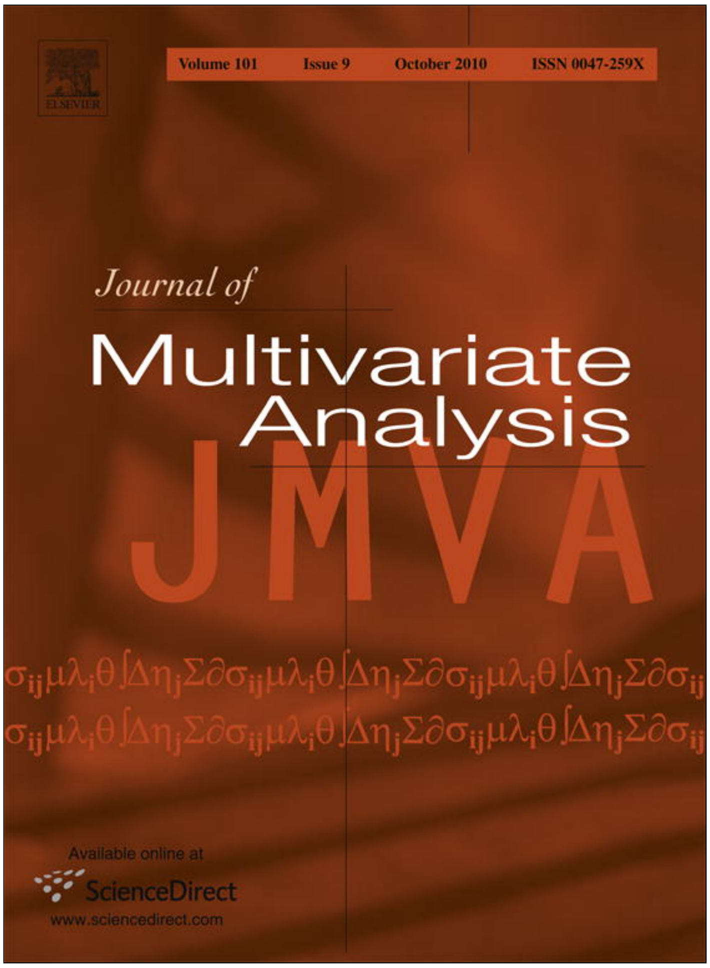


ELSEVIER

Volume 101 Issue 9 October 2010 ISSN 0047-259X

Journal of
Multivariate
Analysis

JMVA

Available online at
ScienceDirect
www.sciencedirect.com

Journal of Multivariate Analysis 101 (2010) 1995–2007

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure

Daniel Peña, Francisco J. Prieto, Júlia Viladomat [*]

Departamento de Estadística, Universidad Carlos III de Madrid, c/Madrid 126, 28903 Getafe, Spain

ARTICLE INFO

Article history:
Received 3 April 2009
Available online 12 May 2010

AMS subject classifications:
62H30
62H12

Keywords:
Cluster analysis
Dimension reduction
Fisher subspace
Kurtosis matrix
Multivariate kurtosis
Projection Pursuit

ABSTRACT

In this paper we study the properties of a kurtosis matrix and propose its eigenvectors as interesting directions to reveal the possible cluster structure of a data set. Under a mixture of elliptical distributions with proportional scatter matrix, it is shown that a subset of the eigenvectors of the fourth-order moment matrix corresponds to Fisher's linear discriminant subspace. The eigenvectors of the estimated kurtosis matrix are consistent estimators of this subspace and its calculation is easy to implement and computationally efficient, which is particularly favourable when the ratio n/p is large.

1. Introduction

Given a multivariate sample in $\mathbb{R}^p$ drawn from a mixture of k populations, cluster analysis attempts to partition the sample into homogeneous groups, according to the populations that generate them. Projection Pursuit finds subspaces of low dimension that show interesting views of the data according to some criteria, see [6,5]. Projection Pursuit can be useful in cluster analysis. One may first reduce the dimensionality of the sample by projecting it on a lower-dimensional subspace and then finding the clusters there. The curse of dimensionality can thus be avoided, but care needs to be taken to make sure that the projected data preserve the cluster structure of the original sample. Non-normality is one of the criteria used to find the projections. Huber [8] emphasized that interesting projections are those that produce non-normal distributions. However, non-normality is a general condition, and we need to specify how to measure it.

The idea of maximizing the kurtosis has also been used in cluster analysis, see [9]. Peña and Prieto [19] showed that for clustering the directions that minimize the kurtosis can be more useful than the ones that maximize it. The reason is that the kurtosis can be seen as the variance of the squared standardized differences between the variable and its mean. Consequently, if all observations of the sample are approximately at the same distance to the mean, the variance of these distances is near zero, and the kurtosis will have a small value. This would be the case with two well-separated clusters of the same size. Therefore, directions that minimize the kurtosis could reveal the cluster structure. The method proposed by Peña and Prieto [19] (and Projection Pursuit methods in general) needs to perform numerical optimization in order to find the optimal directions. This is computationally intensive and its efficacy may depend on the choice of the optimization algorithm to be used.

An alternative to this approach is to find a matrix whose eigenvectors are directly directions of maximum or minimum kurtosis. In this paper we study a kurtosis matrix and show that under a mixture of two elliptical distributions with the same

[*] Corresponding author.
E-mail addresses: daniel.pena@uc3m.es (D. Peña), franciscojavier.prieto@uc3m.es (F.J. Prieto), julia.viladomat@uc3m.es, juliaviladomat@gmail.com (J. Viladomat).

0047-259X/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2010.04.014

scatter matrices, the eigenvector associated to the eigenvalue different from the others coincides with the direction that optimizes the kurtosis coefficient, which is Fisher's linear discriminant function. The kurtosis matrix, thus, has similarities to the nonlinear cluster algorithm in [19]. Based on this result, we explore the general case of k groups and we prove that the subspace orthogonal to the eigenspace associated to an eigenvalue with multiplicity $p - k + 1$ is Fisher's linear discriminant subspace. Similar results are found in [3,4], where it is shown that Fisher's subspace can be estimated using the k largest eigenvectors of some Generalized Principal Components matrix based on W-estimates of dispersion. Recently, Tyler et al. [24] prove that a subset of eigenvectors of $S_1^{-1}S_2$ generate Fisher's subspace, S_1 and S_2 being any pair of affine equivariant scatter matrices.

The kurtosis matrix, however, is based on an existent kurtosis-based algorithm which can always be used. The advantage of using the eigenvectors of a kurtosis matrix instead of the univariate kurtosis directions is dependent on the ratio n/p, where n is the sample size and p the dimension. If this ratio is large, the estimation of the kurtosis matrix of dimension p is reliable and therefore the estimation of its eigenvectors becomes accurate and useful. Also, in this case numerical optimization is computationally intensive. However, when n/p is small the estimation of the elements of the matrix has very low precision and we have found that the eigenvalues are not useful. We will illustrate in which situations is more adequate to use one approach or another. Also, we will show that these eigenvectors are consistent estimators of Fisher's subspace, which ensures their convergence.

Note that, similarly to the procedure described in [20] for methods based on the univariate kurtosis extreme directions, the algorithm proposed in this paper, based on the eigenvectors of the kurtosis matrix, can be complemented with additional directions to improve its efficiency on those cases when the kurtosis values are similar for all directions and the extreme directions are not informative.

This paper is organized as follows. Section 2 is a review of multivariate kurtosis coefficients and matrices defined in the literature. In Section 3 we study the theoretical properties of the eigenvectors of a kurtosis matrix for cluster analysis and present results regarding the convergence of their estimators. In Section 4 the behaviour of the eigenvectors to perform cluster analysis is analyzed through a simulation study. We finish with some final remarks in Section 5.

2. Multivariate kurtosis coefficients and matrices

Let X be a multivariate $p \times 1$ random vector, μ its mean vector, Σ its covariance matrix and $Z = \Sigma^{-1/2}(X - \mu)$ the corresponding standardized vector. For $p = 1$ the univariate kurtosis coefficient is $E(z^4)$, where $z = (x - \mu)/\sigma$, and a natural extension of kurtosis to multivariate samples is to consider the second moment of the Mahalanobis distances, $\beta_{2,p} = E(Z^{\mathrm{T}}Z)^2$, which is Mardia's multivariate kurtosis coefficient, see [14]. Since $\beta_{2,p}$ can also be expressed as $\beta_{2,p} = \mathrm{var}(Z^{\mathrm{T}}Z) + E(Z^{\mathrm{T}}Z)^2$ and $E(Z^{\mathrm{T}}Z) = p$, then $\beta_{2,p} \geq p^2$. The sample counterpart of $\beta_{2,p}$ is $b_{2,p} = 1/n \sum_{i=1}^{n} [(x_i - \bar{x})^{\mathrm{T}} S^{-1} (x_i - \bar{x})]^2$, where $\bar{x}$ and S are the mean and covariance matrix of $x_1 \cdots x_n$, a random sample of X. Mardia proposes to use $b_{2,p}$ when testing for normality. Under a Gaussian distribution $\beta_{2,p} = p(p + 2)$, therefore values of $b_{2,p}$ differing significantly from $p(p + 2)$ indicate non-normality.

Koziol [12] proposes as measure of multivariate kurtosis $\tilde{\beta}_{2,p} = \sum_{i,j,k,l}^{p} E(Z_iZ_jZ_kZ_l)^2$. The difference between $\tilde{\beta}_{2,p}$ and $\beta_{2,p}$ is that $\beta_{2,p}$ is the sum of only the fourth-order moments of the type $E(Z_i^4)$ and $E(Z_i^2Z_j^2)$, while $\tilde{\beta}_{2,p}$ is the sum of squares of all fourth-order moments of Z. Oja [17] defines a multivariate kurtosis coefficient considering the volume of the simplex in a p-dimensional space determined by $p + 1$ points as $\beta_{2,p}^{*} = E[\Delta(X_1, \ldots, X_p, \mu)]^4 / [E[\Delta(X_1, \ldots, X_p, \mu)]^2]^2$, being $X_1, \ldots, X_p$ independent random vectors distributed as X and Δ is the volume of this simplex:

$$\Delta\left(X_1, \ldots, X_{p+1}\right) = \mathrm{abs}\left(\frac{1}{p!}\begin{vmatrix} 1 & \cdots & 1 \\ X_{11} & \cdots & X_{p+1,1} \\ \vdots & & \vdots \\ X_{1p} & \cdots & X_{p+1,p} \end{vmatrix}\right).$$

Finally, Malkovich and Afifi [13] define the multivariate kurtosis as the maximum univariate kurtosis produced by any projection of the p-dimensional distribution onto a direction d; $\beta_2^M = \max_d |\beta_2(d) - 3|$, where $\beta_2(d) = E[(d^{\mathrm{T}}X - d^{\mathrm{T}}\mu)^4 / d^{\mathrm{T}}\Sigma d]$. The measures $\beta_{2,p}$, $\beta_{2,p}^{*}$ and β_2^M are invariant under nonsingular affine transformations and reduce to the univariate kurtosis when $p = 1$, which is not the case for $\tilde{\beta}_{2,p}$.

Let $M_4 = E(ZZ^{\mathrm{T}} \otimes ZZ^{\mathrm{T}})$ be the $p^2 \times p^2$ matrix that collects all $p^2 \times p^2$ multivariate fourth-order moments, where $\otimes$ denotes the Kronecker product. We have found two multivariate $p \times p$ kurtosis matrices in the literature. The first one is due to Cardoso [2] and Móri et al. [16], who define

$$K = I_p * M_4 = E(Z^{\mathrm{T}}ZZZ^{\mathrm{T}}), \tag{1}$$

whereas Kollo [11] defines a kurtosis matrix as

$$B = \mathbf{1}_{p\times p} * M_4 = E\left[(Z^{\mathrm{T}}\mathbf{1})^2 ZZ^{\mathrm{T}}\right], \tag{2}$$

where $*$ denotes the star product defined in [15]. The two matrices can be seen as weighted scatter matrices with weights $Z^{\mathrm{T}}Z$ and $(Z^{\mathrm{T}}\mathbf{1})^2$ respectively. Both matrices are positive semidefinite and reduce to the univariate kurtosis coefficient when $p = 1$. Also, the trace of K is $\beta_{2,p}$, Mardia's kurtosis coefficient.

We are interested in projecting the multivariate sample onto the (sub)space generated by (some of) the eigenvectors of a kurtosis matrix, expecting that this new coordinate system will give us insight on the cluster structure of the data. The matrix K in (1) has an important invariant property which is not present in B in (2). Let E be an orthogonal matrix whose columns are eigenvectors of K, the new coordinate system $E^{\mathrm{T}}Z$ is invariant under affine transformations of X. In effect, if $Y = AX + b$ with A nonsingular, then $K_Y = UKU^{\mathrm{T}}$, where U is some orthogonal matrix. This is true because the standardizations of X and Y are the same up to a rotation, $Z_Y = UZ$, where $Z_Y = \Sigma_Y^{-1/2}(Y - \mu_Y)$. That implies that the eigenvalues of K and K_Y are the same and the eigenvectors are rotated versions of each other (the eigenvectors of K_Y are UE). When applying the same transformation to Z_Y, we obtain the same coordinates $E^{\mathrm{T}}U^{\mathrm{T}}UZ = E^{\mathrm{T}}Z$. The matrix B, however, does not have this desirable property because its weights are not invariant under orthogonal transformations.

Our intention is to continue to explore the properties of the matrix K to perform cluster analysis. Up to now, only the univariate kurtosis has been explored for clustering.

3. The eigenvectors of a kurtosis matrix and cluster properties of the data

Let X follow a mixture of k elliptical distributions such that, with probability $\pi_i > 0$, X_i has density

$$f_{X_i}(x) = |V_i|^{-1/2} h_i[(x - \mu_i)^{\mathrm{T}} V_i^{-1}(x - \mu_i)],$$

with parameters μ_i, V_i and for some nonnegative function h_i, $i = 1, \ldots, k$ and $\sum_{i=1}^{k} \pi_i = 1$. The matrix K in (1) can be expressed as (see Appendix A for details)

$$K = \sum_{i=1}^{k} \pi_i[\mathrm{tr}\, W_i(\tilde{k}_i W_i + \delta_i\delta_i^{\mathrm{T}}) + \bar{k}_i W_i^2] + \sum_{i=1}^{k} \pi_i \left[2(\delta_i\delta_i^{\mathrm{T}} W_i + W_i\delta_i\delta_i^{\mathrm{T}}) + \delta_i^{\mathrm{T}}\delta_i(W_i + \delta_i\delta_i^{\mathrm{T}})\right], \tag{3}$$

where $\delta_i = \Sigma^{-1/2}(\mu_i - \mu)$ and $W_i = \Sigma^{-1/2} V_i \Sigma^{-1/2}$ are the means and scatter matrices of the standardized $Z_i = \Sigma^{-1/2}(X_i - \mu)$. This explicit expression for the matrix gives insight on the structure of the problem. Some terms depend on the variability between clusters, the δ_i's, and others on the variability within clusters, the W_i's. We need the eigenstructure of K to capture the cluster structure, which is found in the δ_i's.

3.1. Proportional scatter matrices

If the scatter matrices of the groups are proportional, it is seen in Theorem 1 that the eigenvectors of K reveal some desirable properties for clustering.

Theorem 1. *Suppose X is a mixture of elliptical distributions as stated above with $V_i = V$, for $i = 1, \ldots, k$. The matrix K is*

$$K = \alpha I + \sum_{i=1}^{k}\sum_{j=1}^{k} \beta_{ij}\delta_i\delta_j^{\mathrm{T}}, \tag{4}$$

with $\alpha = \tilde{k}p + (1 - \tilde{k})\sum_{i=1}^{k} \pi_i\delta_i^{\mathrm{T}}\delta_i + \bar{k}$ and where

$$\beta_{ij} = \begin{cases} \gamma\pi_i + (\pi_i + \eta\pi_i^2)\delta_i^{\mathrm{T}}\delta_i & \text{if } i = j \\ \eta\pi_i\pi_j\delta_i^{\mathrm{T}}\delta_j & \text{if } i \neq j \end{cases}$$

with $\gamma = (1 - \tilde{k})p - 2\bar{k} + 4$, $\eta = \tilde{k} + \bar{k} - 6$, $\tilde{k} = \sum_{i=1}^{k} \pi_i\tilde{k}_i$ and $\bar{k} = \sum_{i=1}^{k} \pi_i\bar{k}_i$.

We name $\Delta = \langle\delta_1, \ldots, \delta_k\rangle$ the subspace spanned by the δ_i's, where $\dim \Delta = q \leq k - 1$. If $u \in \Delta^{\perp}$, $Ku = \alpha u$ holds, and α is an eigenvalue of K with multiplicity $p - q$ associated to the eigenspace $\Delta^{\perp}$. The remaining q eigenvectors of K are found in the Δ subspace.

Let $\Phi = \langle\phi_1, \ldots, \phi_k\rangle$ be the subspace spanned by Fisher's directions, $\phi_i = V^{-1}(\mu_i - \mu)$.

Then, the subspaces Φ and Δ_X are the same

$$\Phi = \Delta_X, \tag{5}$$

where Δ_X is the Δ-subspace expressed in the space of the original variables, $\Delta_X = \Sigma^{-1/2}\Delta$.

Under the assumption of proportional scatter matrices the best discriminant procedure is linear and Fisher's linear discriminant subspace is optimal in the sense that the relative separation between means is maximized. The theorem states that an identifiable subset of q eigenvectors of the kurtosis matrix K give the subspace on which the clusters appear more separated. Some details of the theorem are found in Appendix B.

Corollary 1. *In the particular case of a mixture of normal distributions, the constants are respectively $\tilde{k}_i = 1$ and $\bar{k}_i = 2$ and the eigenvalue associated to $\Delta^{\perp}$ has known value $\alpha = p + 2$. Also, if there are no clusters, from (4) we have $K = \alpha I$.*

Also, in the particular case of a mixture of two normal distributions, the matrix K simplifies to

$$K = (p+2)I + \beta\varphi^{\mathrm{T}}\varphi\varphi\varphi^{\mathrm{T}}, \tag{6}$$

where $\beta = \pi_1\pi_2(1 - 6\pi_1\pi_2)$ and $\varphi = \Sigma^{-1/2}(\mu_2 - \mu_1)$. The vector φ is an eigenvector of K with associated eigenvalue $\lambda = p + 2 + \beta(\varphi^{\mathrm{T}}\varphi)^2$, the rest of eigenvalues are equal to $p + 2$. Also, $\operatorname{tr}(K) = p(p+2) + \beta(\varphi^{\mathrm{T}}\varphi)^2$ and $\det(K) = (p+2)^p + \beta(p+2)^{p-1}(\varphi^{\mathrm{T}}\varphi)^2$. Note that φ is Fisher's best linear discriminant function in the Z-space. The eigenvalue λ is the largest if $\beta > 0$ and the smallest otherwise. The parameter β is positive if $\pi_1 \in (0, (\sqrt{3}-1)/(2\sqrt{3}))$ and negative if $\pi_1 \in ((\sqrt{3}-1)/(2\sqrt{3}), 0.5]$. Therefore, if we have homogeneous clusters, the eigenvector associated with the smallest eigenvalue will be the one that better separates the clusters, while when the two clusters have very different sizes, the largest eigenvalue is the one that identifies the significant eigenvector. These values are the same values that arise in Corollary 2 in [19], where they prove that the direction that optimizes the univariate kurtosis coefficient corresponds to Fisher's best linear discriminant function, maximizing it if $\pi_1 \in (0, (\sqrt{3}-1)/(2\sqrt{3}))$ and minimizing it if $\pi_1 \in ((\sqrt{3}-1)/(2\sqrt{3}), 0.5]$. Both approaches give estimations of Fisher's linear discriminant function, and the question is in which circumstances one way is more appropriate than the other. On one hand, the estimation of eigenvectors can suffer from lack of precision when the sample size is small, on the other hand a nonlinear computationally intensive algorithm is needed to solve the optimization problem of finding the direction of kurtosis. We will address the issue in the next section with the help of some simulations.

Theorem 1 is in agreement with Theorem 5.2 in [24] and is similar to Proposition 1 in [3]. In the former the authors present a general method to generate an affine invariant coordinate system to reveal interesting departures from an elliptical distribution by using the eigenvectors of $S_1^{-1}S_2$, one scatter matrix relative to another. The idea is to first 'standardize' the data with respect to one scatter matrix S_1, and then perform generalized principal components on the 'standardized' data using a different scatter statistic S_2. Calculating the eigenvectors of the kurtosis matrix K is equivalent to choosing $S_1 = \Sigma$, and $S_2 = E[Z^{\mathrm{T}}Z(X-\mu)(X-\mu)^{\mathrm{T}}]$. In this case $S_1^{-1}S_2 = \Sigma^{-1/2}K\Sigma^{1/2}$, and the eigenvalues of $S_1^{-1}S_2$ and K are the same while the eigenvectors are $\Sigma^{-1/2}u$ and u respectively. As a matter of fact, these choices are the ones proposed in Caussinus and Ruiz-Gazen [3], where more generally they study $S_2 = E[\omega(\beta Z^{\mathrm{T}}Z)(X-\mu)(X-\mu)^{\mathrm{T}}]/E[\omega(\beta Z^{\mathrm{T}}Z)]$, being ω a positive decreasing function and β a positive parameter.

The general case of different scatter matrices, however, is not considered in these references. In particular, the use of just any pair of robust scatter matrices in [24] does not guarantee the identification of the clusters, while the kurtosis has already proven to be effective in this situation. Also, the calculation of most robust matrices is computationally very expensive. For further details, we wrote a contribution to the discussion of Tyler et al. [24].

Under the same assumptions considered when calculating (4) plus normality for the components of the mixture, the matrix B in (2) is

$$B = pI + 2\mathbf{1}\mathbf{1}^{\mathrm{T}} + \sum_{i=1}^{k}\sum_{j=1}^{k}\gamma_{ij}\delta_i\delta_j^{\mathrm{T}}\mathbf{1}\mathbf{1}^{\mathrm{T}},$$

where

$$\gamma_{ij} = \begin{cases} (\pi_i - 3\pi_i^2)\delta_i^{\mathrm{T}}\delta_i & \text{if } i = j \\ -3\pi_i\pi_j\delta_i^{\mathrm{T}}\delta_j & \text{if } i \neq j. \end{cases}$$

Let $\Delta_{\mathbf{1}} = \langle \Delta, \mathbf{1} \rangle$ be the subspace spanned by the $\mathbf{1}$ and the δ_i's and suppose we are in the general case $\mathbf{1} \notin \Delta$ and $\mathbf{1} \not\perp \Delta$. If $u \in \Delta_{\mathbf{1}}^{\perp}$, $Bu = pu$ holds, and p is an eigenvalue of B with multiplicity $p - k$ associated to the eigenspace $\Delta_{\mathbf{1}}^{\perp}$. The remaining k eigenvectors are found in the $\Delta_{\mathbf{1}}$ subspace. When using the matrix K, the Δ subspace can be identified by selecting the eigenvectors with eigenvalues different from $p + 2$. Instead, if we were to use the matrix B, we could only isolate the $\Delta_{\mathbf{1}}$ subspace, which has a redundant non-informative dimension. The procedure thus becomes dependent on the position of the δ_i's with respect to the vector $\mathbf{1}$. This dependency is the reason why the matrix B is not invariant under affine transformations. In the two special cases where $\mathbf{1} \in \Delta$ or $\mathbf{1} \perp \Delta$, the Δ subspace can still be identified using eigenvectors of B. In effect, if $\mathbf{1} \in \Delta$ then we can choose $p - k + 1$ orthogonal eigenvectors from $\Delta^{\perp}$ with eigenvalues equal to p. And if $\mathbf{1} \perp \Delta$ then $\mathbf{1}$ is an eigenvector itself with eigenvalue $3p$, which also brings the total number of eigenvectors in $\Delta^{\perp}$ with known eigenvalues to $p - k + 1$. The remaining $k - 1$ eigenvectors are therefore an orthogonal basis of Δ.

3.2. *Consistency of the eigenvectors of the estimated matrix K_n*

Let $\mu_{r_1,\ldots,r_p} = E(\prod_{j=1}^{p} X_j^{r_j})$ be a k-order moment of X, $r_1 + \cdots + r_p = k$, then $\hat{\mu}_{r_1,\ldots,r_p}$ converges to $\mu_{r_1,\ldots,r_p}$ in probability and, since K is a continuous function of the moments, K_n converges to K in probability and therefore the matrix K_n is a consistent estimator of K. The spectral set of K, denoted Λ, is the set of all eigenvalues of K. The eigenspace of K associated with λ is $V(\lambda) = \{x \in \mathbb{R}^p \mid Kx = \lambda x\}$, whose dimension is the algebraic multiplicity of λ. Since K is symmetric, then $\mathbb{R}^p = \operatorname{span}\left(\sum_{\lambda\in\Lambda} V(\lambda)\right)$ holds. The eigenprojection of K associated with λ, denoted $P(\lambda)$, is the projection operator onto $V(\lambda)$ with respect to the decomposition of $\mathbb{R}^p$. If υ is any subset of the spectral set Λ, then the total eigenprojection for K associated with the eigenvalues in υ is defined to be $\sum_{\lambda\in\upsilon} P(\lambda)$. The following lemma [23] states that, for any subset υ of

Table 1
Two groups and equal scatter matrices. Angle between Fisher's direction and: 1. The direction (kurt) that maximizes $|\log(\kappa_d)-\log(3)|$ and 2. The eigenvector of K_n (*eig K*) whose eigenvalue maximizes $|\lambda_i-(p+2)|$.

p	kurt	*eig K*	kurt	*eig K*	kurt	*eig K*	kurt	*eig K*
4	16.03	35.39	10.10	21.45	6.91	15.08	3.64	8.01
8	16.03	36.44	12.93	21.74	6.88	18.15	4.36	7.52
15	11.25	42.86	8.96	25.92	14.82	19.61	9.60	10.28
30	24.99	50.30	12.41	26.37	8.32	19.95	4.77	8.70
	$n=100p$		$n=500p$		$n=1000p$		$n=5000p$	

eigenvalues of Λ, we can identify the corresponding subset υ_n (because of the relative position of the eigenvalues), and the subspace sum of subspaces span $\left(\sum_{\lambda\in\upsilon_n} V_n(\lambda)\right)$ will converge in probability to the subspace span $\left(\sum_{\lambda\in\upsilon} V(\lambda)\right)$. That is, the subspace generated by eigenvectors of K_n associated to an eigenvalue or a subset of them, is a consistent estimator for the subspace generated by eigenvectors of K associated to the corresponding eigenvalues υ.

Lemma 1. *Let K_n be a $p\times p$ symmetric matrix with eigenvalues $\lambda_1^n\geq\cdots\geq\lambda_p^n$. Let $P_{j,t}^n$ represent the subspace generated by the eigenvectors of K_n associated with $\lambda_j^n,\ldots,\lambda_t^n$ for $t\geq j$. If K_n converges to K in probability, then*

1. *λ_j^n converges to λ_j in probability,*
2. *$P_{j,t}^n$ converges to $P_{j,t}$ in probability, provided $\lambda_{j-1}\neq\lambda_j$ and $\lambda_t\neq\lambda_{t+1}$.*

The distance between two subspaces is measured using $\|P_1-P_2\|_2$, the matrix spectral norm, and the proof of the lemma can be found in Section VIII–Section 3.5 of [10]. A corollary of the lemma is that, when the scatter matrices are the same, the subspace orthogonal to the eigenspace associated to the eigenvalue equal to α of multiplicity q, is a consistent estimator for Fisher's subspace.

In order to study this convergence we will generate samples from mixtures of normal distributions with equal scatter matrices. Throughout the paper, the mixtures are generated as sets of $100p$ random observations, with dimensions $p=2,4,8,15,30$, from a mixture of k multivariate normal distributions. The number of observations in each population is determined randomly, but ensuring that each cluster contains a minimum of $p+1$ observations. The means for each normal distribution are chosen as values from a multivariate normal distribution $N_p(0,fI)$, for a factor f selected to be as small as possible whereas ensuring that the probability of overlapping between groups is roughly equal to 1%, see Table 1 in Peña and Prieto [19] for the values of f. The covariance matrices are generated as $S=UDU^{\mathrm{T}}$, using a random orthogonal matrix U and a diagonal matrix D with entries from a uniform distribution on $[10^{-3},5\sqrt{p}]$.

In Table 1 we deal with the case of a mixture of two normal distributions and present the angle between Fisher's discriminant function $V^{-1}(\mu_2-\mu_1)$ and the eigenvector of K_n associated to the eigenvalue that differs most from the value $p+2$. Also, we compare the results with the angle between Fisher's and the direction of kurtosis that maximizes $|\log(\kappa_d)-\log(3)|$ among the $2p$ considered in [19], where κ_d is the univariate kurtosis coefficient of the direction d.

The results for small sample sizes are better for the kurtosis directions due to the limited precision of the eigenvectors and therefore it is advised to use the optimization algorithm in these circumstances. However, the angles become more similar as the sample size increases, as expected.

We generate now mixtures of three normal distributions. In this case the subspace of interest is a plane and we want to measure how close Fisher's plane, and the plane generated by the two eigenvectors associated to eigenvalues that differ most from the value $p+2$, are. Again, in order to compare the results with the kurtosis directions, we will also consider the plane generated by the two directions that maximize $|\log(\kappa_{d_i})-\log(3)|$. When comparing directions, the angle between them was a convenient measure. As a measure of distance between subspaces we will compute the angle between two planes, which is defined as stated in Section 12.4.3 of [7]. Section 16.5 of [18] provides a geometrical interpretation of the angle. Let F and G be planes in $\mathbb{R}^p$, the angle between F and G is defined as the angle θ^* between u^* and v^*, the vectors that maximize $\cos\theta=u^{\mathrm{T}}v$, where $u\in F$ and $v\in G$, subject to $\|u\|=\|v\|=1$. Geometrically, u^* is collinear with the projection of v^* into F and v^* is collinear with the projection of u^* into G. In practice, to obtain θ^* we perform the singular value decomposition of $Q_F^{\mathrm{T}}Q_G$, where the columns of the $p\times 2$ matrices Q_F and Q_G define orthonormal bases for F and G respectively. The smallest singular value is the cosine of θ^*. The angles in Table 2 are calculated using this decomposition. This case is an example of the benefit of using the matrix K_n. For three groups we know that the optimal direction is a combination of the directions δ_1 and δ_2, the ones related to the cluster structure, but we cannot identify the directions that would define the best plane. Instead, the eigenvectors do identify the optimal subspace. The angles in both approaches are similar for small samples, but as soon as the sample size increases, the eigenvectors reduce the distance to Fisher's subspace, as expected from the results in Lemma 1, while the convergence of the directions has slower rates.

Another factor in consideration when comparing both approaches is related to the time needed for the kurtosis directions and the eigenvectors to be calculated. We did compute the running times for the p eigenvectors of K_n and the two extreme kurtosis directions. Their increase with n is similar for both approaches, slightly faster than linear. This agrees with the fact that the main effort affected by n is the computation of the kurtosis matrix and the evaluation of the kurtosis coefficient, respectively. Regarding increases in p, the matrix K_n presents a clear advantage, as the time ratios grow from values in the order of 4 for small dimensions to values in the order of 13–20 for the largest dimension under consideration ($p=30$).

Table 2
Three groups and equal scatter matrices. Angle between Fisher's plane and: 1. The plane generated by the directions (kurt) that maximize $|\log(\kappa_d) - \log(3)|$ and 2. The plane generated by the two eigenvectors of K_n (*eig K*) whose eigenvalues maximize $|\lambda_i - (p+2)|$.

p	kurt	*eig K*	kurt	*eig K*	kurt	*eig K*	kurt	*eig K*
4	44.90	44.53	37.76	26.75	30.68	19.03	33.47	10.21
8	43.55	51.28	39.69	27.66	31.34	20.47	25.71	12.77
15	51.62	56.05	42.65	35.94	42.10	30.78	35.86	16.54
30	62.79	63.76	45.59	41.80	40.63	33.12	35.94	19.47
	$n = 100p$		$n = 500p$		$n = 1000p$		$n = 5000p$	

This growth is associated with the use of Newton's method in the optimization of the kurtosis coefficient, and the need to factorize the corresponding second-derivative matrix in each iteration, as opposed to a single eigenvalue computation for the matrix K_n. In summary, the proposed algorithm seems to be computationally more efficient, particularly for the case of higher-dimensional data.

3.3. *Different scatter matrices*

In order to study the general case of different scatter matrices in a mixture of elliptical distributions, we will start by studying a perturbation of the simpler model, a mixture of two normal distributions with equal scatter matrices. We perturb the covariance matrix of one of the mixtures in order to see the effect that the relaxation in the condition of equal covariances causes in both the eigenvectors of K and the directions that optimize the kurtosis coefficient.

After standardization and using the same notation as in previous sections, the mixture is characterized as $\pi_1 N(\delta_1, W) + \pi_2 N(\delta_2, W + \Delta W)$, where ΔW is the perturbation added to the model. Consider now the equations that define the solutions for both approaches, an eigenvector of K and the optimum univariate kurtosis direction. For the kurtosis matrix, an eigenvector d is such that $Kd = \lambda d$, which in our case can be formulated as

$$(a_0 - \lambda)d + a_1 \Delta W d + a_2 \Delta W^2 d = -b_1 \delta_1 - b_2 \Delta W \delta_1. \tag{7}$$

For the kurtosis direction, the equivalent equation comes from $\nabla \kappa_d = \lambda d$, and reduces to

$$(c_0 - \lambda)d + c_1 \Delta W d = -f_1 \delta_1. \tag{8}$$

Details of the derivations are found in Appendix C.

When the scatter matrices are the same, the solution to both approaches is $d = c\delta_1$, for some constant c. Deviations from this solution appear as terms related to the perturbation such as ΔW and ΔW^2, the latter found only in (7). Consequently, in addition to ΔW, the eigenvectors of K differ from Fisher's discriminant function also in a quadratic term that does not arise in (8). Nevertheless, as we will see in simulation studies, the use of K is helpful when the sample size is not small, as in these cases the nonlinear algorithm for finding the optimal directions is time consuming and the results are similar to the ones obtained using K.

Moreover, this result gives hints on how one might modify the matrix K in order to improve the performance when the scatter matrices are different, which has not been addressed yet in the literature. Further research will we focus in finding a matrix that could manage to reduce the impact of the terms ΔW and ΔW^2.

4. Computational results

We perform a set of simulations to evaluate the properties of the eigenvectors of K_n for cluster analysis. The measure chosen to assess the performance is the proportion of total projected variance explained by the projected clusters, given by $\phi = d^{\mathrm{T}} B d / (d^{\mathrm{T}} \Sigma d)$, where $B = \pi_1 \pi_2 (\mu_2 - \mu_1)(\mu_2 - \mu_1)^{\mathrm{T}}$. The larger the gap between the projected means, the more separated the clusters are. Note that in the case when the scatter matrices are different other alternative measures may have better properties. In particular, and related to the ideas in [1], $d^{\mathrm{T}} B d$ could be replaced for example by $[(\mu_2 - \mu_1)^{\mathrm{T}} d - z_\alpha(\sqrt{d^{\mathrm{T}} V_1 d} + \sqrt{d^{\mathrm{T}} V_2 d})]^2$, a measure of the gap between the α quantiles of the distributions, (V_1 and V_2 are the covariance matrices of each population). In our simulation studies the variability is relatively small compared to the distance between means, and we use ϕ as our separation measure.

We are interested in the directions that return a large value of ϕ. It is well known that Fisher's direction $d = (\pi_1 V_1 + \pi_2 V_2)^{-1}(\mu_2 - \mu_1)$ satisfies the condition $\delta\phi/\delta d = 0$, maximizing ϕ. We generate random samples from a mixture of two p-variate normal populations, and estimate ϕ for the eigenvectors of K_n and for the directions of minimum and maximum kurtosis as follows. We assume we do not know the parameters of the two populations, although we know of the existence of the two clusters. For each eigenvector and direction, we need a procedure to assign the projected observations to clusters. Since the clustering in this case reduces to one dimension, the problem of finding the optimal assignment (the one that maximizes $\hat{\phi}$) reduces to determine n_1 such that $\hat{\phi} = d^{\mathrm{T}} \hat{B} d / [(n-1) d^{\mathrm{T}} S d]$ is maximized, where $\hat{B} = n_1 n_2 / (n_1 + n_2)(\bar{x}_2 - \bar{x}_1)(\bar{x}_2 - \bar{x}_1)^{\mathrm{T}}$ and $\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{(i)}$. There are exactly n different ways of assigning observations to the two clusters, since in a one-dimensional space the observations can be sorted. Also, in order to have an idea of how close we are to the optimum, we will include the value $\hat{\phi}$ for Fisher's direction.

Table 3
Two groups and equal scatter matrices. Proportion of variance explained by the clusters ($\hat{\phi}$) for the optimum direction (d. opt), the eigenvector of K_n associated with the max/min eigenvalue (max/min *eig K*), the max/min kurtosis direction (max/min kurt), the best eigenvector of K_n (best *eig K*) and the best kurtosis direction (best kurt).

p	n	d. opt	max/min *eig K*	max/min kurt	best *eig K*	best kurt
2	200	0.80	0.77	0.77	0.77	0.79
4	400	0.86	0.79	0.77	0.79	0.83
8	800	0.89	0.79	0.82	0.79	0.84
15	1 500	0.93	0.78	0.86	0.78	0.87
30	3 000	0.95	0.75	0.87	0.75	0.88
2	1 000	0.78	0.78	0.76	0.78	0.78
4	2 000	0.84	0.80	0.79	0.81	0.82
8	4 000	0.89	0.85	0.85	0.85	0.88
15	7 500	0.94	0.87	0.90	0.87	0.92
30	15 000	0.96	0.86	0.92	0.86	0.93
2	2 000	0.82	0.81	0.79	0.81	0.81
4	4 000	0.84	0.82	0.82	0.83	0.83
8	8 000	0.88	0.85	0.85	0.85	0.86
15	15 000	0.93	0.86	0.89	0.86	0.90
30	30 000	0.96	0.87	0.92	0.87	0.93
Average		0.88	0.82	0.84	0.82	0.86

Table 4
Two groups and equal scatter matrices. Percentage (%) of misclassified observations for the optimum direction (d. opt), the eigenvector of K_n associated with the max/min eigenvalue (max/min *eig K*), the max/min kurtosis direction (max/min kurt), the best eigenvector of K_n (best *eig K*) and the best kurtosis direction (best kurt).

p	n	d. opt	max/min *eig K*	max/min kurt	best *eig K*	best kurt
2	200	2.0	3.9	5.1	3.9	2.7
4	400	0.7	4.9	6.4	3.9	1.5
8	800	0.1	6.1	7.0	5.2	3.4
15	1 500	0.0	6.9	6.1	6.2	4.2
30	3 000	0.0	8.4	7.6	8.1	5.6
2	1 000	2.8	3.7	4.6	3.7	3.2
4	2 000	0.7	4.0	5.4	2.3	2.0
8	4 000	0.1	2.5	3.7	2.2	0.9
15	7 500	0.0	3.4	3.5	2.7	1.8
30	15 000	0.0	2.8	3.3	2.6	2.3
2	2 000	1.9	2.3	3.5	2.3	2.1
4	4 000	0.9	2.0	3.2	1.6	1.5
8	8 000	0.1	2.7	3.6	1.7	1.5
15	15 000	0.0	3.4	4.2	2.9	2.2
30	30 000	0.0	3.0	3.3	2.9	2.5
Average		0.6	4.0	4.7	3.5	2.5

4.1. *Proportional scatter matrices*

We start analyzing the results when the scatter matrices are the same. The mixtures are generated as stated above in Section 3.2. Table 3 presents the measure $\hat{\phi}$ for the optimum direction $V^{-1}(\mu_2 - \mu_1)$, the eigenvector of K_n ('max/min *eig K*') that maximizes $\hat{\phi}$ among the two eigenvectors corresponding to the maximum and minimum eigenvalue, the univariate kurtosis direction ('max/min kurt') that maximizes $\hat{\phi}$ among the maximum and minimum univariate kurtosis directions, the eigenvector of K_n ('best *eig K*') that maximizes $\hat{\phi}$ among the p existing eigenvectors and the univariate kurtosis direction ('best kurt') that maximizes $\hat{\phi}$ among the $2p$ directions considered in Peña and Prieto [19]. In Table 4 we present the proportion of misclassified observations after assigning them to clusters as stated above. Each value has been replicated 100 times.

When considering only two eigenvectors and two kurtosis directions, the results in the two tables are similar. We observe that the extreme eigenvector of K_n performs better when the dimension of the space is small (2, 4, 8), whereas the univariate kurtosis has better results when p is larger. We also observe that the values are very close to the optimum ones, indicating the appropriateness of the two methods. However, when all eigenvectors and kurtosis directions are considered, the results for the eigenvectors are very similar (column 'max/min *eig K*' and 'best *eig K*' are practically identical) whereas there is some improvement in the projected kurtosis directions, especially for large p. Note that, for a given p, the eigenvectors improve as n increases, while the kurtosis directions behave more stable in this sense. Also, if we count the number of times that the selected eigenvector in 'best *eig K*' does not correspond to one of the extreme eigenvalues, we obtain that this number is very small, specially when n is large. Thus we conclude that the maximum/minimum eigenvalue of the kurtosis matrix provides a useful direction for clustering which is very fast to compute. The computation of the matrix K and its eigenvectors is computationally very efficient, while the directions of kurtosis require an optimization algorithm and are computationally more expensive.

Table 5

Two groups and different scatter matrices. Proportion of variance explained by the clusters ($\hat{\phi}$) for the optimum direction (d. opt), the eigenvector of K_n associated with the max/min eigenvalue (max/min *eig K*), the max/min kurtosis direction (max/min kurt), the best eigenvector of K_n (best *eig K*) and the best kurtosis direction (best kurt).

p	n	d. opt	max/min *eig K*	max/min kurt	best *eig K*	best kurt
2	200	0.78	0.74	0.72	0.75	0.77
4	400	0.82	0.74	0.75	0.74	0.76
8	800	0.87	0.72	0.77	0.73	0.78
15	1 500	0.90	0.66	0.77	0.76	0.81
30	3 000	0.93	0.61	0.78	0.69	0.80
2	1 000	0.78	0.75	0.75	0.75	0.77
4	2 000	0.81	0.77	0.77	0.77	0.77
8	4 000	0.87	0.71	0.77	0.79	0.80
15	7 500	0.90	0.68	0.78	0.76	0.79
30	15 000	0.93	0.61	0.79	0.75	0.82
2	2 000	0.77	0.76	0.76	0.76	0.76
4	4 000	0.82	0.75	0.77	0.77	0.77
8	8 000	0.87	0.72	0.77	0.77	0.78
15	15 000	0.90	0.68	0.78	0.80	0.81
30	30 000	0.93	0.60	0.79	0.75	0.81
Average		0.86	0.70	0.77	0.75	0.79

Table 6

Two groups and different scatter matrices. Percentage (%) of misclassified observations for the optimum direction (d. opt), the eigenvector of K_n associated with the max/min eigenvalue (max/min *eig K*), the max/min kurtosis direction (max/min kurt), the best eigenvector of K_n (best *eig K*) and the best kurtosis direction (best kurt).

p	n	d. opt	max/min *eig K*	max/min kurt	best *eig K*	best kurt
2	200	3.20	6.00	8.00	5.30	4.30
4	400	1.10	7.00	7.00	4.00	3.90
8	800	0.30	8.00	8.00	5.00	3.30
15	1 500	0.10	9.00	8.00	4.40	5.20
30	3 000	0.00	1.10	8.00	6.10	5.50
2	1 000	2.80	5.00	6.00	4.80	3.50
4	2 000	1.30	5.00	5.00	4.90	4.10
8	4 000	0.30	7.00	8.00	3.80	3.50
15	7 500	0.10	9.00	8.00	3.30	5.10
30	15 000	0.00	1.10	8.00	3.90	5.00
2	2 000	3.30	5.00	5.00	5.00	4.00
4	4 000	0.90	6.00	5.00	4.40	3.70
8	8 000	0.30	7.00	6.00	4.00	3.10
15	15 000	0.10	9.00	8.00	2.60	4.70
30	30 000	0.00	1.20	1.00	3.80	5.40
Average		0.92	8.00	7.00	4.35	4.29

4.2. Different scatter matrices

In the general case of different scatter matrices, the optimum direction for ϕ is $(\pi_1 V_1+\pi_2 V_2)^{-1}(\mu_2-\mu_1)$. If we compare in Table 5 the columns 'best *eig K*' and 'best kurt' we observe that the kurtosis directions perform slightly better. However, if we look at the same columns in Table 6, the proportion of misclassified observations, the results are very similar. In particular, the eigenvectors perform better when the sample size is large. This behaviour could be due to the lack of precision in the eigenvectors when the sample size is small.

As for the column 'max/min *eig K*', we observe that the extreme eigenvalues not always identify the best eigenvector (in terms of $\hat{\phi}$) particularly when p, the number of eigenvalues, is large. In Table 7 we show the number of times the maximum or minimum eigenvalues correspond to the eigenvector of K_n that maximizes $\hat{\phi}$.

Also, instead of considering only the maximum and minimum eigenvalues, we can extend it to all eigenvalues that are significantly different from the rest, taking for example the median of all eigenvalues as a central measure, and the median absolute deviation as a measure of dispersion. If we do so, the values in Table 7 increase considerably.

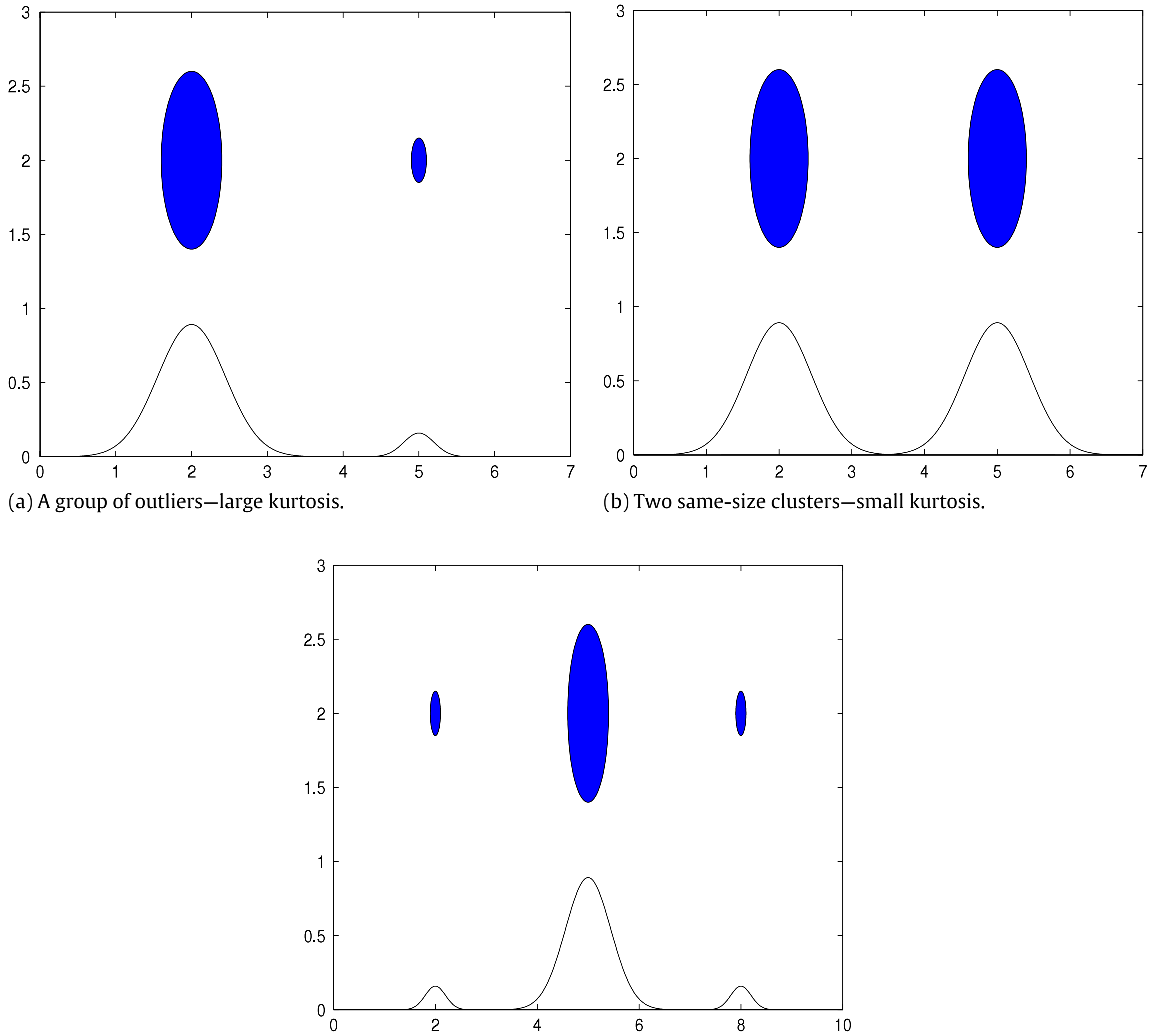
Alternatively, we can use other criteria to decide which eigenvectors to consider, such as a measure of the significative gaps between consecutive observations, as it is suggested in [1,19].

4.3. Comparison with some choices proposed by Tyler et al. [24]

As we have mentioned earlier, Tyler et al. [24] use the eigenvectors of $S_1^{-1}S_2$, S_1 and S_2 being any pair of affine equivariant scatter matrices, in order to find interesting views of the data. Among others, they suggest using as the second scatter matrix S_2 robust estimations of the covariance matrix, such as the Minimum Covariance Determinant (MCD) or the Minimum

Table 7
Two groups and different scatter matrices. Number of times out of 100 where the extreme eigenvalues (maximum or minimum) correspond to the eigenvector of K_n that maximizes $\hat{\phi}$.

p	n	
2	200	100
4	400	92
8	800	80
15	1 500	70
30	3 000	57
2	1 000	100
4	2 000	96
8	4 000	81
15	7 500	68
30	15 000	56
2	2 000	100
4	4 000	94
8	8 000	77
15	15 000	71
30	30 000	54



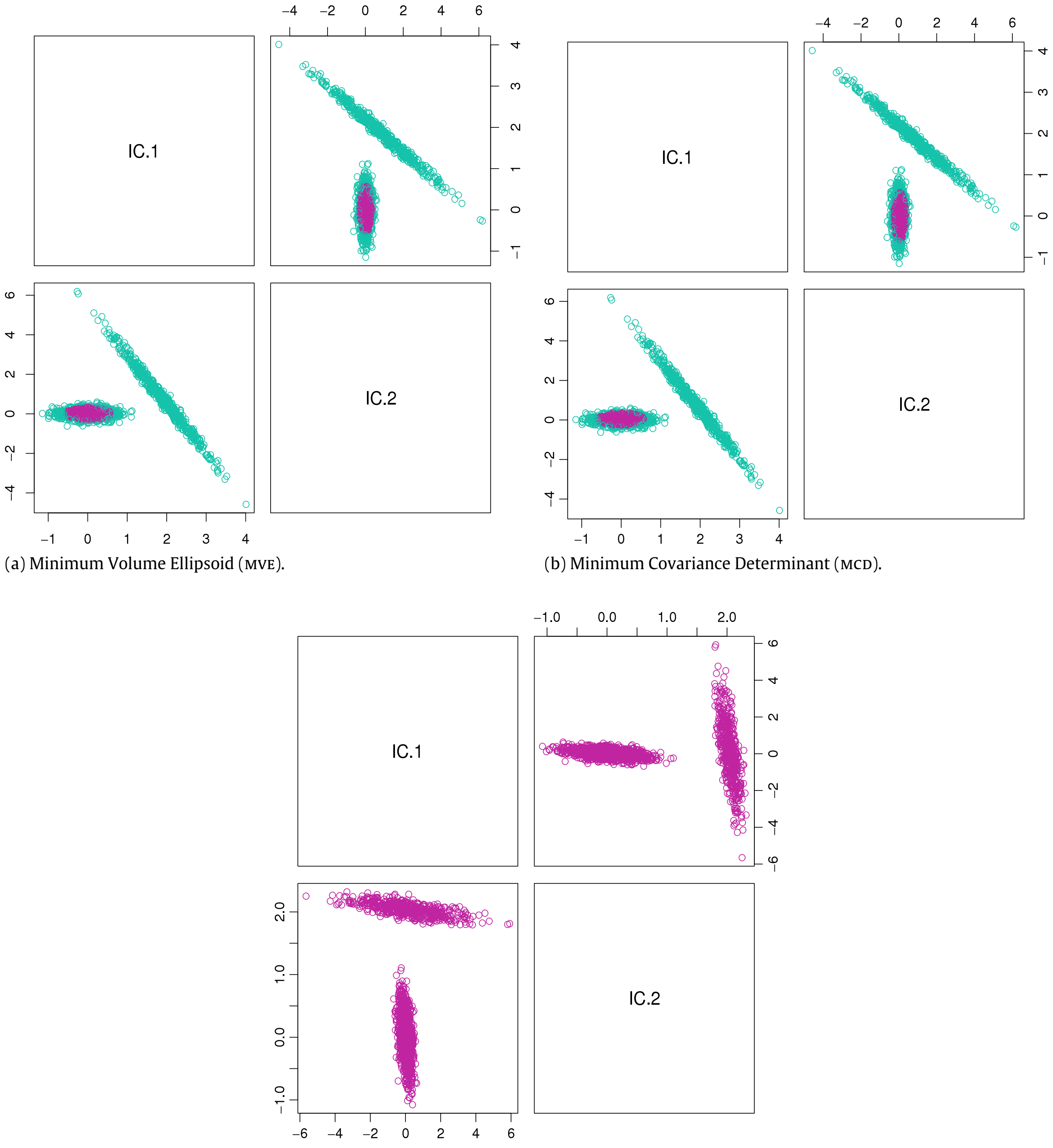
(a) A group of outliers—large kurtosis.

(b) Two same-size clusters—small kurtosis.

(c) Two groups of outliers—small kurtosis.

Fig. 1. The value of the univariate kurtosis coefficient for different scenarios.

Volume Ellipsoid (MVE), see [22]. Although in some cases these matrices perform well to identify clusters, the truth is that robust statistical theory assumes that the observed data are a mixture of good data (or clean data) and contamination, and generally the goal is to ignore the contaminated cases. When performing cluster analysis, we want the opposite: identify all the populations of the mixture. Therefore, it is easy to find examples where the eigenvectors of these robust matrices fail to identify the groups.



(a) Minimum Volume Ellipsoid (MVE).

(b) Minimum Covariance Determinant (MCD).

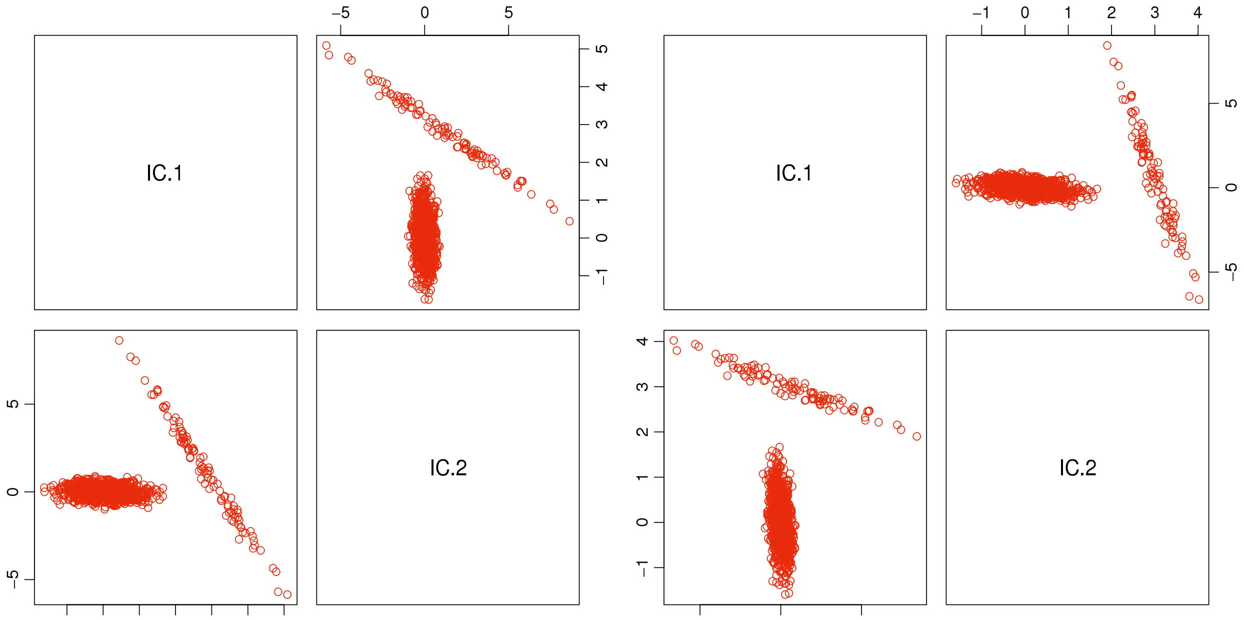
(c) Kurtosis.

Fig. 2. A standardized mixture of two bivariate normals with different scatter matrices projected onto the eigenvectors of the corresponding matrix $S_1^{-1}S_2$, with S_1 being the regular covariance matrix and S_2 being respectively the estimates MVE, MCD and the matrix based on fourth-order moments. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

On the contrary, the properties of the kurtosis are more suitable to identify clusters or outliers. Let $z_i = s^{-1}(x_i - \bar{x})$ be the scores of a univariate distribution, the variance of the squared scores is,

$$s_{z^2} = \frac{1}{n}\sum_{i=1}^{n}(z_i^2 - \bar{x}_{z^2})^2 = \frac{1}{n}\sum_{i=1}^{n} z_i^4 - 1 = k - 1,$$

where $\bar{x}_{z^2} = 1$ is the mean of the squared scores, and k the univariate kurtosis coefficient. The kurtosis can be seen as the variance of the distances between the observations and their mean and therefore as a measure of heterogeneity. Heterogeneity arises in several situations; if for example the sample is given by two clusters of similar size, the mean of the sample will be located in the middle of the two clusters and therefore the distances between the observations and the mean will be similar for all observations, specially if the clusters are well separated and their variances are small. Thus, the kurtosis will have a small value, reaching its minimum in the extreme case of a two point-mass distribution. The same happens under



(a) Caussinus and Ruiz-Gazen [3].

(b) Kurtosis.

Fig. 3. A standardized mixture of two bivariate normals with different scatter matrices projected onto the eigenvectors of the matrices proposed by Caussinus and Ruiz-Gazen [3] and kurtosis respectively.

the presence of three clusters, if the clusters in the extremes have the same size. On the other hand, if we have a sample where most of the observations come from a given distribution except for some outliers, the mean of the sample will be located near or in the larger cluster, and the distances between the outliers and the mean will be large compared to the other observations, returning a large value of kurtosis. Note that these properties do not depend on any assumption on the distribution of the elements of the mixture, and therefore is not restricted to normal (or elliptical) mixture types, (see [19, 21] for details). Fig. 1 illustrates these situations.

Suppose that the sample comes from a mixture of two bivariate normal distributions with parameters $\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\mu_2 = \begin{bmatrix} 4 \\ 4.5 \end{bmatrix}$, $V_1 = \begin{bmatrix} 0.55 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$ and $V_2 = \begin{bmatrix} 1 & -0.95 \\ -0.95 & 1 \end{bmatrix}$, and sample sizes $n_1 = 1000$ and $n_2 = 500$. We will compare the results obtained.

Fig. 2 shows the standardized sample projected onto the eigenvectors of the corresponding matrix $S_1^{-1}S_2$, where S_1 is the regular covariance matrix and S_2 is respectively the robust estimate MVE, the robust estimate MCD, and the scatter matrix based on fourth-order moments. The estimates MVE and MCD are calculated taking into account only half of the observations of the sample appropriately taken from the larger cluster (coloured purple in Fig. 2(a) and (b)). They are robust estimates so they manage to 'successfully' ignore the smaller cluster. Thus, if we project the observations onto any eigenvector of MVE or MCD, which are the principal axes of the larger cluster, the two clusters will overlap. In this case the good direction is $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$, the direction of the means, and the second eigenvector of the kurtosis matrix does capture it; the clusters do not overlap when projected onto it, see Fig. 2(c).

4.4. *Comparison with the method proposed by Caussinus and Ruiz-Gazen [3]*

The method proposed by Caussinus and Ruiz-Gazen [3] can also be seen as a particular case of the general setting described by Tyler et al. [24]. In this case, the choice for S_2 is the following weighted covariance matrix,

$$S_2 = \frac{E[\omega(\beta Z^{\mathrm{T}} Z)(X-\mu)(X-\mu)^{\mathrm{T}}]}{E[\omega(\beta Z^{\mathrm{T}} Z)]}, \tag{9}$$

with $\omega = \exp\left(\frac{-x}{2}\right)$ being a positive decreasing function. The choice proposed in our paper, using the notation in Tyler et al. [24], is $S_2 = E[Z^{\mathrm{T}} Z(X-\mu)(X-\mu)^{\mathrm{T}}]$, which can also be seen as a weighted covariance matrix. However, the weights in both cases are to be interpreted differently. For our choice, the larger the Mahalanobis distance of an observation to the mean, the larger the influence of the observation on the calculation of the eigenvectors. On the other hand, the weights in (9) are robust and give more influence to observations close to the mean, and the eigenvectors might ignore the directions of the clusters. In Fig. 3 we use the same example described above (now with $n_2 = 100$) to compare both approaches. Fig. 3(a) shows the overlapping of the two clusters when projected onto the second component, which is not the case in Fig. 3(b).

Although the proposed method performs quite well in practice, as shown in the preceding results, it should be noted that in some cases, for example when we have two groups with $\pi = (\sqrt{3}-1)/(2\sqrt{3}) = 0.2113$, the matrix K is diagonal (see expression (6)), and therefore the eigenvectors will not identify the direction of the means. In these cases the proposed algorithm can be complemented with additional directions, as mentioned in the introduction and along the lines proposed in [20].

Additionally, if we consider the example presented by Hennig in his contribution to the discussion of Tyler et al. [24], we observe that the kurtosis approach (Fig. 7) does not provide the optimal solution, whereas the approach suggested by Hennig (Fig. 9) performs better in terms of the separation between clusters. Nevertheless, in practice the differences are small and the kurtosis approach is still able to identify the three clusters from the projections onto the extreme eigenvectors. Also, the computational cost for the kurtosis approach is quite low compared to that for the scatter matrices suggested in the discussion, which require computing a matrix of Mahalanobis distances and a covariance matrix for each observation.

5. Discussion

Further research will be focused on modifying the kurtosis matrix to improve the performance when the scatter matrices are different, which has not been addressed yet in the literature. In particular, it would be interesting to explore variations of the kurtosis matrix where the terms in (3) that depend on the scatter matrices W_i have less influence on the eigenstructure of the matrix.

Appendix A. Matrix *K* under a mixture of elliptical distributions

We standardize X using its global mean $\mu = \sum_i \pi_i \mu_i$, and covariance matrix $\Sigma = \sum_i \pi_i V_i + \sum_i \pi_i (\mu_i - \mu)(\mu_i - \mu)^{\mathrm{T}}$. The standardized variable $Z = \Sigma^{-1/2}(X - \mu)$ is also a mixture of elliptical distributions Z_i with means and scatter matrices δ_i and W_i, $\delta_i = \Sigma^{-1/2}(\mu_i - \mu)$ and $W_i = \Sigma^{-1/2} V_i \Sigma^{-1/2}$. Using expectation properties

$$K = E(Z^{\mathrm{T}} Z Z Z^{\mathrm{T}}) = \sum_{i=1}^{k} \pi_i E(Z_i^{\mathrm{T}} Z_i Z_i Z_i^{\mathrm{T}}).$$

The fourth-order moment matrix is

$$E(Z_i^{\mathrm{T}} Z_i Z_i Z_i^{\mathrm{T}}) = E\{(Z_i - \delta_i)^{\mathrm{T}}(Z_i - \delta_i)(Z_i - \delta_i)(Z_i - \delta_i)^{\mathrm{T}}\} + \operatorname{tr} W_i \delta_i \delta_i^{\mathrm{T}} + \delta_i^{\mathrm{T}} \delta_i W_i + 2(\delta_i \delta_i^{\mathrm{T}} W_i + W_i \delta_i \delta_i^{\mathrm{T}}) + \delta_i^{\mathrm{T}} \delta_i \delta_i \delta_i^{\mathrm{T}},$$

where we have used that $Z_i = W_i^{1/2} Y + \delta_i$, with Y following a spherical distribution, the intermediate results $E(Z_i Z_i^{\mathrm{T}}) = W_i + \delta_i \delta_i^{\mathrm{T}}$, $E(Y^{\mathrm{T}} W_i Y) = \operatorname{tr} W_i$, $E(\delta_i^{\mathrm{T}} W_i^{1/2} Y W_i^{1/2} Y \delta_i^{\mathrm{T}}) = E(W_i^{1/2} Y Y^{\mathrm{T}} W_i^{1/2} \delta_i \delta_i^{\mathrm{T}})$ and the fact that all odd moments of Y are equal to zero.

The fourth-order central moment matrix of Z_i is

$$\begin{aligned}
M_4 &= E\{(Z_i - \delta_i)^{\mathrm{T}}(Z_i - \delta_i)(Z_i - \delta_i)(Z_i - \delta_i)^{\mathrm{T}}\} \\
&= |W_i|^{-1/2} \int (z - \delta_i)^{\mathrm{T}}(z - \delta_i)(z - \delta_i)(z - \delta_i)^{\mathrm{T}} h_i((z - \delta_i)^{\mathrm{T}} W_i^{-1}(z - \delta_i)) \mathrm{d}z \\
&= \int y^{\mathrm{T}} W_i y W_i^{1/2} y y^{\mathrm{T}} W_i^{1/2} h_i(y^{\mathrm{T}} y) \mathrm{d}y = W_i^{1/2} U \int t^{\mathrm{T}} \Omega t t t^{\mathrm{T}} h_i(t^{\mathrm{T}} t) \mathrm{d}t U^{\mathrm{T}} W_i^{1/2} \\
&= W_i^{1/2} U \sum_j \omega_j \int t_j^2 t t^{\mathrm{T}} h_i(t^{\mathrm{T}} t) \mathrm{d}t U^{\mathrm{T}} W_i^{1/2} = \sum_j \omega_j \tilde{k}_i W_i + \bar{k}_i W_i^{1/2} U \Omega U^{\mathrm{T}} W_i^{1/2} \\
&= \tilde{k}_i \operatorname{tr} W_i W_i + \bar{k}_i W_i^2,
\end{aligned}$$

where we have introduced $y = W_i^{-1/2}(z - \delta_i)$, $t = U^{\mathrm{T}} y$ and $\int t_j^2 t t^{\mathrm{T}} h_i(t^{\mathrm{T}} t) \mathrm{d}t = \tilde{k}_i I + \bar{k}_i e_j e_j^{\mathrm{T}}$ for $\tilde{k}_i = \int t_j^2 t_k^2 h_i(t^{\mathrm{T}} t) \mathrm{d}t$ where $j \neq k$, and $\bar{k}_i = \int t_j^4 h_i(t^{\mathrm{T}} t) \mathrm{d}t - \tilde{k}_i$. Thus, K reduces to (3).

Appendix B. Proof of Theorem 1

Proof of Theorem 1. The result in (4) is obtained using in expression (3) the result $W_i = \Sigma^{-1/2} V \Sigma^{-1/2} = I - \sum_i \pi_i \delta_i \delta_i^{\mathrm{T}}$, where $V = \Sigma - \sum_i \pi_i (\mu_i - \mu)(\mu_i - \mu)^{\mathrm{T}}$.

Denote $\Sigma = V + MPM^{\mathrm{T}}$, with $M = (\mu_1 - \mu, \ldots, \mu_k - \mu)$ and P diagonal with elements $(\pi_1, \ldots, \pi_k)$, then, from the inverse of the sum property, we have $\Sigma^{-1} = V^{-1} - V^{-1} M (M^{\mathrm{T}} V^{-1} M + P^{-1})^{-1} M^{\mathrm{T}} V^{-1}$, and multiplying by M,

$$\Sigma^{-1} M = V^{-1} M \left\{ I - (M^{\mathrm{T}} V^{-1} M + P^{-1})^{-1} M^{\mathrm{T}} V^{-1} M \right\}.$$

Therefore, $\Sigma^{-1} M = V^{-1} M T$. And, if we add and subtract P^{-1} appropriately, we can see that $T = \{P(M^{\mathrm{T}} V^{-1} M + P^{-1})\}^{-1}$ is invertible. Therefore, the columns of $\Sigma^{-1} M$ and $V^{-1} M$ generate the same subspace and thus $\Phi = \Delta_X$ and (5) is proven. □

Appendix C. Derivations for the case of different scatters

We have that $\delta_2 = -\frac{\pi_1}{\pi_2}\delta_1$ and, from the decomposition of the covariance matrix in the case of mixture distributions, $I = \pi_1 W + \pi_2 W + \pi_2 \Delta W + \sum_i \pi_i \delta_i \delta_i^{\mathrm{T}}$, and thus $W = \bar{W} - \pi_2 \Delta W$, where $\bar{W} = I - \frac{\pi_1}{\pi_2}\delta_1\delta_1^{\mathrm{T}}$ corresponds to the equal scatter matrices case. Also $W + \Delta W = \bar{W} + \pi_1 \Delta W$. Replacing $W_1 = W = \bar{W} - \pi_2 \Delta W$ and $W_2 = W + \Delta W = \bar{W} + \pi_1 \Delta W$ in (3) we obtain

$$
\begin{aligned}
K = {} & \bar{K} + \pi_1\pi_2 \Delta W \operatorname{tr} \Delta W + 2\pi_1\pi_2 \Delta W^2 + \frac{\pi_1}{\pi_2}(\pi_1 - \pi_2)\delta_1\delta_1^{\mathrm{T}} \operatorname{tr} \Delta W \\
& + 2\frac{\pi_1}{\pi_2}(\pi_1 - \pi_2)\left(\delta_1\delta_1^{\mathrm{T}}\Delta W + \Delta W \delta_1\delta_1^{\mathrm{T}}\right) + \frac{\pi_1}{\pi_2}(\pi_1 - \pi_2)\delta_1^{\mathrm{T}}\delta_1 \Delta W,
\end{aligned}
$$

where $\bar{K} = (p + 2)I + \frac{\pi_1}{\pi_2^3}(1 - 6\pi_1\pi_2)\delta_1^{\mathrm{T}}\delta_1\delta_1\delta_1^{\mathrm{T}}$. The kurtosis coefficient on a direction is $\kappa_d = 3\sum_i \pi_i (d^{\mathrm{T}} W_i d)^2 + 6\sum_i \pi_i (d^{\mathrm{T}} W_i d)(\delta_i^{\mathrm{T}} d)^2 + \sum_i \pi_i (\delta_i^{\mathrm{T}} d)^4$, and substituting in our case

$$
\kappa_d = \bar{\kappa}_d + 3\pi_1\pi_2 (d^{\mathrm{T}} \Delta W d)^2 + 6\frac{\pi_1}{\pi_2}(\pi_1 - \pi_2)(\delta_1^{\mathrm{T}} d)^2 d^{\mathrm{T}} \Delta W d,
$$

where $\bar{\kappa}_d = 3(d^{\mathrm{T}} d)^2 + \frac{\pi_1}{\pi_2^3}(1 - 6\pi_1\pi_2)(\delta_1^{\mathrm{T}} d)^4$. The parameters in Eqs. (7) and (8) derived from these results are $a_0 = p + 2$, $a_1 = \pi_1\pi_2 \operatorname{tr} \Delta W + \frac{\pi_1}{\pi_2}(\pi_1 - \pi_2)\delta_1^{\mathrm{T}}\delta_1$, $a_2 = 2\pi_1\pi_2$, $b_1 = \frac{\pi_1}{\pi_2^3}(1 - 6\pi_1\pi_2)\delta_1^{\mathrm{T}}\delta_1\delta_1^{\mathrm{T}} d + \frac{\pi_1}{\pi_2}(\pi_1 - \pi_2)(\delta_1^{\mathrm{T}} d \operatorname{tr} \Delta W + 2\delta_1^{\mathrm{T}} \Delta W d)$, $b_2 = 2\frac{\pi_1}{\pi_2}(\pi_1 - \pi_2)\delta_1^{\mathrm{T}} d$, $c_0 = 12$, $c_1 = 12\pi_1\pi_2 d^{\mathrm{T}} \Delta W d + 12\frac{\pi_1}{\pi_2}(\pi_1 - \pi_2)(\delta_1^{\mathrm{T}} d)^2$ and $f_1 = 4\frac{\pi_1}{\pi_2^3}(1 - 6\pi_1\pi_2)(\delta_1^{\mathrm{T}} d)^3 + 12\frac{\pi_1}{\pi_2}(\pi_1 - \pi_2)\delta_1^{\mathrm{T}} d d^{\mathrm{T}} \Delta W d$.

References

[1] T.W. Anderson, R.R. Bahadur, Classification into two multivariate normal distributions with different covariance matrices, Ann. Math. Statist. 33 (1962) 420–431.

[2] J.F. Cardoso, Source separation using higher order moments, in: Proc. ICASSP, vol. 4, 1989, pp. 2109–2112.

[3] H. Caussinus, A. Ruiz-Gazen, Projection pursuit and generalized principal component analysis, in: S. Morgenthaler, E. Ronchetti, W. Stahel (Eds.), New Directions in Statistical Data Analysis and Robustness, Birkhäuser Verlag, Basel, 1993, pp. 35–46.

[4] H. Caussinus, A. Ruiz-Gazen, Metrics for finding typical structures by means of principal component analysis, in: Y. Escoufier, C. Hayashi (Eds.), Data Science and its Applications, Academy Press, Tokyo, 1995, pp. 177–192.

[5] J.H. Friedman, Exploratory projection pursuit, J. Amer. Statist. Assoc. 82 (1987) 249–266.

[6] J.H. Friedman, J.W. Tukey, A projection pursuit algorithm for exploratory data analysis, IEEE Trans. Comput. C-23 (1974) 881–889.

[7] G.H. Golub, C.F. van Loan, Matrix Computations, The Johns Hopkins University Press, 1996.

[8] P.J. Huber, Projection pursuit, Ann. Statist. 13 (1985) 435–475.

[9] M.C. Jones, R. Sibson, What is projection pursuit? J. R. Stat. Soc. 150 (1987) 1–37.

[10] T. Kato, Perturbation Theory for Linear Operators, Springer, Berlin, 1980.

[11] T. Kollo, Multivariate skewness and kurtosis measures with an application in ICA, J. Multivariate Anal. 99 (2008) 2328–2338.

[12] J.A. Koziol, A note on measures of multivariate kurtosis, Biom. J. 31 (1989) 619–624.

[13] J.F. Malkovich, A.A. Afifi, On tests for multivariate normality, J. Amer. Statist. Assoc. 68 (1973) 176–179.

[14] K.V. Mardia, Measures of multivariate skewness and kurtosis with applications, Biometrika 57 (1970) 519–530.

[15] E.C. McRae, Matrix derivatives with an application to an adaptive linear decision problem, Ann. Statist. 2 (1974) 337–346.

[16] T.F. Móri, V.K. Rohatgi, G.J. Székely, On multivariate skewness and kurtosis, Theory Probab. Appl. 38 (1993) 547–551.

[17] H. Oja, Descriptive statistics for multivariate distributions, Statist. Probab. Lett. 1 (1983) 327–332.

[18] D. Peña, Análisis de datos multivariantes, McGraw-Hill, 2002.

[19] D. Peña, F.J. Prieto, Cluster identification using projections, J. Amer. Statist. Assoc. 96 (2001) 1433–1445.

[20] D. Peña, F.J. Prieto, Combining random and specific directions for outlier detection and robust estimation of high-dimensional multivariate data, J. Comput. Graph. Statist. 16 (2007) 228–254.

[21] D. Peña, F.J. Prieto, Robust covariance matrix estimation and multivariate outlier detection (with discussion), Technometrics 43 (2001) 286–310.

[22] P.J. Rousseeuw, Multivariate estimation with high breakdown point, in: W. Grossman, G. Pflug, I. Vincze, W. Wertz (Eds.), Mathematical Statistics and Applications, Reidel, Dordrecht, 1986, pp. 283–297.

[23] D.E. Tyler, Asymptotic inference for eigenvectors, Ann. Statist. 9 (1981) 725–736.

[24] D.E. Tyler, F. Critchley, L. Dümbgen, H. Oja, Invariate co-ordinate selection (with discussion), J. R. Stat. Soc. Ser. B 71 (3) (2009) 1–27.