WORKING PAPERS

N° 1585

May 2025

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

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

Toulouse School of Economics

ICS FOR COMPLEX DATA WITH APPLICATION TO OUTLIER DETECTION FOR DENSITY DATA

A PREPRINT

Camille Mondon
Mathematics and Statistics
Toulouse School of Economics
Toulouse, 31000
camille.mondon@tse-fr.eu

Huong Thi Trinh
Faculty of Mathematical Economics
Thuongmai University
Hanoi
trinhthihuong@tmu.edu.vn

Anne Ruiz-Gazen
Mathematics and Statistics
Toulouse School of Economics
Toulouse, 31000
anne.ruiz-gazen@tse-fr.eu

Christine Thomas-Agnan
Mathematics and Statistics
Toulouse School of Economics
Toulouse, 31000
christine.thomas@tse-fr.eu

May 20, 2025

ABSTRACT

Invariant coordinate selection (ICS) is a dimension reduction method, used as a preliminary step for clustering and outlier detection. It has been primarily applied to multivariate data. This work introduces a coordinate-free definition of ICS in an abstract Euclidean space and extends the method to complex data. Functional and distributional data are preprocessed into a finite-dimensional subspace. For example, in the framework of Bayes Hilbert spaces, distributional data are smoothed into compositional spline functions through the Maximum Penalised Likelihood method. We describe an outlier detection procedure for complex data and study the impact of some preprocessing parameters on the results. We compare our approach with other outlier detection methods through simulations, producing promising results in scenarios with a low proportion of outliers. ICS allows detecting abnormal climate events in a sample of daily maximum temperature distributions recorded across the provinces of Northern Vietnam between 1987 and 2016.

Keywords Bayes spaces • Distributional data • Extreme weather • Functional data • Invariant coordinate selection • Outlier detection • Temperature distribution • 62H25 • 62R10 • 62G07 • 65D07

1 Introduction

The invariant coordinate selection (ICS) method was introduced in a multivariate data analysis framework by Tyler et al. (2009). 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 Loperfido 2021; Nordhausen and Ruiz-Gazen 2022, 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. 2009) 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 (Archimbaud, Nordhausen, and Ruiz-Gazen 2018) or clustering (Alfons et al. 2024). 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 Nordhausen and Ruiz-Gazen 2022, for more details). ICS has later been adapted to more complex data, namely compositional data (Ruiz-Gazen et al. 2023), functional data (Rendón Aguirre 2017; B. Li et al. 2021, for ICA) and multivariate functional data (Archimbaud, Boulfani, et al. 2022; Virta et al. 2020, for ICA).

A significant contribution of the present work is the formulation of a coordinate-free variant of ICS, considering data objects in an abstract Euclidean space, without having to choose a specific basis. This formulation allows ICS to be consistently defined in a very general framework, unifying its original definition for multivariate data and its past adaptations to specific types of complex data. In the case of compositional data, the coordinate-free approach yields an alternative implementation of ICS that is more computationally efficient. We are also able to propose a new version of invariant coordinate selection adapted to distributional data. Note that a coordinate-free version of ICS has already been mentioned in (Tyler et al. 2009), in the discussion by Mervyn Stone, who proposed to follow the approach of Stone (1987). In their response, Tyler and co-authors agree that this could offer a theoretically elegant and concise view of the topic. A coordinate-free approach of ICA is proposed by B. Li et al. (2021), but to our knowledge, no coordinate-free approach to ICS exists for a general Euclidean space.

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 by Archimbaud, Nordhausen, and Ruiz-Gazen (2018). This method, called ICSOutlier, flags outlying observations and has been implemented for multivariate data by Nordhausen, Archimbaud, and Ruiz-Gazen (2023). It has been adapted to compositional data by Ruiz-Gazen et al. (2023) and to multivariate functional data by Archimbaud, Boulfani, et al. (2022). We propose to extend this detection procedure to complex data and illustrate it on distributional data.

Detecting outliers is already challenging in a classical multivariate context because outliers may differ from the other observations in their correlation pattern (see Aggarwal 2017, for an overview on outlier detection and analysis). Archimbaud, Nordhausen, and Ruiz-Gazen (2018) demonstrate how the ICS procedure outperforms those based on the Mahalanobis distance and PCA (robust or not). For compositional data, the constraints of positivity and constant sum of coordinates must be taken into account as detailed in (Ruiz-Gazen et al. 2023) and further examined in this paper. For univariate functional data, outliers are categorised 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 measures, including the Mahalanobis distance (see, e.g., the recent paper Dai et al. 2020, and the included references). Density data are constrained functional 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 Menafoglio (2021), Lei, Chen, and H. Li (2023) and Murph, Strait, et al. (2024) only. Two types of outliers have been identified for density data: the horizontal-shift outliers and the shape outliers, with shape outliers being again more challenging to detect (see Lei, Chen, and H. Li 2023, for details). The procedure proposed by Menafoglio (2021) is based on an adapted version of functional PCA to density objects in a control chart context. In order to derive a robust distribution-to-distribution regression method, Lei, Chen, and H. Li (2023) propose a transformation tree approach that incorporates many different outlier detection methods adapted to densities. Their methods involve transforming density data into unconstrained data and using standard functional outlier detection methods. Murph, Strait, et al. (2024) 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.

Our coordinate-free definition of ICS enables direct adaptation of the ICSOutlier method to complex data. Through a case study on temperature distributions in Vietnam, we assess the impact of preprocessing parameters and provide practical recommendations for their selection. In addition, the results of a simulation study demonstrate that our method performs favourably compared with other approaches. An original application to Vietnamese data provides a detailed description of the various stages involved in detecting low-proportion outliers using ICS, as well as interpreting them from the dual eigendensities.

Section 2 presents ICS in a coordinate-free framework, states a useful result to link ICS in different spaces, and treats the specific cases of compositional, functional and distributional data. For the latter, we develop a Bayes space approach and discuss the maximum penalised likelihood method to preprocess the original samples of real-valued data into a sample of compositional splines. Section 3 describes the ICS-based outlier detection procedure adapted to complex data, discusses the impact of the preprocessing parameters on outlier detection through a toy example. Simulating data from multiple generating schemes, we compare ICS with other outlier detection methods for density data. Section 4 provides an application of the outlier detection methodology to maximum temperature data in Vietnam over 30 years.

Section 5 concludes the paper and offers some perspectives. Supplementary material on ICS, a reminder on Bayes spaces, as well as proofs of the propositions and corollaries are given in the Appendix.

2 ICS for complex data

A naive approach to ICS for complex data would be to apply multivariate ICS to coordinate vectors in a basis. This not only ignores the metric on the space when the basis is not orthonormal, but also gives a potentially different ICS method for each choice of basis (as in Archimbaud, Boulfani, et al. 2022). Defining a unique coordinate-free ICS problem avoids defining multiple ICS methods and having to discuss the potential links between them, thus making our approach more intrinsic. In particular, it leads to more interpretable invariant components that are of the same nature as the considered complex random objects. In the case of functional or distributional data, the usual framework assumes that the data objects reside in an infinite-dimensional Hilbert space, which leads to non-orthonormal bases and incomplete inner product spaces. We choose to restrict our attention to finite-dimensional approximations of the data in the framework of Euclidean spaces, which are particularly suitable here because ICS is known to fail when the dimension is larger than the sample size (Tyler 2010). This suggests that an ICS method for infinite-dimensional Hilbert spaces would require modifying the core of the method, which is beyond the scope of this work.

2.1 A coordinate-free ICS problem

In order to generalise invariant coordinate selection (Tyler et al. 2009, def. 1) to a coordinate-free framework in a Euclidean space E, we need to eliminate any reference to a coordinate system, which means replacing coordinate vectors by abstract vectors, matrices by linear mappings, bases or quadratic forms, depending on the context. This coordinate emancipation procedure will ensure that our definition of ICS for an E-valued random object X does not depend on any particular choice of basis of E to represent X.

Following this methodology, we are able to immediately generalise the definition of (affine equivariant) scatter operators from random vectors in $E = \mathbb{R}^p$ (as defined in Tyler et al. 2009, eq. 3) to random objects in a Hilbert space E. This is a perfect example of how the coordinate-free framework can be used to extend existing work to infinite-dimensional spaces. For further details, see Definition 3 in the Appendix. A notable difference from (Tyler et al. 2009) is that we work directly with random objects instead of their underlying distributions. In particular, we introduce an affine invariant space $\mathcal{E}$ of random objects on which the scatter operators are defined and to which we assume that X belongs. For example, $\mathcal{E} = L^p(\Omega, E)$ corresponds to assuming the existence of the p first moments of $\|X\|$.

Again, emancipating from coordinates allows us to naturally generalise ICS to complex random objects in a Euclidean space.

Definition 1 (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 objects, S_1 and S_2 two scatter operators on $\mathcal{E}$ and $X \in \mathcal{E}$. The invariant coordinate selection problem $\mathrm{ICS}(X, S_1, S_2)$ is to find a basis $H = (h_1, \ldots, h_p)$ of E and a finite non-increasing real sequence $\Lambda = (\lambda_1 \geq \ldots \geq \lambda_p)$ such that

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

where $\delta_{jj'}$ equals 1 if $j = j'$ and 0 otherwise. 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

$$z_j = \langle X - \mathbb{E}X, h_j\rangle \tag{2}$$

is called the j-th invariant coordinate (in the eigenbasis H).

In Definition 1, our coordinate emancipation procedure does not yield a generalisation to infinite-dimensional Hilbert spaces, where a basis H would not be properly defined as it is not necessarily orthonormal.

Remark (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 Tyler et al. (2009).

In the ICS problem Equation 1, the scatter operators S_1 and S_2 do not play symmetrical roles. This is because the usual method of solving $\mathrm{ICS}(X, S_1, S_2)$ is to use the associated inner product of $S_1[X]$, which requires $S_1[X]$ to be injective. In that case, Proposition 2 in the Appendix proves the existence of solutions to the ICS problem.

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 a corresponding ICS problem in the second. This is the object of the following proposition, which will be used in Section 2.3.

Proposition 1. *Let $\varphi : (E, \langle\cdot,\cdot\rangle_E) \to (F, \langle\cdot,\cdot\rangle_F)$ 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 objects, $S_1^{\mathcal{E}}$ 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 objects, 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 object $X^{\mathcal{E}} \in \mathcal{E}$, any basis $H^{\mathcal{E}} = (h_1^{\mathcal{E}}, \ldots, h_p^{\mathcal{E}})$ of E, and any finite non-increasing real sequence $\Lambda = (\lambda_1 \geq \ldots \geq \lambda_p)$ 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.

2.2 The case of weighted covariance operators

A difficulty in ICS is to find interesting scatter operators that capture the non-ellipticity of the random object. Usually, for multivariate data, we use the pair of scatter matrices $(\mathrm{Cov}, \mathrm{Cov}_4)$. In this section, we define an important family of scatter operators, namely the weighted covariance operators, which contains both Cov and Cov_4. They are explicitly defined by coordinate-free formulas which allow us to relate ICS problems using weighted covariance operators between any two Euclidean spaces. We denote by $\mathcal{GL}(E)$ the group of linear automorphisms of E and by $A^{1/2}$ the unique non-negative square root of a linear mapping A.

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

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

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

$$\forall (x, y) \in E^2, \langle \mathrm{Cov}_w[X]x, y\rangle = \mathbb{E}\left[w^2\left(\left\|\mathrm{Cov}[X]^{-1/2}(X - \mathbb{E}X)\right\|\right)\langle X - \mathbb{E}X, x\rangle\langle X - \mathbb{E}X, y\rangle\right]. \tag{3}$$

When necessary, we will also write Cov_w^E for the w-weighted covariance operator on E to avoid any ambiguity. It is easy to check that weighted covariance operators are affine equivariant scatter operators in the sense of Definition 3.

Example 1. If $w = 1$, we retrieve Cov, the usual covariance operator on $L^2(\Omega, E)$.

Example 2. If for $x \in \mathbb{R}^+$, $w(x) = (p+2)^{-1/2}x$, we obtain the fourth-order moment operator Cov_4 (as in Nordhausen and Ruiz-Gazen 2022, for the case $E = \mathbb{R}^p$) on $\mathcal{E}_w = \left\{X \in L^4(\Omega, E) \mid \mathrm{Cov}[X] \in \mathcal{GL}(E)\right\}$.

The following corollary applies Proposition 1 to the pair of w_ℓ-weighted covariance operators $S_\ell^{\mathcal{E}} = \mathrm{Cov}_{w_\ell}, \ell \in \{1, 2\}$, for which the corresponding $S_\ell^{\mathcal{F}}$ are exactly the w_ℓ-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 object $X \in \mathcal{E}_{w_1} \cap \mathcal{E}_{w_2}$ (with the notations from Definition 2), the equality*

$$\mathrm{Cov}_{w_\ell}^F[\varphi(X)] = \varphi \circ \mathrm{Cov}_{w_\ell}^E[X] \circ \varphi^{-1} \tag{4}$$

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 $\Lambda = (\lambda_1 \geq \ldots \geq \lambda_p)$:

(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.

2.3 Implementation

In order to implement coordinate-free ICS in any Euclidean space E, we restrict our attention to the pair $(\mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ of weighted covariance operators defined in Section 2.2. Note that we could also transport other known scatter matrices, such as the Minimum Covariance Determinant (defined in Rousseeuw 1985), back to the space E using Proposition 1, but this approach would no longer be coordinate-free.

We now choose a basis $B = (b_1, \ldots, b_p)$ of E in order to represent each element x of E by its coordinate vector $[x]_B = ([x]_{b_1} \ldots [x]_{b_p})^\top \in \mathbb{R}^p$. Then, the following corollary of Proposition 1 allows one to relate the coordinate-free approach in E to three different multivariate approaches applied to the coordinate vectors in any basis B of E, where the Gram matrix $G_B = (\langle b_j, b_{j'} \rangle)_{1 \leq j,j' \leq p}$ appears, accounting for the non-orthonormality of B. Notice that, since the ICS problem has been defined in Section 2.1 without any reference to a particular basis, it is obvious that the basis B has no influence on ICS.

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, $G_B = (\langle b_j, b_{j'} \rangle)_{1 \leq j,j' \leq p}$ 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 $\Lambda = (\lambda_1 \geq \ldots \geq \lambda_p)$ the following assertions are equivalent:*

(1) (H, Λ) *solves* $\mathrm{ICS}(X, \mathrm{Cov}^E_{w_1}, \mathrm{Cov}^E_{w_2})$ *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* $\mathbb{R}^p$*{#eq-third}*

(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 practice, we prefer Assertion (3) (transforming the data by the Gram matrix of the basis) because it is the only one that does not require inverting the Gram matrix in order to recover the eigenobjects. Then, the problem is reduced to multivariate ICS, already implemented in the R package `ICS` using the QR decomposition (Archimbaud, Drmač, et al. 2023). This QR approach enhances stability compared to methods based on a joint diagonalisation of two scatter matrices, which can be numerically unstable in some ill-conditioned situations.

After we obtain the ICS eigenelements, we can use them to reconstruct the original random object, in order to interpret the contribution of each invariant component. Proposition 3 in the Appendix generalises the multivariate reconstruction formula to complex data. In order to implement this reconstruction, 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

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

Remark (Empirical ICS and estimation). In order to work with samples of complex random objects, we can study the particular case of a finite E-valued random object X where we have a fixed sample $D_n = (x_1, \ldots, x_n)$ and we assume that X follows the empirical probability distribution P_{D_n} of $(x_1, \ldots, x_n)$. 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 object X_0. We can estimate solutions of the problem $\mathrm{ICS}(X_0, S_1, S_2)$ from Definition 1 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. Since the population version of ICS for a complex random object $X \in E$ is more concise than its sample counterpart for $D_n = (X_1, \ldots, X_n)$, we shall use the notations of the former in the next sections.

2.4 ICS for compositional data

The specific case of coordinate-free ICS for compositional data is equivalent to the approach of Ruiz-Gazen et al. (2023). To see this, let us consider the simplex $E = (\mathcal{S}^{p+1}, \oplus, \odot, \langle \cdot, \cdot \rangle_{\mathcal{S}^{p+1}})$ of dimension p in $\mathbb{R}^{p+1}$ with the Aitchison structure (Pawlowsky-Glahn, Juan José Egozcue, and Tolosana-Delgado 2015). The results from 5.1 (resp. 5.2) in (Ruiz-Gazen et al. 2023) can be recovered by applying Corollary 1 to any isometric log-ratio transformation (see Pawlowsky-Glahn, Juan José Egozcue, and Tolosana-Delgado 2015, for a definition) (resp. the centred log-ratio transformation).

Corollary 2 gives a new characterisation of the problem $\mathrm{ICS}(X, \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ using additive log-ratio transformations. 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:

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

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

1. (H, Λ) solves $\mathrm{ICS}(X, \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ in the space $\mathcal{S}^{p+1}$
2. $(\mathrm{alr}_j(H), \Lambda)$ 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 the centred log-ratio transform of $x \in \mathcal{S}^{p+1}$ from which the j-th coordinate has been removed. This suggests a new and fastest implementation of invariant coordinate selection for compositional data, in an unconstrained space and only requiring the choice of an index j instead of a full contrast matrix.

2.5 ICS for functional data

The difficulty of functional data (in the broader sense, encompassing density data) is twofold: first, functions are usually analysed within the infinite-dimensional Hilbert space $L^2(a, b)$, second, a random function is almost never observed for every argument, but rather on a discrete grid. This grid can be regular or irregular, deterministic or random, dense (the grid spacing goes to zero) or sparse. We describe a general framework for adapting coordinate-free ICS to functional data, solving both difficulties at the same time by smoothing the observed values into a random function u that belongs to a Euclidean subspace E of $L^2(a, b)$.

2.5.1 Choosing an approximating Euclidean subspace

We usually choose polynomial spaces, spline spaces with given knots and order, or spaces spanned by a truncated Hilbert basis of $L^2(a, b)$. In practice, this choice also depends on the preprocessing method that we have in mind to smooth discrete observations into functions.

2.5.2 Preprocessing the observations into the approximating space

Considering a dense, deterministic grid $(t_1, \ldots, t_N)$, we need to reconstruct an E-valued random function u from its noisy observed values $(u(t_1) + \varepsilon_1, \ldots, u(t_N) + \varepsilon_N)$. There are many well-documented approximation techniques to carry out this preprocessing step, such as interpolation, spline smoothing, or Fourier methods (for a detailed presentation, see Eubank 2014).

2.5.3 Solving ICS in the approximating space

Once we have obtained an E-valued random function u, we can apply the method described in Section 2.3 to reduce $\mathrm{ICS}(u, \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ to a multivariate problem on the coordinates in a basis of E. In particular, for an orthonormal basis B of E (such as a Fourier basis or a Hermite polynomial basis), Corollary 2 gives the equivalence between the following two assertions:

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

If E is a finite-dimensional spline space, we usually work with the coordinates of u in a B-spline basis of E, but then we should take into account its Gram matrix, as in Corollary 2.
ICS has previously been defined for multivariate functional data by Archimbaud, Boulfani, et al. (2022), who define a pointwise method and a global method. Unlike the pointwise approach, which is specific to multivariate functional data, the global method can also be applied to univariate functional data in $L^2(a, b)$, as it corresponds to applying multivariate ICS to truncated coordinate vectors in a Hilbert basis of $L^2(a, b)$. The above framework retrieves the global method in (Archimbaud, Boulfani, et al. 2022) as a particular case when taking a Hilbert basis B of $L^2(a, b)$ and solving coordinate-free ICS in the space E spanned by the p first elements of B.

2.6 ICS for distributional data

A first option to adapt ICS to density data would be to consider it as constrained functional data and directly follow the approach of Section 2.5. However, distributional data does not reduce to density data (such as absorbance spectra studied in Ferraty and Vieu 2002), as it can also be histogram data or sample data (such as the dataset of temperature samples analysed in Section 4). Moreover, the framework of Bayes Hilbert spaces, described by (Van Den Boogaart, Juan José Egozcue, and Pawlowsky-Glahn 2014) and recalled in the Appendix, is specifically adapted to the study of distributional data. Taking into account the infinite-dimensional nature of distributional data, we follow a similar framework as the one of Section Section 2.5, restricting our attention to finite-dimensional subspaces E of the Bayes space $B^2(a,b)$ with the Lebesgue measure as reference.

2.6.1 Choosing an approximating Euclidean space

Following smoothing splines methods, adapted to Bayes spaces by Machalová, Hron, and Monti (2016) and recalled in the Appendix, we choose to work in the space $E = \mathcal{C}_d^{\Delta\gamma}(a,b)$ of compositional splines on (a,b) of order $d+1$ with knots $\Delta\gamma = (\gamma_1, \ldots, \gamma_k)$. Note that the centred log-ratio transform clr is an isometry between E and the space $F = \mathcal{Z}_d^{\Delta\gamma}(a,b)$ of zero-integral splines on (a,b) of order $d+1$ (degree less than or equal to d) and with knots $\Delta\gamma = (\gamma_1, \ldots, \gamma_k)$. They both have dimension $p = k + d$.

2.6.2 Preprocessing the observations into the approximating space

We consider the special cases of histogram data and of sample data. In the former, we follow (Machalová, Talská, et al. 2021) to smooth each histogram into a compositional spline in E. In the latter, we assume that a random density is observed through a finite random sample $(X_1, \ldots, X_N)$ drawn from it. The preprocessing step consists in estimating the density from the observed sample. To perform the estimation, we need a nonparametric estimation procedure that yields a compositional spline belonging to E. That is why we opt for maximum penalised likelihood (MPL) density estimation, introduced by Silverman (1982). The principle of MPL is to maximise a penalised 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 likelihood term alone is unbounded above, hence the penalty term is necessary. In our case of densities on an interval (a,b), we select the value $m = 3$ so that (according to Silverman 1982, Theorem 2.1) when the smoothing parameter tends to infinity, 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. This family comprises the uniform density, exponential and Gaussian densities truncated to (a,b). 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. Note that compositional splines verify these conditions. With Theorem 4.1 in (Silverman 1982), the optimisation problem has at least a solution. Since the estimate f of the density of $(X_1, \ldots, X_N)$ needs to belong to the chosen finite-dimensional subspace $E = \mathcal{C}_d^{\Delta\gamma}(a,b)$, we restrict MPL to E, using the R function `fda::density.fd`, designed by Ramsay, Hooker, and Graves (2024). This function returns the coordinates of $\log(f)$ in the B-spline basis with knots $\Delta\gamma$ and order $d+1$, that we project onto $\mathcal{Z}_d^{\Delta\gamma}(a,b)$ and to which we apply clr^{-1} so that we obtain an element of $\mathcal{C}_d^{\Delta\gamma}(a,b)$.

2.6.3 Solving ICS in the approximating space

We have now obtained an E-valued random compositional spline f. In order to work with two weighted covariance operators Cov_{w_1} and Cov_{w_2}, where $w_1, w_2 : \mathbb{R}^+ \to \mathbb{R}$ are two measurable functions, we assume that $f \in \mathcal{E}_{w_1} \cap \mathcal{E}_{w_2}$, using the notations of Definition 2. Now, we refer to Section 2.3 to reduce the problem $\mathrm{ICS}(f, \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ to a multivariate ICS problem on the coordinates of f in the CB-spline basis of $\mathcal{C}_d^{\Delta\gamma}(a,b)$ (defined in Machalová, Talská, et al. 2021), transformed by the Gram matrix of said CB-spline basis. Note that Corollary 1 applied to the centred log-ratio isometry between $\mathcal{C}_d^{\Delta\gamma}(a,b)$ and $\mathcal{Z}_d^{\Delta\gamma}(a,b)$ gives the equivalence between:

1. (H, Λ) solves $\mathrm{ICS}(f, \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ in the space $E = \mathcal{C}_d^{\Delta\gamma}(a,b)$
2. $(\mathrm{clr}(H), \Lambda)$ solves $\mathrm{ICS}(\mathrm{clr}(f), \mathrm{Cov}_{w_1}, \mathrm{Cov}_{w_2})$ in the space $F = \mathcal{Z}_d^{\Delta\gamma}(a,b)$.

Then, it is completely equivalent, and useful for implementation, to work with the coordinates of $\mathrm{clr}(f)$ in the ZB-spline basis of $\mathcal{Z}_d^{\Delta\gamma}(a,b)$.

3 Outlier detection for complex data using ICS

3.1 Implementation of ICS on complex data for outlier detection

We propose using ICS to detect outliers in complex data, specifically in scenarios with a small proportion of outliers (typically 1 to 2%). For this, we follow the three-step procedure defined by Archimbaud, Nordhausen, and Ruiz-Gazen (2018), modifying the first step based on the implementation of coordinate-free ICS in Section 2.3.

3.1.1 Computing the invariant coordinates

For the scatter operators, we follow the recommendation of Archimbaud, Nordhausen, and Ruiz-Gazen (2018) who compare several pairs of more or less robust scatter estimators in the context of a small proportion of outliers, and conclude that $(\mathrm{Cov}, \mathrm{Cov}_4)$ is the best choice. Thus, we use the empirical scatter pair $(\mathrm{Cov}, \mathrm{Cov}_4)$ (see Example 1 and Example 2), and compute the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$, and the invariant coordinates $z_{ji}, 1 \leq j \leq p$, for each observation $X_i, 1 \leq i \leq n$. As outlined in Section 2.3, for a given sample of random complex objects $D_n = \{X_1, \ldots, X_n\}$ in a Euclidean space E, solving the empirical version of ICS is equivalent to solving an ICS problem in a multivariate framework (see Tyler et al. 2009) with the coordinates of the objects in a basis B of E. In order to choose a basis, we follow the specific recommendations for each type of data from Section 2.4 and Section 2.6.

3.1.2 Selecting the invariant components

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 observations, 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 $(\mathrm{Cov}, \mathrm{Cov}_4)$, 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 by Archimbaud, Nordhausen, and Ruiz-Gazen (2018), 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 by Archimbaud, Nordhausen, and Ruiz-Gazen (2018), 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 Archimbaud, Nordhausen, and Ruiz-Gazen 2018). 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.

3.1.3 Choosing a cut-off

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 Archimbaud, Nordhausen, and Ruiz-Gazen (2018), 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 in several examples (see Archimbaud, Nordhausen, and Ruiz-Gazen 2018), and is implemented in the R package `ICSOutlier` (see Nordhausen, Archimbaud, and Ruiz-Gazen 2023). However, in the context of densities, the impact of preprocessing parameters (see Section 2.6) on the ICSOutlier procedure emerges as a crucial question that needs to be examined.

3.2 Influence of the preprocessing parameters for the density data application

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 Section 4.1.

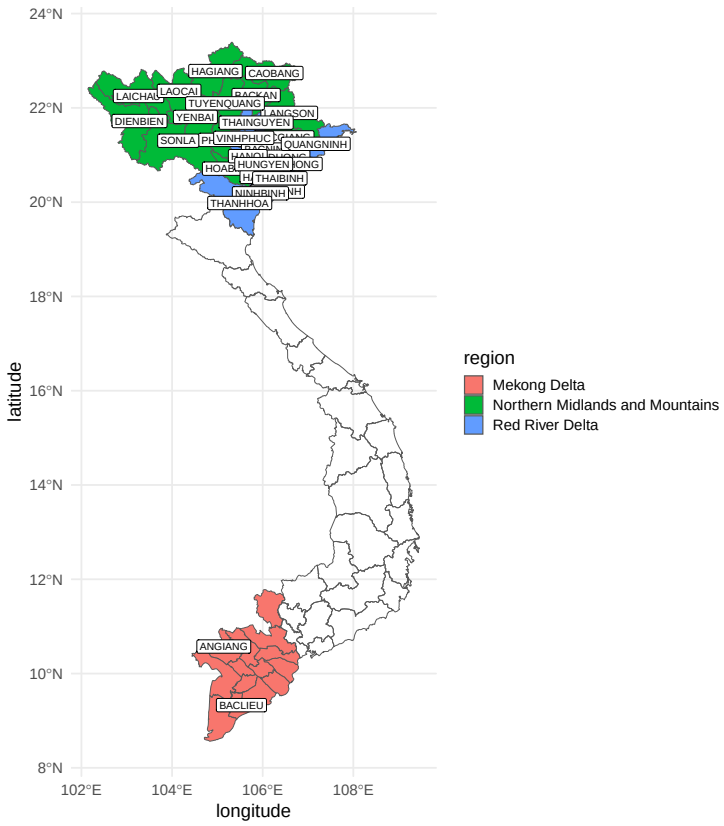
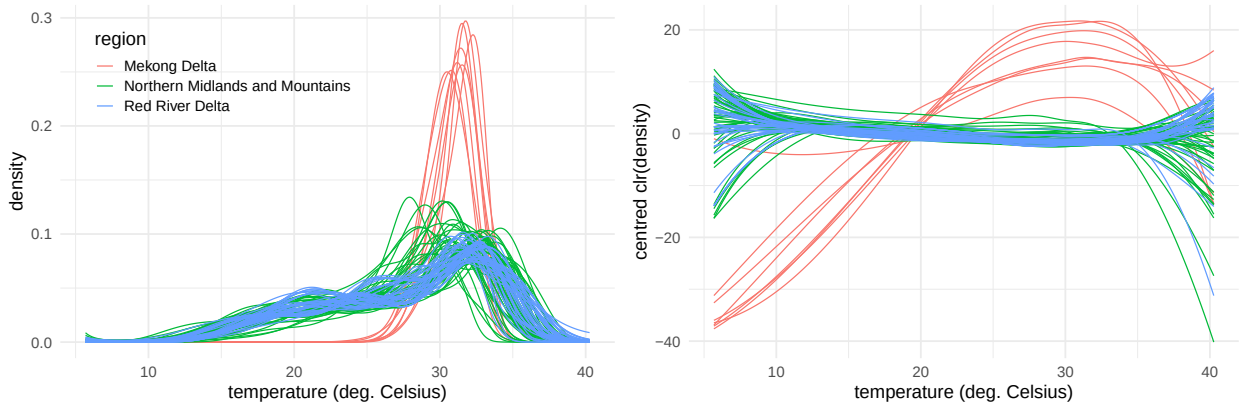


Figure 1: Map of Vietnam showing the 63 provinces, with the three regions under study colour-coded. The 28 provinces included in the toy example are labelled.



(a) (b)

Figure 2: Plots of the 112 densities (left panel) and clr densities (right panel), colour-coded by region for the toy example.

Figure 1 displays a map of Vietnam with the contours of all provinces and coloured 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 red) 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 Section 2.6.2). Our goal is to analyse how the

detected outliers vary depending on the preprocessing when using the maximum penalised 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 Section 3.1. For each of the experimental scenarios detailed below, we compute the squared ICS distances of the 112 observations as defined in Section 3.1, using $\kappa = 4$. Observations are classified as outliers when their squared ICS distance exceeds the threshold defined in Section 3.1, 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 dark squares. The eight densities from Southern Vietnam are in red and 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),
- from the left to the right of Figure 3, the number of knots varies from 0 to 14 by increments of 2, and then takes the values 25 and 35 (overall 10 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 $2 \times 10 \times 17 = 340$ 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 color-coded by region. 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 white 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 Section 2.6.2), 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 6), especially when looking at the densities from the south of Vietnam in red. 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 6 (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, Boulfani, et al. 2022), 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.

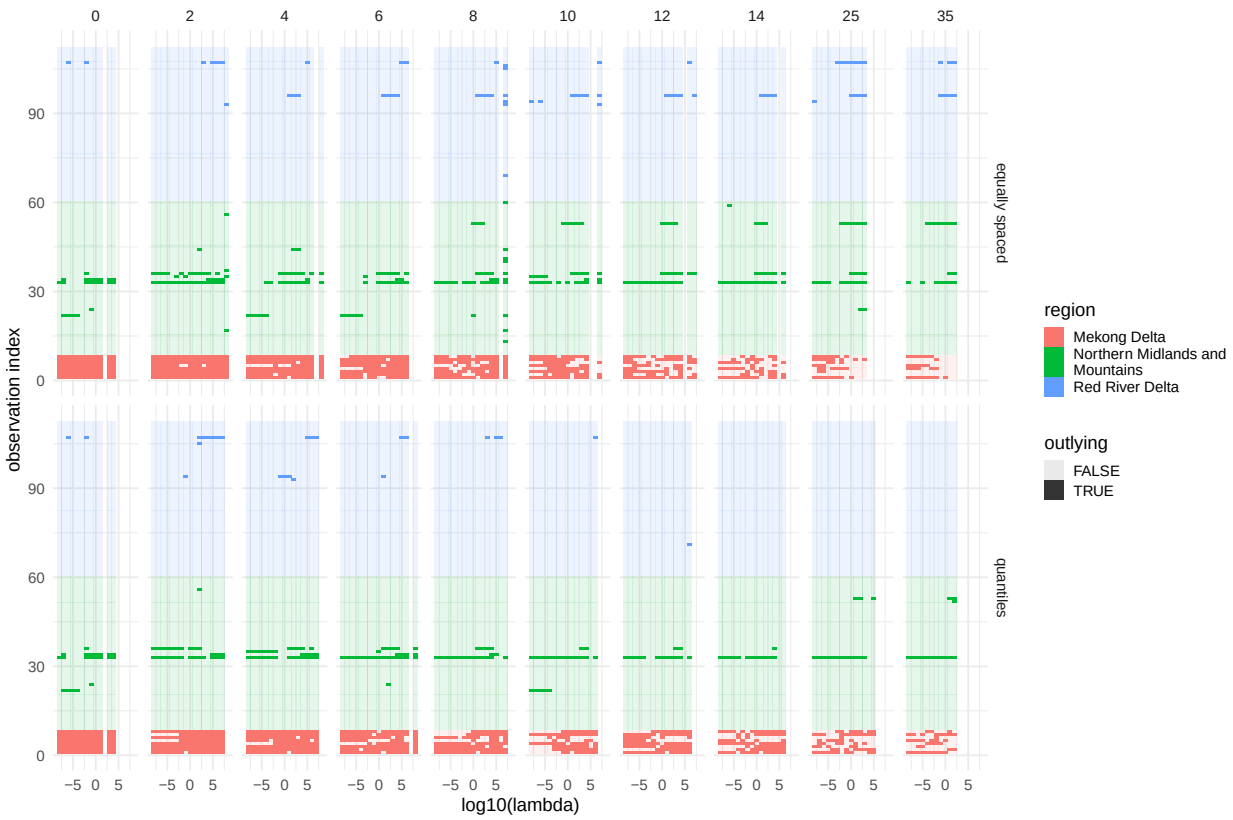


Figure 3: Outlier detection by ICS across smoothing parameters for the Vietnam toy example. *Top:* knots at quantiles; *Bottom:* equally spaced knots. y*-axis:* observation indices; x*-axis:* λ parameter. Columns correspond to knot numbers (0-35). Outliers are dark and colour-coded by region.

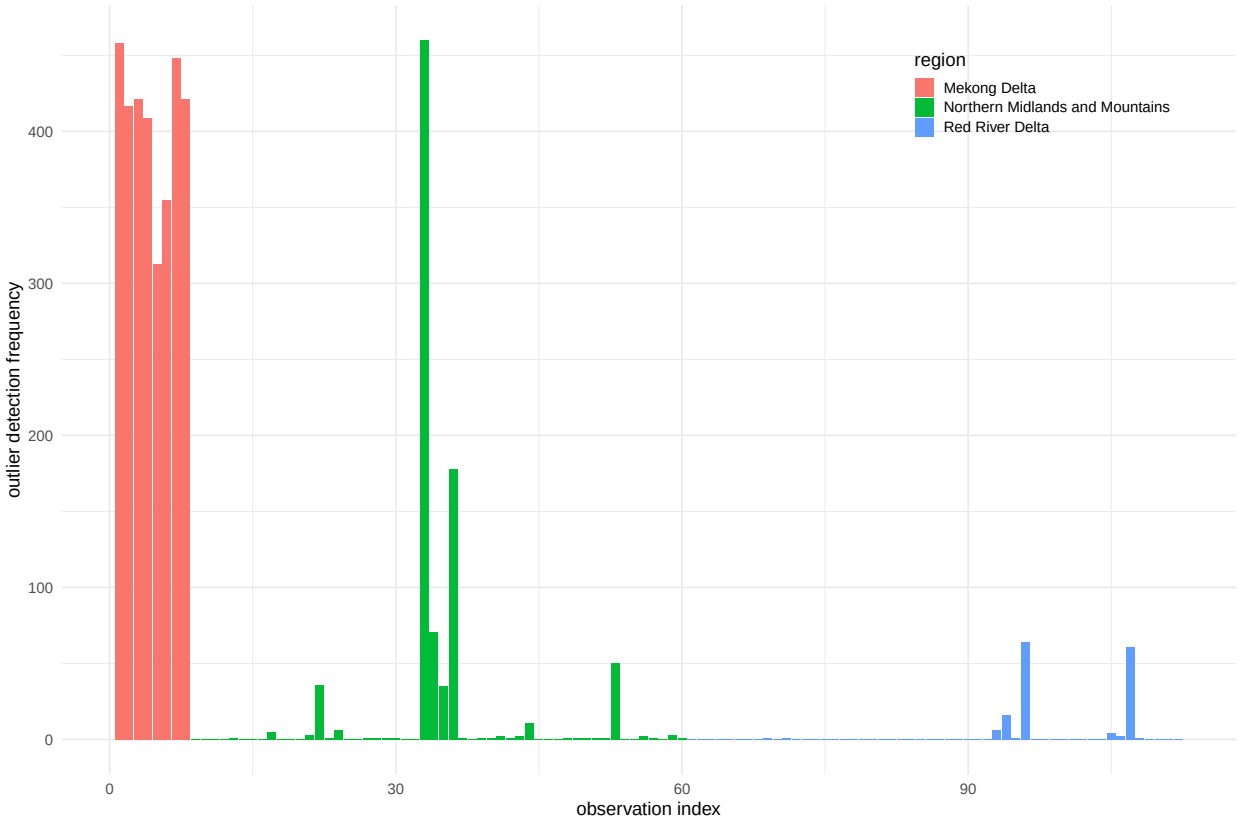


Figure 4: Frequency of outlier detection by ICS across 340 scenarios with varying smoothing parameters, for each observation in the Vietnam toy example.

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 Section 4.3.

3.3 Comparison with other methods

We now compare ICS for functional data (presented in Section 2.5) to eight outlier detection methods already existing in the literature, such as median-based approaches (Murph, Strait, et al. 2024), the modified band depth method (Sun and Genton 2011) and MUOD indices (Ojo, Fernández Anta, et al. 2022).

Our simulation uses three density-generating processes with 2% of outliers. The scheme named `GP_clr`, based on model 4 of the `fdaoutlier` package (Ojo, Fernández Anta, et al. 2022, section 4.1), first simulates a discretised random function in $L^2(0,1)$ from a mixture of two Gaussian processes with different means, and applies the inverse clr transformation to obtain a random density in the Bayes space $B^2(0,1)$. The scheme named `GP_margin` first simulates a discretised random function in $L^2(0,1)$ using model 5 of the `fdaoutlier` package, which consists in a mixture of two Gaussian processes with different covariance operators. Then, the random density is obtained as a kind of marginal distribution of the discrete values of the random function, where the x-axis is discarded: theses values are considered as a random sample and smoothed using MPL (see Section 2.6 with parameters $\lambda = 1$, 10 basis functions and knots (as well as interval bounds) at quantiles of the full sample. This scheme is similar to the data generating process of the Vietnamese climate dataset. Finally, the `Gumbel` scheme first draws parameters from a mixture of two Gaussian distributions in $\mathbb{R}^2$ and computes the Gumbel density functions corresponding to these parameters (it generates shift outliers as described in (Murph, Strait, et al. 2024)). Note that the output of all the schemes is a set of discretised densities on a regular grid of size $p = 100$ that covers an interval (a, b) (which is $(0, 1)$ for `GP_clr` and `Gumbel` and the range of the full sample for `GP_margin`). In each sample, there are $n = 200$ densities.

For the outlier detection methods, we denote them as `<Approach>_<Metric>` so that for instance, `ICS_B2` refers to ICS for density data in the Bayes space $B^2(a,b)$. The steps of the `ICS_B2` method are as follows. After applying the discrete clr transformation to each discretised density function, we approximate the underlying clr transformed smooth density by a smoothing spline in $L_0^2(a,b)$ using the preprocessing described in (Machalová, Hron, and Monti 2016). During this process, densities should not take values too close to 0 to avoid diverging clr, so we replace by 10^{-8} all density values below this threshold. The parameters of the compositional spline spaces are chosen by the function `fda.usc::fdata2fd`. Then, we solve ICS in the chosen compositional spline space, automatically selecting the components with tests as before. The `ICS_L2` method first smooths each discretised density using splines in $L^2(a,b)$ treating the densities as ordinary functional parameters. In the second step, we apply ICS in the chosen spline space, selecting the components automatically through D'Agostino normality tests. The MBD (López-Pintado and Romo 2009) and MUOD (Azcorra et al. 2018) approaches are implemented using the package `fdaoutlier` (Ojo, Lillo, and Anta 2023), either directly (`<Approach>_L2`) or after transforming the densities into log quantile densities (`<Approach>_LQD`) or into quantile functions (`<Approach>_QF`). The median-based methods such as `Median_LQD` and `Median_Wasserstein` are described in (Murph, Strait, et al. 2024) and implemented in the `DeBoinR` package from (Murph and Strait 2023) using the recommended default parameters.

For each combination between a generating scheme and a method, we average the TPR (True Positive Rate, or sensitivity) and the FPR (False Positive Rate, one minus specificity) over $N = 200$ repetitions, for each value of PP (the number of predicted positive) which scales from 0 to n. We also compute point-wise confidence bounds using the standard deviation of the TPR over the N repetitions and the standard Gaussian quantile of order 97.5%. The ROC curves together with their confidence bands are represented in Figure 5, separately for the three density-generating processes. Table 1 summarises the performance of the methods across the schemes, by means of the average area under the curve (AUC).

We can see that both ICS methods give quite similar results except for the `GP_clr` generating process where `ICS_B2` outperforms `ICS_L2`. Together with `MUOD_L2` and `MUOD_QF`, these methods are the best in terms of average AUC, although ICS-based methods perform more consistently across the different generating schemes. The `Median_LQD` and `MBD_LQD` methods are worse than the others for all generating schemes. Overall, we can recommend ICS versus the other outlier detection methods in this situation where the proportion of outliers is small.

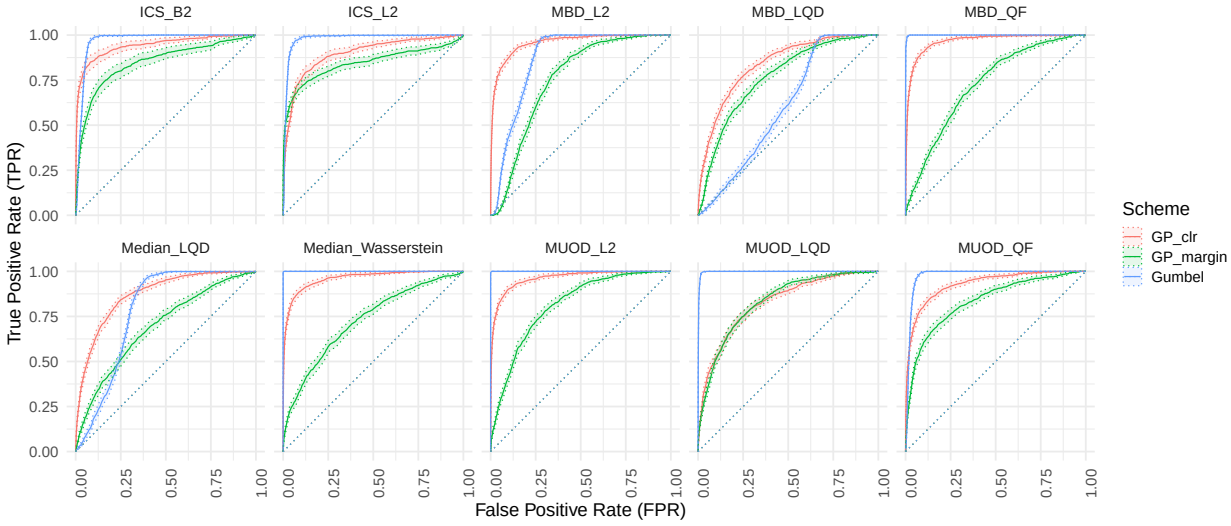


Figure 5: ROC curves of 10 different outlier detection methods for density data with 3 generating schemes.

Table 1: AUC for the 10 outlier detection methods, averaged across the 3 generating schemes.

Approach	Metric	Average AUC
MUOD	L2	0.92
ICS	B2	0.92
MUOD	QF	0.91

Table 1: AUC for the 10 outlier detection methods, averaged across the 3 generating schemes.

Approach	Metric	Average AUC
ICS	L2	0.91
MBD	QF	0.90
Median	Wasserstein	0.90
MUOD	LQD	0.88
MBD	L2	0.86
Median	LQD	0.78
MBD	LQD	0.74

4 An application to Vietnamese climate data

4.1 Data description and preprocessing

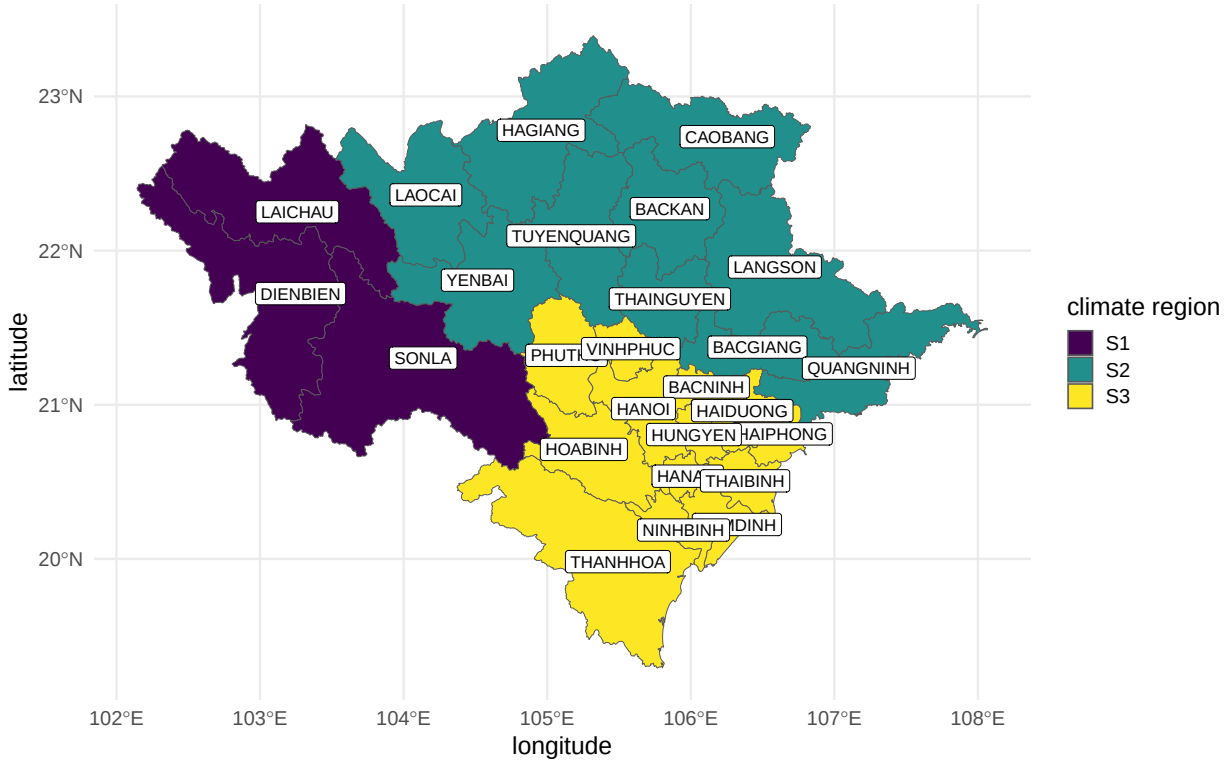


Figure 6: The three climate regions of Northern Vietnam.

In this application, we study 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, Thomas-Agnan, and Simioni 2023). 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 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. (2020). Figure 6 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. 2020).

Figure 7 shows the maximum temperature densities for the 13 provinces of S3, plotted by year, using the preprocessing detailed in Section 2.6.2 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.

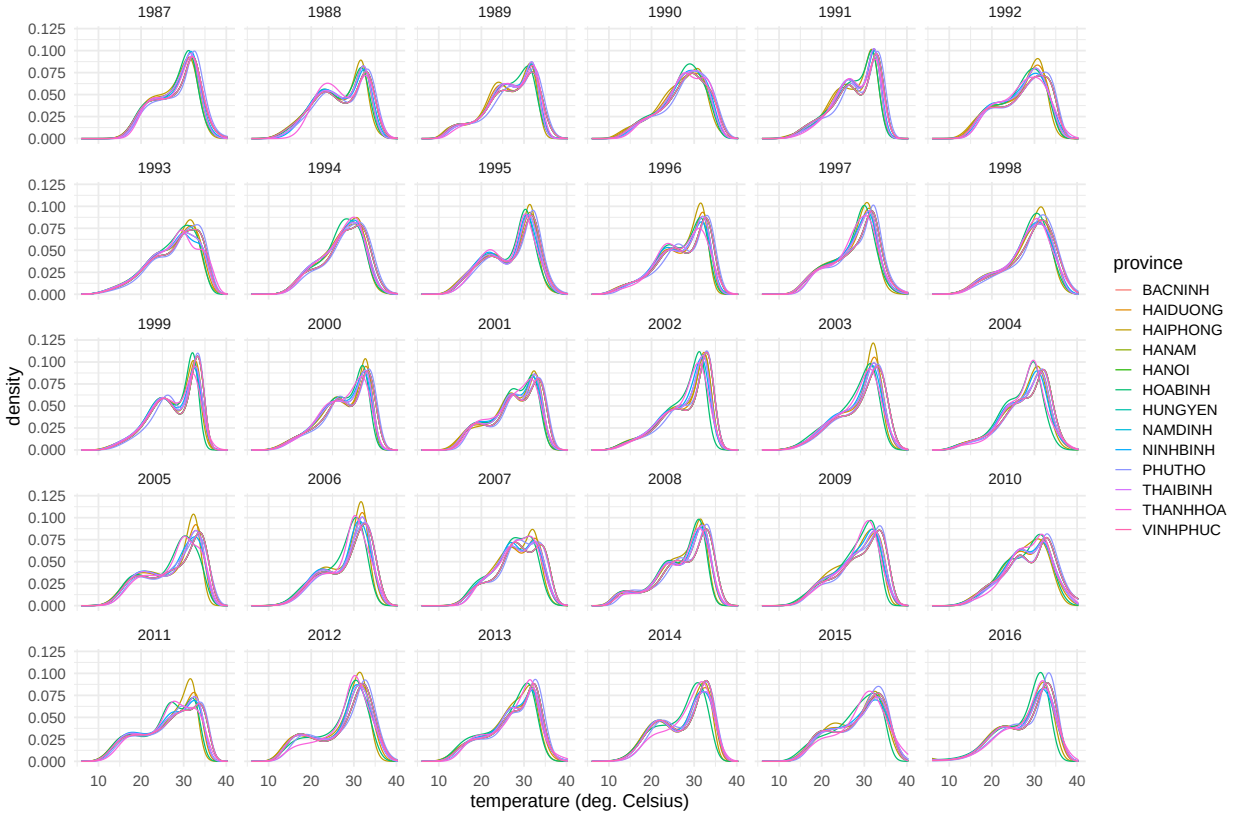


Figure 7: Maximum temperature densities for the 13 provinces in the S3 climate region of Northern Vietnam, 1987-2016, colour-coded by province.

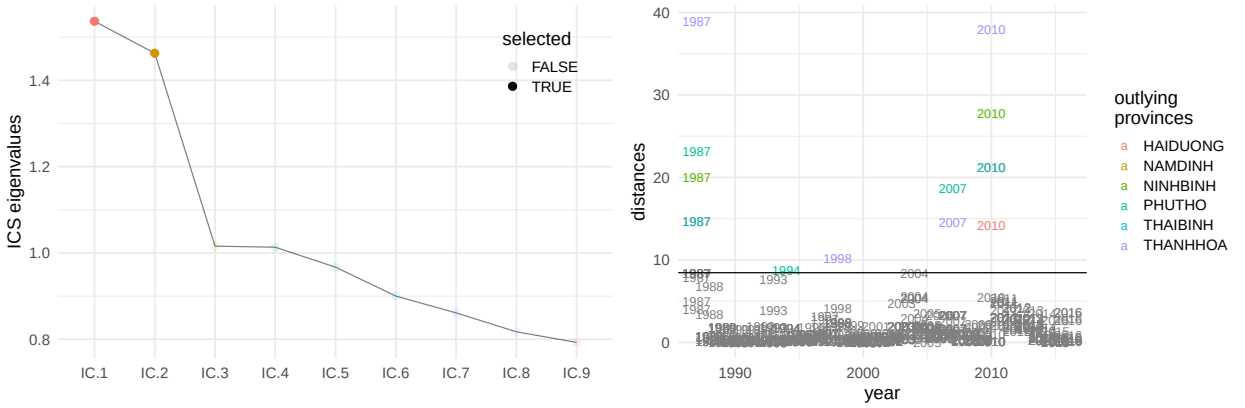
4.2 Outlier detection using ICS for the S3 climate region of Vietnam

We follow the different steps described in Section 3.1, and examine the results of ICS outlier detection using the scatter pair $(\mathrm{Cov}, \mathrm{Cov}_4)$ 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 8 clearly indicates that we should retain the first two invariant components. The right panel of Figure 8 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.

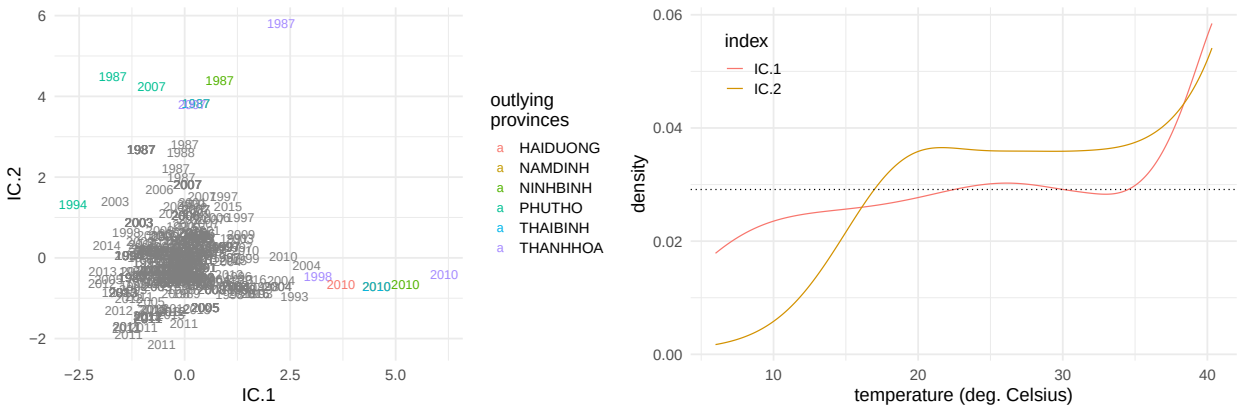
The left panel of Figure 9 displays the scatter plot of the first two components, labelled by year. The densities are coloured by province for the outliers and coloured in grey for the other provinces. This plot reveals that the outliers are either densities from 2010 (and one density from 1998) that 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 dual eigendensities plotted in the right panel of Figure 9 together with Figure 10, which represents the densities and their centred log-ratio transformation, colour-coded by year for the outliers and in grey for the other observations. This is justified by the reconstruction formula of Proposition 3 in the Appendix. The horizontal line on the eigendensities plot (right plot of Figure 9) 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 9). 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



(a) (b)

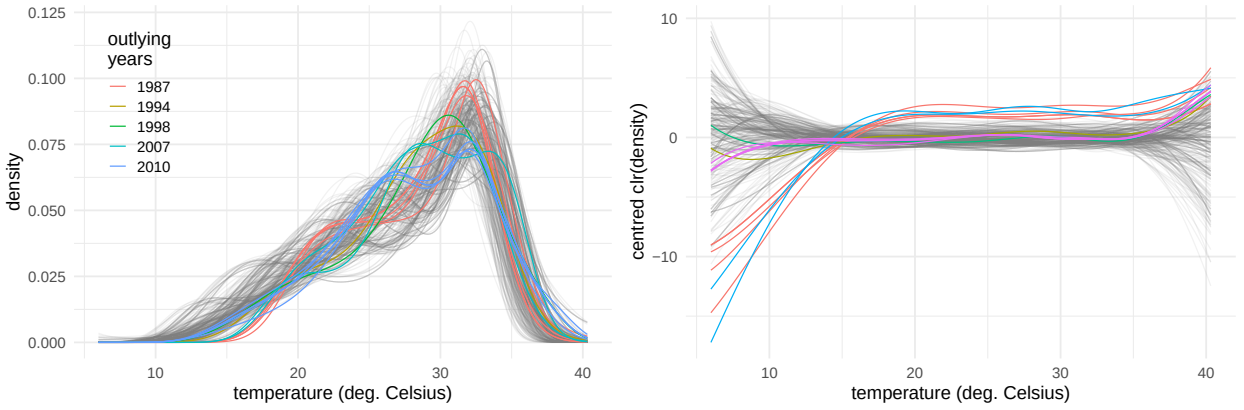
Figure 8: Scree plot of the ICS eigenvalues (left panel), and the ICS distances based on the first two components (right panel) for maximum temperature densities for the 13 provinces in the S3 climate region of Northern Vietnam, 1987-2016.



(a) (b)

Figure 9: Scatter plot of the first two invariant components (left panel) labelled by year and coloured by province, and the first two ICS dual eigendensities (right panel) of the maximum temperature densities for the 13 provinces in the S3 climate region of Northern Vietnam, 1987-2016.

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 10. Compared to the other densities, these four densities 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 9). 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 IC.2, and correspond to densities with very few days with maximum temperature less than 15 degrees Celsius compared to other densities.



(a) (b)

Figure 10: Maximum temperature densities (left panel) and their centred log-ratio transforms (right panel) for the 13 provinces in the S3 climate region of Northern Vietnam, 1987-2016, outlying densities are colour-coded by year.

4.3 Influence of the preprocessing parameters

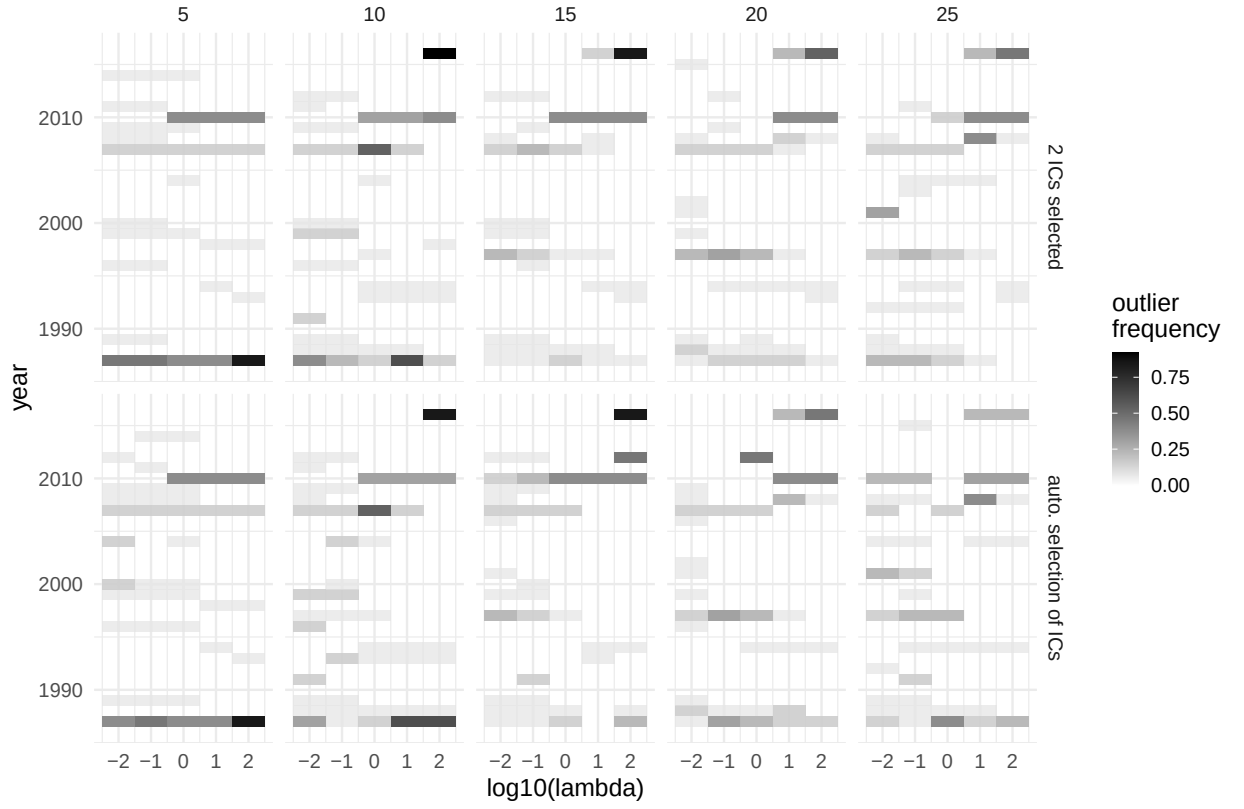


Figure 11: Outlier detection by ICS across smoothing parameters for the Vietnam climate data. *Top:* 2 invariant components selected; *Bottom:* automatic selection through D'Agostino tests. y-*axis:* year; x-*axis:* λ parameter. Columns correspond to knot numbers (5-25). Outliers are marked as light gray to black squares depending on their detection frequency.

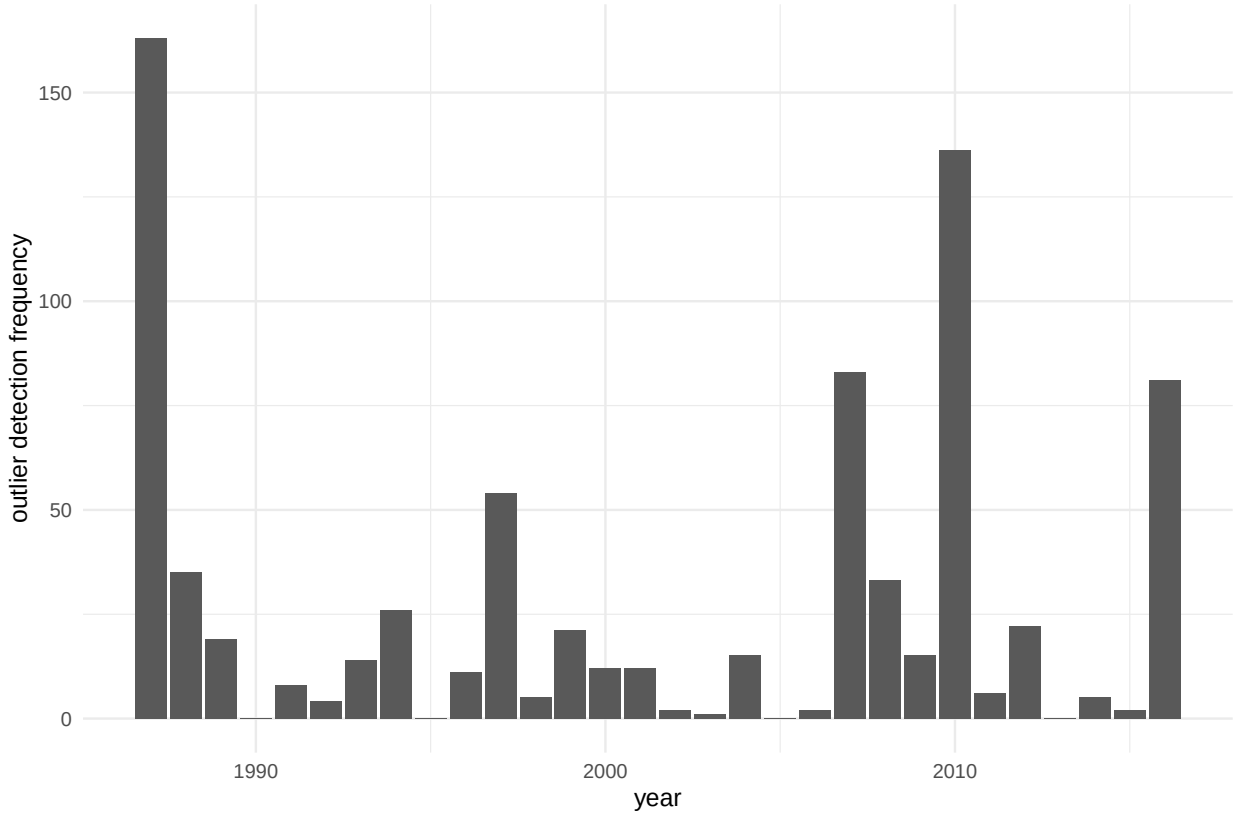


Figure 12: Frequency of outlier detection by ICS across all 25 scenarios with varying smoothing parameters and all 13 provinces, for each year in the Vietnamese climate dataset.

As mentioned in Section 3.2, 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 Section 3.1. 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 Section 3.1.

We plot in Figure 11 the years on the y-axes for the 25 smoothing parameter setups, indicating outlying years with light gray to black squares depending on their detection frequency. Figure 12 displays a bar plot of the frequency of outlier detection (across the 25 setups and the 13 provinces) for each year. Note that the choice of the number of selected invariant components has minimal impact. Both Figure 11 and Figure 12 confirm the results of the previous section. 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 (Stojanovic et al. 2020), 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 characterised by a behaviour that differs from the other densities in the tails of the distribution. This is not surprising because the Bayes inner product defined by equation Equation 9 involves the ratio of densities which can be large when a density is small (at the tails of the distribution).

5 Conclusion and perspectives

We propose a coordinate-free presentation of ICS that allows ICS to be applied to more complex objects than the coordinates vectors of multivariate analysis. We focus on the case of distributional data and describe an outlier detection procedure based on ICS. However, one of the limitations of the coordinate-free approach is that it is mainly adapted to pairs of weighted covariance operators, because they have a coordinate-free definition. 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 Archimbaud (2018) for a small proportion of outliers. But it is unclear how we could generalise 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 Alfons et al. 2024).

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 generalise the ICSOutlier procedure as in (Archimbaud, Boulfani, et al. 2022) 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, Tyler (2010) 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.

Acknowledgments

The major part of this work was completed while the authors were visiting the Vietnam Institute for Advanced Study in Mathematics (VIASM) in Hanoi 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. We also thank the two reviewers who gave us constructive comments that allowed us to improve our article.

Appendix

Scatter operators for random objects in a Hilbert space

Let us first discuss some definitions relative to scatter operators in the framework of a Hilbert space $(E, \langle\cdot,\cdot\rangle)$. We consider an E-valued random object $X : \Omega \to E$ where Ω is a probability space and E is a Hilbert 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

$$\forall (x, y) \in E^2, \langle \mathrm{Cov}[X]x, y\rangle = \mathbb{E}\left[\langle X - \mathbb{E}X, x\rangle\langle X - \mathbb{E}X, y\rangle\right], \tag{5}$$

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 Hilbert norm of X is assumed to be square-integrable, and A^* is the adjoint linear operator of A in the Hilbert space E, represented by the transpose of the matrix that represents A.

Definition 3 (Scatter operators)**.** Let $(E, \langle\cdot,\cdot\rangle)$ be a Hilbert space of dimension p, $\mathcal{E}$ an affine invariant set of E-valued random objects, i.e. that verifies:

$$\forall X \in \mathcal{E}, \forall A \in \mathcal{GL}(E), \forall b \in E, AX + b \in \mathcal{E}. \tag{6}$$

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:
$$\forall (X, Y) \in \mathcal{E}^2, X \sim Y \Rightarrow S[X] = S[Y].$$
2. Affine equivariance:
$$\forall X \in \mathcal{E}, \forall A \in \mathcal{GL}(E), \forall b \in E, S[AX + b] = AS[X]A^*.$$

We do not know whether there exist other scatter operators than the covariance when the Hilbert space has infinite dimension.

Details on coordinate-free ICS

The problem $\mathrm{ICS}(X, S_1, S_2)$ defined by Equation 1 is equivalent to assuming that $S_1[X]$ is injective and finding an orthonormal basis H that diagonalises the non-negative 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 2 (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 objects, 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 problem $\mathrm{ICS}(X, S_1, S_2)$, and Λ is a uniquely determined non-increasing sequence of positive real numbers.*

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

$$\begin{aligned}\forall (x, y) \in E^2, \langle S_1[X]S_1[X]^{-1}S_2[X]x, y\rangle &= \langle S_2[X]x, y\rangle = \langle S_2[X]y, x\rangle \\ &= \langle S_1[X]S_1[X]^{-1}S_2[X]y, x\rangle.\end{aligned}$$

Thus, the spectral theorem guarantees that there exists an orthonormal basis H of $(E, \langle S_1[X]\cdot, \cdot\rangle)$ in which $S_1[X]^{-1}S_2[X]$ is diagonal. □

This methodology does not generalise to the infinite-dimensional case, because the inner product space $(\mathcal{H}, \langle\cdot, S_1[X]\cdot\rangle)$ is not necessarily complete, so the spectral theorem does not apply.

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

$$h_j \in \underset{h\in E, \langle S_1[X]h, h_{j'}\rangle = 0 \text{ if } 0<j'<j}{\operatorname{argmax}} \frac{\langle S_2[X]h, h\rangle}{\langle S_1[X]h, h\rangle} \text{ and } \lambda_j = \max_{h\in E, \langle S_1[X]h, h_{j'}\rangle = 0 \text{ if } 0<j'<j} \frac{\langle S_2[X]h, h\rangle}{\langle S_1[X]h, h\rangle}. \tag{7}$$

The following reconstruction formula, extended from multivariate to complex data, is useful to interpret the ICS dual eigenbasis $H^* = (h_j^*)_{1\leq j\leq p}$, which is defined as the only basis of the space E that satisfies

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

Proposition 3 (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 objects, 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*

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

where the $z_j, 1 \leq j \leq p$ are defined as in Equation 2 and $H^* = (h_j^*)_{1\leq j\leq p} = (S_1[X]h_j)_{1\leq j\leq p}$ is the dual basis of H.

Reminder on Bayes spaces

The most recent and complete description of the Bayes spaces approach can be found in (Van Den Boogaart, Juan José Egozcue, and Pawlowsky-Glahn 2014). For the present work, we will identify the elements of a Bayes space, as defined by Van Den Boogaart, Juan José Egozcue, and Pawlowsky-Glahn (2014), 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.

Note that the simplex $\mathcal{S}^p$ used in compositional data analysis can be seen as a Bayes space when considering, instead of an interval (a, b) equipped with the Lebesgue measure, the finite set $\{1, \ldots, p+1\}$ equipped with the counting measure (see Example 2 in Van Den Boogaart, Juan José Egozcue, and Pawlowsky-Glahn 2014).

Let us first briefly recall 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

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

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 $u \in L_0^2(a,b)$,

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

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, Juan José Egozcue, and Pawlowsky-Glahn (2014). 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 J. J. Egozcue, Díaz–Barrero, and Pawlowsky–Glahn (2006) and of Van Den Boogaart, Juan José Egozcue, and Pawlowsky-Glahn (2014) 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 $L_0^2(a,b)$

$$\langle u, v \rangle_{L_0^2} = \int_a^b u(t) v(t) d\lambda(t), \tag{8}$$

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

$$\langle f, g \rangle_{B^2} = \frac{1}{2\lambda(a,b)} \int_a^b \int_a^b (\log f(t) - \log f(s))(\log g(t) - \log g(s)) d\lambda(t) d\lambda(s). \tag{9}$$

This normalization yields an inner product which is homogeneous to the measure λ whereas the Van Den Boogaart, Juan José Egozcue, and Pawlowsky-Glahn (2014) 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:

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

where $\ominus$ is the negative perturbation defined by $f \ominus g = f \oplus [(-1) \odot g]$.

Reminder on compositional splines

Following (Machalová, Talská, et al. 2021), in order to construct a basis of $E = \mathcal{C}_d^{\Delta\gamma}(a,b)$, which is required in practice, 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, Machalová, Hron, and Monti (2016) 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 characterised 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 $p = k + d$. Let $\mathcal{C}_d^{\Delta\gamma}$ 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

$$f(t) = \bigoplus_{j=1}^{p} [f]_{C_j} C_j(t), \tag{10}$$

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

$$\mathrm{clr}\, f(t) = \sum_{j=1}^{p} [f]_{C_j} Z_j(t). \tag{11}$$

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, Hron, and Monti (2016), the basis functions of $\mathcal{Z}_d^{\Delta\gamma}$ can be written in a B-spline basis, see Schumaker (1981), which is convenient to allow using existing code for their computation.

Proofs

Proposition 1. First, let us verify that the problem $\mathrm{ICS}(X^{\mathcal{F}}, S_1^{\mathcal{F}}, S_2^{\mathcal{F}})$ 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

$$\|\varphi(X)\|_F = \|X\|_E$$

and

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

(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 $\ell \in \{1, 2\}$,

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

and

$$\begin{aligned} S_\ell^{\mathcal{F}}[A\varphi(X) + b] &= \varphi \circ S_\ell^{\mathcal{E}}[\varphi^{-1} \circ A \circ \varphi(X) + \varphi^{-1}(b)] \circ \varphi^{-1} \\ &= A \circ \varphi \circ S_\ell^{\mathcal{E}}[X] \circ \varphi^{-1} \circ A^* = A S_\ell^{\mathcal{F}}[\varphi(X)] A^*. \end{aligned}$$

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

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

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

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

Corollary 1. Let $\ell \in \{1, 2\}$ and $\tilde{X} = X - \mathbb{E}X$. In order to prove the equation Equation 4, we will need to prove that, for any $(x, y) \in F^2$,

$$\begin{aligned} \langle \varphi \circ \mathrm{Cov}_{w_\ell}^E[X] \circ \varphi^{-1}(x), y \rangle_F &= \langle \mathrm{Cov}_{w_\ell}^E[X] \varphi^{-1}(x), \varphi^{-1}(y) \rangle_E \\ &= \mathbb{E}[w_\ell(\|\mathrm{Cov}^E[X]^{-1/2} \tilde{X}\|_E)^2 \langle \tilde{X}, \varphi^{-1}(x) \rangle_E \langle \tilde{X}, \varphi^{-1}(y) \rangle_E] \\ &= \mathbb{E}[w_\ell(\|\mathrm{Cov}^F[\varphi(X)]^{-1/2} \varphi(\tilde{X})\|_F)^2 \langle \varphi(\tilde{X}), x \rangle_F \langle \varphi(\tilde{X}), y \rangle_F] \\ \langle \varphi \circ \mathrm{Cov}_{w_\ell}^E[X] \circ \varphi^{-1}(x), y \rangle_F &= \langle \mathrm{Cov}_{w_\ell}^F[\varphi(X)] x, y \rangle_F. \end{aligned} \tag{12}$$

It is enough to show the equality between Equation 12 (2) and Equation 12 (3), 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 Equation 4 holds for the covariance operator. If $w_\ell \neq 1$, we now know from the case $w_\ell = 1$ that

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

so that

$$\|\mathrm{Cov}^E[X]^{-1/2} \tilde{X}\|_E = \|\mathrm{Cov}^F[\varphi(X)]^{-1/2} \varphi(\tilde{X})\|_F \tag{13}$$

Once the equation Equation 4 is proved, one only needs to apply Proposition 1 to finish the proof. □

Corollary 2. Applying Corollary 1 to the isometry

$$\varphi_B : \begin{cases} (E, \langle \cdot, \cdot \rangle_E) & \to & (\mathbb{R}^p, \langle \cdot, \cdot \rangle_{\mathbb{R}^p}) \\ x & \mapsto & G_B^{1/2}[x]_B, \end{cases}$$

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 $(x, y) \in E^2$:

$$\begin{aligned} \langle \mathrm{Cov}_{w_\ell}^E[X]x, y\rangle_E &= \langle \mathrm{Cov}_{w_\ell}(G_B^{1/2}[X]_B)G_B^{1/2}[x]_B, G_B^{1/2}[y]_B\rangle_{\mathbb{R}^p} \\ &= \langle \mathrm{Cov}_{w_\ell}(G_B[X]_B)[x]_B, [y]_B\rangle_{\mathbb{R}^p} \\ &= \langle \mathrm{Cov}_{w_\ell}([X]_B)G_B[x]_B, G_B[y]_B\rangle_{\mathbb{R}^p}, \end{aligned} \tag{14}$$

where Equation 14 (1) comes from the equation Equation 4, and the equalities Equation 14 (2) and Equation 14 (3) come from the affine equivariance of Cov_{w_ℓ}. □

Proposition 3. Let us decompose $S_1[X]^{-1}(X - \mathbb{E}X)$ over the basis H, which is orthonormal in $(E, \langle \cdot, S_1[X]\cdot\rangle)$:

$$\begin{aligned} S_1[X]^{-1}(X - \mathbb{E}X) &= \sum_{j=1}^{p} \langle S_1[X]^{-1}(X - \mathbb{E}X), S_1[X]h_j\rangle h_j \\ &= \sum_{j=1}^{p} \langle X - \mathbb{E}X, h_j\rangle h_j \\ S_1[X]^{-1}(X - \mathbb{E}X) &= \sum_{j=1}^{p} z_j h_j. \end{aligned}$$

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

Code & reproducibility

In order to implement coordinate-free ICS, we created the R package `ICSFun`, which is used to generate the figures (see the code in this HTML version of the article).

References

Aggarwal, Charu C. (2017). *Outlier Analysis*. en. Cham: Springer International Publishing. ISBN: 978-3-319-47577-6 978-3-319-47578-3. DOI: 10.1007/978-3-319-47578-3. (Visited on 08/05/2024) (cit. on p. 2).

Alfons, Andreas et al. (Mar. 2024). "Tandem clustering with invariant coordinate selection". In: *Econometrics and Statistics*. ISSN: 2452-3062. DOI: 10.1016/j.ecosta.2024.03.002. (Visited on 07/18/2024) (cit. on pp. 2, 19).

Archimbaud, Aurore (2018). "Détection non-supervisée d'observations atypiques en contrôle de qualité: un survol". In: *Journal de la Société Française de Statistique* 159.3, pp. 1–39 (cit. on p. 19).

Archimbaud, Aurore, Feriel Boulfani, et al. (Mar. 2022). "ICS for multivariate functional anomaly detection with applications to predictive maintenance and quality control". In: *Econometrics and Statistics*. ISSN: 2452-3062. DOI: 10.1016/j.ecosta.2022.03.003. (Visited on 01/10/2024) (cit. on pp. 2, 3, 6, 10, 19).

Archimbaud, Aurore, Zlatko Drmač, et al. (Mar. 2023). "Numerical Considerations and a new implementation for invariant coordinate selection". en. In: *SIAM Journal on Mathematics of Data Science* 5.1, pp. 97–121. ISSN: 2577-0187. DOI: 10.1137/22M1498759. (Visited on 05/03/2024) (cit. on p. 5).

Archimbaud, Aurore, Klaus Nordhausen, and Anne Ruiz-Gazen (Dec. 2018). "ICS for multivariate outlier detection with application to quality control". en. In: *Computational Statistics & Data Analysis* 128, pp. 184–199. ISSN: 01679473. DOI: 10.1016/j.csda.2018.06.011. (Visited on 10/13/2022) (cit. on pp. 2, 8).

Azcorra, A. et al. (May 2018). "Unsupervised Scalable Statistical Method for Identifying Influential Users in Online Social Networks". en. In: *Scientific Reports* 8.1. Publisher: Nature Publishing Group, p. 6955. ISSN: 2045-2322. DOI: 10.1038/s41598-018-24874-2. URL: https://www.nature.com/articles/s41598-018-24874-2 (visited on 04/03/2025) (cit. on p. 13).

Dai, Wenlin et al. (Sept. 2020). "Functional outlier detection and taxonomy by sequential transformations". In: *Computational Statistics & Data Analysis* 149, p. 106960. ISSN: 0167-9473. DOI: 10.1016/j.csda.2020.106960. (Visited on 08/05/2024) (cit. on p. 2).

Egozcue, J. J., J. L. Díaz–Barrero, and V. Pawlowsky–Glahn (July 2006). "Hilbert Space of Probability Density Functions Based on Aitchison Geometry". en. In: *Acta Mathematica Sinica, English Series* 22.4, pp. 1175–1182. ISSN: 1439-8516, 1439-7617. DOI: 10.1007/s10114-005-0678-2. (Visited on 04/08/2024) (cit. on p. 21).

Eubank, Randall L. (Apr. 2014). *Nonparametric Regression and Spline Smoothing*. 2nd ed. Boca Raton: CRC Press. ISBN: 978-0-429-18267-9. DOI: 10.1201/9781482273144 (cit. on p. 6).

Ferraty, Frédéric and Philippe Vieu (Dec. 2002). "The Functional Nonparametric Model and Application to Spectrometric Data". en. In: *Computational Statistics* 17.4, pp. 545–564. ISSN: 1613-9658. DOI: 10.1007/s001800200126. (Visited on 12/30/2024) (cit. on p. 7).

Lei, Xinyi, Zhicheng Chen, and Hui Li (July 2023). "Functional Outlier Detection for Density-Valued Data with Application to Robustify Distribution-to-Distribution Regression". In: *Technometrics* 65.3, pp. 351–362. ISSN: 0040-1706. DOI: 10.1080/00401706.2022.2164063. (Visited on 03/27/2024) (cit. on p. 2).

Li, Bing et al. (2021). *Functional independent component analysis : an extension of fourth-order blind identification*. URL: https://sites.google.com/site/germainvanbever/publica (visited on 10/18/2023) (cit. on p. 2).

Loperfido, Nicola (Nov. 2021). "Some theoretical properties of two kurtosis matrices, with application to invariant coordinate selection". In: *Journal of Multivariate Analysis* 186, p. 104809. ISSN: 0047-259X. DOI: 10.1016/j.jmva.2021.104809. (Visited on 03/13/2024) (cit. on p. 1).

López-Pintado, Sara and Juan Romo (June 2009). "On the Concept of Depth for Functional Data". In: *Journal of the American Statistical Association* 104.486. Publisher: ASA Website _eprint: https://doi.org/10.1198/jasa.2009.0108, pp. 718–734. ISSN: 0162-1459. DOI: 10.1198/jasa.2009.0108. URL: https://doi.org/10.1198/jasa.2009.0108 (visited on 04/03/2025) (cit. on p. 13).

Machalová, J., K. Hron, and G.S. Monti (June 2016). "Preprocessing of centred logratio transformed density functions using smoothing splines". en. In: *Journal of Applied Statistics* 43.8, pp. 1419–1435. ISSN: 0266-4763, 1360-0532. DOI: 10.1080/02664763.2015.1103706. (Visited on 03/12/2024) (cit. on pp. 7, 13, 21).

Machalová, J., Renáta Talská, et al. (June 2021). "Compositional splines for representation of density functions". en. In: *Computational Statistics* 36.2, pp. 1031–1064. ISSN: 0943-4062, 1613-9658. DOI: 10.1007/s00180-020-01042-7. (Visited on 03/12/2024) (cit. on pp. 7, 21).

Menafoglio, Alessandra (Apr. 2021). *Anomaly detection for density data based on control charts*. IASC-ERS Course. URL: https://iasc-isi.org/events/iasc-ers-course-an-introduction-to-functional-data-analysis-for-density-functions-in-bayes-spaces/ (cit. on p. 2).

Murph, Alexander C. and Justin D. Strait (Dec. 2023). *DeBoinR: Box-Plots and Outlier Detection for Probability Density Functions*. URL: https://cran.r-project.org/web/packages/DeBoinR/ (visited on 03/27/2024) (cit. on p. 13).

Murph, Alexander C., Justin D. Strait, et al. (2024). "Visualisation and outlier detection for probability density function ensembles". en. In: *Stat* 13.2, e662. ISSN: 2049-1573. DOI: 10.1002/sta4.662. (Visited on 04/12/2024) (cit. on pp. 2, 12, 13).

Nordhausen, Klaus, Aurore Archimbaud, and Anne Ruiz-Gazen (Dec. 2023). *ICSOutlier: Outlier Detection Using Invariant Coordinate Selection*. URL: https://cran.r-project.org/web/packages/ICSOutlier/ (visited on 07/19/2024) (cit. on pp. 2, 8).

Nordhausen, Klaus and Anne Ruiz-Gazen (Mar. 2022). "On the usage of joint diagonalization in multivariate statistics". In: *Journal of Multivariate Analysis*. 50th Anniversary Jubilee Edition 188, p. 104844. ISSN: 0047-259X. DOI: 10.1016/j.jmva.2021.104844. (Visited on 07/02/2024) (cit. on pp. 1, 2, 4).

Ojo, Oluwasegun Taiwo, Antonio Fernández Anta, et al. (Sept. 2022). "Detecting and classifying outliers in big functional data". en. In: *Advances in Data Analysis and Classification* 16.3, pp. 725–760. ISSN: 1862-5355. DOI: 10.1007/s11634-021-00460-9. (Visited on 12/27/2024) (cit. on p. 12).

Ojo, Oluwasegun Taiwo, Rosa Elvira Lillo, and Antonio Fernandez Anta (Sept. 2023). *fdaoutlier: Outlier Detection Tools for Functional Data Analysis*. URL: https://cran.r-project.org/web/packages/fdaoutlier/index.html (visited on 03/19/2025) (cit. on p. 13).

Pawlowsky-Glahn, Vera, Juan José Egozcue, and Raimon Tolosana-Delgado (Mar. 2015). *Modeling and Analysis of Compositional Data*. en. 1st ed. Wiley. ISBN: 978-1-118-44306-4 978-1-119-00314-4. DOI: 10.1002/9781119003144. (Visited on 07/19/2024) (cit. on p. 5).

Ramsay, James, Giles Hooker, and Spencer Graves (Mar. 2024). *fda: Functional Data Analysis*. URL: https://cran.r-project.org/web/packages/fda/ (visited on 03/16/2024) (cit. on p. 7).

Rendón Aguirre, Janeth Carolina (May 2017). "Clustering in high dimension for multivariate and functional data using extreme kurtosis projections". eng. PhD thesis. Universidad Carlos III de Madrid. DOI: 10016/25286. (Visited on 03/14/2024) (cit. on p. 2).

Rousseeuw, Peter (1985). "Multivariate Estimation with High Breakdown Point". en. In: *Mathematical Statistics and Applications*. Ed. by Wilfried Grossmann et al. Dordrecht: Springer Netherlands, pp. 283–297. ISBN: 978-94-010-8901-2 978-94-009-5438-0. DOI: 10.1007/978-94-009-5438-0_20. (Visited on 12/30/2024) (cit. on p. 5).

Ruiz-Gazen, Anne et al. (2023). "Detecting Outliers in Compositional Data Using Invariant Coordinate Selection". en. In: *Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler*. Ed. by Mengxi Yi and Klaus Nordhausen. Cham: Springer International Publishing, pp. 197–224. ISBN: 978-3-031-22687-8. DOI: 10.1007/978-3-031-22687-8_10. (Visited on 10/12/2023) (cit. on pp. 2, 5).

Schumaker, Larry (1981). *Spline Functions: Basic Theory*. 3rd ed. Cambridge University Press. ISBN: 978-0-521-70512-7 978-0-511-61899-4. DOI: 10.1017/CBO9780511618994. (Visited on 10/20/2023) (cit. on p. 21).

Silverman, B. W. (Sept. 1982). "On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method". In: *The Annals of Statistics* 10.3, pp. 795–810. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176345872. (Visited on 12/05/2023) (cit. on p. 7).

Stojanovic, Milica et al. (Mar. 2020). "Trends and Extremes of Drought Episodes in Vietnam Sub-Regions during 1980–2017 at Different Timescales". en. In: *Water* 12.3. Number: 3, p. 813. ISSN: 2073-4441. DOI: 10.3390/w12030813. (Visited on 06/26/2024) (cit. on pp. 14, 18).

Stone, Mervyn (1987). *Coordinate-Free Multivariable Statistics: An Illustrated Geometric Progression from Halmos to Gauss and Bayes*. eng. Oxford statistical science series 2. Oxford: Clarendon Pr. ISBN: 978-0-19-852210-2 (cit. on p. 2).

Sun, Ying and Marc G. Genton (Jan. 2011). "Functional Boxplots". In: *Journal of Computational and Graphical Statistics* 20.2. Publisher: ASA Website _eprint: https://doi.org/10.1198/jcgs.2011.09224, pp. 316–334. ISSN: 1061-8600. DOI: 10.1198/jcgs.2011.09224. URL: https://doi.org/10.1198/jcgs.2011.09224 (visited on 04/03/2025) (cit. on p. 12).

Trinh, Thi Huong, Christine Thomas-Agnan, and Michel Simioni (Feb. 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 (cit. on p. 14).

Tyler, David E. (Sept. 2010). "A note on multivariate location and scatter statistics for sparse data sets". In: *Statistics & Probability Letters* 80.17, pp. 1409–1413. ISSN: 0167-7152. DOI: 10.1016/j.spl.2010.05.006. (Visited on 03/13/2024) (cit. on pp. 3, 19).

Tyler, David E. et al. (June 2009). "Invariant co-ordinate selection". en. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.3, pp. 549–592. ISSN: 13697412, 14679868. DOI: 10.1111/j.1467-9868.2009.00706.x. (Visited on 10/13/2022) (cit. on pp. 1–3, 8).

Van Den Boogaart, Karl Gerald, Juan José Egozcue, and Vera Pawlowsky-Glahn (June 2014). "Bayes Hilbert Spaces". en. In: *Australian & New Zealand Journal of Statistics* 56.2, pp. 171–194. ISSN: 13691473. DOI: 10.1111/anzs.12074. (Visited on 09/15/2023) (cit. on pp. 7, 20, 21).

Virta, Joni et al. (2020). "Independent component analysis for multivariate functional data". In: *Journal of Multivariate Analysis* 176, p. 104568. DOI: 10.1016/j.jmva.2019.104568 (cit. on p. 2).