


September 2024

“Invariant Coordinate Selection and Fisher Discriminant Subspace Beyond The Case of Two Groups”


Colombe Becquart, Aurore Archimbaud, Anne Ruiz-Gazen,
Luka Prilé and Klaus Nordhausen


INVARIANT COORDINATE SELECTION AND FISHER DISCRIMINANT SUBSPACE BEYOND THE CASE OF TWO GROUPS


A PREPRINT

 **Colombe Becquart**
Toulouse School of Economics
France
Université de Toulouse
France

 **Aurore Archimbaud***
TBS Business School
France

 **Anne Ruiz-Gazen**
Toulouse School of Economics
France

 **Luka Prilć**
Toulouse School of Economics
France

 **Klaus Nordhausen**
Department of Mathematics and Statistics
University of Jyväskylä
Finland

September 26, 2024

ABSTRACT

Invariant Coordinate Selection (ICS) is a multivariate technique that relies on the simultaneous diagonalization of two scatter matrices. It serves various purposes, including its use as a dimension reduction tool prior to clustering or outlier detection. Unlike methods such as Principal Component Analysis, ICS has a theoretical foundation that explains why and when the identified subspace should contain relevant information. These general results have been examined in detail primarily for specific scatter combinations within a two-cluster framework. In this study, we expand these investigations to include more clusters and scatter combinations. The case of three clusters in particular is studied at length. Based on these expanded theoretical insights and supported by numerical studies, we conclude that ICS is indeed suitable for recovering Fisher’s discriminant subspace under very general settings and cases of failure seem rare.

Keywords Dimension reduction · Simultaneous diagonalization · Mixture of elliptical distributions · Scatter matrix · Subspace estimation

1 Introduction

In many fields, the number of variables is increasing while it is often assumed that the actual information of interest remains contained within a low-dimensional subspace. This is the fundamental idea behind dimension reduction (DR): that it is possible to estimate this lower-dimensional space without losing crucial information, and that analyzing data within this subspace simplifies the process. Clustering and outlier detection are two unsupervised multivariate methods that can benefit significantly from prior dimension reduction. However, the justifications for why a specific DR method is suitable to recover an effective subspace for clustering and outlier detection are often heuristic in nature. Typical dimension reduction methods include principal component analysis (PCA, Hotelling [1933], Jolliffe [2002]), projection pursuit (PP, Huber [1985], Jones and Sibson [1987], Fischer et al. [2021]), and invariant coordinate selection (ICS, Tyler et al. [2009], Caussinus and Ruiz [1990]).

*Corresponding author: aurore.archimbaud@tbs-education.fr

Recent research has sought to provide justifications for using these methods as preprocessing steps, often considering specific settings. The most popular framework involves the Gaussian mixture model, with the benchmark being whether the DR method can estimate Fisher’s discriminant subspace (FDS) Fisher [1936] in an unsupervised, or blind, manner. Outlier detection approaches are incorporated into this framework when some clusters are rare.

In the context of PCA, also known as tandem PCA Arabie and Hubert [1994] when it is used in conjunction with clustering, it is well-documented that FDS is seldom obtained Radojčić et al. [2021]. PP is considered here when skewness or kurtosis are used as PP indices, and conditions under which these estimate the FDS are detailed in Radojčić et al. [2021] in the two-cluster case. ICS is the most recent DR method among those mentioned. ICS simultaneously diagonalizes two scatter matrices and has been considered for outlier detection in Archimbaud et al. [2018] and as tandem clustering with ICS in Alfons et al. [2024]. Tyler et al. [2009] actually provide very general results for ICS with respect to FDS estimation. However, detailed results exist only for very specific scatter combinations in very specific settings. The goal of this paper is to extend these results to broader settings. The structure of the paper is as follows. In Section 2 we recall ICS in more detail, particularly focusing on its application in estimating FDS. Section 3 examines the properties of ICS for DR when the clusters in the mixtures exhibit no variability. In this context, and when the dimension of the FDS is equal to the number of groups minus one, we prove that for any scatter combination, the ICS eigenvalues do not depend on the cluster locations but only on the cluster proportions. Section 4 then gives closer attention to this model for a specific scatter combination and studies the behavior of the ICS eigenvalues when the cluster proportions vary. Section 5 explores the same scatter combination within a certain Gaussian mixture model consisting of three clusters. Section 6 validates the theoretical results and investigates the behavior of previously unexamined scatter combinations in such settings through a simulation study. The paper concludes with Section 7. Calculations and proofs are provided in the appendix together with some details on the parameters for our experiments.

2 What is already known about ICS in relation with the Fisher discriminant subspace

2.1 General principle of ICS

Following the notations in Tyler et al. [2009], let Y be a random vector of dimension p , with distribution function F_Y . Let \mathcal{P}_p be the set of all positive definite symmetric matrices of order p . A scatter matrix of Y , denoted by $V(F_Y)$, is a function of the distribution of Y that is uniquely defined at F_Y , is such that $V(F_Y)$ belongs to \mathcal{P}_p , and is affine equivariant, i.e. $V(F_{AY+b}) = AV(F_Y)A^\top$ for all non-singular matrices A of dimension $p \times p$ and for all $b \in \mathbb{R}^p$, and where $^\top$ denotes the transpose operation. In what follows, we will drop the dependence on F_Y and simply denote by V the scatter $V(F_Y)$ when the context is obvious.

If the data distribution is elliptical, then all affine equivariant scatter matrices are proportional Nordhausen and Tyler [2015]. This property is not true in general, outside the context of elliptical distributions, and ICS exploits the difference between two scatter matrices to detect non-elliptical structures such as clusters [Alfons et al., 2024] or outliers [Archimbaud et al., 2018]. To perform this comparison, ICS relies on the simultaneous diagonalization of two scatter matrices V_1 and V_2 :

$$H^\top V_1 H = D_1 \quad \text{and} \quad H^\top V_2 H = D_2,$$

where D_1 and D_2 are diagonal matrices such that $D_1^{-1}D_2 = \text{diag}(\rho_1, \dots, \rho_p)$, $\rho_1 \geq \dots \geq \rho_p$ being the eigenvalues of $V_1^{-1}V_2$ sorted in descending order, and $H = (h_1, \dots, h_p)$ is a $p \times p$ non singular matrix containing the corresponding eigenvectors. Usually H is scaled such that $D_1 = I_p$, where I_p denotes the identity matrix of dimension p . The term “generalized eigendecomposition” (and correspondingly, “generalized eigenvalue” and “generalized eigenvector”) is sometimes used in this context, but for simplicity, we avoid this terminology in the present paper.

The $Z = H^\top Y$ transformation of Y leads to new variables, that are invariant under affine transformation in the sense of Theorems 1 and 2 in Tyler et al. [2009], and are called invariant coordinates or components. Apart from being invariant by affine transformations, ICS has many applications and useful properties, as investigated for example in Nordhausen et al. [2008], Peña et al. [2010], Nordhausen et al. [2011], Loperfido [2013], Alashwali and Kent [2016], Loperfido [2021], Nordhausen and Ruiz-Gazen [2022], Archimbaud et al. [2023a]. Of particular interest to us, is the following property for finite mixture models of the form:

$$f_Y(y) = \det(\Gamma)^{-1/2} \sum_{j=1}^k \alpha_j g_j((y - \mu_j)^\top \Gamma^{-1} (y - \mu_j)), \quad (1)$$

where $\alpha_j > 0$ for $j \in \{1, \dots, k\}$, with $\sum_{j=1}^k \alpha_j = 1$, are the mixture proportions, the μ_j are distinct location vectors (also called group centers or group means), $\Gamma \in \mathcal{P}_p$ is the within-group scatter matrix parameter, and g_1, \dots, g_k are

non-negative functions. Thus, each of the k components of the mixture is elliptical and the standard Gaussian mixture model is a special case. For such a model, Theorem 4 from Tyler et al. [2009] is fundamental and justifies the use of ICS as a dimension reduction method when the objective is clustering or anomaly detection. In short, this result proves that for model (1) and under some conditions, the ICS components associated with the largest and/or smallest eigenvalues span the FDS which is the one obtained by using the linear discriminant analysis in a supervised context. In the next subsection, we recall this theorem and also discuss some of its limitations.

2.2 ICS and the Fisher discriminant subspace

In a supervised context, when the distribution of Y is a mixture of distributions, Fisher [1936] suggested to look for the linear function of the p variables in Y which maximize the ratio of the between-group variability to the within-group variability. This function corresponds to the projection of Y onto the eigenvector associated with the largest eigenvalue of the simultaneous decomposition of the between and within-group covariance matrices. In the case of a mixture of two Gaussian groups with means μ_1 and μ_2 and equal within-group covariance matrices Γ , this eigenvector is equal to $\Gamma^{-1}(\mu_1 - \mu_2)$ and spans the FDS with dimension $q = k - 1 = 1$. For a mixture of k groups such as model (1), the FDS is spanned by the vectors $\Gamma^{-1}(\mu_j - \mu_k)$ for $j \in \{1, \dots, k - 1\}$.

Let us denote by q the dimension of the vector space associated with the affine space spanned by the group centers μ_j , $j \in \{1, \dots, k\}$. According to Theorem 4 from Tyler et al. [2009], the simultaneous diagonalization of two scatter matrices results in at least one eigenvalue, denoted by ρ^* , with a multiplicity greater than or equal to $p - q$. This eigenvalue is associated with an eigenspace of dimension at least $p - q$ which is the direct sum of the complementary space of the FDS (with dimension $p - q$) and, if the multiplicity of ρ^* is strictly larger than $p - q$, a space that is included in the FDS. If no eigenvalue has multiplicity greater than $p - q$, the subspace spanned by the eigenvectors associated with the other eigenvalues than ρ^* (larger or smaller) is the FDS.

Theorem 4 in Tyler et al. [2009] is proven by using the equivariance property of the scatter matrices V_1 and V_2 , and transforming the random vector Y , with distribution given by (1), in the following way. Let $M = (\mu_1, \dots, \mu_k)$ be the matrix of rank q which contains the location vectors of Y , and let $M_0 = \Gamma^{-\frac{1}{2}}(M - \mu_k \mathbf{1}_k^\top)$, where $\mathbf{1}_k$ is a k -dimensional vector of ones. The QR decomposition of M_0 is:

$$M_0 = PT = P \begin{pmatrix} T_u & 0 \\ 0 & 0 \end{pmatrix},$$

where P is an orthogonal matrix, $T = (t_1, \dots, t_k)$ with t_j a p -dimensional vector for $j \in \{1, \dots, k\}$, and T_u is an upper triangular matrix of dimension $k - 1 \geq 1$ such that the last $k - 1 - q \geq 0$ rows are zero. Note that the dimension of T_u in Tyler et al. [2009] differs slightly from ours, but this does not impact the remainder of the proof. The distribution of the transformed random vector $X = P^\top \Gamma^{-\frac{1}{2}}(Y - \mu_k \mathbf{1}_k^\top)$ is a mixture of spherical distributions with density:

$$f_X(x) = \sum_{j=1}^k \alpha_j g_j((x - t_j)^\top (x - t_j)), \quad (2)$$

where for $j \in \{1, \dots, k\}$, $\alpha_j > 0$, $\sum_{j=1}^k \alpha_j = 1$, the t_j are distinct, t_k is the zero vector, and g_1, \dots, g_k are non-negative functions. By decomposing the scatter matrices $V_1(F_X)$ and $V_2(F_X)$ in four blocks where the left top block has dimension $q \times q$, and using the affine equivariance of the scatter matrices, Tyler et al. [2009] are able to write:

$$V_1(F_X)^{-1} V_2(F_X) = \begin{pmatrix} A_q & 0 \\ 0 & \gamma I_{p-q} \end{pmatrix},$$

for some $\gamma > 0$. This result implies that $V_1(F_X)^{-1} V_2(F_X)$ has an eigenvalue equal to γ with multiplicity at least equal to $p - q$. The associated eigenspace can be written as the direct sum of the complementary space of the FDS (with dimension $p - q$) and, if the multiplicity of γ is strictly larger than $p - q$, a space that is included in the FDS. The result is then derived for Y by noting that the eigenvalues of $V_1(F_Y)^{-1} V_2(F_Y)$ are proportional to the ones of $V_1(F_X)^{-1} V_2(F_X)$. Thus there exists an eigenvalue ρ^* of $V_1(F_Y)^{-1} V_2(F_Y)$ with multiplicity at least $p - q$. Moreover, if the multiplicity of ρ^* is $p - q$, it is proven that the eigenvectors of $V_1(F_Y)^{-1} V_2(F_Y)$ which are not associated with ρ^* span the same subspace that is spanned by $\Gamma^{-1}(M - \mu_k \mathbf{1}_k^\top)$ and which corresponds to the FDS.

This result holds for model (1) and for any pair of scatter matrices, making it very general. The problem, however, is that in order to perform dimension reduction with ICS, the eigenvalues associated with the FDS should be distinct from ρ^* meaning that the multiplicity of ρ^* should not be greater than $p - q$. Tyler et al. [2009] mention that this condition generally holds except for special cases, and that these special cases depend on the scatter pair and on the model parameters. In the following subsection, we recall some special cases for which the behavior of ICS is understood more precisely.

2.3 Mixture of two Gaussian groups with different centers and other special cases

If $k = 2$, model (1) is a mixture of two elliptical distributions and $q = 1$. Thus the multiplicity of ρ^* greater than $p - 1$ corresponds to a multiplicity equal to p and $V_1^{-1}V_2$ is proportional to the identity. In this case, it is not possible to distinguish the Fisher discriminant direction from the others. In order to better understand the conditions under which this situation occurs, we need to specify the scatter pair and the elliptical distributions in the mixture (1). The case of a mixture of two Gaussian distributions has been further explored in Archimbaud et al. [2018] in the context of anomaly detection, where the authors recommend the use of the covariance matrix, denoted COV , for V_1 , and the matrix based on fourth-order moments, denoted COV_4 , for V_2 :

$$\begin{aligned}\text{COV} &= \mathbb{E}[(Y - \mathbb{E}(Y))(Y - \mathbb{E}(Y))^\top], \\ \text{COV}_4 &= \frac{1}{p+2} \mathbb{E}[d^2(Y) (Y - \mathbb{E}(Y))(Y - \mathbb{E}(Y))^\top],\end{aligned}$$

where $d^2(Y)$ is the square of the Mahalanobis distance:

$$d^2(Y) = (Y - \mathbb{E}(Y))^\top \text{COV}^{-1} (Y - \mathbb{E}(Y)).$$

For the scatter pair combination $\text{COV} - \text{COV}_4$, also known as FOBI Cardoso [1989], Nordhausen and Virta [2019], $\gamma = 1$ and the case where all eigenvalues of ICS are equal to one occurs when one of the groups has a proportion exactly equal to $(3 - \sqrt{3})/6$, i.e., approximately 21% [Tyler et al., 2009]. If one group has a proportion below this threshold, there is one eigenvalue that is strictly greater than the others, which are all equal to one. Conversely, if both groups have proportions above this threshold, there will be an eigenvalue strictly lower than the others, which are all equal to one. For outlier detection, it makes sense to assume that the proportion of outliers is less than 21% and thus, selecting the first invariant component is sufficient as proposed in Archimbaud et al. [2018]. However, limiting the theoretical study to two groups only is restrictive and we aim at finding more general results.

In the appendix of Chapter 2 in Archimbaud [2018], the eigenvalues of $\text{COV}^{-1}\text{COV}_4$ are computed for a mixture of three Gaussian distributions: one group with zero mean, and two groups with opposite non-zero mean vectors. The two non-zero group means have the same proportion in the mixture. Due to the symmetry of the mixture, the dimension of the FDS restricts to $q = 1$. Corollary 2 in Archimbaud [2018] addresses the case where all three groups have a covariance matrix equal to the identity. It states that the largest (or smallest) eigenvalue of ICS corresponds to the FDS if and only if the zero-mean group has a proportion greater than (or less than) $1/3$. In the special case where all groups have equal proportions ($1/3$ each), $\text{COV}^{-1}\text{COV}_4$ becomes the identity matrix, and consequently, ICS fails to detect the group structure. While this three-group case extends beyond the two-group case, the assumption of symmetric groups is highly restrictive. Our goal is to develop more general results.

Before going further into the behavior of ICS for mixture models (2) with more than two groups, we point out that ICS can also make it possible to find the FDS in mixture models other than (2). The so-called ‘‘barrow wheel’’ distribution introduced in Hampel et al. [2011] [see also the discussion by Stahel and Mächler in Tyler et al., 2009] is the following two groups mixture distribution:

$$\left(1 - \frac{1}{p}\right) \mathcal{N}(0_p, \text{diag}(\sigma_{11}^2, 1, \dots, 1)) + \frac{1}{p}G,$$

where p is the dimension and G is such that if $Y = (Y_1, Y_2, \dots, Y_p)^\top$ is distributed according to G , Y_1^2 follows a chi-square distribution with ν degrees of freedom, and is independent of Y_2, \dots, Y_p , while $(Y_2, \dots, Y_p)^\top$ follows a $p - 1$ dimensional Gaussian distribution with zero mean and covariance matrix equal to $\sigma_{22}^2 I_{p-1}$. As detailed on page 43 of Tyler et al. [2009], the result of Theorem 4 still applies to the barrow wheel distribution, for any scatter pair matrices (as soon as they are not proportional). Moreover, Proposition 4 in the appendix of Chapter 2 of Archimbaud [2018] gives the eigenvalues of $\text{COV}^{-1}\text{COV}_4$ as functions of the dimension p , the group proportions, and the parameters ν , σ_{11}^2 and σ_{22}^2 .

2.4 Going further with the scatter pair $\text{COV} - \text{COV}_4$ and the Gaussian mixture

Considering the scatter pair $\text{COV} - \text{COV}_4$, it is possible to make calculations for model (2) with Gaussian densities. The COV and COV_4 matrices can be expressed as follows (see Appendix A.1 for details):

$$\text{COV} = \begin{bmatrix} \beta & 0 \\ 0 & I_{p-q} \end{bmatrix}, \quad \text{COV}^{-1} = \begin{bmatrix} B & 0 \\ 0 & I_{p-q} \end{bmatrix}, \quad \text{COV}_4 = \begin{bmatrix} \Psi & 0 \\ 0 & I_{p-q} \end{bmatrix},$$

where $\beta = (\beta_{ms})$, $\Psi = (\psi_{ms})$ and $B = (b_{ms})$ are $q \times q$ matrices, and the product of the two scatter matrices is then:

$$\text{COV}^{-1}\text{COV}_4 = \begin{bmatrix} B\Psi & 0 \\ 0 & I_{p-q} \end{bmatrix}. \quad (3)$$

These computations illustrate Theorem 4 from Tyler et al. [2009] since we get a block diagonal matrix for $\text{COV}^{-1}\text{COV}_4$ with the identity of dimension $p - q$ as the lower diagonal block, meaning that $\text{COV}^{-1}\text{COV}_4$ has an eigenvalue equal to one with multiplicity at least $p - q$ and the associated eigenspace is in direct sum with the FDS which is the space spanned by the columns of T (the group centers matrix). We go further in Appendix A.1 and derive the terms of matrix Ψ :

$$\psi_{ms} = \frac{1}{p+2} \left[\sum_{i=1}^q \sum_{j=1}^q b_{ij} \mathbb{E}[x_m^c x_s^c x_i^c x_j^c] + (p-q) \mathbb{E}[x_m^c x_s^c] \right], \quad (4)$$

for $m, s \in \{1, \dots, q\}$, with $(x_1^c, \dots, x_p^c)^\top = X - \mathbb{E}(X)$. The expectations in ψ_{ms} and the expressions of β_{ms} are also provided in Appendix A.1. However, when using formula (3) and (4) to compute $\text{COV}^{-1}\text{COV}_4$, we do not have easy expressions for COV^{-1} and we propose to compute $\text{COV}^{-1}\text{COV}_4$ and its eigenvalues numerically for particular mixture models (2).

We set the value of p to 6 and the four panels of Figure 1 correspond to different values of k and q . On each panel, we make the group proportions vary on the x -axis. We consider 20 different configurations for the group means (see Table 1 and the details in Appendix B), and draw boxplots of the eigenvalues of $\text{COV}^{-1}\text{COV}_4$ on Figure 1. On most panels, there are $p - q$ eigenvalues equal to one, which is consistent with the block I_{p-q} in expression 3. However, the eigenvalues of the block $B\Psi$ are more difficult to analyze, and vary with the number of groups k , the group centers and the dimension q of the FDS. For the two-group case, we observe the following known results. When no group proportion is less than 21%, the last eigenvalue is the only eigenvalue distinct from and smaller than one. When one group has a proportion of approximately 21%, all eigenvalues are equal or nearly equal to one. In the case of a group with a proportion of less than 21%, only the first eigenvalue differs and is larger than one.

The number of eigenvalues larger (resp. smaller) than one seems to depend on the group proportions with some variability of the eigenvalues depending on the group means. In all panels, when the groups are balanced and have proportions equal to $1/k$, we observe q eigenvalues less than one. When the group proportions are unbalanced, implying that some groups have proportions of less than $1/k$, we observe that some eigenvalues can be larger than one. For the two-group case, we know that the value 21% is a threshold for the group proportions. A change in one of the two group proportions from a value larger than 21% to a value lower than 21% leads to an ICS eigenvalue shifting from smaller than one to larger than one. For more than two groups, one question arises. Is there a threshold value, depending on the number of groups, group centers and group proportions, such that, when groups are unbalanced, changing a group proportion from a value larger than the threshold to a value lower than the threshold leads to a change of an eigenvalue smaller than one to an eigenvalue larger than one? The question is not easy to answer.

Indeed, the experiment illustrates that the eigenvalues of ICS exhibit a complicated dependence on the mixture parameters, requiring a simplified framework to discover interesting results. Theorem 4 in Tyler et al. [2009] ensures that for the elliptical mixture (1), at least $p - q$ eigenvalues of the simultaneous diagonalization of two affine equivariant scatter matrices are equal and are not associated with FDS. With this result established, we now examine the remaining eigenvalues. To simplify the analysis of their behavior, we eliminate the dimensions not associated with FDS, meaning that we focus on the case where p equals q . Using a dimension reduction method is not interesting in practice if $p = q$ but the objective here is only to simplify the calculations. Even in this context, it remains difficult to understand the behavior of the ICS eigenvalues associated with the FDS. We therefore further restrict our study to the case where there is no within-group variability. This means the covariance matrix equals the between-group covariance matrix, and the distribution of Y is simply a mixture of k Dirac distributions in p dimensions. Finally, in the next two sections, we also focus on the case where $q = k - 1$, meaning that the group centers $q \times k$ matrix M is of maximum rank $q = k - 1$. As explained in the following section, this particular case with $p = q = k - 1$ and no within-group variability yields a remarkable result: the ICS eigenvalues depend solely on the proportions of the components of the mixture, and not on the group centers. This property no longer holds when q is less than $k - 1$. The latter case is more complex and will be examined only for $k = 3$ and $q = 1$ in Section 5.

3 Study of $V_1^{-1}V_2$ eigenvalues when $p = q$ with varying group centers, varying group proportions and no within-group variability

Let us consider a mixture of Dirac distributions in $p = q$ dimensions: $Y \sim \sum_{j=1}^k \alpha_j \delta_{\mu_j}$ with distinct group centers μ_j , and proportions $\alpha_j > 0$ such that $\sum_{j=1}^k \alpha_j = 1$, for $j \in \{1, \dots, k\}$. Following Tyler et al. [2009], given the affine

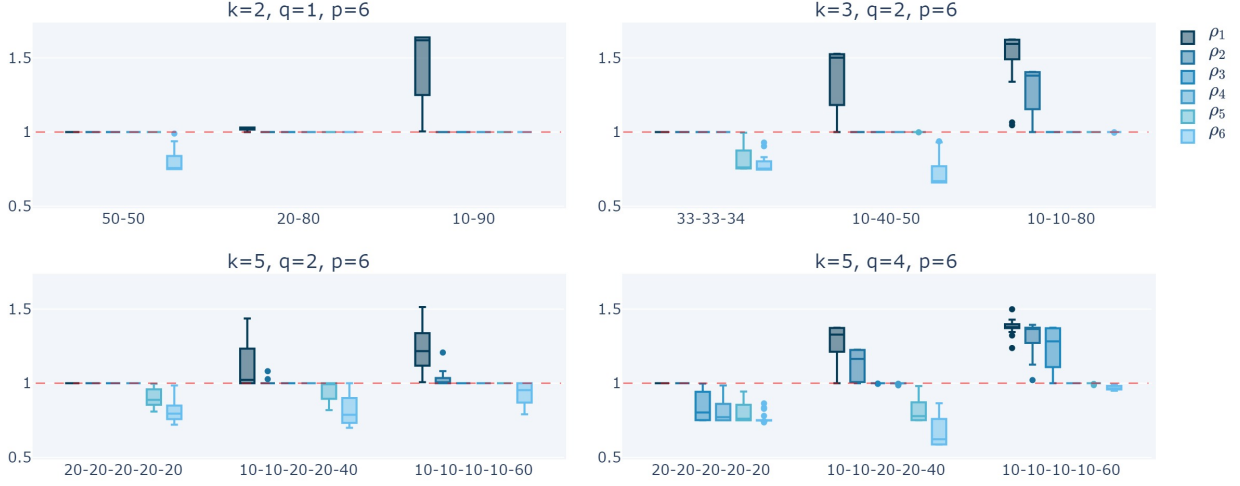


Fig. 1: Boxplots of the eigenvalues of $\text{COV}^{-1}\text{COV}_4$, where the group centers vary across 20 different configurations (see Appendix B for details), with $p = 6$, and 12 different group proportions scenarios ($\alpha_j, j \in \{1, \dots, k\}$) on the x -axis. The values of k and q vary across the panels.

invariance property of ICS, we can simplify this model and consider the following mixture model:

$$X \sim \sum_{j=1}^k \alpha_j \delta_{t_j}, \quad (5)$$

with t_k being the q -dimensional zero vector, and with distinct group centers t_j for $j \in \{1, \dots, k\}$. This model corresponds in some sense to model (2) with $p = q$ and after removing the within-group variability. This is an extreme situation but it will give us indications on the behavior of the eigenvalues of ICS when the noise (measured by the within-group covariance matrix) is “small” compared to the signal (measured by the between-group covariance matrix). Even if the Dirac measures do not have probability density functions, it is possible to compute some scatter matrices such as COV and COV_4 , and simultaneously diagonalize them. Note however that since the between-group covariance is equal to the total covariance matrix for model (5), linear discriminant analysis consists in diagonalizing the identity matrix which is of no interest.

3.1 General theoretical result for $q = k - 1$

For model (5) and under the assumption that the group centers matrix is of full rank $q = k - 1$, we can derive the following proposition that states that the eigenvalues of ICS do not depend on the group centers.

Proposition 1. *For the mixture model (5) with full rank group centers matrix, and for any pair of affine equivariant scatter matrices $V_1(F_X) - V_2(F_X)$ that exist at F_X , the eigenvalues of ICS, i.e. of $V_1(F_X)^{-1}V_2(F_X)$, do not depend on the group centers matrix but only on the proportions $\alpha_j, j \in \{1, \dots, k\}$ of the mixture components.*

Proof. Let us consider two mixtures of the form (5) with the same component proportions α_j for $j \in \{1, \dots, k\}$, but with different $(k - 1) \times k$ full rank group center matrices $T = (t_1, \dots, t_k)$ and $\tilde{T} = (\tilde{t}_1, \dots, \tilde{t}_k)$ respectively, where t_k and \tilde{t}_k are zero vectors. Let T_{k-1} (resp. \tilde{T}_{k-1}) be the $(k - 1) \times (k - 1)$ matrix such that $T = (T_{k-1}, 0)$ (resp. $\tilde{T} = (\tilde{T}_{k-1}, 0)$). The matrices T_{k-1} and \tilde{T}_{k-1} are square matrices with rank $k - 1$ and are thus invertible. We can write $\tilde{T}_{k-1} = AT_{k-1}$ where $A = \tilde{T}_{k-1}T_{k-1}^{-1}$ is a $(k - 1) \times (k - 1)$ non-singular matrix and we also have that $\tilde{T} = AT$. Let X be a $(k - 1)$ -dimensional random vector following the mixture (5) with group centers matrix T . The random vector $Y = AX$ follows the mixture (5) with the same proportions $\alpha_i, i \in \{1, \dots, k\}$ and the group centers matrix \tilde{T} . Using the affine equivariance property of the scatter matrices we have that $V_1(F_Y)^{-1}V_2(F_Y) = (A^T)^{-1}V_1(F_X)^{-1}V_2(F_X)A^T$ and has the same eigenvalues as $V_1(F_X)^{-1}V_2(F_X)$. \square

This result is a kind of generalization of the result for the two-group case detailed in Subsection 2.3. For two groups with distinct group centers, we have $q = 1$ and thus the assumption $q = k - 1$ always holds. The result in Subsection

2.3 however differs from Proposition 1 in several aspects. The result for the two-group case focuses on a Gaussian mixture and accommodates some within-group variability but applies only to the $\text{COV} - \text{COV}_4$ scatter pair. In the next subsection, we investigate numerically the ICS eigenvalues for $\text{COV} - \text{COV}_4$ for model (5) when $p = q = k - 1$.

3.2 Numerical results for $\text{COV} - \text{COV}_4$

For model (5) with $k = 2$, we have $p = q = 1$ and it is possible to derive the exact expression of $\text{COV}^{-1}\text{COV}_4$, which is a positive number. More precisely, we get after some simple computation, $\text{COV} = \alpha_1\alpha_2(t_1 - t_2)^2$ and $\text{COV}_4 = (\alpha_1^3 + \alpha_2^3)(t_1 - t_2)^2/3$ so that $\text{COV}^{-1}\text{COV}_4 = (\alpha_1^3 + \alpha_2^3)/(3\alpha_1\alpha_2)$ does not depend on t_1 and t_2 . We then recover a result similar to the one at the beginning of Subsection 2.3 for $\text{COV} - \text{COV}_4$ and a mixture of two Gaussian distributions: the eigenvalue associated with the FDS is equal to one (which is the value of the other eigenvalues when $p > q$) if and only if one of the two-group proportion is equal to $(3 - \sqrt{3})/6$. More specifically, using the fact that $\alpha_2 = 1 - \alpha_1$ and fixing $\alpha_1 \leq \alpha_2$, $\text{COV}^{-1}\text{COV}_4$ is equal to one (respectively larger or smaller than one) if and only if $\alpha_1 = (3 - \sqrt{3})/6$ (respectively α_1 smaller or larger than $(3 - \sqrt{3})/6$).

When $k > 2$, the eigenvalues of $\text{COV}^{-1}\text{COV}_4$ cannot be easily derived explicitly and numerical computations are necessary. By using the expressions derived in Appendix A.2, alongside numerical computations for the inverse of COV , the eigenvalues of $\text{COV}^{-1}\text{COV}_4$ can be calculated. These calculations are performed for 20 different group centers and 18 different scenarios of group proportions (see Appendices A.2 and B for details). Boxplots in Figure 2 display the outcomes when $p = q = k - 1$, for $k = 2, 3, 5$ and 10 groups (as in Figure 1). The boxplots consist of a single line for each scenario, indicating that varying the group means does not impact the eigenvalues when $q = k - 1$. However, the values of the eigenvalues differ from one scenario to another. Therefore, it can be concluded that the group proportions do have an impact on the eigenvalues. A more detailed examination of this impact will be presented in the next section. When $q < k - 1$, the results are different, as illustrated in Figure 3 for $k = 3$ and $q = 1$, $k = 5$ and $q = 1$, $k = 3$ and $q = 2$, and $k = 3$ and $q = 3$ (see also Figure 10 in Appendix B with $k = 10$ and $q = 1, 3, 5, 7$). For each scenario, the eigenvalues exhibit quite large variability across the different group centers. This variability is such that boxplots cross red horizontal dotted line corresponding to the value one. Unlike the case where $q = k - 1$, we cannot draw any conclusions about the conditions which cause the eigenvalues to be larger or smaller than one. The influence of group proportions on eigenvalues persists, but the influence of the group centers is also crucial.

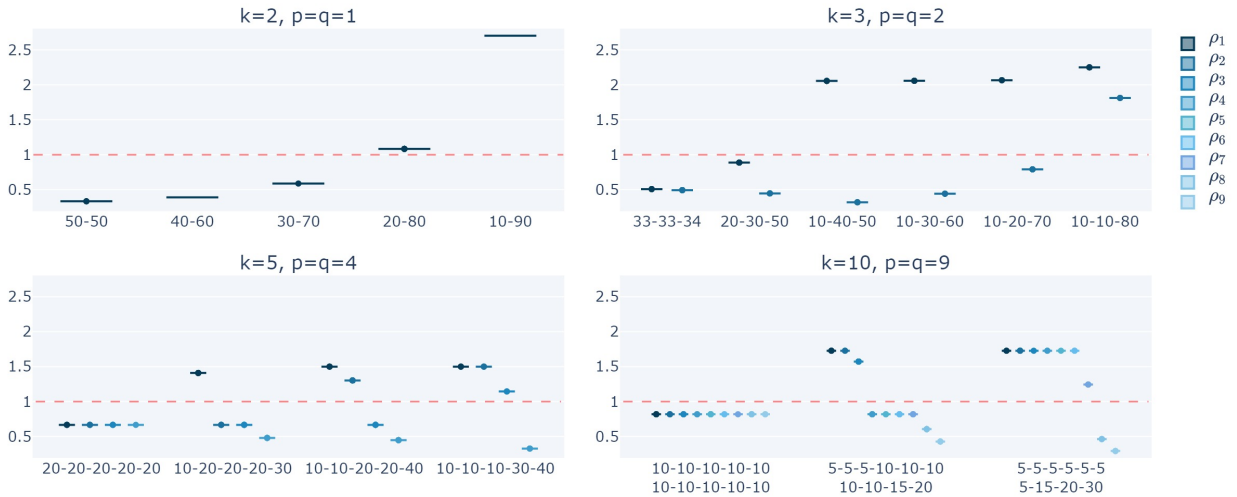


Fig. 2: Boxplots of the eigenvalues of $\text{COV}^{-1}\text{COV}_4$ for $p = q = k - 1$ and with no within-group variability, when the group centers vary across 20 different configurations (see Appendix B for details). The values of k vary across the panels and the x -axis of each panel corresponds to 18 group proportions scenarios ($\alpha_j, j \in \{1, \dots, k\}$).

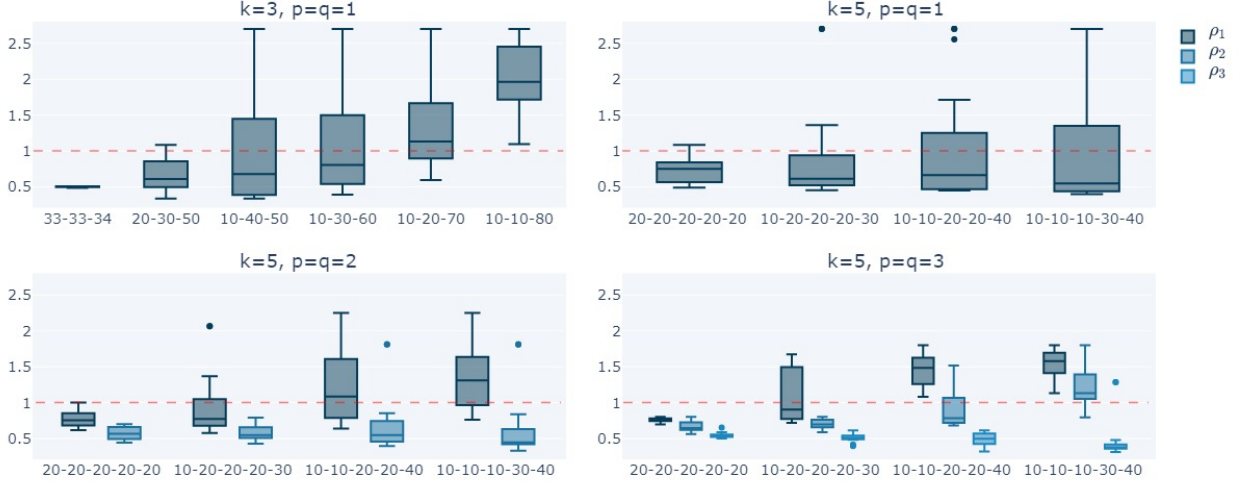


Fig. 3: Boxplots of the eigenvalues of $\text{COV}^{-1}\text{COV}_4$ for $p = q < k - 1$ and with no within-group variability, when the group centers vary across 20 different configurations (see Appendix B for details). The values of k and q vary across the panels and the x -axis of each panel corresponds to 18 group proportions scenarios ($\alpha_j, j \in \{1, \dots, k\}$).

4 Study of $\text{COV} - \text{COV}_4$ eigenvalues for $q = k - 1$ with fixed group centers, varying group proportions, and no within-group variability.

As stated in Section 3, in the context of a mixture of Dirac distributions with $q = k - 1$, only the group proportions influence the eigenvalues of $\text{COV}^{-1}\text{COV}_4$. In this section, the group means are therefore fixed to the first row of the group centers of Table 1 of Appendix B. The relationship between the group proportions and the eigenvalues is further investigated in the same context as in the previous subsection. Theoretical calculations do not permit to measure precisely the impact of group proportions on the eigenvalues of $\text{COV}^{-1}\text{COV}_4$. Therefore, we vary the group proportions on a grid and calculate numerically the corresponding eigenvalues to better understand their behavior. The three-group case is described in Subsection 4.1, while Subsection 4.2 generalises to k groups.

4.1 The case of three groups

One of the most advantageous aspects of studying three groups is the ability to represent the proportions of each group in a ternary diagram, as in Figure 4. This diagram [see Pawlowsky-Glahn et al., 2015] displays the values of three positive variables that sum up to 100%, and can be applied for the three-group proportions. Each point on the ternary diagram represents a scenario of group proportions, and the color of the point is determined by the eigenvalues obtained for that scenario. As $k = 3$, we have $q = 2$ and we focus on the case $p = 2$ in order to avoid irrelevant dimensions associated with eigenvalues equal to one. $\text{COV}^{-1}\text{COV}_4$ has thus two eigenvalues $\rho_1 \geq \rho_2$ and two ternary diagrams are plotted: Figure 4a for ρ_1 , and Figure 4b for ρ_2 .

Figure 4a illustrates that the values of ρ_1 are less than one in the center of the plot, indicating that when none of the three groups has a proportion smaller than roughly 18%, ρ_1 is less than one. Around this region, there is a white area that corresponds (given the continuity of the eigenvalues as functions of the group proportions) to ρ_1 almost equal or equal to one. $\rho_1 = 1$ is thus indistinguishable from the eigenvalues associated with the $p - q$ dimensions that do not span the FDS. It occurs when one group proportion is roughly 18%. The area beyond this white zone in the direction of the vertices is red, which indicates that the value of ρ_1 is greater than one when a group proportion is smaller than 18%. Looking at the areas around the vertices on Figure 4b demonstrates that the value of ρ_2 is greater than one for scenarios where two proportions are smaller than roughly 18%. A white area is also present around each of these regions, which corresponds to values of ρ_2 almost equal or equal to one. It occurs once again when one group has a proportion of roughly 18% and another group has a proportion smaller than or equal to 18%. The rest of the plot is blue, implying ρ_2 is smaller than one. A comparison of Figures 4a and 4b indicates that the white regions do not seem to intersect, thereby suggesting that the two eigenvalues are not equal to one simultaneously. In the two-group case, at the threshold value of 21%, all eigenvalues are equal to one. In the three-group case with $q = k - 1$, our conjecture is that there is at least one eigenvalue that differs from one for any possible group proportion scenario.

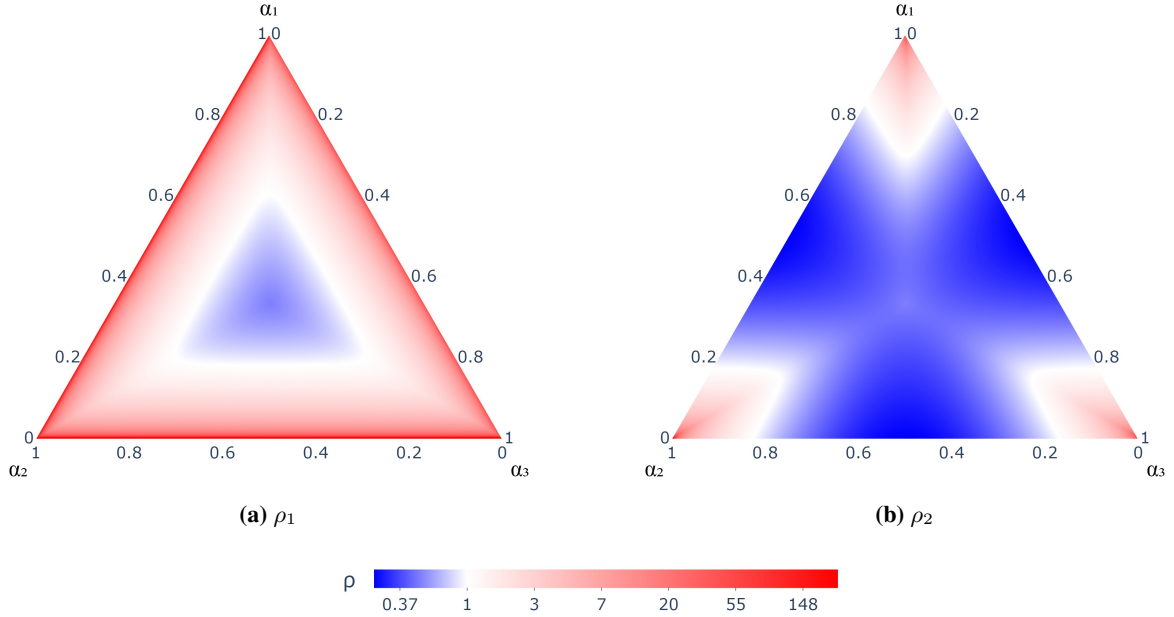


Fig. 4: Ternary diagram of the group proportions ($\alpha_1, \alpha_2, \alpha_3$) color-coded by the values of ρ_1 (a) and ρ_2 (b) when $k = 3$ and with no within-group variability. The color gradient indicates the departure of the eigenvalues from one (white when equal, blue when smaller and red when larger). Group means are fixed to $t_1 = (200, 0)$, $t_2 = (400, 100)$, $t_3 = (0, 0)$ but have no impact on the results. See Appendix C.1 for details.

The eigenvalues representation from Figures 4a and 4b are interesting since the greater the deviation from one, the easier it will be to detect the directions that span the FDS when in practice we will have $p > k$. To facilitate the comparison between the behavior of the two eigenvalues, both are plotted on the same ternary diagram on Figure 5. In the event that two groups exhibit proportions roughly smaller than 18%, both eigenvalues are greater than one, as may be observed in the red area of Figure 5. When a single group proportion is roughly smaller than 18%, the first eigenvalue is greater than one and the second is less than one. This occurs in the purple region in Figure 5. Finally, when all groups are in proportion roughly larger than 18%, in the blue area of Figure 5, both eigenvalues are less than one. As illustrated in Figures 4a and 4b, each boundary between the colored zones should display scenarios for which a single eigenvalue differs from one. However, the analysis has a numerical constraint: the 0.1% step size for the discretization of the group proportions while we use the exact value of one as the reference eigenvalue. A consequence of this experiment limitation is that only three scenarios exhibit an eigenvalue precisely equal to one. They correspond to a proportion of 60% for one group and 20% for the other two groups. These points are not represented in Figure 5.

The findings for $k = 3$ indicate that there is a threshold value (around 18%) such that when a group proportion goes from a value above this threshold to a value below, an eigenvalue of $\text{COV}^{-1}\text{COV}_4$ goes from a value less than one to a value greater than one. For two groups, this threshold is known and is approximately 21%. From Figure 5, establishing the precise threshold for three groups is not easy, but it is clear that the threshold is less than 21% and approximately 18%. Additionally, the rounded boundaries between the zones suggest that this threshold may slightly vary with the other group proportions. The goal now is to increase the number of groups beyond three, and to study whether there also exists a threshold value for the group proportions that would induce the same impact on the behavior of the ICS eigenvalues with $\text{COV} - \text{COV}_4$. Subsection 4.2 also proposes to numerically approximate the values of these thresholds as a function of the number of groups k .

4.2 The general case of k groups

The thresholds described previously are computed in this section for three setups of group proportions and for values of k between 2 and 10. Note that for a given k , the maximum proportion of a group is larger than $1/k$, while the minimum proportion is smaller than $1/k$. The balanced case corresponds to proportions all equal to $1/k$. In Setup 1, we consider scenarios of the mixture proportions such that the first group proportion ranges from 0.001 to $1/k$, the intermediate group proportions are set to $1/k$, and the last group proportion is adjusted to maintain the total sum to one. The eigenvalues of $\text{COV}^{-1}\text{COV}_4$ are calculated for each scenario. The threshold is defined as the first value of the

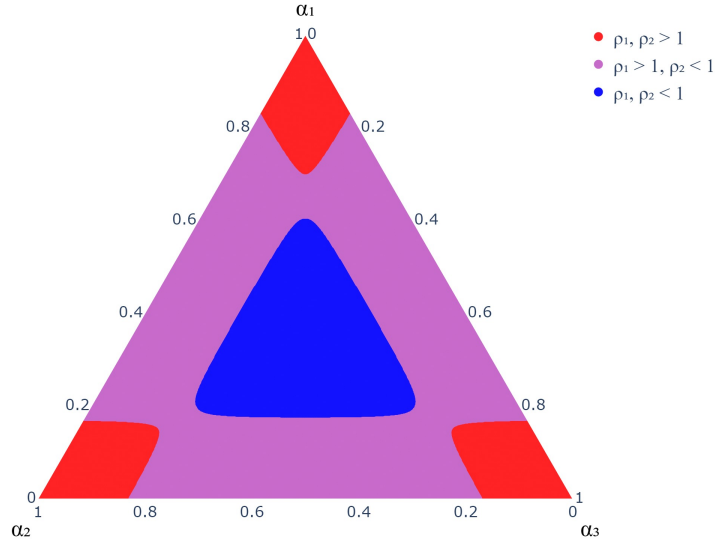


Fig. 5: Ternary diagram of the group proportions $(\alpha_1, \alpha_2, \alpha_3)$ color-coded by the eigenvalues of $\text{COV}^{-1}\text{COV}_4$ for $k = 3$ and $p = q = 2$ (blue when both eigenvalues are smaller than one, purple when one eigenvalue is larger and the other is smaller than one, and red when both eigenvalues are larger than one). Group centers are fixed to $t_1 = (200, 0)$, $t_2 = (400, 100)$, and $t_3 = (0, 0)$ but have no impact on the results. See Appendix C.1 for more details.

first group proportion for which an eigenvalue exceeds one when increasing the first group proportion. In Setup 2, the second group proportion is set to the threshold value identified in Setup 1 plus 2%. The first group proportion ranges from 0.001 to $1/k$. The remaining groups maintain a proportion of $1/k$, with the exception of the last group, whose proportion is adjusted as in Setup 1 to maintain the total sum to one. The objective of this setup is to examine the influence of a group located in proximity to the threshold. It requires at least three groups. The threshold is determined similarly to Setup 1. The grid in Setup 3 is identical to the one in Setup 2, with the exception that the second group proportion is fixed at 5%, and the last group proportion is adjusted accordingly. This setup includes a group proportion initially below the threshold for any value of k between 3 and 10, thereby ensuring that at least one eigenvalue exceeds one. The threshold is defined as the first value of the first group proportion for which two eigenvalues exceed one. As with Setup 2, this method requires at least three groups. Further details and examples may be found in Appendix C.2. The thresholds identified through the three setups, for $k = 2$ to 10 groups, are plotted in Figure 6. Since the grid generation in Setups 2 and 3 requires more than three groups, there is only one point when $k = 2$ for Setup 1. This point is roughly equal to 21% which is the result already known theoretically. For $k = 3$, the threshold value found for Setup 1 is around 18%, the one for Setup 2 is 20%, and the one for Setup 3 is 17%. Fixing a group proportion near the first threshold or below the threshold has indeed an impact. Nonetheless, the thresholds remain around 18%. As k increases, the thresholds for the three setups appear to converge toward the same value and to depend solely on k . Table 4 in Appendix C.2 gives the specific values of the thresholds for $k \in \{3, \dots, 10\}$. It is worth noting that the presented setups could be applied to a larger number of groups, although we have limited our analysis to 10 groups.

Once the thresholds for three to ten groups have been identified, the boxplots from the previous section can be reproduced for adjusted scenarios. It is of particular interest to visualize the eigenvalues when one group proportion is at the identified threshold, and to select two or three of the previous mixture proportion scenarios to vary the number of groups with proportions below this threshold. Figure 7 illustrates this experiment. Although boxplots are plotted, the group centers remain fixed and equal to the values of the first row of the group centers of Table 1 of Appendix B.

The results presented in Figure 7 validate the thresholds of Setup 1 presented in Figure 6 for $k = 2, 3, 5$, and 10 groups. In each boxplot, the scenario where a group is at the threshold level has an eigenvalue almost equal to one. This is the first eigenvalue, given that these scenarios do not have proportions below the respective threshold. The results are consistent with the logic described in Figure 5, which states that the number of eigenvalues greater than one is linked to the number of groups with proportion below the threshold. For the scenarios of Figure 7, when $k = 5$ (resp. $k = 10$), the number of eigenvalues larger than one is equal to the number of groups with proportions equal to 10% (resp. 5%). However, this relationship needs further investigation in cases involving more than three groups.

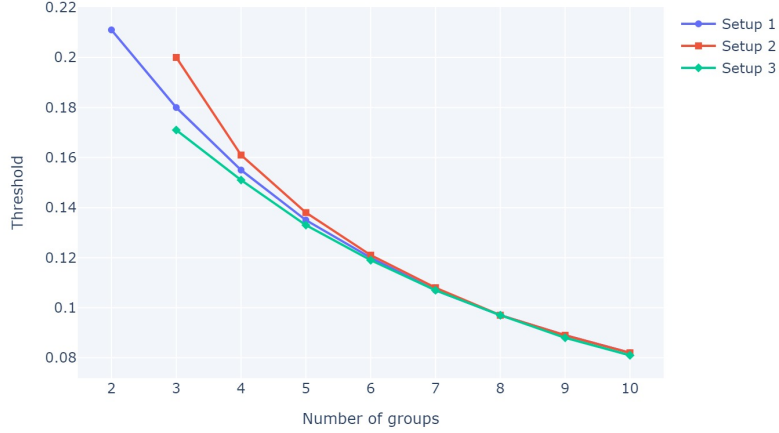


Fig. 6: Thresholds computed in Setups 1, 2 and 3, for different values of k .

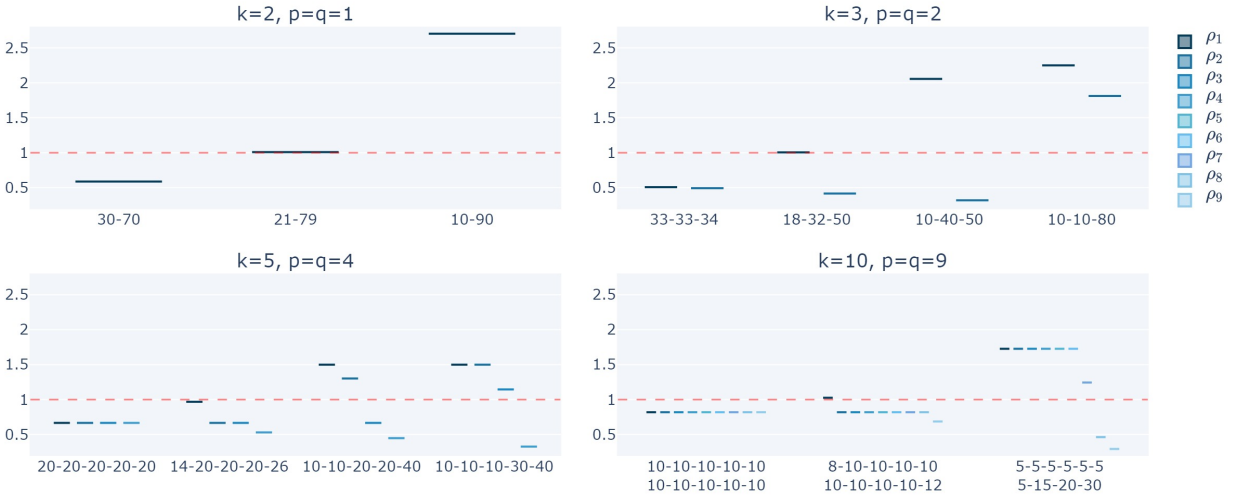


Fig. 7: Boxplots of the eigenvalues of $\text{COV}^{-1}\text{COV}_4$ with no within-group variability, with $p = q = k - 1$, and 14 different group proportions scenarios ($\alpha_j, j \in \{1, \dots, k\}$) on the x -axis. The values of k vary across panels. In each panel, the second scenario (on the x -axis) contains a group at the identified threshold value (see Appendices B and C.2 for details).

5 Study of $\text{COV}^{-1}\text{COV}_4$ eigenvalues for a Gaussian mixture of three groups with aligned group centers

In Section 3, Figures 2 and 3 display the eigenvalues of ICS using $\text{COV} - \text{COV}_4$ for various Gaussian mixture models described by equation (5). Figure 2 illustrates that, if $q = k - 1$, these eigenvalues do not depend on the group means, and Proposition 1 in Section 4 confirms the result theoretically. However, as shown in Figure 3, this result does not hold if $q < k - 1$, leading to a more complex analysis of the eigenvalues. In the present section, we examine the eigenvalues of $\text{COV}^{-1}\text{COV}_4$ for model (2) with three Gaussian groups ($k = 3$) and $q = 1 < k - 1 = 2$. This corresponds to the following p -dimensional Gaussian mixture:

$$\alpha_1 \mathcal{N}_p(t_1, I_p) + \alpha_2 \mathcal{N}_p(t_2, I_p) + (1 - \alpha_1 - \alpha_2) \mathcal{N}_p(0_p, I_p), \quad (6)$$

with $t_1 = (t_{11}, 0, \dots, 0)^\top$, $t_2 = (t_{21}, 0, \dots, 0)^\top$, $t_{11} \neq 0$, $t_{21} \neq 0$, and $t_{11} \neq t_{21}$. We can prove the following proposition.

Proposition 2. *Let us consider model (6) and let r_{α_1, α_2} be the following polynomial of degree 4:*

$$r_{\alpha_1, \alpha_2}(x) = \alpha_1(-1 + 7\alpha_1 - 12\alpha_1^2 + 6\alpha_1^3)x^4 + 4\alpha_1\alpha_2(1 - 6\alpha_1 + 6\alpha_1^2)x^3 + 6\alpha_1\alpha_2(1 - 2\alpha_2 + \alpha_1(-2 + 6\alpha_2))x^2 + 4\alpha_1\alpha_2(1 - 6\alpha_2 + 6\alpha_2^2)x + \alpha_2(-1 + 7\alpha_2 - 12\alpha_2^2 + 6\alpha_2^3).$$

All eigenvalues of ICS for the scatter pair $\text{COV} - \text{COV}_4$ are equal to one if and only if the first coordinates t_{11} and t_{21} of the group centers are such that

$$r_{\alpha_1, \alpha_2}(t_{11}/t_{21}) = 0. \quad (7)$$

The proof is given in Appendix D. Proposition 2 demonstrates that the behavior of the ICS eigenvalues for a mixture of three Gaussian groups with aligned group centers differs from the two-group case (see Subsection 2.3). For model (6), the condition under which all eigenvalues of $\text{COV}^{-1}\text{COV}_4$ equal one depends not only on the group proportions but also on the group centers, as determined by the fourth-order equation (7). This equation has four solutions in the complex plane but, between zero and four solutions in the real space, depending on the group proportions. Using Mathematica [Wolfram Research, Inc., 2022] to simplify the computations, we found that for $\alpha_1 = \alpha_2 \leq (3 - \sqrt{3})/12$ ($\simeq 10.6\%$), or $\alpha_1 = \alpha_2$ equal to $1/3$ or $1/4$, equation (7) has no real solution. This implies that one and only one eigenvalue of ICS differs from one, and is associated with an eigenvector that spans the FDS (which is only one-dimensional for model (6)). However, we can also identify group proportions for which, given specific group centers, all eigenvalues of ICS equal one, indicating that ICS does not work in all situations. For instance, when $\alpha_1 = \alpha_2 = 1/6$, equation (7) has three real solutions, and all the eigenvalues of ICS equal one if t_{21}/t_{11} equal -1 , $(2\sqrt{6} + 7)/5$ or $(-2\sqrt{6} + 7)/5$. While these situations are highly specific, they illustrate that analysing the eigenvalues of $\text{COV}^{-1}\text{COV}_4$ becomes more complex when q is less than $k - 1$, and that this phenomenon is already true for $k = 3$.

6 Empirical study

In Section 3, we derive a theoretical understanding of the behavior of the eigenvalues of $V_1^{-1}V_2$ in the context of a mixture of Dirac distributions (model (5)). Firstly, for a mixture model with full rank group centers matrix and in the absence of within-group variability, the eigenvalues of $V_1^{-1}V_2$, depend only on the group proportions (see Proposition 1). Secondly, in Subsection 4.2, the thresholds for the proportion of groups that result in an eigenvalue transitioning from less than the eigenvalue of multiplicity $p - q$ to greater than it, have been identified for different number of groups. Both analyses are restricted to the case where the dimension of the FDS is $q = k - 1$. To confirm those results in a more general context but still with $q = k - 1$, we perform simulations of a mixture of Gaussian distributions, including within-group variability and noise, and we also focus on different scatter pairs. Subsection 6.1 describes the simulations settings, Subsection 6.2 studies the eigenvalues of $\text{COV}^{-1}\text{COV}_4$ when group centers vary in the presence of within-group variability, while Subsection 6.3 focuses on the behavior of eigenvalues for different scatter pairs.

6.1 Simulation design

We generate $n = 1000$ observations from a particular case of model (2) of a mixture of Gaussian distributions with k different group means and equal covariance matrix, under the assumption that the group centers matrix is of full rank $q = k - 1$ as in Alfons et al. [2024]:

$$X \sim \sum_{j=1}^k \alpha_j \mathcal{N}(t_j, I_p), \quad (8)$$

where $\alpha_j > 0$ are mixture weights such that $\sum_{j=1}^k \alpha_j = 1$, for $j \in \{1, \dots, k\}$, $t_1 = 0$, and $t_{l+1} = \delta e_l$, for $l \in \{1, \dots, k-1\}$, e_l is a p -dimensional vector with one in the l -th coordinate and zero elsewhere, for different numbers of clusters $k \in \{2, 3, 5, 10\}$, on $p = 5k$ variables. In this setting, the cluster structure lies in a low-dimensional subspace of dimension $q = k - 1$. We consider 17 scenarios of group proportions $(\alpha_j, j \in \{1, \dots, k\})$:

- $k = 2$ clusters: 50–50, 40–60, 30–70, 21–79, 10–90,
- $k = 3$ clusters: 33–33–34, 18–35–50, 10–40–50, 10–30–60, 10–80–10,
- $k = 5$ clusters: 20–20–20–20–20, 14–20–20–20–26, 10–20–20–20–30, 10–10–20–20–40,
- $k = 10$ clusters: 10–10–10–10–10–10–10–10–10–10, 8–10–10–10–10–10–10–10–12, 5–5–5–5–5–5–15–20–30.

We consider $\delta \in \{1, 5, 10, 50, 100\}$ for the study on different group centers in Subsection 6.2. For the study on different scatters in Subsection 6.3, we fix $\delta = 10$ and we focus on the following scatter pairs: COV – COV₄, COVAXIS – COV, TCOV – COV and MCD _{τ} – COV with $\tau \in \{0.25, 0.5, 0.75\}$. COVAXIS is a one-step M-estimator using the inverse weight function of the squared Mahalanobis distance, TCOV is also a one-step M-estimator but with weights based on pairwise Mahalanobis distances and MCD _{τ} are the (raw) minimum covariance determinant estimators based on $n_\tau = \lceil \tau n \rceil$ observations for which the sample covariance matrix has the smallest determinant. See Subsection 3.1 of Alfons et al. [2024] for details on those scatter matrices. Note that we follow Alfons et al. [2024] and take V_1 being more robust than V_2 . In addition, MCD _{τ} are computed using the FAST-MCD algorithm of [Rousseeuw and Driessen, 1999]. For the selection of components, we choose the *med criterion* as introduced in Alfons et al. [2024], which selects the $k - 1$ components whose eigenvalues deviate the most from the median of all eigenvalues. This test relies on the assumption that the dimension of interest q is low compared to the number of variables p : $q \leq p/2$, which implies that the majority of the eigenvalues should be equal, which is met in our context. We simulate 50 data sets for each of the different settings.

6.2 COV⁻¹COV₄ eigenvalues when group centers vary in the presence of within-group variability and $q = k - 1$

As proven in Proposition 1, theoretically, for a mixture model with full rank group centers matrix and in the absence of within-group variability, the eigenvalues of $V_1^{-1}V_2$ depend only on the group proportions $\alpha_j, j \in \{1, \dots, k\}$ of the mixture components. Here, instead of computing the eigenvalues theoretically, we simulate a mixture of Gaussian distributions with some within-group variability to evaluate the sensitivity of this proposition and we compare with the results in Figure 7 for COV⁻¹COV₄. In Figure 8, we display the boxplots of the $k - 1$ first and the $k - 1$ last eigenvalues of COV⁻¹COV₄ over 50 replications when the group centers vary in the presence of within-group variability. Grey-shaded areas expose the eigenvalues which are theoretically different from one. The cases for which the eigenvalue one has multiplicity greater than $p - q$ are identified by the red-shaded areas.

Contrary to Figure 7, we can see some variability for the eigenvalues over the replications and between the different group centers, determined by the values of δ . In addition, given the presence of within-group variability and noise with $p > q = k - 1$, none of the eigenvalues are strictly equal to one. However, the eigenvalues of interest for highlighting the groups' structure (in grey-shaded areas), are still easily identifiable as the ones further away from one. For example, for the first row with scenarios of a mixture of two components, only the first or the last eigenvalue should be different from one. More specifically, as illustrated in Subsection 4.2, for a mixture with group proportion of 50 – 50, only the last eigenvalue is different from one, as we can see in the first subplot. If one group proportion decreases until being below 21% then it is now the first eigenvalue which is different from one. For this threshold of 21%, all eigenvalues are equal to one as highlighted by the red-shaded area. We can identify the same behavior when the number of groups increases. For example, in the second row, if all the three group proportions are “large enough” then the last two eigenvalues are different from one. When only one group proportion is “low”, as in the scenario 10 – 40 – 50, the first and the last components are different from one. When two group proportions are “low”, like in the scenario 10 – 10 – 80, then only the first two components are different from one. This pattern repeats itself for mixtures of 5 or 10 groups as illustrated in the third and fourth rows. The thresholds are clearly identifiable when one group proportion is equal to: 18% for 3 groups, 14% for 5 groups and 8% for 10 groups and confirms the ones identified in Subsection 4.2 in case of absence of within-group variability.

It is also important to note that the variability of the eigenvalues between group centers, i.e for different values of δ , is smaller for the eigenvalues which are theoretically equal to one (in white or red-shaded areas) no matter the value of δ . This supports evidence for some stability of the eigenvalues associated with no structure of the data, no matter the group centers. Clearly, at the thresholds, identified by red-shaded areas, there is almost no variability of the eigenvalues, as we can see in the first row, in the well-known case of a mixture of two components with group proportions 21 – 79.

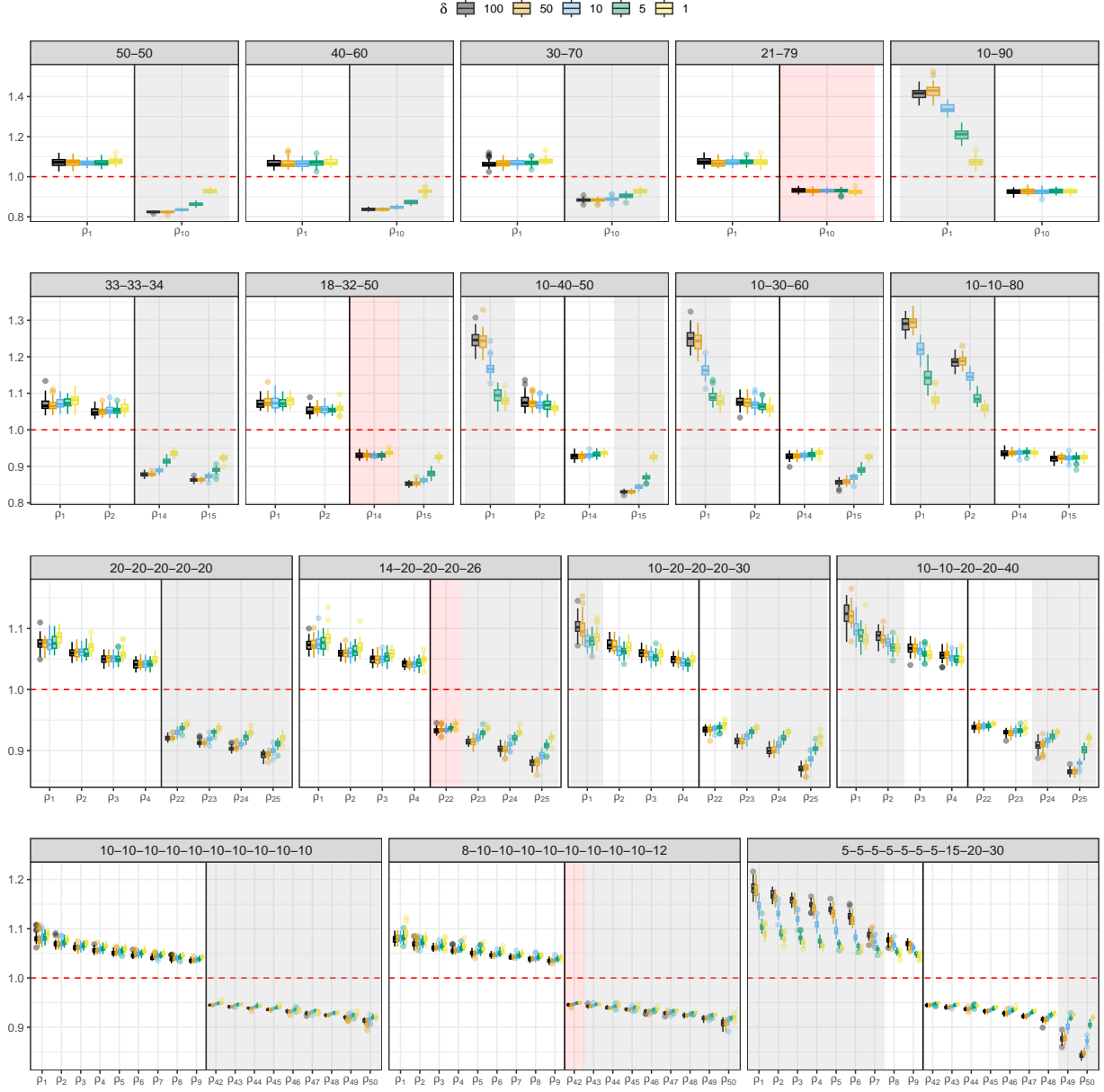


Fig. 8: Boxplots of the $k - 1$ first and the $k - 1$ last eigenvalues of $\text{COV}^{-1}\text{COV}_4$ when the group centers vary in the presence of within-group variability, for $q = k - 1$, $p = 5k$, $n = 1000$, and different values of k over 50 replications. Results for different group proportions are shown in different panels. Grey-shaded areas expose the eigenvalues which are theoretically different from one. Red-shaded areas highlight the cases for which the eigenvalue one has multiplicity greater than $p - q$.

For the eigenvalues of interest (in grey-shaded areas), some variability is visible but they seem to quickly converge as δ increases. In addition, the eigenvalues are clearly different from one as soon as $\delta \geq 5$. For $\delta = 1$, the ratio signal over noise is too low and so, all eigenvalues are close to one. On the contrary, if $\delta \geq 50$, the ratio signal over noise is high and the noise is almost null so, we are almost in the same context as for the Proposition 1, and the eigenvalues with $\delta = 50$ or $\delta = 100$ are almost equal. In fact, it appears that the eigenvalues converge more slowly in the presence of a low group proportion, as observed in scenarios where at least one group has a proportion of 5% or 10%. In this context, the eigenvalues are also higher in absolute value for any δ so it is one of the favorable cases for ICS since it is easy to

identify it is different from one. So, for the next study, we consider that the group centers have a negligible effect, and we focus on the case $\delta = 10$.

6.3 $V_1^{-1}V_2$ eigenvalues for different scatter pairs in the presence of within-group variability and for $q = k - 1$

In this section, we extend our study to the behavior of the eigenvalues of $V_1^{-1}V_2$ for different affine equivariant scatter pairs and not only to $\text{COV} - \text{COV}_4$. In such a context, and in the absence of within-group variability, only the group proportions should impact the eigenvalues. Here, we introduce some within-group variability and we analyze if the general behavior for the eigenvalues of $\text{COV}^{-1}\text{COV}_4$ is the same for different scatter pairs. More specifically, we want to know if the thresholds are linked to the same group proportions and if only “low” group proportions are associated with first eigenvalues being different from each other. Deriving theoretical eigenvalues for more complex scatter pairs is a difficult task so, we focus on the analysis of the invariant components instead. Indeed, to highlight the data structure, we have to project the data onto the subspace spanned by the eigenvectors associated with eigenvalues which are not equal to each other. So we can hypothesize that the invariant components to select are the ones associated with the eigenvalues which are not equal to each other. To identify the components of interest, we followed the recommendations of Alfons et al. [2024] and we chose the $k - 1$ invariant components based on the *med criterion*, i.e the ones associated with eigenvalues which deviate the most from the median of all eigenvalues.

In Figure 9, we display heatmaps of the percentage of selections for the $k - 1$ first and the $k - 1$ last invariant components of $V^{-1}V_2$ by the *med criterion*, for different numbers of groups, for group proportions and for different scatter pairs. If the value is of 100%, it indicates that the value of the eigenvalue associated with this component is stable and different from each others. If the percentage is less than 100%, it means that over the replications, the criterion has not chosen consistently the same eigenvalue which suggests that the eigenvalue might be not significantly different from the other ones. For example, in the first subplot, for the scatter pair $\text{COV} - \text{COV}_4$, the value is 100% for the case of a mixture of two Gaussian distributions with group proportions of 10-90. We can deduce that the first component IC_1 is always chosen over the 50 replications. This is in line with the previous conclusion which states that if one group proportion is “low” then the first eigenvalue is different from one. In case of 50-50, it is always the last one which is chosen and of interest. For group proportions in between, the choice is not as clear mainly for the scenario 21-79. In this case, IC_1 is selected instead of IC_{10} only in 75% of cases. When the selection is not clear, this is a sign that the eigenvalues are not as far away from the others and it might indicate a threshold. Precisely, this result confirms the theory recalled in Subsection 2.3 that, when a group proportion is equal to 21% and in presence of two groups, then all the eigenvalues are equal to one with $\text{COV} - \text{COV}_4$. In practice, the group proportion associated with the threshold is not as precise as in the theoretical situation since, even when a group proportion is around 30%, it can be difficult to have a clear separation between eigenvalues in 15% of cases.

Overall, for $\text{COV} - \text{COV}_4$ we can corroborate the same thresholds as mentioned in the previous section: approximately 21% for 2 groups, 18% for 3 groups, 14% for 5 groups and 8% for 10 groups. However, because we infer the behavior of the eigenvalues based on the selection of components and we are in the context of within-group variability, noise and simulated data, the values of the thresholds are not as precise as in the theoretical cases illustrated in Figure 7. Nevertheless, this analysis globally confirms the results already observed and demonstrated in Subsection 4.2. It is also quite easy to visually identify on which components the structure of interest relies depending on the group proportions and to extend our understanding of such behavior for the other scatter pairs.

The scatter pair $\text{COVAXIS} - \text{COV}$ is almost exhibiting the same behavior as $\text{COV} - \text{COV}_4$ as well as for the thresholds. This is not surprising since COVAXIS is also a one-step M-estimator as COV_4 but uses an inverse weight function of the squared Mahalanobis distance instead. For $\text{TCOV} - \text{COV}$, the choice of the components seems to be easy and always associated with the first eigenvalues, except for one case with 5 groups (10-20-20-20-30). For the scatter pairs based on MCD_τ , the patterns are conditional on the value τ . It is interesting to note that the behavior for $\text{MCD}_{0.25} - \text{COV}$ looks almost the same as for $\text{TCOV} - \text{COV}$, apart from one case with three balanced groups. For $\text{MCD}_{0.50} - \text{COV}$ then the situation is no longer perfect and sometimes the last eigenvalue is chosen, e.g. in the scenario of group proportions of 50-50. Finally, for $\text{MCD}_{0.75} - \text{COV}$ the situation looks worse and the choice is not clear in a lot of cases, even more than for the scatter pair $\text{COV} - \text{COV}_4$. These results support the recommendations of using $\text{TCOV} - \text{COV}$ or $\text{MCD}_{0.25} - \text{COV}$ when clustering is the goal. Alfons et al. [2024].

To conclude, it appears that those two scatter pairs, $\text{TCOV} - \text{COV}$ or $\text{MCD}_{0.25} - \text{COV}$, seem to behave in a simpler manner than $\text{COV} - \text{COV}_4$ by finding groups only on the first components no matter the group proportions and with fewer thresholds. In addition, the more groups are present, the easier it looks to find them. However, we have shown that in a few scenarios, those scatter pairs might fail to identify some groups. Since our analysis is not exhaustive on the number of scenarios, we can suppose it might happen in other situations. One idea to overcome such a potential issue, without knowing the theoretical eigenvalues, is to run ICS multiple times, with different scatter pairs and to combine and

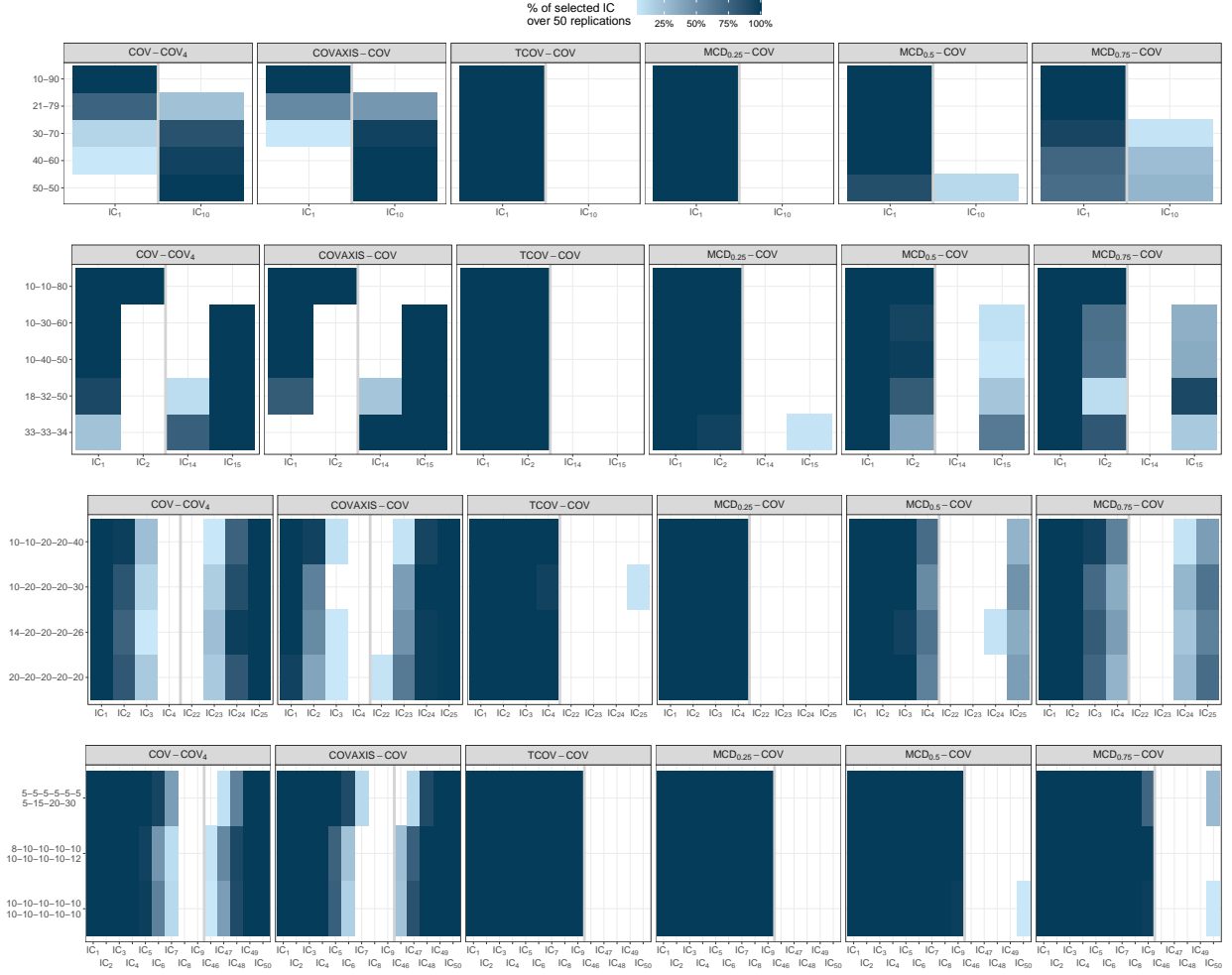


Fig. 9: Heatmaps of the percentage of selection for the $k - 1$ first and $k - 1$ last selected invariant components of $V^{-1}V_2$ by the med criterion, in the presence of within-group variability, for $q = k - 1$, $p = k/0.2$, $n = 1000$, different values of k over 50 replications. Results for different numbers of groups are shown in separate rows, for different group proportions in the y-axis and for different scatter pairs in different panels.

analyze the selected components from multiple pairs to be sure to detect the entire group structure. Another possibility is to perform localized PP after ICS to see which directions are interesting as suggested in Dümbsen et al. [2023].

7 Conclusions and perspectives

Dimension reduction is becoming increasingly important. PCA is probably the most utilized method in practice due to its simplicity, despite lacking guarantees for its effectiveness as a preprocessing tool for clustering or outlier detection. In these contexts, PP appears much more natural and has theoretical justification Radojičić et al. [2021]. However, PP is computationally expensive. From that perspective, ICS is a promising alternative – it is computationally less demanding, and theoretical properties can be derived in quite general mixture model frameworks. As shown in the seminal paper Tyler et al. [2009], ICS can recover the FDS. It essentially reduces to an eigenvalue problem where the noise space has identical eigenvalues. The crucial question is then whether all eigenvalues belonging to the space spanning the FDS are distinct from the noise value. In the two-component Gaussian mixture model using the combination $COV - COV_4$, it was known that this generally works except for one specific mixing proportion. In this work, we extended this result to cover richer mixture models and different scatter combinations, though theoretical studies seem mainly feasible only for $COV - COV_4$. Based on the findings, it seems that ICS is indeed a natural dimension reduction method for clustering and outlier detection, and with an increasing number of clusters (k still smaller than p), the FDS will be estimated.

Based on the current paper, it would be worthwhile to pursue extending these results to even richer mixture models and also consider in more detail other scatter combinations. It could also be investigated whether, in cases where one scatter combination fails, there are other combinations that will work, or if there exists a global worst-case scenario.

In practice, it seems customary to compare the performance of different scatter combinations and choose the best one, which is, however, still done heuristically, and corresponding tools for a comparison could be developed. One crucial issue in practice is then also to establish what the noise space eigenvalue is, and which eigenvalues are distinct from that value. Some heuristic rules are, for example, discussed in Archimbaud et al. [2018], Alfons et al. [2024], Radojicic and Nordhausen [2020], but inferential tools are still missing. So far, only Kankainen et al. [2007] propose some tests using $\text{COV} - \text{COV}_4$ if all eigenvalues are equal in the Gaussian case (i.e., testing for multivariate normality) and Luo and Li [2016], Radojicic and Nordhausen [2020], Nordhausen et al. [2022] in a non-Gaussian component analysis framework for the equality of eigenvalues for components belonging to Gaussian components. Similar tests might be of interest also in model (1).

Computational details

All computations in Sections 2, 3, 4 are performed with Python version 3.10.13 [Van Rossum and Drake, 2009], notable packages include NumPy [Harris et al., 2020] for numerical computations, Pandas [The pandas development team, 2020, McKinney, 2010] for data manipulation and Plotly [Plotly Technologies Inc., 2015] for data visualization in Subsection 2.4, Sections 3 and 4. Furthermore, all simulations in Section 6 are performed with R version 4.3.3 [R Core Team, 2023] and uses the R packages ICS [Nordhausen et al., 2008, 2023] for ICS, ICSClust [Archimbaud et al., 2023b] for selecting the components, rrcov [Todorov and Filzmoser, 2009] for the MCD scatter matrix. Replication files for the theoretical computations and simulations are available upon request.

Acknowledgments

We thank Camille Mondon for stimulating discussions about ICS. Anne Ruiz-Gazen acknowledges funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d’Avenir) program, grant ANR-17-EURE-0010. Klaus Nordhausen was supported by the HiTEc COST Action (CA21163).

Appendix A. Calculation details for the scatter pair $\text{COV} - \text{COV}_4$

Appendix A.1. The case of Gaussian mixture in Subsection 2.4

Because of the invariance property of ICS, we consider wlog [see A.2 in Tyler et al., 2009] the mixture :

$X \sim \sum_{j=1}^k \alpha_j \mathcal{N}_p(t_j, I_p)$, where the $t_j = (t_{ji}) \in \mathbb{R}^p$ are distinct for $j \in \{1, \dots, k\}$ and can be written $\sum_{i=1}^j t_{ji} e_i$ (with

e_i the p -dimensional vector with one in the i -th coordinate and zero elsewhere for $j \in \{1, \dots, k-1\}$), and t_k is the p -dimensional zero vector. Furthermore, to ease our computations, we define $X^c := X - \mathbb{E}(X)$ whose distribution is

$\sum_{j=1}^k \alpha_j \mathcal{N}_p(t_j^c, I_p)$, where $t_j^c = t_j - \mathbb{E}(X)$ has coordinates t_{ji}^c , for $j \in \{1, \dots, k\}$ and $i \in \{1, \dots, q\}$. The covariance

matrix of X is $\text{COV} = \Gamma_W + \Gamma_B$ where $\Gamma_W = I_p$ is the within-group covariance matrix and $\Gamma_B = \sum_{j=1}^k \alpha_j t_j^c (t_j^c)^\top$ is the between-group covariance matrix. This yields:

$$\text{COV} = \begin{bmatrix} \beta & 0 \\ 0 & I_{p-q} \end{bmatrix} \quad \text{and} \quad \text{COV}^{-1} = \begin{bmatrix} B & 0 \\ 0 & I_{p-q} \end{bmatrix},$$

where the terms of β are $\beta_{ms} = \sum_{j=1}^k \alpha_j t_{jm}^c t_{js}^c$ for distinct $m, s \in \{1, \dots, q\}$, and $\beta_{mm} = 1 + \sum_{j=1}^k \alpha_j (t_{jm}^c)^2$ for $m \in \{1, \dots, q\}$. $B = (b_{ij})$ is $q \times q$ matrix whose terms are difficult to express. Concerning COV_4 , we get:

$$\text{COV}_4 = \frac{1}{p+2} \times \mathbb{E}((X^c)^\top \text{COV}^{-1} X^c X^c (X^c)^\top) = \frac{1}{p+2} \times \mathbb{E} \left[\left(\sum_{i=1}^q \sum_{j=1}^q x_i^c x_j^c b_{ij} + \sum_{i=q+1}^p (x_i^c)^2 \right) X^c (X^c)^\top \right],$$

where x_i^c are the coordinates of X^c , for $i \in \{1, \dots, p\}$. We proceed to compute the moments of the coordinates of X^c to obtain the final form of COV_4 . Let $f_X(\mu) = \mu$, $f_{X^2}(\mu, \sigma) = \mu^2 + \sigma^2$, $f_{X^3}(\mu, \sigma) = \mu^3 + 3\mu\sigma^2$, and

$f_{X^4}(\mu, \sigma) = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$ be the functions that give the moments of order 1 to 4 of the univariate normal distribution with mean μ and variance σ^2 . The moments of the mixture are equal to the linear combination of the moments of its components. For $i, j \in \{1, \dots, q\}$, $x_i^{\mathcal{C}}$ and $x_j^{\mathcal{C}}$ are not independent, but for each mixture component, the coordinates are independent and we get the following expressions:

$$\begin{aligned}\mathbb{E}[x_a^{\mathcal{C}^2}] &= \sum_{j=1}^k \alpha_j f_{X^2}(t_{ja}^{\mathcal{C}}, 1), & \mathbb{E}[x_a^{\mathcal{C}} x_b^{\mathcal{C}}] &= \sum_{j=1}^k \alpha_j f_X(t_{ja}^{\mathcal{C}}) f_X(t_{jb}^{\mathcal{C}}), & \mathbb{E}[x_a^{\mathcal{C}^4}] &= \sum_{j=1}^k \alpha_j f_{X^4}(t_{ja}^{\mathcal{C}}, 1), \\ \mathbb{E}[x_a^{\mathcal{C}^3} x_b^{\mathcal{C}}] &= \sum_{j=1}^k \alpha_j f_{X^3}(t_{ja}^{\mathcal{C}}, 1) f_X(t_{jb}^{\mathcal{C}}), & \mathbb{E}[x_a^{\mathcal{C}^2} x_b^{\mathcal{C}^2}] &= \sum_{j=1}^k \alpha_j f_{X^2}(t_{ja}^{\mathcal{C}}, 1) f_{X^2}(t_{jb}^{\mathcal{C}}, 1), \\ \mathbb{E}[x_a^{\mathcal{C}^2} x_b^{\mathcal{C}} x_c^{\mathcal{C}}] &= \sum_{j=1}^k \alpha_j f_{X^2}(t_{ja}^{\mathcal{C}}, 1) f_X(t_{jb}^{\mathcal{C}}) f_X(t_{jc}^{\mathcal{C}}) \\ \mathbb{E}[x_a^{\mathcal{C}} x_b^{\mathcal{C}} x_c^{\mathcal{C}} x_d^{\mathcal{C}}] &= \sum_{j=1}^k \alpha_j f_X(t_{ja}^{\mathcal{C}}) f_X(t_{jb}^{\mathcal{C}}) f_X(t_{jc}^{\mathcal{C}}) f_X(t_{jd}^{\mathcal{C}}),\end{aligned}$$

for distinct $a, b, c, d \in \{1, \dots, q\}$. For $i > q$ and $j \in \{1, \dots, p\}$, $x_i^{\mathcal{C}}$ and $x_j^{\mathcal{C}}$ are independent which leads to many elements of COV_4 being equal to 0 and to the following expression: $\text{COV}_4 = \begin{bmatrix} \Psi & 0 \\ 0 & I_{p-q} \end{bmatrix}$ where $\Psi = (\psi_{ms})$ is $q \times q$ matrix whose terms for $m, s \in \{1, \dots, q\}$ are given by:

$$\psi_{ms} = \frac{1}{p+2} \mathbb{E} \left[\sum_{i=1}^q \sum_{j=1}^q x_m^{\mathcal{C}} x_s^{\mathcal{C}} x_i^{\mathcal{C}} x_j^{\mathcal{C}} b_{ij} + x_m^{\mathcal{C}} x_s^{\mathcal{C}} \sum_{i=q+1}^p (x_i^{\mathcal{C}})^2 \right] = \frac{1}{p+2} \left[\sum_{i=1}^q \sum_{j=1}^q b_{ij} \mathbb{E}[x_m^{\mathcal{C}} x_s^{\mathcal{C}} x_i^{\mathcal{C}} x_j^{\mathcal{C}}] + (p-q) \mathbb{E}[x_m^{\mathcal{C}} x_s^{\mathcal{C}}] \right].$$

Appendix A.2. The case of Dirac mixture in Section 3

Upon the removal of the noise and the dimensions not associated with FDS, the Gaussian distributions are simplified to Dirac distributions with parameter t_j , denoted by δ_{t_j} : $X \sim \sum_{j=1}^k \alpha_j \delta_{t_j}$, where $t_j = \sum_{i=1}^j t_{ji} e_i$, and e_i is a q -dimensional vector with one in the i -th coordinate and zero elsewhere for $j \in \{1, \dots, k-1\}$, t_k is the zero vector, and the t_j are distinct for $j \in \{1, \dots, k\}$.

The calculations are very similar to the ones in Appendix A.1 with the within-group matrix Γ_W equals to the zero matrix, which removes some terms in the scatter expressions. In accordance with the previously established notation, it yields $\text{COV} = [\beta_{ms}]$ where $\beta_{ms} = \sum_{j=1}^k \alpha_j t_{jm}^{\mathcal{C}} t_{js}^{\mathcal{C}}$ for $m, s \in \{1, \dots, q\}$, and $\text{COV}_4 = \frac{1}{q+2} \times \mathbb{E} \left[\left(\sum_{i=1}^q \sum_{j=1}^q x_i^{\mathcal{C}} x_j^{\mathcal{C}} b_{ij} \right) X^{\mathcal{C}} (X^{\mathcal{C}})^{\top} \right]$. The moments of a random variable following a Dirac distribution centered at $t^{\mathcal{C}}$ is given by $m_{\nu} = (t^{\mathcal{C}})^{\nu}$. All the moments can be expressed as $\mathbb{E}[x_a^{\mathcal{C}} x_b^{\mathcal{C}} x_c^{\mathcal{C}} x_d^{\mathcal{C}}] = \sum_{j=1}^k \alpha_j t_{ja}^{\mathcal{C}} t_{jb}^{\mathcal{C}} t_{jc}^{\mathcal{C}} t_{jd}^{\mathcal{C}}$ for $a, b, c, d \in \{1, \dots, q\}$. It yields $\text{COV}_4 = [\psi_{ms}]$ where $\psi_{ms} = \frac{1}{q+2} \left[\sum_{l=1}^k \alpha_l t_{lm}^{\mathcal{C}} t_{ls}^{\mathcal{C}} (t_l^{\mathcal{C}})^{\top} \text{COV}^{-1}(t_l^{\mathcal{C}}) \right]$ for $m, s \in \{1, \dots, q\}$.

$$(\text{COV}^{-1} \text{COV}_4)_{ms} = \frac{1}{q+2} \left[\sum_{l=1}^k \alpha_l (t_l^{\mathcal{C}})^{\top} \text{COV}^{-1}(t_l^{\mathcal{C}}) \sum_{i=1}^q b_{mi} t_{li}^{\mathcal{C}} t_{is}^{\mathcal{C}} \right] \text{ for } m, s \in \{1, \dots, q\}.$$

Appendix B. Details for the study of $\text{COV}^{-1} \text{COV}_4$ eigenvalues in Subsections 2.4, 3.2, 4.2 and additional figure

Subsection 2.4 and Sections 3 and 4 study the eigenvalues behavior of $\text{COV}^{-1} \text{COV}_4$. Appendix B.1 provides details about the analysis in the case of a Gaussian mixture (Subsection 2.4) whereas Appendix B.2 explains the study in the case of a Dirac mixture (Sections 3 and 4).

Appendix B.1. Gaussian mixture

In Subsection 2.4, the eigenvalues of $\text{COV}^{-1}\text{COV}_4$ are analysed in the case of a mixture of a Gaussian distributions. Figure 1 displays the eigenvalues obtained with numerical computations which are based on the theoretical ones described in Appendix A.1. In order to comprehend the behavior of eigenvalues for any number of groups k and any value of the dimension spanned by the group centers q , we will examine in particular the following cases: (i) $k = 2$: $q = 1$, (ii) $k = 3$: $q = 2$, (iii) $k = 5$: $q = 2$ and $q = 4$. There are two cases with $k = 5$ to illustrate the two subcases $q = k - 1$ and $q < k - 1$. To perform such numerical computations, we set to 6 the number of variables p , and we specify values for the group proportions and the group means. The latter are described in Table 1. We recall from Subsection 2.2 that:

$$T = \begin{pmatrix} T_u & 0 \\ 0 & 0 \end{pmatrix},$$

where $T = (t_1, \dots, t_k)$ with t_j a p -dimensional vector for $j \in \{1, \dots, k\}$, and T_u is an upper triangular matrix of dimension $k - 1 \geq 1$ such that the last $k - 1 - q \geq 0$ rows are zero. In Table 1, only the elements above the diagonal of T_u are mentioned. The values are selected manually, ensuring that the matrix T_u is full rank and that different configurations are represented. Some group means exhibit a particular structure. The first configuration of group means has a regular increment δ_i for each dimension, i.e. $t_{(j+1)i} = t_{ji} + \delta_i$ if $t_{ji} \neq 0$ for $j \in \{1, \dots, k - 1\}$, $i \in \{1, \dots, p\}$ and t_{ji} being the i -th element of t_j . The second configuration is equal to the first one multiplied by 10, i.e. $T_2 = 10 \times T_1$ where T_1 (T_2) are the group mean matrices of the first (second) configuration. The third configuration is a diagonal matrix. The remaining group mean configurations were randomly selected to ensure a variety of configurations. The dimensions of the matrix containing the group means are determined by the number of groups. Consequently, the initial $(k - 1)$ columns of Table 1 are selected, and a column of zeros is appended to the end. Only the first q -th rows are selected and zeros must be added up to the p -th row in order to fill the elements below the diagonal of T_u . This methodology yields the desired structure, which is consistent with the number of groups selected. For example, if $k = 3$ and $p = 3$, we obtain from the first row of Table 1 $T_{1,q=2}$ if $q = 2$ and $T_{1,q=1}$ if $q = 1$:

$$T_{1,q=2} = \begin{pmatrix} 200 & 200 & 0 \\ 0 & 400 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad T_{1,q=1} = \begin{pmatrix} 200 & 400 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

For each value of k , we select mixture proportions, i.e., the values of α_j for $j \in \{1, \dots, k\}$. The values have been selected to represent a range of scenarios and are listed in Table 2. For each number of groups, there is a scenario in which all groups have the same proportion. Then, the number of “low” and “large” groups is varied.

Next, for each mixture proportions scenario and each group centers configuration, the eigenvalues of $\text{COV}^{-1}\text{COV}_4$ are calculated. The results are plotted on a graph with the group proportions on the x -axis and a boxplot per eigenvalue on the y -axis, representing the distribution of the eigenvalue across the group centers for a given scenario. This representation permits the comparison of eigenvalues across scenarios and across group centers for a given number of groups.

Appendix B.2. Dirac Mixture

Section 3 analyses the eigenvalues of $V_1^{-1}V_2$ in the case of a mixture of a Dirac distributions where $p = q$. Figures 2, 10 and the additional Figure 10 show the eigenvalues of $\text{COV}^{-1}\text{COV}_4$ obtained with numerical calculations which are based on the theoretical ones described in Appendix A.2. To understand the behavior of eigenvalues for any number of groups, we will examine the cases of a mixture of 2, 3, 5, and 10 groups in particular. For each of these cases, we select mixture proportions scenarios that are described in Table 3. For 2, 3 and 5 groups, we added more scenarios to the one from Table 2. For $k = 10$, we created scenarios following the same logic. The centers of the groups are the same as those in Appendix B.1, described in Table 1. In Figure 2, $q = k - 1$. In Figures 3 and 10, $q < k - 1$. More precisely, we looked at: (i) $k = 3$: $q = 1$, (ii) $k = 5$: $q = 1$, $q = 2$ and $q = 3$, (iii) $k = 10$: $q = 1$, $q = 3$, $q = 5$ and $q = 7$.

In Subsection 4.2, we still consider the case of a Dirac mixture where $p = q$. However, the subsection focuses on the case $q = k - 1$ and does not address the impact of group centers. The group centers are then fixed to the configuration given by the first row of Table 1. Subsection 4.2 studies the idea of threshold values in the group proportions that result in an eigenvalue transitioning from less than one to more than one. Figure 7 illustrates these thresholds. Its configuration is very similar to the one of Figure 2: the represented number of groups k are 2, 3, 5 and 10, and $p = q = k - 1$. The proportions scenarios are described in Table 3. For each value of k , there is one scenario in which the proportion of the first group is at threshold level. The threshold is defined with the Setup 1 explained in Appendix C.2.

Table 1: Group centers used to compute the theoretical eigenvalues

(a) Groups 1 to 5

	t_1	t_2	t_3	t_4	t_5
1	[200]	[400, 100]	[600, 300, 200]	[800, 500, 300, 300]	[1000, 700, 400, 500, 200]
2	[2000]	[4000, 1000]	[6000, 3000, 2000]	[8000, 5000, 3000, 3000]	[10000, 7000, 4000, 5000, 2000]
3	[2]	[0, 2]	[0, 0, 6]	[0, 0, 0, 7]	[0, 0, 0, 0, 5]
4	[1]	[5, 7]	[9, 10, 2]	[13, 13, 3, 4]	[17, 16, 4, 8, 6]
5	[6]	[-3, 7]	[-10, 6, 4]	[1, 2, 8, 6]	[3, -3, 7, 9, 4]
6	[60]	[-30, 70]	[5, 15, 45]	[75, 23, 54, 66]	[59, 86, 38, 29, -32]
7	[18]	[9, 14]	[15, 22, 150]	[65, 42, 32, 15]	[22, 45, 12, 28, 41]
8	[-2]	[11, 4]	[0, 0, 10]	[0, 3, 0, 13]	[0, 0, 11, 34, 14]
9	[1200]	[180, 910]	[320, 112, 1000]	[550, 321, 875, 200]	[1000, 710, 593, 340, 900]
10	[36]	[39, 66]	[15, 88, 18]	[48, 67, 36, 83]	[22, 59, 48, -43, 36]
11	[300]	[460, 180]	[250, 80, 110]	[410, 100, 230, 260]	[230, 200, 160, 420, 318]
12	[12]	[50, 27]	[5, 13, 31]	[25, 25, 31, 42]	[49, 37, 21, 39, 45]
13	[16]	[-30, 18]	[1, 18, 50]	[11, -18, 59, -36]	[11, -18, 64, -12, 39]
14	[63]	[-35, 12]	[22, -12, 19]	[55, -71, 22, 38]	[54, -60, 31, 23, 45]
15	[22]	[9, 33]	[61, 42, 41]	[86, 42, 71, 30]	[48, 67, 36, 83, 22]
16	[-2]	[1, 2]	[3, 4, 6]	[8, 1, 5, 24]	[3, -3, 7, 9, 5]
17	[3]	[0, 6]	[0, 0, 9]	[0, 0, 0, 4]	[0, 0, 0, 0, 5]
18	[400]	[200, 550]	[280, -450, 356]	[312, 718, 462, 425]	[410, 100, 230, 0, 300]
19	[1080]	[2100, 1800]	[1870, 3200, 2200]	[1200, 2690, 3850, 4086]	[4800, -1700, 600, 2300, 2200]
20	[54]	[32, 41]	[12, 33, 81]	[43, 12, 79, 84]	[11, 18, 59, 36, 23]

(b) Groups 6 and 7

	t_6	t_7
1	[1200, 900, 500, 700, 300, 100]	[1400, 1100, 600, 900, 400, 200, 100]
2	[12000, 9000, 5000, 7000, 3000, 1000]	[14000, 11000, 6000, 9000, 4000, 2000, 1000]
3	[0, 0, 0, 0, 0, 8]	[0, 0, 0, 0, 0, -3]
4	[21, 19, 5, 12, 12, 10]	[25, 22, 6, 16, 18, 20, 9]
5	[-4, 2, 8, -5, 12, 7]	[4, 12, 5, -7, 23, 3, 8]
6	[77, 43, -55, 63, 0, 48]	[54, 92, 19, 74, -23, 44, 32]
7	[12, 33, 16, 32, 46, 33]	[17, 33, 21, 38, 57, 49, 67]
8	[0, 0, 8, 28, 17, 16]	[0, 0, 17, 22, 31, 2, 16]
9	[370, 212, 2000, 560, 230, 410]	[220, 230, 900, 120, 222, 714, 600]
10	[65, 45, 32, 56, 74, 43]	[8, 43, 59, -23, 31, 32, 54]
11	[530, 210, 310, 450, 530, 310]	[50, 100, 300, 536, 430, 320, 120]
12	[45, 27, 32, 65, 27, 33]	[32, 37, 26, 34, 45, 43, 21]
13	[12, 37, 21, 22, 7, 12]	[23, 37, 21, 34, 45, 32, 12]
14	[22, 37, 21, 38, 45, 69]	[22, 33, 64, 62, 45, 23, 36]
15	[49, 37, 21, 39, 45, 13]	[39, 45, 22, 45, 76, 23, 34]
16	[5, 2, 6, -4, 3, 8]	[1, 15, 4, 5, 6, 7, -3]
17	[0, 0, 0, 0, 0, 6]	[0, 0, 0, 0, 0, 2]
18	[390, 0, 450, 360, -300, 380]	[690, -212, 500, 560, -120, 410, 340]
19	[2100, 1000, -1000, 1389, 2300, 4500]	[500, 1000, 0, 4362, 4300, 3200, 1200]
20	[49, 37, 21, 39, 7, 65]	[45, 37, 32, 65, 28, 31, 66]

Table 1: Group centers used to compute the theoretical eigenvalues – *Continued*

(c) Groups 8 and 9

	t_8	t_9
1	[1600, 1300, 700, 1100, 500, 300, 200, 100]	[1800, 1500, 800, 1300, 600, 400, 300, 200, 100]
2	[16000, 13000, 7000, 11000, 5000, 3000, 2000, 1000]	[18000, 15000, 8000, 13000, 6000, 4000, 3000, 2000, 1000]
3	[0, 0, 0, 0, 0, 0, 0, 16]	[0, 0, 0, 0, 0, 0, 0, 20]
4	[29, 25, 7, 20, 24, 30, 14, 5]	[34, 27, 8, 24, 30, 40, 19, 10, 4]
5	[2, 32, 14, 3, -3, 0, 12, 10]	[6, 21, 19, 8, 7, 11, 26, 20, 16]
6	[34, 56, 44, 22, 0, 38, 0, 32]	[51, 32, 39, 30, 10, 78, 40, 28, 29]
7	[12, 20, 11, 31, 37, 9, 12, 30]	[0, 5, 31, 16, 7, 11, 19, 20, 17]
8	[0, 0, 7, 17, 27, 13, 9, 13]	[0, 0, 17, 32, 41, 5, 3, 19, 22]
9	[320, 530, 1100, 1200, 333, 743, 800, 900]	[120, 130, 2200, 3200, 444, 743, 2100, 200, 300]
10	[18, 32, 14, 45, 89, 21, 43, 32]	[20, 22, 44, 56, 61, 41, 61, 53, 33]
11	[110, 300, 500, 136, 230, 320, 0, 220]	[210, 0, 650, 536, 560, 400, 30, 100, 300]
12	[12, 32, 25, 35, 51, 54, 31, 50]	[32, 54, 76, 36, 22, 62, 66, 13, 56]
13	[37, 6, 54, 32, 0, 66, 10, 12]	[19, 26, 32, 54, 0, 55, 34, 17, 31]
14	[-39, 32, 22, 81, 56, 38, 74, 12]	[55, 34, 69, 65, 26, 38, 64, 42, 22]
15	[32, 38, 44, 52, 19, 42, 51, 21]	[25, 8, 79, 63, 43, 72, 61, 51, 32]
16	[2, 20, 11, 3, 3, 9, 12, 10]	[22, 2, 1, 4, 0, 5, 1, 8, 13]
17	[0, 0, 0, 0, 0, 0, 0, 8]	[0, 0, 0, 0, 0, 0, 0, 5]
18	[0, 89, 345, 280, 0, 270, 500, 220]	[610, 0, 385, 680, 0, 470, 76, 290, 300]
19	[0, 3100, 1400, 3800, 2300, -1100, 4200, 1800]	[3900, 100, 4200, 1900, 3700, 1100, 3100, -300, 2400]
20	[41, 32, 24, 74, 51, 23, 44, 65]	[27, 6, 51, 42, 61, 56, 33, 55, 52]

Table 2: Scenarios analysed in Figure 1

k	Scenarios
2	50-50, 20-80, 10-90
3	33-33-34, 10-40-50, 10-10-80
5	20-20-20-20-20, 10-10-20-20-40, 10-10-10-10-60

Table 3: Scenarios analysed in Figures 2, 3, 7 and 10

k	Scenarios for Figures 2, 3, 10	Scenarios for Figure 7
2	50-50, 40-60, 30-70, 20-80, 10-90	30-70, 21-79, 10-90
3	33-33-34, 20-30-50, 10-40-50, 10-30-60, 10-20-70, 10-10-80	33-33-34, 18-32-50, 10-40-50, 10-10-80
5	20-20-20-20-20, 10-20-20-20-30, 10-10-20-20-40, 10-10-10-30-40	20-20-20-20-20, 14-20-20-20-26, 10-10-20-20-40, 10-10-10-30-40
10	10-...-10, 5-5-5-10-10-10-10-15-20, 5-5-5-5-5-5-5-15-20-30	10-...-10, 8-10-10-10-10-10-10-10-12, 5-5-5-5-5-5-5-15-20-30

Appendix C. Details for the study of $\text{COV}^{-1}\text{COV}_4$ eigenvalues when group proportions vary in the absence of within-group variability in Section 4

In Section 4, we consider the case of a Dirac mixture, the scatter pair $\text{COV} - \text{COV}_4$, and $p = q = k - 1$, where p is the number of variables and q is the dimension of the space spanned by the group centers. We know that in this case the group centers have no effect on the eigenvalues of $\text{COV} - \text{COV}_4$. Therefore, we only vary the proportions of the groups. The eigenvalues of $\text{COV}^{-1}\text{COV}_4$ are obtained with numerical calculations using the theoretical developments described in Appendix A.2.

Appendix C.1. The case of three groups

Subsection 4.1 focuses on the case where the number of groups k is 3, and thus the number of variables p is 2. The eigenvalues are represented in ternary diagrams in Figures 4 and 5. The construction of such figures is achieved through

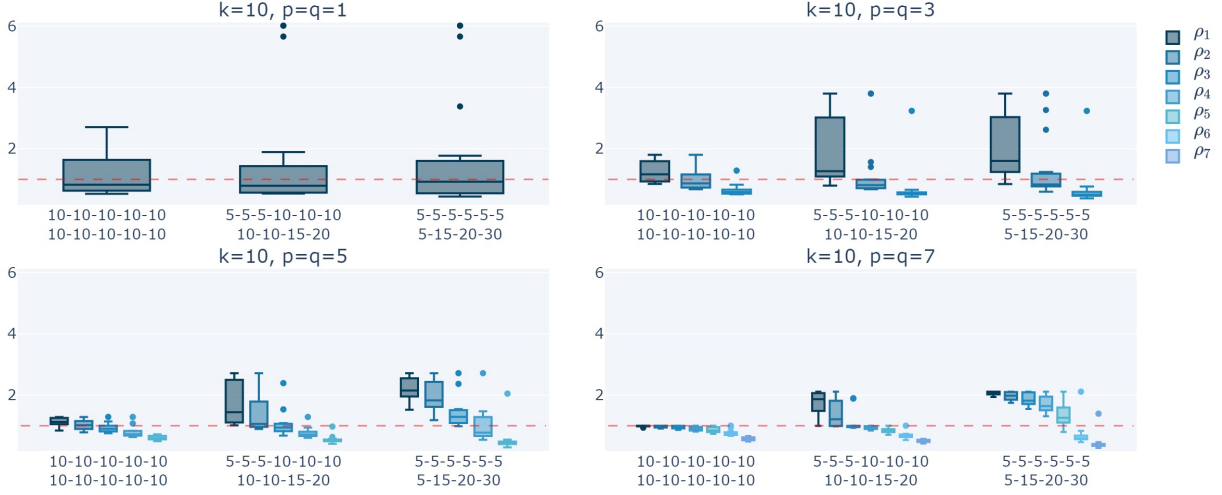


Fig. 10: Boxplots of the eigenvalues of $\text{COV}^{-1}\text{COV}_4$ with no within-group variability, where the group centers vary across 20 different configurations, with $k = 10$ and $p = q < k - 1$. The values of q vary across the panels. The 12 group proportions scenarios ($\alpha_j, j \in \{1, \dots, k\}$) vary on the x -axis.

the generation of a grid comprising all possible combinations of the group proportions α_1, α_2 and α_3 , with values between zero and one, such that their sum is equal to one. The step size is 0.001 (0.1%). This approach ensures exhaustive coverage of the proportion space and guarantees that every point on the ternary diagram represents a valid combination of α_1, α_2 , and α_3 . The eigenvalues of $\text{COV}^{-1}\text{COV}_4, \rho_1$ and ρ_2 , are calculated for each combination of the grid. To perform this calculation, it is necessary to have values for the centers of the three groups. As these values will have no impact on the result, they may be assigned randomly. For the sake of simplicity, we choose the first row of Table 1.

In Figure 4, it is of interest to highlight the case in which the eigenvalue is equal to one (white), as this is the case in which ICS does not “work”. Subsequently, two situations are distinguished: when the eigenvalue is smaller than one (blue), and when the eigenvalue is greater than one (red). Both blue and red parts follow a color gradient to illustrate the distance of the value from one. The logarithm of the eigenvalue was used in the coloring of the plot, because it allows for a more balanced distribution of colors. The logarithmic scale reduces the range difference between the blue and red parts, and since $\log(1) = 0$, it conveniently positions the transition color at a meaningful point. The scale of the gradient ranges from blue at the minimum of $\log(\rho)$ to red at the maximum of $\log(\rho)$, $\log(\rho) = 0$ being white. The gradient effectively illustrates the variations in the eigenvalue across the group proportions grid and highlights instances where ICS does not “work”. However, the color bar at the bottom of the plot represents the original eigenvalue, enhancing the plot’s interpretability.

Both eigenvalues are shown qualitatively in Figure 5. Since the information of interest is whether the eigenvalues are greater or less than one (as explained in Subsection 4.2, three cases are distinguished: ρ_1 and $\rho_2 > 1$ (red), $\rho_1 > 1$ and $\rho_2 < 1$ (purple), and ρ_1 and $\rho_2 < 1$ (blue)).

Appendix C.2. The case of k groups

Subsection 4.2 generalizes the context to a number k of groups and focuses on the notion of “thresholds”. Figure 6 is an illustration of the thresholds found using three different setups, for 2 to 10 groups. The threshold values are described in Table 4.

The first setup (Setup 1 in Figure 6) is the simplest. It is known that the sum of all group proportions must be equal to one. The most balanced scenario is the one in which all groups have the same proportion. In this case, each proportion is equal to $1/k$, where k is the number of groups. Thus, the smallest group proportion cannot exceed $1/k$ and the largest group proportion cannot be strictly smaller than $1/k$. A grid with group proportions is created for a given number of groups. The proportions are ordered in ascending order. For the first group proportion α_1 , the values range from 0.001 to $1/k$ with a step of 0.001. For each intermediate group proportion $\alpha_j, j \in \{2, \dots, k-1\}$, the value is set to $1/k$. The last group proportion α_k is the remaining proportion: $\alpha_k = 1 - \sum_{j=1}^{k-1} \alpha_j$. This setup ensures that the sum of group

Table 4: Thresholds computed for each setup

k	Setup 1	Setup 2	Setup 3
3	0.18	0.2	0.171
4	0.155	0.161	0.151
5	0.135	0.138	0.133
6	0.12	0.121	0.119
7	0.107	0.108	0.107
8	0.097	0.097	0.097
9	0.089	0.089	0.088
10	0.082	0.082	0.081

proportions is exactly one. Table 5 provides an example of the grid obtained with the first setup when $k = 4$. For each scenario in the grid, the eigenvalues of $\text{COV}^{-1}\text{COV}_4$ are calculated. Initially, all proportions are equal to $1/k$, and all eigenvalues are less than one. Subsequently, the proportion of the first group is the smallest and decreases by 0.001 at each iteration. The threshold of Setup 1 is the first value of the first group proportion for which one eigenvalue exceeds one. This procedure is repeated for 2 to 10 groups.

Table 5: Grids used in Setups 1, 2, and 3 for $k = 4$

Setup 1				Setup 2				Setup 3			
α_1	α_2	α_3	α_4	α_1	α_2	α_3	α_4	α_1	α_2	α_3	α_4
0.25	0.25	0.25	0.25	0.25	0.175	0.25	0.325	0.25	0.05	0.25	0.45
0.249	0.25	0.25	0.251	0.249	0.175	0.25	0.326	0.249	0.05	0.25	0.451
0.248	0.25	0.25	0.252	0.248	0.175	0.25	0.327	0.248	0.05	0.25	0.452
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
0.001	0.25	0.25	0.499	0.001	0.175	0.25	0.574	0.001	0.05	0.25	0.699

The second setup (Setup 2 in Figure 6) uses a grid based on the results obtained with the first setup. From the Setup 1 grid, the value of the second group proportion is updated. It is set to the threshold found in Setup 1 to which 2% is added. The proportion of the first group is the same, ranging from 0.001 to $1/k$ with a step of 0.001. The intermediate groups have a proportion of $1/k$ only for group 3 through $k - 1$, since the proportion of the second group is already set. The value of the last group proportion is also updated, but it is still equal to one minus the other group weights. This procedure ensures that the sum of each row is exactly one. An example of the grid obtained with the second setup when $k = 4$ is shown in Table 5. However, it requires at least three groups: one that is set to the first threshold plus 2%, and two groups with variable proportions. For each configuration of group weights, the eigenvalues of $\text{COV}^{-1}\text{COV}_4$ are calculated. Initially, the first group is not necessarily the smallest one; it could be the second group. However, the value of the second group proportion is set such that it is above the threshold, implying that all eigenvalues are initially below one. Thus, the threshold of Setup 2 is, as in Setup 1, the first value of the first group weight for which one eigenvalue exceeds one. The concept in this procedure revolves around examining the impact of a group that is positioned near the threshold to determine if this proximity affects it. This process is repeated for a number of groups ranging from 3 to 10.

Finally, the third procedure (Setup 3 in Figure 6) uses a grid similar to Setup 2. The second group proportion is no longer dependent on the threshold of Setup 1 but is set to 0.05. Again, the last group proportion is updated to one minus the combined weights of the other groups. All other group proportions remain the same as in Setups 1 and 2. Table 5 provides an example of the third setup grid with $k = 4$. As for the grid in Setup 2, it requires at least three groups: one for the fixed proportion and two to vary. The objective of this procedure is to include a small group initially, specifically below the threshold for k from 3 to 10. This implies that, from the initial point, there is already an eigenvalue greater than one. Thus, the threshold of Setup 3 is the first value of the first group weight for which two eigenvalues exceed one. This is aimed at identifying the point where the second eigenvalue surpasses one. This setup is repeated for 3 to 10 groups.

Appendix D. Proof of Proposition 2

Proof. By adapting the calculations in Appendix A (with the same notations) to the Gaussian mixture model (9), we obtain

$$\text{COV} = \begin{bmatrix} \beta_{11} & 0 \\ 0 & I_{p-1} \end{bmatrix}, \quad \text{COV}^{-1} = \begin{bmatrix} b_{11} & 0 \\ 0 & I_{p-1} \end{bmatrix}, \quad \text{and} \quad \text{COV}_4 = \begin{bmatrix} a_{11} & 0 \\ 0 & I_{p-q} \end{bmatrix}$$

where $\beta_{11} = 1 + \sum_{j=1}^3 \alpha_j (t_{j1}^{\mathbf{c}})^2$, $b_{11} = \beta_{11}^{-1}$ and $a_{11} = \frac{1}{p+2} \left[b_{11} \mathbb{E}[(x_1^{\mathbf{c}})^4] + (p-1) \mathbb{E}[(x_1^{\mathbf{c}})^2] \right]$. We have

$$\text{COV}^{-1} \text{COV}_4 = \begin{bmatrix} b_{11} a_{11} & 0 \\ 0 & I_{p-1} \end{bmatrix}$$

All eigenvalues of $\text{COV}^{-1} \text{COV}_4$ equal one if and only if

$$b_{11} a_{11} = 1, \tag{9}$$

(9) is equivalent to $a_{11} = b_{11}^{-1} = \beta_{11}$. Noting that $\mathbb{E}[(x_1^{\mathbf{c}})^2] = \beta_{11}$, and expanding the expression of a_{11} , (9) is equivalent to $3\beta_{11}^2 = \mathbb{E}[(x_1^{\mathbf{c}})^4]$. Using the formula of β_{11} and the formula of $\mathbb{E}[(x_1^{\mathbf{c}})^4]$ from Appendix A, we obtain that (9) is equivalent to:

$$\begin{aligned} & \alpha_1(3\alpha_1 - 1)(t_{11}^{\mathbf{c}})^4 + \alpha_2(3\alpha_2 - 1)(t_{21}^{\mathbf{c}})^4 + \alpha_3(3\alpha_3 - 1)(t_{31}^{\mathbf{c}})^4 \\ & + 6\alpha_1\alpha_2(t_{11}^{\mathbf{c}})^2(t_{21}^{\mathbf{c}})^2 + 6\alpha_1\alpha_3(t_{11}^{\mathbf{c}})^2(t_{31}^{\mathbf{c}})^2 + 6\alpha_2\alpha_3(t_{21}^{\mathbf{c}})^2(t_{31}^{\mathbf{c}})^2 = 0. \end{aligned}$$

For the mixture (6), we know from Theorem 4 in Tyler et al. [2009] and using the computations of Subsection 2.4, that at least $p-1$ eigenvalues of $\text{COV}^{-1} \text{COV}_4$ are equal to one. As for the remaining eigenvalue, using the computations of Subsection 2.4 and Mathematica Wolfram Research, Inc. [2022], we can prove that it is equal to one if and only if

$$p_{\alpha_1, \alpha_2}(t_{11}, t_{12}) = 0, \tag{10}$$

where

$$\begin{aligned} p_{\alpha_1, \alpha_2}(t_{11}, t_{12}) = & \alpha_1(-1 + 7\alpha_1 - 12\alpha_1^2 + 6\alpha_1^3)t_{11}^4 + 4\alpha_1\alpha_2(1 - 6\alpha_1 + 6\alpha_1^2)t_{11}^3 t_{12} + 6\alpha_1\alpha_2(1 - 2\alpha_2 + \\ & \alpha_1(-2 + 6\alpha_2))t_{11}^2 t_{12}^2 + 4\alpha_1\alpha_2(1 - 6\alpha_2 + 6\alpha_2^2)t_{11} t_{12}^3 + \alpha_2(-1 + 7\alpha_2 - 12\alpha_2^2 + 6\alpha_2^3)t_{12}^4. \end{aligned}$$

Since $t_{21} \neq 0$, we have

$$p_{\alpha_1, \alpha_2}(t_{11}, t_{12}) = t_{21}^4 r_{\alpha_1, \alpha_2}(t_{11}/t_{21}),$$

for the polynomial r_{α_1, α_2} of degree 4 given in Proposition 2, which concludes the proof. \square

References

- H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.
- I. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, New York, 2nd edition, 2002. ISBN 978-0-387-95442-4.
- Peter J Huber. Projection pursuit. *Annals of Statistics*, 13:435–475, 1985.
- M. C. Jones and R. Sibson. What is projection pursuit? *Journal of the Royal Statistical Society: Series A (General)*, 150(1):1–18, 1987.
- Daniel Fischer, Alain Berro, Klaus Nordhausen, and Anne Ruiz-Gazen. REPlab: An R package for detecting clusters and outliers using exploratory projection pursuit. *Communications in Statistics-Simulation and Computation*, 50(11): 3397–3419, 2021. doi:10.1080/03610918.2019.1626880.
- David E. Tyler, Frank Critchley, Lutz Dümbgen, and Hannu Oja. Invariant co-ordinate selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):549–592, June 2009. ISSN 13697412, 14679868. doi:10.1111/j.1467-9868.2009.00706.x.
- H. Caussinus and A. Ruiz. Interesting Projections of Multidimensional Data by Means of Generalized Principal Component Analyses. In Konstantin Momirović and Vesna Mildner, editors, *Compstat*, pages 121–126. Physica-Verlag HD, 1990. doi:10.1007/978-3-642-50096-1_19.

- R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188, 1936. ISSN 2050-1439. doi:10.1111/j.1469-1809.1936.tb02137.x.
- P. Arabie and L. J. Hubert. Cluster Analysis in Marketing Research. In R. P. Bagozzi, editor, *Advanced methods in marketing research*, pages 160–189. Blackwell, Oxford, 1994.
- U. Radojičić, K. Nordhausen, and J. Virta. Large-sample properties of blind estimation of the linear discriminant using projection pursuit. *Electronic Journal of Statistics*, 15(2), 2021.
- Aurore Archimbaud, Klaus Nordhausen, and Anne Ruiz-Gazen. ICS for multivariate outlier detection with application to quality control. *Computational Statistics & Data Analysis*, 128:184–199, December 2018. ISSN 01679473. doi:10.1016/j.csda.2018.06.011.
- Andreas Alfons, Aurore Archimbaud, Klaus Nordhausen, and Anne Ruiz-Gazen. Tandem clustering with invariant coordinate selection. *Econometrics and Statistics*, March 2024. ISSN 2452-3062. doi:10.1016/j.ecosta.2024.03.002.
- Klaus Nordhausen and David E. Tyler. A cautionary note on robust covariance plug-in methods. *Biometrika*, 102(3): 573–588, 2015. ISSN 0006-3444. doi:10.1093/biomet/asv022.
- Klaus Nordhausen, Hannu Oja, and David E. Tyler. Tools for Exploring Multivariate Data: The Package ICS. *Journal of Statistical Software*, 28:1–31, 2008. doi:10.18637/jss.v028.i06.
- Daniel Peña, Francisco J Prieto, and Júlia Viladomat. Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure. *Journal of Multivariate Analysis*, 101(9):1995–2007, 2010.
- Klaus Nordhausen, Hannu Oja, and Esa Ollila. Multivariate models and the first four moments. In *Nonparametric Statistics and Mixture Models*, pages 267–287. World Scientific, Hackensack, 2011. doi:10.1142/9789814340564_0016.
- N. Loperfido. Skewness and the linear discriminant function. *Statistics & Probability Letters*, 83(1):93–99, 2013.
- Fatimah Alashwali and John T. Kent. The use of a common location measure in the invariant coordinate selection and projection pursuit. *Journal of Multivariate Analysis*, 152:145–161, 2016. doi:10.1016/j.jmva.2016.08.007.
- Nicola Loperfido. Some theoretical properties of two kurtosis matrices, with application to invariant coordinate selection. *Journal of Multivariate Analysis*, 186:104809, 2021. doi:https://doi.org/10.1016/j.jmva.2021.104809.
- Klaus Nordhausen and Anne Ruiz-Gazen. On the usage of joint diagonalization in multivariate statistics. *Journal of Multivariate Analysis*, 188:104844, 2022. doi:https://doi.org/10.1016/j.jmva.2021.104844.
- Aurore Archimbaud, Zlatko Drmač, Klaus Nordhausen, Una Radojičić, and Anne Ruiz-Gazen. Numerical Considerations and a new implementation for invariant coordinate selection. *SIAM Journal on Mathematics of Data Science*, 5(1):97–121, 2023a. doi:10.1137/22M1498759.
- J.-F. Cardoso. Source Separation Using Higher Order Moments. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2109–2112. IEEE, 1989.
- Klaus Nordhausen and Joni Virta. An overview of properties and extensions of FOBI. *Knowledge-Based Systems*, 173: 113–116, 2019. Publisher: Elsevier.
- Aurore Archimbaud. *Méthodes statistiques de détection d’observations atypiques pour des données en grande dimension*. Thèse, Toulouse 1 Capitole, January 2018. URL <http://www.theses.fr/2018TOU10001>.
- F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley Series in Probability and Statistics. Wiley, 2011. ISBN 978-1-118-15068-9. URL <https://books.google.fr/books?id=XK3uhrVefXQC>.
- Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. *Modeling and Analysis of Compositional Data*. Wiley, 1 edition, March 2015. ISBN 978-1-118-44306-4 978-1-119-00314-4. doi:10.1002/9781119003144.
- Wolfram Research, Inc. Mathematica, Version 13.2, 2022. URL <https://www.wolfram.com/mathematica>. Champaign, IL.
- Peter J. Rousseeuw and Katrien Van Driessen. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41(3):212–223, August 1999. ISSN 0040-1706. doi:10.1080/00401706.1999.10485670.
- Lutz Dümbgen, Katrin Gysel, and Fabrice Perler. Refining Invariant Coordinate Selection via Local Projection Pursuit. In Mengxi Yi and Klaus Nordhausen, editors, *Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler*, pages 121–136. Springer International Publishing, Cham, 2023. ISBN 978-3-031-22687-8. doi:10.1007/978-3-031-22687-8_6.
- U. Radojicic and K. Nordhausen. Non-Gaussian Component Analysis: Testing the Dimension of the Signal Subspace. In M. Maciak, M. Pesta, and M. Schindler, editors, *Analytical Methods in Statistics. AMISTAT 2019*, pages 101–123. Springer, Cham, 2020.

- Annaliisa Kankainen, Sara Taskinen, and Hannu Oja. Tests of multinormality based on location vectors and scatter matrices. *Statistical Methods & Applications*, 16:357–379, 2007.
- Wei Luo and Bing Li. Combining Eigenvalues and Variation of Eigenvectors for Order Determination. *Biometrika*, 103(4):875–887, 2016.
- K. Nordhausen, H. Oja, and D. E. Tyler. Asymptotic and Bootstrap Tests for Subspace Dimension. *Journal of Multivariate Analysis*, 188:104830, 2022. doi:<https://doi.org/10.1016/j.jmva.2021.104830>.
- Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.
- Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi:10.1038/s41586-020-2649-2.
- The pandas development team. `pandas-dev/pandas`: Pandas, February 2020.
- Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. doi:10.25080/ajora-92bf1922-00a.
- Plotly Technologies Inc. Collaborative data science, 2015. URL <https://plot.ly>.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL <https://www.R-project.org/>.
- Klaus Nordhausen, Andreas Alfons, Aurore Archimbaud, Hannu Oja, Anne Ruiz-Gazen, and David E. Tyler. ICS: Tools for Exploring Multivariate Data via ICS/ICA, September 2023. URL <https://cran.r-project.org/web/packages/ICS/index.html>. R package 1.4-1.
- Aurore Archimbaud, Andreas Alfons, Klaus Nordhausen, and Anne Ruiz-Gazen. *ICSClust: Tandem Clustering with Invariant Coordinate Selection*. 2023b. URL <https://CRAN.R-project.org/package=ICSClust>.
- Valentin Todorov and Peter Filzmoser. An Object-Oriented Framework for Robust Multivariate Analysis. *Journal of Statistical Software*, 32(3):1–47, 2009. doi:10.18637/jss.v032.i03.