log-ratio of the density curves in a subspace of $L_0^2(a, b)$ of zero-integral spline functions and conduct ICS in this finite dimensional subspace. We describe the different steps of the procedure for outlier detection and study the impact of some parameters of this procedure on the results. The methodology is then illustrated on a sample of daily maximum temperatures densities recorded across northern Vietnamese provinces between 1987 and 2016.

Keywords: Bayes spaces, distributional data, functional data, invariant coordinate selection, outlier detection, Vietnam temperature densities
2020 MSC: 62H25, 62R10, 62G07, 65D07

1. Introduction

The invariant coordinate selection (ICS) method was introduced in a multivariate data analysis framework by Tyler et al. [29]. ICS is one of the dimension reduction methods that extend beyond Principal Component Analysis (PCA) and second moments. ICS seeks projection directions associated with the largest and/or smallest eigenvalues of the simultaneous diagonalisation of two scatter matrices [see 11, 16, for recent references]. This approach enables ICS to uncover underlying structures, such as outliers and clusters, that might be hidden in high-dimensional spaces. ICS is termed "invariant" because it produces components, linear combinations of the original features of the data, that remain invariant (up to their sign and some permutation) under affine transformations of the data, including translations, rotations and scaling. Moreover, Theorem 4 in Tyler et al. [29] demonstrates that, for a mixture of elliptical distributions, the projection directions of ICS associated with the largest or smallest eigenvalues usually generate the Fisher discriminant subspace, regardless of the chosen pair of scatter matrices and without prior knowledge of group assignments. Once the pair of scatter matrices is chosen, invariant components can be readily computed, and dimension reduction is achieved by selecting the components that reveal the underlying structure. Recent articles have examined in detail the implementation of ICS in a multivariate framework, focusing on objectives such as anomaly detection [6] or clustering [2]. These studies particularly address the choice of pairs of scatter matrices and the selection of relevant invariant components. Note that this idea of joint diagonalisation of scatter matrices is also used in the context of blind source separation and more precisely for Independent Component Analysis (ICA) which is a model-based approach as opposed to ICS [see 16, for more details].

A significant contribution of the present work is the formulation of a coordinate-free variant of ICS, assuming the data almost surely resides in a Euclidean subspace. This formulation allows ICS to be defined in a very general

*Camille Mondon. Email address: camille.mondon@tse-fr.eu

framework, facilitating its adaptation to more complex data than multivariate data. Examples of complex data for which there already exists an adaptation of ICS (or ICA) include compositional data [22], functional data [21, 10, for ICA] or multivariate functional data Archimbaud et al. [4]. Additionally, in this paper, we also extend the multivariate ICS method to distributional data, an application for which, to our knowledge, no adaptation of ICS currently exists. Note that a coordinate-free version of ICS was already mentioned in [29], in the discussion by Professor Mervyn Stone, who proposed to follow the approach by [26]. The response by Tyler and co-authors agreed that this could offer a theoretically elegant and concise view of the topic. A coordinate-free approach of ICA is proposed in Li et al. [10], but to our knowledge, no coordinate-free approach to ICS exists for a general Euclidean space like the one we derive here.

As mentioned above, a possible application of ICS is outlier detection. In the context of a small proportion of outliers, a complete detection procedure integrating a dimension reduction step based on the selection of invariant coordinates is described in [6]. This method, which is called ICSOutlier and which flags outlying observations, has been implemented for multivariate data in [15], and adapted to compositional data in [22], and multivariate functional data in [4]. We propose to extend and illustrate this detection procedure for density data. Detecting outliers is already challenging in a classical multivariate context because outliers may differ from the other observations in their correlation pattern [see 1, for an overview on outlier detection and analysis]. [6] demonstrate how the ICS procedure outperforms the Mahalanobis distance and PCA (non-robust and robust). For compositional data, the constraints of positivity and constant sum of the components must be taken into account as detailed in [22] and further examined in this paper. For univariate functional data, outliers are categorized as either magnitude or shape outliers, with shape outliers being more challenging to detect because they are hidden among the other curves. Many existing detection methods for functional data rely on depth measurements, including the Mahalanobis distance [see, e.g., the recent paper 7, and the included references]. Density data are constrained function data, and thus combine the challenges associated with both compositional and functional data. The literature on outlier detection for density data is very sparse and recent with, as far as we know, the papers by [13], [9], [14] only. Two types of outliers exist for density data: the horizontal-shift outliers and the shape outliers, with shape outliers being again more challenging to detect [see 9, for details]. The procedure proposed by [13] is based on an adapted version of functional PCA to density objects in a control chart context. [9] proposes a transformation tree approach that incorporates many different outlier detection methods adapted to densities. The density data are generally transformed to an unconstrained data type and functional outlier detection methods are used. The objective of the authors is to propose a robust distribution regression method. [14] continue the work of the previously cited article by comparing more methods through simulations, and give an application to gas transport data. ICS is not mentioned in these references. The coordinate-free definition of ICS we propose allows to directly apply the ICSOutlier method to density data in a context of a small proportion of outliers.

Section 2 presents the coordinate-free version of ICS and highlights its significance with examples. From Section 3, we focus on the application to density objects. After recalling some basic facts on the theory of Bayes spaces, we detail the preprocessing method used (maximum penalized likelihood) to transform the original samples of real-valued data into a sample of density functions. We describe the finite dimensional subspaces of densities generated by specific spline functions adapted to density objects. Section 4 describes the ICS-based outlier detection procedure adapted to density objects and discusses the impact of the preprocessing parameters on outlier detection through a toy example. Section 5 provides an application of the outlier detection methodology to Vietnam maximum temperature data over 30 years. Section 6 concludes the paper and offers some perspectives. The proofs of the propositions and corollaries are given in the appendix.

2. Coordinate-free presentation of ICS

Invariant coordinate selection (ICS) is a data exploration method that solves a simultaneous reduction problem of two scatter operators in order to reveal interesting projections that highlight the lack of ellipticity of the distribution of the data. ICS has been defined by Tyler et al. [29] in the Euclidean space $E = \mathbb{R}^p$ using the coordinates in the canonical basis. However, in order to apply ICS to more general objects such as random densities, we propose the following generalisation to any Euclidean space E without relying on a particular choice of basis. A coordinate-free approach to ICS was suggested by Professor Mervyn Stone in the discussion section of Tyler et al. [29] and applied to Independent Component Analysis by Li et al. [10].

2.1. Scatter operators

Let us first discuss some definitions relative to scatter operators in the framework of a general Euclidean space $(E, \langle\cdot,\cdot\rangle)$. We consider E-valued random variable $X : \Omega \to E$ where Ω is a probability space and E is a Euclidean space equipped with the Borel σ-algebra. In order to define ICS, we need at least two scatter operators, which generalise the covariance operator defined on E by

$$\forall (x, y) \in E^2, \langle \mathrm{Cov}[X]x, y\rangle = \mathbb{E}\left[\langle X - \mathbb{E}X, x\rangle\langle X - \mathbb{E}X, y\rangle\right], \tag{1}$$

while keeping its affine equivariance property

$$\forall A \in \mathcal{GL}(E), \forall b \in E, \mathrm{Cov}[AX + b] = A\,\mathrm{Cov}[X]A^*,$$

where the Euclidean norm of X is assumed to be square-integrable, $\mathcal{GL}(E)$ denotes the group of linear automorphisms of E and A^* is the adjoint linear operator of A in the Euclidean space E, represented by the transpose of the matrix that represents A.

Definition 1 (Scatter operators)**.** Let $(E, \langle\cdot,\cdot\rangle)$ be a Euclidean space of dimension p, $\mathcal{E}$ an affine invariant set of E-valued random variables, i.e. that verifies:

$$\forall X \in \mathcal{E}, \forall A \in \mathcal{GL}(E), \forall b \in E, AX + b \in \mathcal{E}. \tag{2}$$

An operator $S : \mathcal{E} \to \mathcal{S}^+(E)$ (where $\mathcal{S}^+(E)$ is the space of non-negative symmetric operators on E) is called an (affine equivariant) scatter operator (defined on $\mathcal{E}$) if it satisfies the following two properties:

1. Invariance by equality in distribution:
$$\forall (X, Y) \in \mathcal{E}^2, X \sim Y \Rightarrow S[X] = S[Y]. \tag{3}$$
2. Affine equivariance:
$$\forall X \in \mathcal{E}, \forall A \in \mathcal{GL}(E), \forall b \in E, S[AX + b] = AS[X]A^*, \tag{4}$$
where $\mathcal{GL}(E)$ denotes the group of linear automorphisms of E.

Let $L^p(\Omega, E)$ denote the space of E-valued random variables whose Euclidean norm's p-th power is integrable. If A is a linear operator, we denote by $A^{1/2}$ the unique non-negative square root of A.

Definition 2 (Weighted covariance operators)**.** For any measurable function $w : \mathbb{R}^+ \to \mathbb{R}$, let

$$\mathcal{E}_w = \left\{ X \in L^1(\Omega, E) \,\middle|\, \mathrm{Cov}[X] \in \mathcal{GL}(E) \text{ and } w\left(\left\|\mathrm{Cov}[X]^{-1/2}(X - \mathbb{E}X)\right\|\right)\|X - \mathbb{E}X\| \in L^2(\Omega, \mathbb{R}) \right\}. \tag{5}$$

Note that $\mathcal{E}_w$ is an affine invariant set of integrable E-valued random variables. For $X \in \mathcal{E}_w$, we define the w-weighted covariance operator $\mathrm{Cov}_w[X]$ by

$$\forall (x, y) \in E^2, \langle \mathrm{Cov}_w[X]x, y\rangle = \mathbb{E}\left[w^2\left(\left\|\mathrm{Cov}[X]^{-1/2}(X - \mathbb{E}X)\right\|\right)\langle X - \mathbb{E}X, x\rangle\langle X - \mathbb{E}X, y\rangle\right]. \tag{6}$$

When it necessary, we will also write Cov_w^E for the w-weighted covariance operator on E to avoid any ambiguity. It is easy to verify that weighted covariance operators are affine equivariant scatter operators.

Example 1.

- If $w = 1$, we retrieve Cov, the usual covariance operator on $L^2(\Omega, E)$.
- If for $x \in \mathbb{R}^+$, $w(x) = (p + 2)^{-1/2}x$, we obtain the fourth-order moment covariance operator Cov_4 [as in 16, for the case $E = \mathbb{R}^p$] on $\mathcal{E}_w = \left\{X \in L^4(\Omega, E) \,|\, \mathrm{Cov}[X] \in \mathcal{GL}(E)\right\}$.

2.2. The ICS problem

We define a coordinate-free version of the ICS problem in a Euclidean space E. This general framework ensures that ICS does not depend on any particular choice of basis of E to represent the E-valued random object X. Then we prove a proposition that will allow us to relate coordinate-free ICS to multivariate ICS applied to the coordinate vectors in a basis of E. Our practical implementation of coordinate-free ICS will be based on this result.

In what follows, $\delta_{jj'} = \begin{cases} 1 & \text{if } j = j' \\ 0 & \text{if } j \neq j' \end{cases}$ is the Kronecker delta of two integers j, j'.

Definition 3 (Coordinate-free ICS)**.** Let $(E, \langle \cdot, \cdot \rangle)$ be a Euclidean space of dimension p, $\mathcal{E} \subseteq L^1(\Omega, E)$ an affine invariant set of integrable E-valued random variables, and S_1 and S_2 two scatter operators on $\mathcal{E}$.

For any $X \in \mathcal{E}$, any basis $H = (h_1, \ldots, h_p)$ of E, and any finite non-increasing real sequence $\Lambda = (\lambda_1 \geq \ldots \geq \lambda_p)$, we define invariant coordinate selection (ICS) as the following problem:

$$\mathrm{ICS}(X, S_1, S_2) : \begin{cases} \langle S_1[X]h_j, h_{j'} \rangle &= \delta_{jj'} \text{ for all } 1 \leq j, j' \leq p \\ \langle S_2[X]h_j, h_{j'} \rangle &= \lambda_j \delta_{jj'} \text{ for all } 1 \leq j, j' \leq p. \end{cases} \tag{7}$$

Such a basis H is called an $\mathrm{ICS}(X, S_1, S_2)$ eigenbasis, whose elements are $\mathrm{ICS}(X, S_1, S_2)$ eigenobjects. Such a Λ is called an $\mathrm{ICS}(X, S_1, S_2)$ spectrum, whose elements are called $\mathrm{ICS}(X, S_1, S_2)$ eigenvalues or generalised kurtosis. Given an $\mathrm{ICS}(X, S_1, S_2)$ eigenbasis H and $1 \leq j \leq p$, the real number $z_j = \langle X - \mathbb{E}X, h_j \rangle$ is called the j-th invariant coordinate (in the eigenbasis H).

Intuitively, the objective of invariant coordinate selection is to find a basis that simplifies the comparison between the non-negative symmetric operators $S_1[X]$ and $S_2[X]$. In Chapter 15 of [17], this pair $(S_1[X], S_2[X])$ is called a pencil and the problem (7) is a general (or generalised) eigenvalue problem. Because the linear operators in question are symmetric, this is equivalent to simultaneous reduction to diagonal form.

In the ICS problem (7), the scatter operators S_1 and S_2 do not exactly play symmetrical roles. This is because in order to prove the existence of solutions in Proposition 1, we will assume that the first scatter operator evaluated at X is a bijection.

Remark 1 (Multivariate case)**.** If $E = \mathbb{R}^p$, we identify S_1 and S_2 with their associated $(p \times p)$-matrices in the canonical basis, and we identify an ICS eigenbasis H with the $(p \times p)$-matrix of its vectors stacked column-wise, so that we retrieve the classical formulation of invariant coordinate selection by [29].

Remark 2 (Courant-Fischer variational principle)**.** The ICS problem (7) can be stated as a maximisation problem. If $1 \leq j \leq p$, the following equalities hold:

$$h_j = \operatorname*{argmax}_{h \in E, \langle S_1[X]h, h_{j'} \rangle = 0 \text{ if } 0 < j' < j} \frac{\langle S_2[X]h, h \rangle}{\langle S_1[X]h, h \rangle} \text{ and } \lambda_j = \max_{h \in E, \langle S_1[X]h, h_{j'} \rangle = 0 \text{ if } 0 < j' < j} \frac{\langle S_2[X]h, h \rangle}{\langle S_1[X]h, h \rangle}. \tag{8}$$

Whenever $S_1[X]$ is an automorphism, the $\mathrm{ICS}(X, S_1, S_2)$ problem boils down to finding an orthonormal basis H that diagonalises the symmetric operator $S_1[X]^{-1}S_2[X]$ in the Euclidean space $(E, \langle S_1[X] \cdot, \cdot \rangle)$. The $\mathrm{ICS}(X, S_1, S_2)$ spectrum Λ is unique and is simply the spectrum of $S_1[X]^{-1}S_2[X]$.

Proposition 1 (Existence of solutions)**.** *Let $(E, \langle \cdot, \cdot \rangle)$ be a Euclidean space of dimension p, $\mathcal{E} \subseteq L^1(\Omega, E)$ an affine invariant set of integrable E-valued random variables, S_1 and S_2 two scatter operators on $\mathcal{E}$.*

For any $X \in \mathcal{E}$ such that $S_1[X]$ is an automorphism, there exists at least one solution (H, Λ) to the $\mathrm{ICS}(X, S_1, S_2)$ problem, and Λ is a uniquely determined non-increasing sequence of positive real numbers.

In the following reconstruction formula, we will denote $H^* = (h_j^*)_{1 \leq j \leq p}$ the dual basis of a basis $H = (h_1, \ldots, h_p)$ of E, i.e. the only basis of the Euclidean space E that satisfies

$$\langle h_j, h_{j'}^* \rangle = \delta_{jj'} \text{ for all } 1 \leq j, j' \leq p. \tag{9}$$

Proposition 2 (Reconstruction formula)**.** *Let $(E, \langle\cdot,\cdot\rangle)$ be a Euclidean space of dimension p, $\mathcal{E} \subseteq L^1(\Omega, E)$ an affine invariant set of integrable E-valued random variables, S_1 and S_2 two scatter operators on $\mathcal{E}$.*

For any $X \in \mathcal{E}$ such that $S_1[X]$ is an automorphism and any $\mathrm{ICS}(X, S_1, S_2)$ *eigenbasis $H = (h_1, \ldots, h_p)$ of E, we have*

$$X = \mathbb{E}X + \sum_{j=1}^{p} z_j h_j^*, \tag{10}$$

where $H^* = (h_j^*)_{1\le j\le p} = (S_1[X]h_j)_{1\le j\le p}$ is the dual basis of H.

Another way to understand the coordinate-free nature of this ICS problem is to work with data isometrically represented in two spaces and to understand how we can relate a given ICS problem in the first space to an ICS problem in the second. This is the object of the following proposition.

Proposition 3. *Let $(E, \langle\cdot,\cdot\rangle_E) \xrightarrow{\varphi} (F, \langle\cdot,\cdot\rangle_F)$ be an isometry between two Euclidean spaces of dimension p, $\mathcal{E} \subseteq L^1(\Omega, E)$ an affine invariant set of integrable E-valued random variables, $S_1^{\mathcal{E}}$ and $S_2^{\mathcal{E}}$ two affine equivariant scatter operators on $\mathcal{E}$. Then:*

(a) *$\mathcal{F} = \varphi(\mathcal{E}) = \{\varphi(X^{\mathcal{E}}), X^{\mathcal{E}} \in \mathcal{E}\}$ is an affine invariant set of integrable F-valued random variables, and we denote $X^{\mathcal{F}} = \varphi(X^{\mathcal{E}}) \in \mathcal{F}$ whenever $X^{\mathcal{E}} \in \mathcal{E}$;*

(b) *$S_\ell^{\mathcal{F}} : X^{\mathcal{F}} \in \mathcal{F} \mapsto \varphi \circ S_\ell^{\mathcal{E}}[X^{\mathcal{E}}] \circ \varphi^{-1}$, $\ell \in \{1, 2\}$, are two affine equivariant scatter operators on $\mathcal{F}$;*

(c) *$H^{\mathcal{F}} = \varphi(H^{\mathcal{E}}) = (\varphi(h_1^{\mathcal{E}}), \ldots, \varphi(h_p^{\mathcal{E}}))$ is a basis of F whenever $H^{\mathcal{E}} = (h_1^{\mathcal{E}}, \ldots, h_p^{\mathcal{E}})$ is a basis of E.*

For any E-valued random variable $X^{\mathcal{E}} \in \mathcal{E}$, any basis $H^{\mathcal{E}} = (h_1^{\mathcal{E}}, \ldots, h_p^{\mathcal{E}})$ of E, and any finite non-increasing real sequence $\Lambda = (\lambda_1 \ge \ldots \ge \lambda_p)$ the following assertions are equivalent:

(i) *$(H^{\mathcal{E}}, \Lambda)$ solves* $\mathrm{ICS}(X^{\mathcal{E}}, S_1^{\mathcal{E}}, S_2^{\mathcal{E}})$ *in the space E*

(ii) *$(H^{\mathcal{F}}, \Lambda)$ solves* $\mathrm{ICS}(X^{\mathcal{F}}, S_1^{\mathcal{F}}, S_2^{\mathcal{F}})$ *in the space F.*

The following corollary is an important application to the case of weighted covariance operators $S_\ell^{\mathcal{E}} = \mathrm{Cov}_{w_\ell}$, $\ell \in \{1, 2\}$, for which the definition of $S_\ell^{\mathcal{F}}$ coincides with that of weighted covariance operators on F.

Corollary 1. *Let $(E, \langle\cdot,\cdot\rangle_E) \xrightarrow{\varphi} (F, \langle\cdot,\cdot\rangle_F)$ be an isometry between two Euclidean spaces of dimension p and $w_1, w_2 : \mathbb{R}^+ \to \mathbb{R}$ two measurable functions. For any integrable E-valued random variable $X \in \mathcal{E}_{w_1} \cap \mathcal{E}_{w_2}$ (with the notations from Definition 2), the equality*

$$\mathrm{Cov}_{w_\ell}^F[\varphi(X)] = \varphi \circ \mathrm{Cov}_{w_\ell}^E[X] \circ \varphi^{-1} \tag{11}$$

holds for $\ell \in \{1, 2\}$, as well as the equivalence between the following assertions, for any basis $H = (h_1, \ldots, h_p)$ of E, and any finite non-increasing real sequence $\Lambda = (\lambda_1 \ge \ldots \ge \lambda_p)$:

(i) *(H, Λ) solves* $\mathrm{ICS}(X, \mathrm{Cov}_{w_1}^E, \mathrm{Cov}_{w_2}^E)$ *in the space E.*

(ii) *$(\varphi(H), \Lambda)$ solves* $\mathrm{ICS}(\varphi(X), \mathrm{Cov}_{w_1}^F, \mathrm{Cov}_{w_2}^F)$ *in the space F.*

In order to implement coordinate-free ICS, we will represent the elements of E by their coordinates in a basis B of E. For a basis $B = (b_1, \ldots, b_p)$ of E, let us denote by $[x]_B = ([x]_{b_1} \ldots [x]_{b_p})^\top$ the coordinate vector of $x \in E$ in the basis B and $G_B = (\langle b_j, b_{j'}\rangle)_{1\le j,j'\le p}$ the Gram matrix of the basis B. Then, Corollary 1, in the special case where φ is the isometry

$$\varphi_B : \left\{ \begin{array}{ccc} E & \to & \mathbb{R}^p \\ x & \mapsto & G_B^{1/2}[x]_B \end{array} \right. \tag{12}$$

allows one to relate the coordinate-free approach in E to the multivariate approach applied to the coordinate vectors in any basis of E, when we consider weighted covariance operators (in $\mathbb{R}^p$, they correspond to weighted covariance matrices of the coordinates). This is made clear by the following corollary.

Corollary 2. *Let $(E, \langle\cdot,\cdot\rangle)$ be a Euclidean space of dimension p, $w_1, w_2 : \mathbb{R}^+ \to \mathbb{R}$ two measurable functions. Let B be any basis of E, $G_B = (\langle b_j, b_{j'}\rangle)_{1\le j,j'\le p}$ its Gram matrix and $[\cdot]_B$ the linear map giving the coordinates in B.*

For any $X \in \mathcal{E}_{w_1} \cap \mathcal{E}_{w_2}$ (with the notations from Definition 2), any basis $H = (h_1, \ldots, h_p)$ of E, and any finite non-increasing real sequence $\Lambda = (\lambda_1 \ge \ldots \ge \lambda_p)$ the following assertions are equivalent:

(1) (H, Λ) *solves* $\mathrm{ICS}(X, \mathrm{Cov}_{w_1}^E, \mathrm{Cov}_{w_2}^E)$ *in the space* E
(2) $(G_B^{1/2}[H]_B, \Lambda)$ *solves* $\mathrm{ICS}(G_B^{1/2}[X]_B, \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ *in the space* $\mathbb{R}^p$
(3) $([H]_B, \Lambda)$ *solves* $\mathrm{ICS}(G_B[X]_B, \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ *in the space* $\mathbb{R}^p$
(4) $(G_B[H]_B, \Lambda)$ *solves* $\mathrm{ICS}([X]_B, \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ *in the space* $\mathbb{R}^p$

where $[H]_B$ *denotes the non-singular* $p \times p$ *matrix representing the basis* $([h_1]_B, \ldots, [h_p]_B)$ *of* $\mathbb{R}^p$.

In order to reconstruct the original random object, we need the coordinates of the elements of the dual ICS eigenbasis. Identifying the basis $[H]_B$ with the matrix whose columns are its vectors, the dual basis $[H^*]_B$ is the matrix

$$[H^*]_B = \left([H]_B^\top G_B\right)^{-1}. \tag{13}$$

Remark 3 (Empirical ICS and estimation)**.** We can study the particular case of a finite E-valued random variable X where we have a fixed sample $D_n = (x_1, \ldots, x_n)$ and we assume that X follows the empirical probability distribution of $(x_1, \ldots, x_n)$ denoted by

$$P_{D_n} = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(\cdot). \tag{14}$$

In that case, the expressions (in Definition 2 for instance) of the form $\mathbb{E}f(X)$ for any function f are discrete and equal to $\frac{1}{n}\sum_{i=1}^n f(x_i)$.

Now, let us assume that we observe an i.i.d. sample $D_n = (X_1, \ldots, X_n)$ following the distribution of an unknown E-valued random variable X_0. We can estimate solutions to the $\mathrm{ICS}(X_0, S_1, S_2)$ problem from Definition 3 by working conditionally on the data $(X_1, \ldots, X_n)$ and taking the particular case where X follows the empirical probability distribution P_{D_n}. This defines estimates of the $\mathrm{ICS}(X_0, S_1, S_2)$ eigenobjects as solutions of an ICS problem involving empirical scatter operators.

Example 2 (Applications)**.**

1. **Compositional data** [22]: Let $E = (\mathcal{S}^{p+1}, \oplus, \odot, \langle\cdot,\cdot\rangle_{\mathcal{S}^{p+1}})$ denote the simplex of dimension p in $\mathbb{R}^{p+1}$ with the Aitchison structure [18].
   In Section 5.1, [22] propose a first definition of $\mathrm{ICS}(X, \mathrm{Cov}, \mathrm{Cov}_4)$ for a random composition $X \in \mathcal{S}^{p+1}$ by (i) choosing a contrast matrix V, (ii) defining the problem $\mathrm{ICS}(\mathrm{ilr}_V(X), \mathrm{Cov}, \mathrm{Cov}_4)$ in $\mathbb{R}^p$ using the approach of [29] and (iii) proving that this problem does not rely on V [for a presentation of contrast matrices and ilr transformations, see 18]. We can easily retrieve and generalise this result in the light of Corollary 1, applied to the isometry $\varphi = \mathrm{ilr}_V$ that gives the equivalence between these two ICS problems:
   (a) (H, Λ) solves $\mathrm{ICS}(X, \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ in the space $\mathcal{S}^{p+1}$
   (b) $(\mathrm{ilr}_V(H), \Lambda)$ solves $\mathrm{ICS}(\mathrm{ilr}_V(X), \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ in the space $\mathbb{R}^p$
   for any measurable functions $w_1, w_2 : \mathbb{R}^+ \to \mathbb{R}$.
   In Section 5.2, [22] propose to work in the zero-sum hyperplane $\mathcal{H}^{p+1} = \{x \in \mathbb{R}^{p+1} | \sum_{j=1}^{p+1} x_j = 0\}$ of $\mathbb{R}^{p+1}$ using the centred log-ratio isometry

$$\mathrm{clr} : \begin{cases} (\mathcal{S}^{p+1}, \oplus, \odot, \langle\cdot,\cdot\rangle_{\mathcal{S}^{p+1}}) & \to (\mathcal{H}^{p+1}, +, \cdot, \langle\cdot,\cdot\rangle_{\mathbb{R}^{p+1}}) \\ x = (x_1, \ldots, x_{p+1})^\top & \mapsto \left(\log(x_1) - \frac{1}{p+1}\sum_{j=1}^{p+1}\log(x_j), \ldots, \log(x_{p+1}) - \frac{1}{p+1}\sum_{j=1}^{p+1}\log(x_j)\right)^\top \end{cases} \tag{15}$$

   in order to establish a definition of $\mathrm{ICS}(X, \mathrm{Cov}, \mathrm{Cov}_4)$ directly in the simplex. Again, Corollary 1 applied to the isometry $\varphi = \mathrm{clr}$ gives the equivalence between the two following ICS problems:
   (a) (H, Λ) solves $\mathrm{ICS}(X, \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ in the space $\mathcal{S}^{p+1}$
   (b) $(\mathrm{clr}(H), \Lambda)$ solves $\mathrm{ICS}(\mathrm{clr}(X), \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ in the space $\mathcal{H}^{p+1}$.
   A third way to characterise $\mathrm{ICS}(X, \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ is to choose a basis of the simplex $\mathcal{S}^{p+1}$ and apply Corollary 2. For a given index $1 \leq j \leq p$, let $B_j = (b_1, \ldots, b_p)$ denote the basis of $\mathcal{S}^{p+1}$ corresponding to the alr_j

transformation, i.e. obtained by taking the canonical basis of $\mathbb{R}^{p+1}$, removing the j-th vector and applying the exponential. In that case, it is easy to compute the $p \times p$ Gram matrix of B_j

$$G_{B_j} = I_p - \frac{1}{p+1}\mathbf{1}_p\mathbf{1}_p^\top = \begin{pmatrix} 1-\frac{1}{p+1} & -\frac{1}{p+1} & \cdots & -\frac{1}{p+1} \\ -\frac{1}{p+1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\frac{1}{p+1} \\ -\frac{1}{p+1} & \cdots & -\frac{1}{p+1} & 1-\frac{1}{p+1} \end{pmatrix}. \tag{16}$$

Then, from Corollary 2 we get the equivalence between the two following ICS problems:

(a) (H, Λ) solves $\mathrm{ICS}(X, \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ in the space $\mathcal{S}^{p+1}$

(b) $(\mathrm{alr}_j(H), \Lambda)$ solves $\mathrm{ICS}(\mathrm{clr}(X)^{(j)}, \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ in the space $\mathbb{R}^p$

where $\mathrm{clr}(x)^{(j)} = G_{B_j}\,\mathrm{alr}_j(x)$ is centred log-ratio transformation of $x \in \mathcal{S}^{p+1}$ from which the j-th coordinate has been removed. This suggests a new implementation of invariant coordinate selection for compositional data, in an unconstrained space and only having to choose an index j instead of a full contrast matrix.

2. **Functional data** [4, 31]: Let $B = (b_j)_{j\geq 1}$ be a Hilbert basis of $L^2(a, b)$, for instance a Fourier basis or a Hermite polynomial basis. For a given integer p, let us consider the space E spanned by the truncated orthonormal basis $B_p = (b_1, \ldots, b_p)$. Then for any E-valued random variable X, we get the equivalence between the following two conditions:

   (a) (H, Λ) solves $\mathrm{ICS}(X, \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ in the space E

   (b) $([H]_B, \Lambda)$ solves $\mathrm{ICS}([X]_B, \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ in the space $\mathbb{R}^p$.

   We could also choose a B-spline basis, but then we should take into account its Gram matrix.

3. **Density data** (see Section 3 for further details): Let $E = C_d^{\Delta\gamma}(a, b)$ be a space of compositional splines on (a, b) of order $d + 1$ and with knots $\Delta\gamma = (\gamma_1, \ldots, \gamma_k)$. We can use either the centred log-ratio isometry between $C_d^{\Delta\gamma}(a, b)$ and the corresponding zero-integral spline space $\mathcal{Z}_d^{\Delta\gamma}(a, b)$ or the CB-spline basis of $C_d^{\Delta\gamma}(a, b)$ to obtain characterisations of the problem $\mathrm{ICS}(X, \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ for a random density $X \in C_d^{\Delta\gamma}(a, b)$.

3. ICS for density objects

3.1. Reminder on Bayes spaces

The most recent and complete description of the Bayes spaces approach can be found in Van Den Boogaart et al. [30]. For the present work, we will identify the elements of a Bayes space, as defined by Van Den Boogaart et al. [30], with their Radon–Nikodym derivative with respect to a reference measure λ. This leads to the following framework: let (a, b) be a given interval of the real line equipped with the Borel σ-algebra, let λ be a finite reference measure on (a, b). Let $B^2(a, b)$ be the space of square-log integrable probability densities $\frac{d\mu}{d\lambda}$, where μ is a measure that is equivalent to λ, which means that μ and λ are absolutely continuous with respect to each other. Let us first briefly remind the construction of the Hilbert space structure of $B^2(a, b)$.

For a density f in $B^2(a, b)$, the clr transformation is defined by

$$\mathrm{clr}\, f(.) = \log f(.) - \frac{1}{\lambda(a, b)} \int_a^b \log f(t) d\lambda(t). \tag{17}$$

The clr transformation maps an element of $B^2(a, b)$ into an element of the space $L_0^2(a, b)$ of functions which are square integrable with respect to λ on (a, b) and whose integral is equal to zero. The clr inverse of a function u of $L_0^2(a, b)$ is B^2-equivalent to $\exp(u)$. More precisely, if $u \in L_0^2(a, b)$,

$$\mathrm{clr}^{-1}(u)(.) = \frac{\exp u(.)}{\int_a^b \exp u(t) d\lambda(t)}. \tag{18}$$

A vector space structure on $B^2(a, b)$ is readily obtained by transporting the vector space structure of $L_0^2(a, b)$ to $B^2(a, b)$ using the clr transformation and its inverse, see for example Van Den Boogaart et al. [30]. Its operations, denoted by $\oplus$ and $\odot$, are called perturbation (the "addition") and powering (the "scalar multiplication").

For the definition of the inner product, we adopt a normalization different from that of Egozcue et al. [8] and of Van Den Boogaart et al. [30] in the sense that we choose the classical definition of inner product in $L_0^2(a,b)$, for two functions u and v in $L_0^2(a,b)$

$$\langle u, v\rangle_{L_0^2} = \int_a^b u(t)v(t)d\lambda(t), \tag{19}$$

so that the corresponding inner product between two densities f and g in the Bayes space $B^2(a,b)$ is given by

$$\langle f, g\rangle_{B^2} = \frac{1}{2\lambda(a,b)} \int_a^b \int_a^b (\log f(t) - \log f(s))(\log g(t) - \log g(s))d\lambda(t)d\lambda(s). \tag{20}$$

This normalization yields an inner product which is homogeneous to the measure λ whereas the Van Den Boogaart et al. [30] normalization is unitless. Note that, for clarity and improved readability, the interval over which the spaces L_0^2 and B^2 are defined are omitted from some notations.

For a random density $f(.)$ in the infinite dimensional space $B^2(a,b)$, the expectation and covariance operators can be defined as follows, whenever they exist:

$$\begin{aligned}
\mathbb{E}^{B^2}[f] &= \operatorname{clr}^{-1} \mathbb{E}[\operatorname{clr} f] \in B^2(a,b) && (21)\\
\operatorname{Cov}^{B^2}[f]g &= \mathbb{E}^{B^2}\left[\langle f \ominus \mathbb{E}^{B^2}[f], g\rangle_{B^2} \odot (f \ominus \mathbb{E}^{B^2}[f])\right] \\
&= \operatorname{clr}^{-1} \mathbb{E}[\langle f, g\rangle_{B^2} \operatorname{clr} f] \\
&= \operatorname{clr}^{-1} \mathbb{E}[\langle \operatorname{clr} f, \operatorname{clr} g\rangle_{L_0^2} \operatorname{clr} f] \quad \text{for any } g \in B^2(a,b), && (22)
\end{aligned}$$

where $\ominus$ is the negative perturbation defined by $f \ominus g = f \oplus [(-1) \odot g]$. In what follows, we will focus on the restriction Cov^E of the covariance operator to a finite dimensional subspace E of the Bayes space $B^2(a,b)$.

3.2. Densities preprocessing

We consider a framework where some random density objects $D_n = \{f_1, \ldots, f_n\}$ are observed through samples. The preprocessing step consists, for each density object, in transforming its sample into a functional object using some nonparametric estimation technique, yielding an estimated realization of this random density. Note that in the ensuing treatment, we will forget this layer and consider that we have observed the preprocessed random density functions f_i, $i = 1, \ldots, n$.

To perform the estimation, we use nonparametric maximum penalized likelihood (hereafter MPL) as introduced by Silverman [24] and implemented in the R package `fda` by Ramsay et al. [20]. The principle of MPL is to maximize a penalized version of the log-likelihood over an infinite dimensional space of densities without parametric assumptions. The penalty is the product of a smoothing parameter λ by the integral over the interval of interest of the square of the m^{th} derivative of the log-density. Therefore the objective functional is a functional of the log-density. Due to the infinite dimension of the ambient space, the penalty term is necessary because the likelihood term alone is unbounded above. As explained in Silverman [24], the value of $m = 3$, which we select, has an attractive property. In our case of densities on an interval (a,b), when the smoothing parameter tends to infinity, by Theorem 2.1 of Silverman [24], the estimated density converges to the parametric maximum likelihood estimate in the exponential family of densities whose logarithm is a polynomial of degree less than or equal to 2, comprising the uniform density, the exponential densities and the Gaussian densities truncated to (a,b).

We consider adapting the MPL method to a Bayes space $B^2(a,b)$ with the Lebesgue measure as reference. In order to use MPL in $B^2(a,b)$ we need to add extra smoothness conditions and therefore we restrict attention to the densities of $B^2(a,b)$ whose log belongs to the Sobolev-space of order m on (a,b), thus ensuring the existence of the penalty term. With Theorem 4.1 in Silverman [24], the optimization problem has at least a solution. When implementing MPL, one needs to solve the minimization problem in a finite-dimensional approximating space. [19] use polynomial spline spaces (whose degree is a priori unrelated to m).

3.3. ICS in Bayes spaces

In order to apply the approach of Section 2, we need to restrict attention to a finite dimensional subspace of $B^2(a,b)$. Following Machalová et al. [12], it is convenient to first construct a basis of a finite dimensional spline subspace of $L_0^2(a,b)$, which we then transfer to $B^2(a,b)$ by the inverse clr transformation. More precisely, they propose a basis of zero-integral splines in $L_0^2(a,b)$ that are called ZB-splines. The corresponding inverse images of these basis functions by clr are called CB-splines.

A ZB-spline basis, denoted by $Z = \{Z_1, \ldots, Z_{k+d-1}\}$, is characterized by the spline of degree less than or equal to d (order $d+1$), the number k and the positions of the so-called inside knots $\Delta\gamma = \{\gamma_1, \ldots, \gamma_d\}$ in (a,b). The dimension of the resulting subspace $\mathcal{Z}_d^{\Delta\gamma}$ is $p = k + d - 1$. Let $C_d^{\Delta\gamma}$ be the subspace generated by $C = \{C_1, \ldots, C_p\}$ in $B^2(a,b)$, where $C_j = \mathrm{clr}^{-1}(Z_j)$ are the back-transforms in $B^2(a,b)$ of the basis functions of the subspace $\mathcal{Z}_d^{\Delta\gamma}$. The expansion of a density f in $B^2(a,b)$ is then given by

$$f(t) = \bigoplus_{j=1}^{p} [f]_{C_j} C_j(t), \tag{23}$$

so that the corresponding expansion of its clr in $L_0^2(a,b)$ is given by

$$\mathrm{clr}\, f(t) = \sum_{j=1}^{p} [f]_{C_j} Z_j(t). \tag{24}$$

Note that the coordinates of f in the basis C are the same as the coordinates of $\mathrm{clr}(f)$ in the basis Z, for $j = 1, \ldots, p$, $[f]_{C_j} = [\mathrm{clr}\, f]_{Z_j}$. Following Machalová et al. [12], the basis functions of $\mathcal{Z}_d^{\Delta\gamma}$ can be written in a B-spline basis, see Schumaker [23], which is convenient to allow using existing code for their computation.

We follow the methodology of Section 2 for a space $E = C_d^{\Delta\gamma}$ of compositional splines. Let us consider an E-valued random probability density f. In order to work with two weighted covariance operators $\mathrm{Cov}_{w_1}^E$ and $\mathrm{Cov}_{w_2}^E$, where $w_1, w_2 : \mathbb{R}^+ \to \mathbb{R}$ are two measurable functions, we assume that $f \in \mathcal{E}_{w_1} \cap \mathcal{E}_{w_2}$. From Definition 2, we can write for any $g \in E$:

$$\begin{aligned} \mathrm{Cov}_{w_\ell}^E[f]g &= \mathbb{E}^{B^2}[w_\ell^2(\|f^{st}\|_{B^2})\langle \tilde{f}, g\rangle_{B^2} \odot \tilde{f}] && (25)\\ &= \mathrm{clr}^{-1}\, \mathbb{E}[w_\ell^2(\|f^{st}\|_{B^2})\langle \mathrm{clr}\, \tilde{f}, \mathrm{clr}\, g\rangle_{L_0^2}\, \mathrm{clr}\, \tilde{f}], \quad \ell \in \{1,2\}, && (26)\end{aligned}$$

where $\tilde{f} = f \ominus \mathbb{E}^{B^2}[f]$ is the centred version of f and $f^{st} = \mathrm{Cov}^E[f]^{-1/2}\tilde{f}$ is the sphered version of f.

Then, solving $\mathrm{ICS}(f, \mathrm{Cov}_{w_1}^E, \mathrm{Cov}_{w_2}^E)$ is equivalent to solving $\mathrm{ICS}(\mathrm{clr}\, f, \mathrm{Cov}_{w_1}^F, \mathrm{Cov}_{w_2}^F)$ where $\mathrm{clr}\, f$ is an F-valued random function with $F = \mathcal{Z}_d^{\Delta\gamma}$. It simply follows from Corollary 1 applied to the clr isometry between E and F. Corollary 2 implies that solving $\mathrm{ICS}(f, \mathrm{Cov}_{w_1}^E, \mathrm{Cov}_{w_2}^E)$ in $C_d^{\Delta\gamma}$ is equivalent to a multivariate ICS problem on the coordinates of the clr of the densities in the ZB-basis which can be done using multivariate techniques. For implementation, we will use the third formulation in Corollary 2.

For a sample of densities $D_n = \{f_1, \ldots, f_n\}$, the empirical version of ICS that will be performed in the next section for outlier detection follows the framework described in Remark 3.

4. ICS for density outlier detection

4.1. Implementation of ICS on density data for outlier detection

As outlined in Section 3.3, for a given sample of random density objects $D_n = \{f_1, \ldots, f_n\}$ in a space of compositional splines $E = C_d^{\Delta\gamma}$, solving the empirical version of ICS is equivalent to solving an ICS problem in a multivariate framework [see 29] with the coordinates of the clr of the densities in the chosen ZB-basis. Following the three-step procedure defined in [6], we propose to use ICS for density outlier detection in a context of a small proportion of outliers.

The first step consists in choosing a pair of scatter matrices and calculating the eigenvalues and invariant coordinates of ICS. Following the recommendation of [6], we use the empirical scatter pair Cov-Cov_4 (see Example 1), and compute the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$, and the invariant components z_{ji}, $j = 1, \ldots, p$, for each density f_i,

$i = 1, \ldots, n$. The algorithm we consider, described in [5], uses the QR decomposition. This approach enhances stability compared to methods based on a joint diagonalization of two scatter matrices, which can be numerically unstable in some ill-conditioned situations.

The second step of the outlier detection procedure based on ICS is the selection of the $\kappa < p$ relevant invariant components and the computation of the "ICS distances." For each of the n densities, the ICS distance is equal to the Euclidean norm of the reconstructed data using the κ selected invariant components. In the case of a small proportion of outliers and for the scatter pair Cov-Cov$_4$, the invariant components of interest are associated with the largest eigenvalues and the squared ICS distances are equal to $\sum_{j=1}^{\kappa} z_{ji}^2$. As noted in [6], there exist several methods for the selection of the number of invariant components. One approach is to examine the scree plot, as in PCA. This method, recommended in [6], is not automatic. Alternative automatic selection methods apply univariate normality tests to each component, starting with the first one, and using some Bonferroni correction [for further details see page 13 of 6]. In the present paper, we use the scree plot approach when there is no need of an automatic method, and we use the D'Agostino normality test for automatic selection. The level for the first test (before Bonferroni correction) is 5%. Dimension reduction involves retaining only the first κ components of ICS instead of the original p variables. Note that when all the invariant components are retained, the ICS distance is equal to the Mahalanobis distance.

The computation of ICS distances allows to rank the observations in decreasing order, with those having the largest distances potentially being outliers. However, in order to identify the outlying densities, we need to define a cut-off, and this constitutes the last step of the procedure. Following [6], we derive cut-offs based on Monte Carlo simulations from the standard Gaussian distribution. For a given sample size and number of variables, we generate 10,000 standard Gaussian samples and compute the empirical quantile of order 97.5% of the ICS-distances using the three steps previously described. An observation with an ICS distance larger than this quantile is flagged as an outlier.

The procedure described above has been illustrated on several examples [see 6], and is implemented in the R package `ICSOutlier` [see 15]. However, in the context of densities, the impact of preprocessing parameters on the ICSOutlier procedure emerges as a crucial question that needs to be examined.

4.2. Impact of the preprocessing parameters

As a toy example, consider the densities of the maximum daily temperatures for the 26 provinces of the two regions Red River Delta and Northern Midlands and Mountains in northern Vietnam between 2013 and 2016. We augment this data set made of 104 densities by adding the provinces AN GIANG and BAC LIEU from southern Vietnam in the same time period. The total number of observations is thus 112. Details on the original data and their source are provided in Subsection 5.1. Figure 1 displays a map of Vietnam with the contours of all provinces and colored according to their administrative region, allowing the reader to locate the 26 provinces in the North and the two in the South. As shown on the left panel of Figure 2, the eight densities of the two provinces from the South for the four years exhibit a very different shape (in yellow) compared to the northern provinces (in blue and green), with much more concentrated maximum temperatures. These two provinces should be detected as outliers when applying the ICSOutlier methodology. However, the results may vary depending on the choice of preprocessing parameters (see Subsection 3.2). Our goal is to analyze how the detected outliers vary depending on the preprocessing when using the maximum penalized likelihood method with splines of degree less than or equal to $d = 4$. Specifically, we study the influence on the results of ICSOutlier of the smoothing parameter λ, the number of inside knots k, and the location of the knots defining the spline basis.

The number κ of selected invariant components is fixed at four in all experiments to facilitate interpretation. This value has been chosen after viewing the scree plots of the ICS eigenvalues following the recommendations in Subsection 4.1. For each of the experimental scenarios detailed below, we compute the squared ICS distances of the 112 observations as defined in Subsection 4.1, using $\kappa = 4$. Observations are classified as outliers when their squared ICS distance exceeds the threshold defined in Subsection 4.1, using a level of 2.5%. For each experiment, we plot on Figure 3 the indices of the observations from 1 to 112 on the y-axis, marking outlying observations with black squares. The eight densities from southern Vietnam correspond to indices 1 to 8. We consider the following scenarios:

- the knots are either located at the quantiles of the temperature values (top panel on Figure 3) or equally spaced (bottom panel on Figure 3),

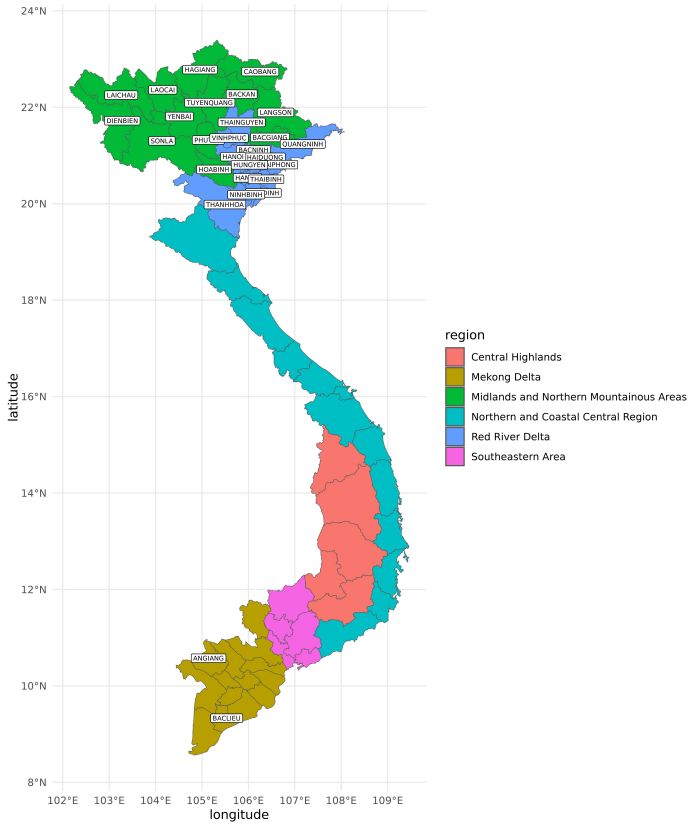

Figure 1: Map of Vietnam showing the 63 provinces, color-coded by region. The 28 provinces included in the toy example are labeled.

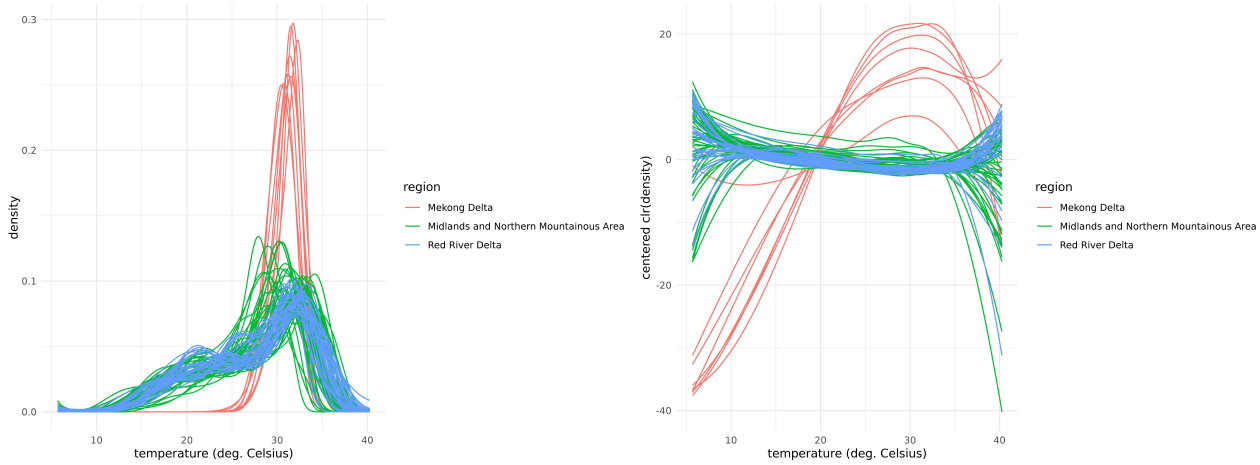

Figure 2: Plots of the 28 densities (left panel) and clr densities (right panel), color-coded by region for the toy example.

- from the left to the right of Figure 3, the number of knots varies from 0 to 15 by increments of 1, then from 15 to 35 by increments of 10 (overall 18 different values). Note that when increasing the number of knots beyond 35, the code returns more and more errors due to multicollinearity issues and the results are not reported.
- the base-10 logarithm of the parameter λ varies from -8 to 8 with an increment of 1 on the x-axis of each plot.

Altogether we have $2 \times 18 \times 17 = 612$ scenarios. Figure 4 is a bar plot showing the observations indices on the x-axis and the frequency of outlier detection across scenarios on the y-axis. The eight densities from the two southern provinces (AN GIANG and BAC LIEU) across the four years are most frequently detected as outliers, along with the province of LAI CHAU (indices 33 to 36), which is located in a mountainous region in northwest of Vietnam. On the original data, we can see that the LAI CHAU province corresponds to densities with low values for high maximum temperatures (above 35°C) coupled with relatively high density values for maximum temperatures below 35°C. A few other observations are detected several times as outliers, but less frequently: indices 53 (TUYEN QUANG in 2013), 96 (QUANG NINH in 2016), and 107 (THANH HOA in 2015).

Looking at Figure 3, we examine the impact of the preprocessing parameters on the detection of outlying observations. First, note that the ICS algorithm returns an error when the λ parameter is large (shown as grey bands in some plots). This is due to a multicollinearity problem. Even though the QR version of the ICS algorithm is quite stable, it may still encounter problems when multicollinearity is severe. Indeed, when λ is large, the estimated densities converge to densities whose logarithm is a polynomial of degree less than or equal to 2 (see details in Subsection 3.2), and belongs to a 3-dimensional affine subspace of the Bayes space, potentially with a dimension smaller than that of the approximating spline space. If we compare the top and the bottom plots, we do not observe large differences in the outlying pattern, except for a few observations rarely detected as outliers. Thus, the knot location has a rather small impact on the ICS results for this data set. Regarding the impact of the λ parameter, the outlier pattern remains relatively stable when the number of knots is small (less than or equal to 5). For a large number of knots, the observations detected as outliers vary with λ. The number of knots has more impact than their location or the λ parameter. When the number of knots is smaller than or equal to 5 (corresponding to $p = 10$ variables), the plots are very similar. However, as p increases, some observations from southern Vietnam are not detected for all λ values, while another density (QUANG NINH in 2016) is detected for large λ values with equally spaced knots, and to a lesser extent for knots at temperature quantiles. In Archimbaud et al. [4], ICS is applied to multivariate functional data with B-splines preprocessing. Based on their empirical experience, the authors recommend using a dimension p (in their case, the number of functional components times the number of B-splines coefficients) no larger than the number of observations divided by 10. Typically in multivariate analysis, the rule of thumb is that the dimension should not exceed the number of observations divided by 5. For functional or distributional data, it appears that even more observations per variable are needed. The reason for this is not entirely clear, but in the case of ICS, we can suspect that the presence of multicollinearity, even approximate, degrades the results. By increasing the number of knots, we precisely increase the multicollinearity problem, especially for large values of λ.

From this experimentation, we recommend using knots located at the quantiles of the measured variable, and a number of knots such that the number of observations is around 10 times the dimension p (here: the dimension of the B-spline basis). The base-10 logarithm of parameter λ can be chosen between -2 and 2 to avoid extreme cases and multicollinearity problems. Moreover, the idea of launching ICS multiple times with different preprocessing parameter values to confirm an observation's atypical nature by its repeated detection is a strategy we retain for real applications, as detailed in Subsection 5.3.

5. Outlier detection: an application to Vietnamese climate data

5.1. Data and preprocessing description

The temperature data used in this application comprises daily maximum temperatures for each of the $I = 63$ Vietnamese provinces over a $T = 30$-year period (1987-2016). Originally from the Climate Prediction Center (CPC) database, developed and maintained by the National Oceanic and Atmospheric Administration (NOAA), the data underwent a preliminary treatment presented in Trinh et al. [27]. From the daily 365 or 366 values for each year, we derive the yearly maximum temperature distribution for each of the 1,890 province-year units. We assume that the temperature samples are independent across years and spatially across provinces, which is a simplifying assumption.

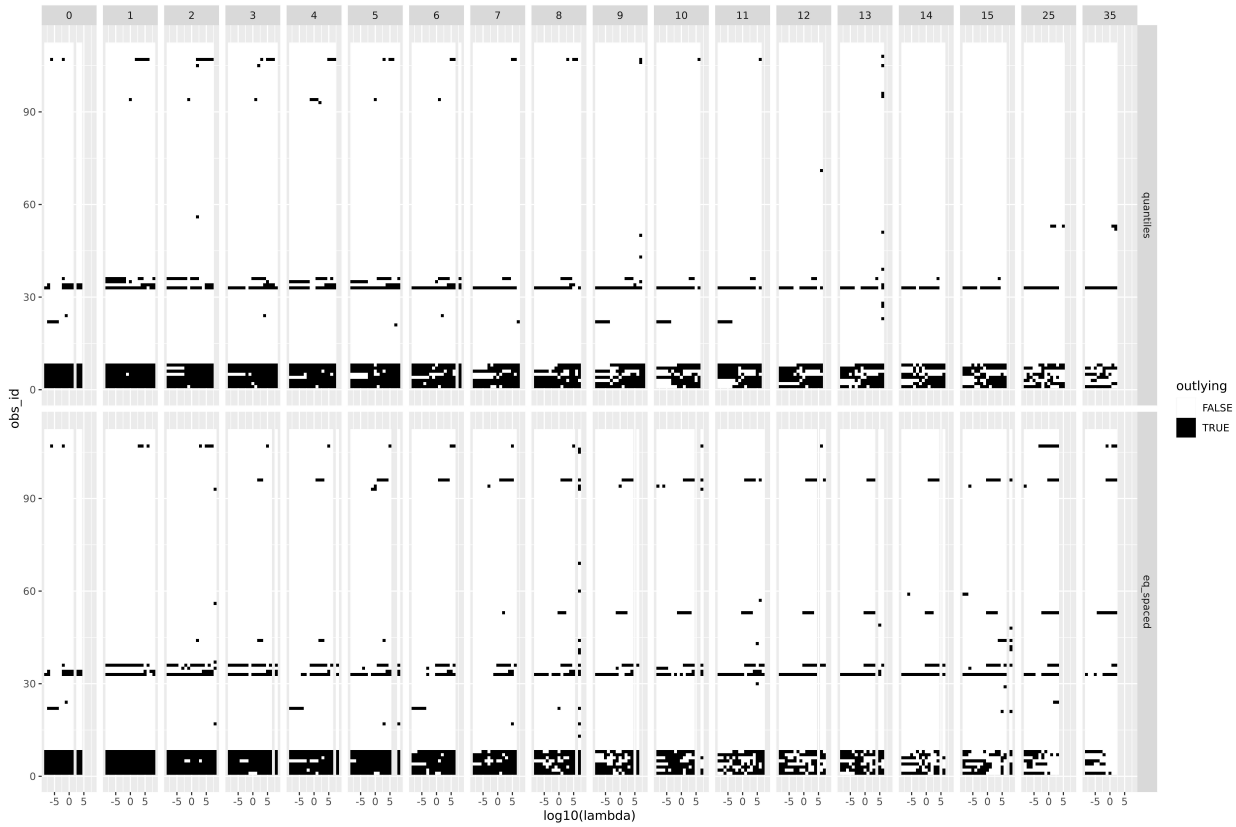

Figure 3: Outlier detection by ICS across smoothing parameters for the Vietnam toy example. Top: knots at quantiles; Bottom: equally spaced knots. y-axis: observation indices; x-axis: λ parameter. Columns correspond to knot numbers (0-35). Outliers are marked as black squares.

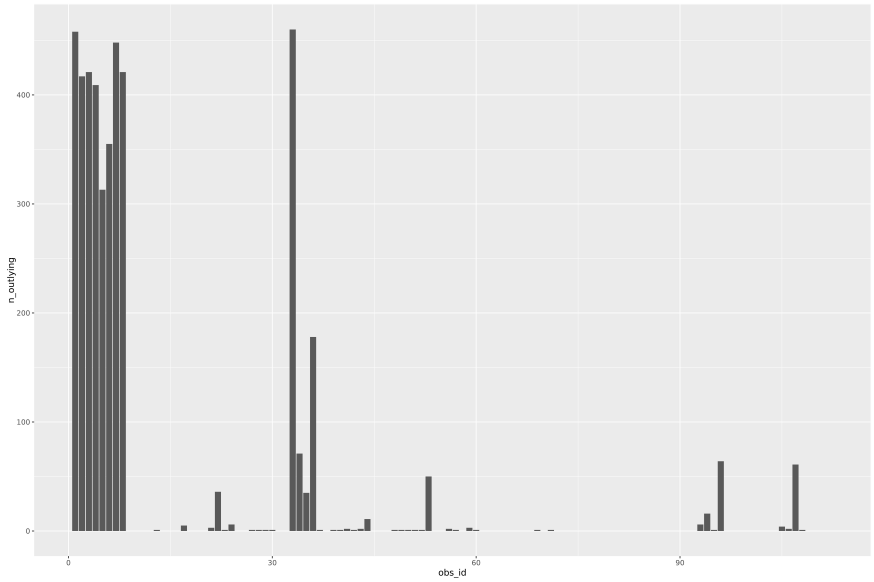

Figure 4: Frequency of outlier detection by ICS across 612 scenarios with varying smoothing parameters, for the Vietnam toy example.

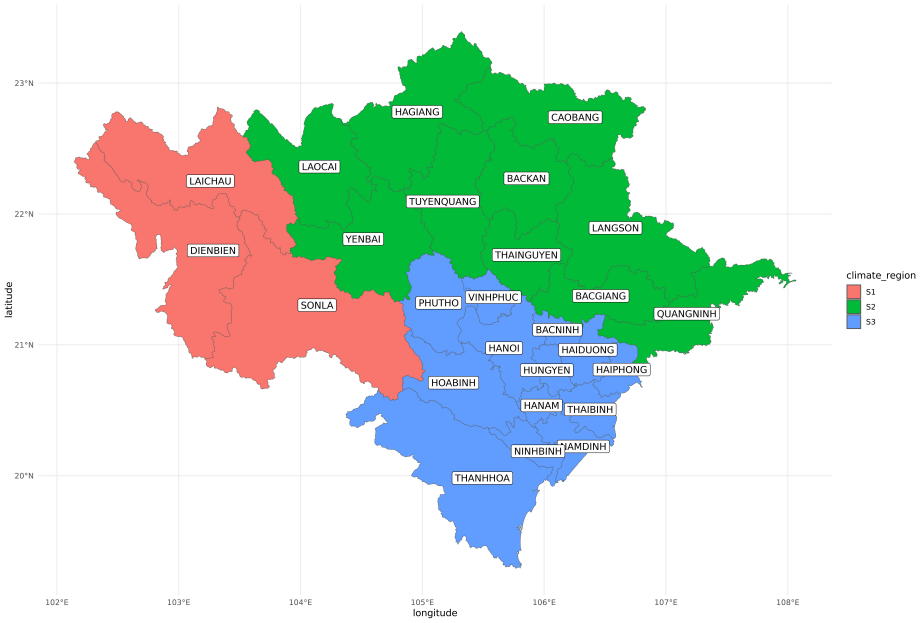

Figure 5: The three climate regions of northern Vietnam.

Figure 1 depicts the six administrative regions of Vietnam, and the corresponding provinces. However, these regions cover areas with varied climates. To achieve more climatically homogeneous groupings, we use clusters of provinces based on climatic regions as defined by Stojanovic et al. [25]. Figure 5 displays the three climatic regions covering northern Vietnam. We focus on region S3, composed of 13 provinces, by similarity with the North Plain (Red River Delta) (S3) in Stojanovic et al. [25]. Figure 6 shows the maximum temperature densities for the 13 provinces of S3, plotted by year, using the preprocessing detailed in Subsection 3.2 with degree less than or equal to $d = 4$, smoothing parameter $\lambda = 10$ and $k = 10$ knots located at quantiles of the pooled sample (across space and time). We observe more variability across time than across space which confirms that the spatial homogeneity objective is achieved.

5.2. ICS outlier detection for the S3 northern Vietnam provinces

We propose to follow the different steps described in Subsection 4.1, and examine the results of ICS outlier detection using the scatter pair $\text{Cov} - \text{Cov}_4$ on the 390 (13 provinces $\times$ 30 years) densities from region S3, obtained after the preprocessing detailed above.

The scree plot on the left panel of Figure 7 clearly indicates that we should retain the first two invariant components. The right panel of Figure 7 shows the squared ICS distances based on these first two components, with the observations index on the x-axis and with a threshold (horizontal line) corresponding to a significance level of 2.5%. This plot reveals that several observations are distinctly above this threshold, especially for the years 1987 and 2010.

The left panel of Figure 8 displays the scatterplot of the first two components, labelled by year. The densities are colored by province for the outliers and colored in pink for the other provinces. This plot reveals that the outliers are either densities from 2010 (and one density from 1998) that are outlying on the first component, or densities from 1987 and 2007 that are outlying on the second component.

To interpret the outlyingness, we can use the eigendensities plotted in the right panel of Figure 8 together with the plots of the densities and their centred log-ratio transformation, color-coded by year for the outliers and in pink for the other observations in Figure 9. The horizontal line on the eigendensities plot (right plot of Figure 8) corresponds to the uniform density on the interval $[5; 40]$. Four provinces in 2010 are outlying with large positive values on the first invariant component (see the left panel of Figure 8). The first eigendensity IC.1 is characterised by a smaller mass of the temperature values on the interval $[5; 20]$, compared to the uniform distribution, a mass similar to the uniform on $[20; 35]$, and a much larger mass than the uniform on the interval $[35; 40]$. These four observations correspond to the four blue curves on the left and right panels of Figure 9. Compared to the other densities, these four densities

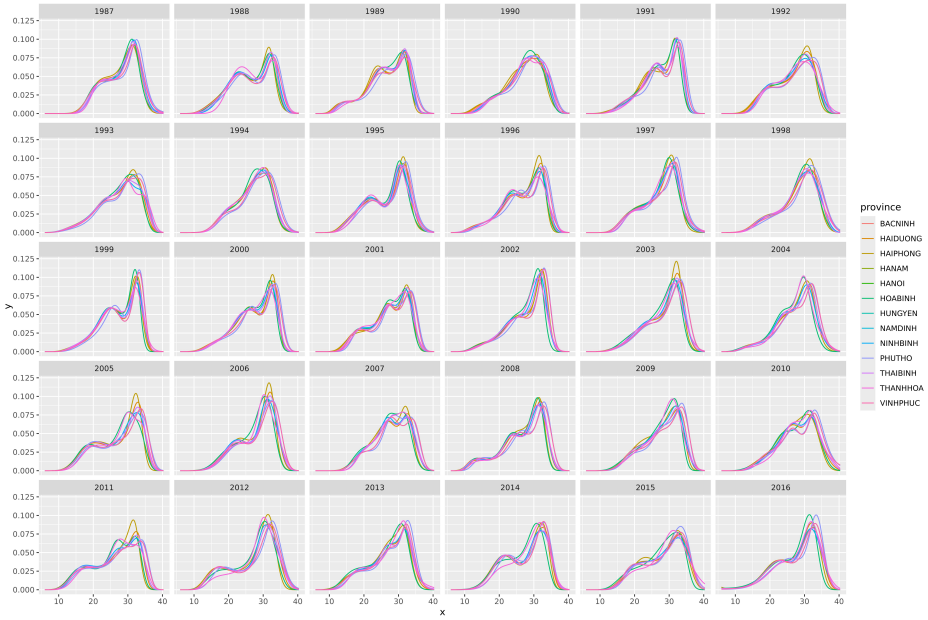

Figure 6: Maximum temperature densities for the 13 provinces in northern Vietnam's S3 region, 1987-2016, color-coded by province.

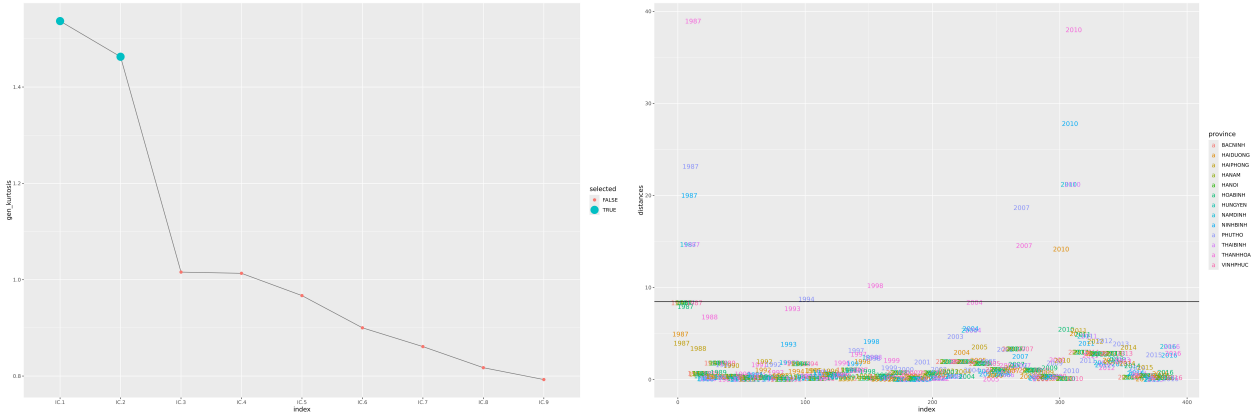

Figure 7: Screeplot of ICS (left panel), and ICS distances based on the first two components (right panel) for maximum temperature densities for the 13 provinces in northern Vietnam's S3 region, 1987-2016.

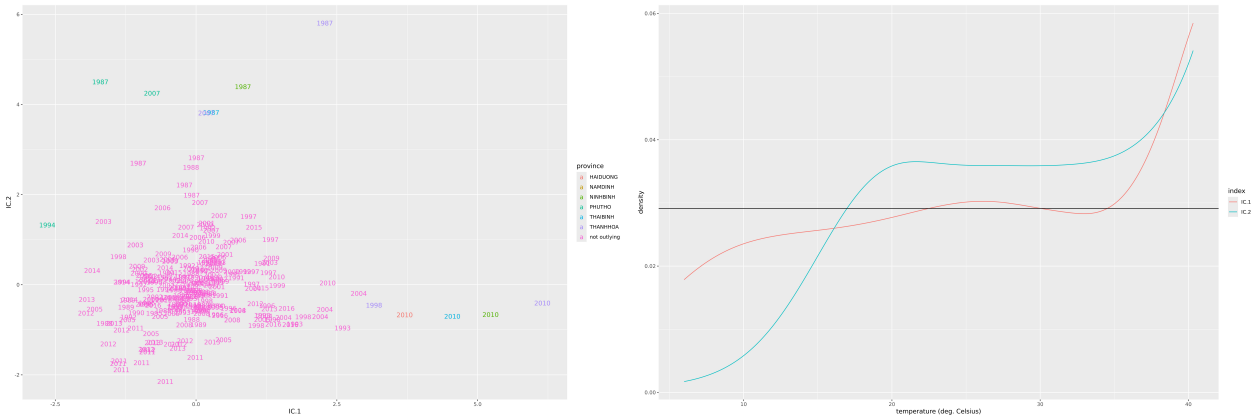

Figure 8: Scatterplot of the first two invariant components (left panel) labeled by year and colored by province, and the first two ICS eigendensities (right panel) of the maximum temperature densities for the 13 provinces in northern Vietnam's S3 region, 1987-2016.

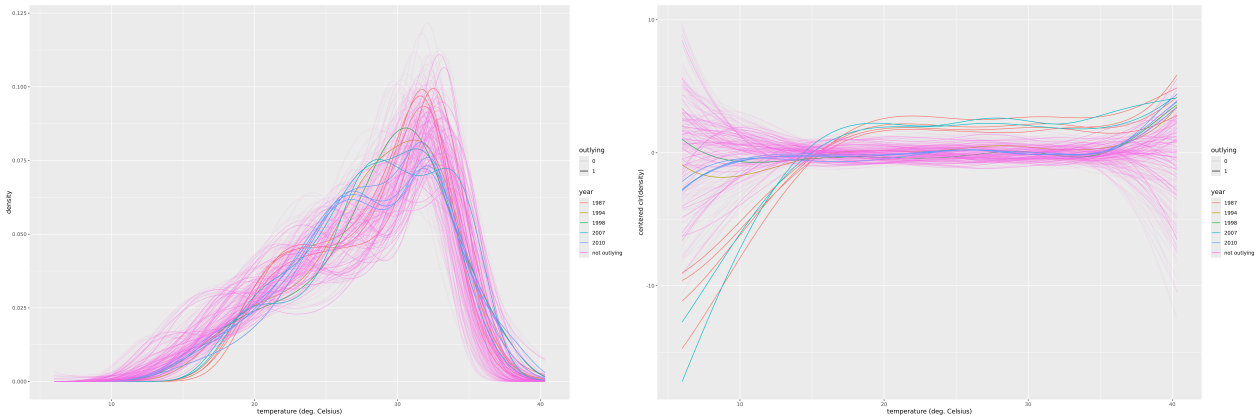

Figure 9: Maximum temperature densities (left panel) and centred log-ratio transformed (right panel) of maximum temperature densities for the 13 provinces in northern Vietnam's S3 region, 1987-2016, color-coded by outlyingness status.

exhibit relatively lighter tails on the lower end of the temperature spectrum and heavier tails on the higher end. For temperature values in the medium range, these four observations fall in the middle of the cloud of densities and of clr transformed densities. On the second invariant component, six observations take large values and are detected as outliers. They correspond to four provinces in 1987 and three in 2007 (see the left panel of Figure 8). The second eigendensity IC.2 differs greatly from the uniform distribution on the whole interval of temperature values. The left tail is much lighter while the right tail is much heavier. Besides the six observations flagged as outliers, other provinces in 1987 and 2007 take large values on ICS.2, and correspond to densities with very few days with maximum temperature less than 15 degrees Celsius compared to other densities.

5.3. Varying the smoothing parameters

As mentioned in Subsection 4.2, we can validate the atypical nature of observations by running the ICSOutlier procedure multiple times with varying smoothing parameter values. Following the rule of thumb of one dimension per 10 observations, with 390 observations, we should consider less than 35 interior knots. In what follows, we take 5, 10, 15, 20 and 25 interior knots and we consider base-10 logarithm values for λ equal to -2, -1, 0, 1 and 2. The number of selected ICS components is either fixed equal to 2, or is automatically determined using the D'Agostino normality test described in Subsection 4.1. We compute the squared ICS distances of the 390 observations, and observations are classified as outliers when their squared distance exceeds the threshold based on a 2.5% level as detailed in Subsection 4.1.

Figure 10: Outlier detection by ICS across smoothing parameters for the Vietnam climate data. Top: 2 invariant components selected; Bottom: automatic selection through D'Agostino tests. y-axis: observation indices; x-axis: λ parameter. Columns correspond to knot numbers (5-25). Outliers are marked as black squares.

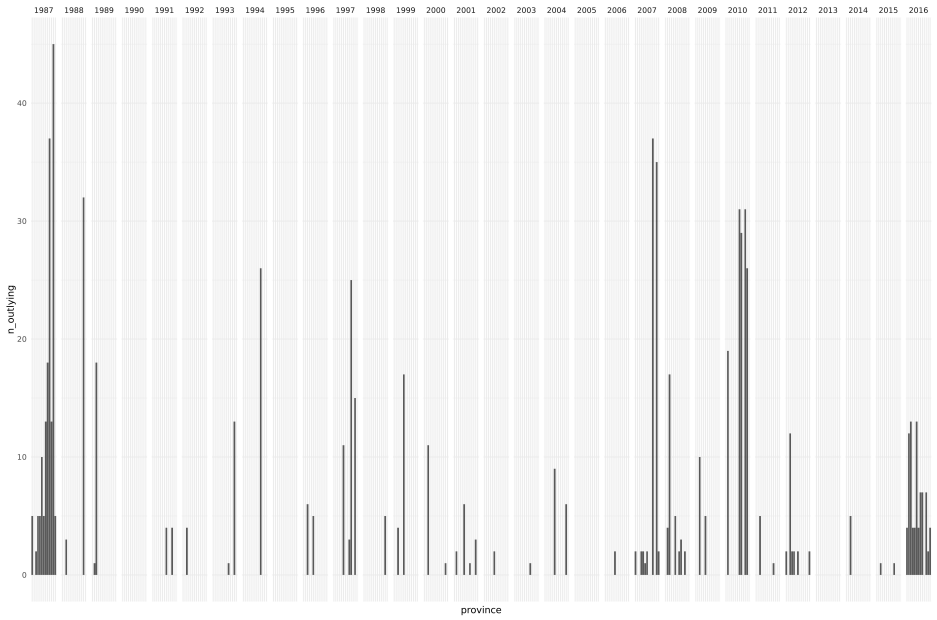

Figure 11: Frequency of outlier detection by ICS across smoothing parameters, for the Vietnam climate data.

We plot in Figure 10 the observations indices from 1 to 390 on the y-axes for the 25 smoothing parameter setups, marking outlying observations with black squares. Figure 11 displays a bar plot showing the observations indices on the x-axis and the frequency of outlier detection on the y-axis across the 25 setups. Note that the choice of the number of selected invariant components has minimal impact. Both Figures 10 and 11 confirm the results of the previous subsection. Most provinces are outlying in 1987 and several are also outlying in 2007 and 2010. For large values of λ, many provinces are also detected as outliers in 2016. Some provinces are detected quite often over the years: THANH HOA, HAI PHONG and HOA BINH. Note that in [25], the province of THANH HOA extends across two climatic regions (S3 and S4) which could explain why it is very often detected as an outlier.

An overall comment regarding the outlier detection procedure that we use in the present application is that, from our experience on other data sets, an outlying density is often characterized by a behavior that differs from the other densities in the tails of the distribution. This is not surprising because the Bayes inner product defined by equation (20) involves the ratio of densities which can be large when a density is small (at the tails of the distribution).

6. Conclusion and perspectives

We propose a coordinate-free presentation of ICS with application to density-objects and a focus on outlier detection. This presentation is very general and allows ICS to be applied to more complex objects than the coordinates vectors of multivariate analysis. However, one of the limitations of the coordinate-free approach is that it is adapted to scatter operator pairs which are made of weighted covariance operators. These pairs of operators include the well-known $(\mathrm{Cov}, \mathrm{Cov}_4)$ pair. Its scatter counterpart in the multivariate context is the one recommended by [3] for a small proportion of outliers. But it is unclear how we could generalize to scatter operators other well-known scatter matrices (such as M-estimators, pairwise-based weighted estimators, or Minimum Covariance Determinant estimators) which are useful when using ICS as a preprocessing step for clustering [see 2].

Concerning a further adaptation of ICSOutlier to density objects, one perspective to our work is to take into account different settings for the preprocessing parameters and aggregate the results in a single outlyingness index. Another perspective is to consider multivariate densities (e.g., not only maximum density temperature but also minimum density temperature, precipitation,...) and generalize the ICSOutlier procedure as in Archimbaud et al. [4] for multivariate functional data.

This coordinate-free framework for ICS lays the groundwork for a generalisation to infinite dimensional Hilbert spaces. Many difficulties arise, such as the compactness of the covariance operator which makes it non surjective, so that one cannot easily define a Mahalanobis distance, on which our definition of weighted covariance operators relies. Moreover, the existence of solutions and other properties of ICS proved in this paper come from the fact that one of the scatter operators is an automorphism, so it cannot be compact (in particular not the covariance). Finally, [28] proved that, whenever the dimension p is larger than the number of observations n, all affine equivariant scatter operators are proportional, which is a bad omen for a straight generalisation to infinite dimensional Hilbert spaces. One can partially circumvent these difficulties by assuming that the data is almost surely in a deterministic finite-dimensional subspace E of H (which is the case for density data after our preprocessing) and applying coordinate-free ICS. Another option could be to alleviate the affine equivariance assumption.

7. Acknowledgments

The major part of this work was completed while the authors were visiting the Vietnam Institute for Advanced Study in Mathematics (VIASM) in Hanoï and the authors express their gratitude to VIASM. This paper has also been funded by the Agence Nationale de la Recherche under grant ANR-17-EURE-0010 (Investissements d'Avenir program). We thank Thibault Laurent for attracting our attention on the climate regions partition of Vietnam.

Appendix

Proof of Proposition 1. Since $S_1[X]$ is non-singular, $S_1[X]^{-1}S_2[X]$ exists and is symmetric in the Euclidean space $(E, \langle \cdot, S_1[X]\cdot \rangle)$, because

$$\forall (x, y) \in E^2, \langle S_1[X]^{-1}S_2[X]x, S_1[X]y\rangle = \langle S_2[X]x, y\rangle = \langle S_2[X]y, x\rangle \quad (27)$$

$$= \langle S_1[X]^{-1}S_2[X]y, S_1[X]x\rangle. \quad (28)$$

Thus, the spectral theorem guarantees that there exists an orthonormal basis H in which $S_1[X]^{-1}S_2[X]$ is diagonal. □

Proof of Proposition 2. Let us decompose $S_1[X]^{-1}(X-\mathbb{E}X)$ over the basis H, which is orthonormal in $(E, \langle\cdot, S_1[X]\cdot\rangle)$:

$$S_1[X]^{-1}(X-\mathbb{E}X) = \sum_{j=1}^{p}\langle S_1[X]^{-1}(X-\mathbb{E}X), S_1[X]h_j\rangle h_j \tag{29}$$

$$= \sum_{j=1}^{p}\langle X-\mathbb{E}X, h_j\rangle h_j \tag{30}$$

$$S_1[X]^{-1}(X-\mathbb{E}X) = \sum_{j=1}^{p} z_j h_j. \tag{31}$$

The dual basis H^* of H is the one that satisfies $\langle h_j, h_{j'}^*\rangle = \delta_{jj'}$ for all $1 \le j, j' \le p$ and we know from the definition of ICS that this holds for $(S_1[X]h_j)_{1\le j\le p}$. □

Proof of Proposition 3. First, let us verify that the ICS$(X^{\mathcal{F}}, S_1^{\mathcal{F}}, S_2^{\mathcal{F}})$ problem is well defined on F:

(a) The application φ is linear so it is measurable. Moreover, if $X \in \mathcal{E}$, $A \in \mathcal{GL}(F)$ and $b \in F$, then

$$\|\varphi(X)\|_F = \|X\|_E \tag{32}$$

and

$$A\varphi(X)+b = \varphi\left(\varphi^{-1}\circ A\circ\varphi(X)+\varphi^{-1}(b)\right) \text{ where } \varphi^{-1}\circ A\circ\varphi(X)+\varphi^{-1}(b) \in \mathcal{E}. \tag{33}$$

(b) If $X \in \mathcal{E}$, $S_\ell^{\mathcal{F}}[\varphi(X)] = \varphi\circ S_\ell^{\mathcal{E}}[X]\circ\varphi^{-1}$ is a non-negative symmetric operator and if $Y \in \mathcal{E}$ verifies $\varphi(X) \sim \varphi(Y)$, then $X \sim Y$ (because the Borel σ-algebra on E is the pullback by φ of that on F) so that, for $\ell \in \{1,2\}$,

$$S_\ell^{\mathcal{F}}[\varphi(X)] = \varphi\circ S_\ell^{\mathcal{E}}[X]\circ\varphi^{-1} = \varphi\circ S_\ell^{\mathcal{E}}[Y]\circ\varphi^{-1} = S_\ell^{\mathcal{F}}[\varphi(Y)] \tag{34}$$

and

$$S_\ell^{\mathcal{F}}[A\varphi(X)+b] = \varphi\circ S_\ell^{\mathcal{E}}[\varphi^{-1}\circ A\circ\varphi(X)+\varphi^{-1}(b)]\circ\varphi^{-1}$$
$$= A\circ\varphi\circ S_\ell^{\mathcal{E}}[X]\circ\varphi^{-1}\circ A^* = AS_\ell^{\mathcal{F}}[\varphi(X)]A^*. \tag{35}$$

(c) The isometry φ preserves the linear rank of any finite sequence of vectors of E.

Now, $(H^{\mathcal{E}}, \Lambda)$ solves ICS$(X^{\mathcal{E}}, S_1^{\mathcal{E}}, S_2^{\mathcal{E}})$ in the space E if and only if

$$\begin{cases}\langle S_1^{\mathcal{E}}[X]h_j^{\mathcal{E}}, h_{j'}^{\mathcal{E}}\rangle_E = \delta_{jj'} \text{ for all } 1\le j,j'\le p\\ \langle S_2^{\mathcal{E}}[X]h_j^{\mathcal{E}}, h_{j'}^{\mathcal{E}}\rangle_E = \lambda_j\delta_{jj'} \text{ for all } 1\le j,j'\le p\end{cases} \tag{36}$$

$$\iff \begin{cases}\langle \varphi(S_1^{\mathcal{E}}[X]h_j^{\mathcal{E}}), \varphi(h_{j'}^{\mathcal{E}})\rangle_F = \delta_{jj'} \text{ for all } 1\le j,j'\le p\\ \langle \varphi(S_2^{\mathcal{E}}[X]h_j^{\mathcal{E}}), \varphi(h_{j'}^{\mathcal{E}})\rangle_F = \lambda_j\delta_{jj'} \text{ for all } 1\le j,j'\le p\end{cases} \tag{37}$$

$$\iff \begin{cases}\langle S_1^{\mathcal{F}}[X]h_j^{\mathcal{F}}, h_{j'}^{\mathcal{F}}\rangle_F = \delta_{jj'} \text{ for all } 1\le j,j'\le p\\ \langle S_2^{\mathcal{F}}[X]h_j^{\mathcal{F}}, h_{j'}^{\mathcal{F}}\rangle_F = \lambda_j\delta_{jj'} \text{ for all } 1\le j,j'\le p,\end{cases} \tag{38}$$

which is equivalent to the fact that $(H^{\mathcal{F}}, \Lambda)$ solves ICS$(X^{\mathcal{F}}, S_1^{\mathcal{F}}, S_2^{\mathcal{F}})$ in the space F. □

Proof of Corollary 1. Let $\ell \in \{1,2\}$ and $\tilde{X} = X - \mathbb{E}X$. In order to prove the equation (11), we will need to prove that, for any $(x,y) \in F^2$,

$$\langle\varphi\circ\mathrm{Cov}_{w_\ell}^E[X]\circ\varphi^{-1}(x), y\rangle_F = \langle\mathrm{Cov}_{w_\ell}^E[X]\varphi^{-1}(x), \varphi^{-1}(y)\rangle_E \tag{39}$$

$$= \mathbb{E}[w_\ell(\|\mathrm{Cov}^E[X]^{-1/2}\tilde{X}\|_E)^2\langle\tilde{X}, \varphi^{-1}(x)\rangle_E\langle\tilde{X}, \varphi^{-1}(y)\rangle_E] \tag{40}$$

$$= \mathbb{E}[w_\ell(\|\mathrm{Cov}^F[\varphi(X)]^{-1/2}\varphi(\tilde{X})\|_F)^2\langle\varphi(\tilde{X}), x\rangle_F\langle\varphi(\tilde{X}), y\rangle_F] \tag{41}$$

$$\langle\varphi\circ\mathrm{Cov}_{w_\ell}^E[X]\circ\varphi^{-1}(x), y\rangle_F = \langle\mathrm{Cov}_{w_\ell}^F[\varphi(X)]x, y\rangle_F. \tag{42}$$

It is enough to show the equality between (40) and (41), for which we treat differently the cases $w_\ell = 1$ and $w_\ell \neq 1$. If $w_\ell = 1$, there is nothing to prove, so that the equation (11) holds for the covariance operator. If $w_\ell \neq 1$, we now know from the case $w_\ell = 1$ that

$$\mathrm{Cov}^F[\varphi(X)]^{-1/2} = \varphi \circ \mathrm{Cov}^E[X]^{-1/2} \circ \varphi^{-1} \tag{43}$$

so that

$$\|\mathrm{Cov}^E[X]^{-1/2}\tilde{X}\|_E = \|\mathrm{Cov}^F[\varphi(X)]^{-1/2}\varphi(\tilde{X})\|_F \tag{44}$$

Once the equation (11) is proved, one only needs to apply Proposition 3 to finish the proof. □

Proof of Corollary 2. Applying Corollary 1 to the isometry

$$\varphi_B : \left\{ \begin{array}{ccc} (E, \langle\cdot,\cdot\rangle_E) & \to & (\mathbb{R}^p, \langle\cdot,\cdot\rangle_{\mathbb{R}^p}) \\ x & \mapsto & G_B^{1/2}[x]_B, \end{array} \right. \tag{45}$$

we obtain the equivalence between the following assertions:

(i) (H, Λ) solves $\mathrm{ICS}(X, \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ in the space E
(ii) $(G_B^{1/2}[H]_B, \Lambda)$ solves $\mathrm{ICS}(G_B^{1/2}[X]_B, \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ in the space $\mathbb{R}^p$,

which gives the equivalence between the assertions (1) and (2). The equivalence between the other assertions are deduced from the fact that for any $\ell \in \{1, 2\}$ and any $(x, y) \in E^2$:

$$\begin{aligned} \langle \mathrm{Cov}^E_{w_\ell}[X]x, y\rangle_E &= \langle \mathrm{Cov}_{w_\ell}(G_B^{1/2}[X]_B)G_B^{1/2}[x]_B, G_B^{1/2}[y]_B\rangle_{\mathbb{R}^p} &(46)\\ &= \langle \mathrm{Cov}_{w_\ell}(G_B[X]_B)[x]_B, [y]_B\rangle_{\mathbb{R}^p} &(47)\\ &= \langle \mathrm{Cov}_{w_\ell}([X]_B)G_B[x]_B, G_B[y]_B\rangle_{\mathbb{R}^p}, &(48) \end{aligned}$$

where (46) comes from the equation (11), and the equalities (47) and (48) come from the affine equivariance of Cov_{w_ℓ}. □

References

[1] Aggarwal, C.C., 2017. Outlier Analysis. Springer International Publishing, Cham. doi:10.1007/978-3-319-47578-3.
[2] Alfons, A., Archimbaud, A., Nordhausen, K., Ruiz-Gazen, A., 2024. Tandem clustering with invariant coordinate selection. Econometrics and Statistics doi:10.1016/j.ecosta.2024.03.002.
[3] Archimbaud, A., 2018. Détection non-supervisée d'observations atypiques en contrôle de qualité: un survol. Journal de la Société Française de Statistique 159, 1–39.
[4] Archimbaud, A., Boulfani, F., Gendre, X., Nordhausen, K., Ruiz-Gazen, A., Virta, J., 2022. ICS for multivariate functional anomaly detection with applications to predictive maintenance and quality control. Econometrics and Statistics doi:10.1016/j.ecosta.2022.03.003.
[5] Archimbaud, A., Drmač, Z., Nordhausen, K., Radojičić, U., Ruiz-Gazen, A., 2023. Numerical Considerations and a new implementation for invariant coordinate selection. SIAM Journal on Mathematics of Data Science 5, 97–121. doi:10.1137/22M1498759.
[6] Archimbaud, A., Nordhausen, K., Ruiz-Gazen, A., 2018. ICS for multivariate outlier detection with application to quality control. Computational Statistics & Data Analysis 128, 184–199. doi:10.1016/j.csda.2018.06.011.
[7] Dai, W., Mrkvička, T., Sun, Y., Genton, M.G., 2020. Functional outlier detection and taxonomy by sequential transformations. Computational Statistics & Data Analysis 149, 106960. doi:10.1016/j.csda.2020.106960.
[8] Egozcue, J.J., Díaz–Barrero, J.L., Pawlowsky–Glahn, V., 2006. Hilbert Space of Probability Density Functions Based on Aitchison Geometry. Acta Mathematica Sinica, English Series 22, 1175–1182. doi:10.1007/s10114-005-0678-2.
[9] Lei, X., Chen, Z., Li, H., 2023. Functional Outlier Detection for Density-Valued Data with Application to Robustify Distribution-to-Distribution Regression. Technometrics 65, 351–362. doi:10.1080/00401706.2022.2164063.
[10] Li, B., Van Bever, G., Oja, H., Sabolová, R., Critchley, F., 2021. Functional independent component analysis : an extension of fourth-order blind identification. URL: https://sites.google.com/site/germainvanbever/publica.
[11] Loperfido, N., 2021. Some theoretical properties of two kurtosis matrices, with application to invariant coordinate selection. Journal of Multivariate Analysis 186, 104809. doi:10.1016/j.jmva.2021.104809.
[12] Machalová, J., Hron, K., Monti, G., 2016. Preprocessing of centred logratio transformed density functions using smoothing splines. Journal of Applied Statistics 43, 1419–1435. doi:10.1080/02664763.2015.1103706.
[13] Menafoglio, A., 2021. Anomaly detection for density data based on control charts. URL: https://iasc-isi.org/events/iasc-ers-course-an-introduction-to-functional-data-analysis-for-density-functions-in-bayes-spaces/.
[14] Murph, A.C., Strait, J.D., Moran, K.R., Hyman, J.D., Stauffer, P.H., 2024. Visualisation and outlier detection for probability density function ensembles. Stat 13, e662. doi:10.1002/sta4.662.
[15] Nordhausen, K., Archimbaud, A., Ruiz-Gazen, A., 2023. ICSOutlier: Outlier Detection Using Invariant Coordinate Selection. URL: https://cran.r-project.org/web/packages/ICSOutlier/.

[16] Nordhausen, K., Ruiz-Gazen, A., 2022. On the usage of joint diagonalization in multivariate statistics. Journal of Multivariate Analysis 188, 104844. doi:10.1016/j.jmva.2021.104844.
[17] Parlett, B.N., 1998. The Symmetric Eigenvalue Problem. Classics in Applied Mathematics, Society for Industrial and Applied Mathematics. doi:10.1137/1.9781611971163.
[18] Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015. Modeling and Analysis of Compositional Data. 1 ed., Wiley. doi:10.1002/9781119003144.
[19] Ramsay, J., Hooker, G., Graves, S., 2009. Functional Data Analysis with R and MATLAB. Springer New York, New York, NY. doi:10.1007/978-0-387-98185-7.
[20] Ramsay, J., Hooker, G., Graves, S., 2024. fda: Functional Data Analysis. URL: https://cran.r-project.org/web/packages/fda/.
[21] Rendón Aguirre, J.C., 2017. Clustering in high dimension for multivariate and functional data using extreme kurtosis projections. Ph.D. thesis. Universidad Carlos III de Madrid. doi:10016/25286.
[22] Ruiz-Gazen, A., Thomas-Agnan, C., Laurent, T., Mondon, C., 2023. Detecting Outliers in Compositional Data Using Invariant Coordinate Selection, in: Yi, M., Nordhausen, K. (Eds.), Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler. Springer International Publishing, Cham, pp. 197–224. doi:10.1007/978-3-031-22687-8_10.
[23] Schumaker, L., 1981. Spline Functions: Basic Theory. 3 ed., Cambridge University Press. doi:10.1017/CBO9780511618994.
[24] Silverman, B.W., 1982. On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method. The Annals of Statistics 10, 795–810. doi:10.1214/aos/1176345872.
[25] Stojanovic, M., Liberato, M.L.R., Sorí, R., Vázquez, M., Phan-Van, T., Duongvan, H., Hoang Cong, T., Nguyen, P.N.B., Nieto, R., Gimeno, L., 2020. Trends and Extremes of Drought Episodes in Vietnam Sub-Regions during 1980–2017 at Different Timescales. Water 12, 813. doi:10.3390/w12030813. number: 3.
[26] Stone, M., 1987. Coordinate-Free Multivariable Statistics: An Illustrated Geometric Progression from Halmos to Gauss and Bayes. Number 2 in Oxford statistical science series, Clarendon Pr, Oxford.
[27] Trinh, T.H., Christine, T.A., Simioni, M., 2023. Discrete and Smooth Scalar-on-Density Compositional Regression for Assessing the Impact of Climate Change on Rice Yield in Vietnam. URL: https://www.tse-fr.eu/publications/discrete-and-smooth-scalar-density-compositional-regression-assessing-impact-climate-change-rice.
[28] Tyler, D.E., 2010. A note on multivariate location and scatter statistics for sparse data sets. Statistics & Probability Letters 80, 1409–1413. doi:10.1016/j.spl.2010.05.006.
[29] Tyler, D.E., Critchley, F., Dümbgen, L., Oja, H., 2009. Invariant co-ordinate selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71, 549–592. doi:10.1111/j.1467-9868.2009.00706.x.
[30] Van Den Boogaart, K.G., Egozcue, J.J., Pawlowsky-Glahn, V., 2014. Bayes Hilbert Spaces. Australian & New Zealand Journal of Statistics 56, 171–194. doi:10.1111/anzs.12074.
[31] Virta, J., Li, B., Nordhausen, K., Oja, H., 2020. Independent component analysis for multivariate functional data. Journal of Multivariate Analysis 176, 104568. doi:10.1016/j.jmva.2019.104568.