TEST (2024) 33:453–472
https://doi.org/10.1007/s11749-023-00902-w

ORIGINAL PAPER

Tensor eigenvectors for projection pursuit

Nicola Loperfido[1]

Received: 1 May 2023 / Accepted: 1 November 2023 / Published online: 11 December 2023

Abstract
Tensor eigenvectors naturally generalize matrix eigenvectors to multi-way arrays: eigenvectors of symmetric tensors of order k and dimension p are stationary points of polynomials of degree k in p variables on the unit sphere. Dominant eigenvectors of symmetric tensors maximize polynomials in several variables on the unit sphere, while base eigenvectors are roots of polynomials in several variables. In this paper, we focus on skewness-based projection pursuit and on third-order tensor eigenvectors, which provide the simplest, yet relevant connections between tensor eigenvectors and projection pursuit. Skewness-based projection pursuit finds interesting data projections using the dominant eigenvector of the sample third standardized cumulant to maximize skewness. Skewness-based projection pursuit also uses base eigenvectors of the sample third cumulant to remove skewness and facilitate the search for interesting data features other than skewness. Our contribution to the literature on tensor eigenvectors and on projection pursuit is twofold. Firstly, we show how skewness-based projection pursuit might be helpful in sequential cluster detection. Secondly, we show some asymptotic results regarding both dominant and base tensor eigenvectors of sample third cumulants. The practical relevance of the theoretical results is assessed with six well-known data sets.

Keywords Asymptotics · Data reduction · Model-based clustering · Skewness · Symmetrization · Tensor eigenpairs

Mathematics Subject Classification 15A69 · 58C40 · 62E20 · 62H05

1 Introduction

Skewness-based projection pursuit looks for interesting data projections by means of skewness maximization, where the skewness of a data projection is measured by its

✉ Nicola Loperfido
nicola.loperfido@uniurb.it

[1] Dipartimento di Economia, Società e Politica, Università degli Studi di Urbino Carlo Bo, Via Saffi 42, 61029 Urbino, PU, Italy

third standardized moment. Skewness maximization is often paired with skewness removal, to ease the search for interesting structures. Skewness-based projection pursuit has been used in normality testing (Malkovich and Afifi 1973), point estimation (Loperfido 2010), cluster analysis (Loperfido 2019) and stochastic ordering (Arevalillo and Navarro 2019).

There has been a renewed interest in skewness-based projection pursuit, with focus on its parametric interpretation when the sampled distribution is either a finite mixture (Loperfido 2013, 2015, 2019), a skew-normal (Loperfido 2010; Balakrishnan and Scarpa 2012; Tarpey and Loperfido 2015), or a scale mixture of skew-normal distributions (Kim and Kim 2017; Arevalillo and Navarro 2015, 2020, 2021a, b). Loperfido (2018) used a generalized skew-normal distribution to illustrate the connection between skewness maximization and tensor eigenvectors.

A tensor is symmetric if it remains unchanged when permuting its subscripts. Its dimension is the number of distinct values that a subscript can take. The third moment $\mathcal{M}_{3,\mathbf{x}} = \left\{\mathrm{E}\left(X_i X_j X_k\right)\right\} \in \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^p$ of a p-dimensional random vector $\mathbf{x} = \left(X_1, ..., X_p\right)^\top$ satisfying $\mathrm{E}\left(\left|X_i^3\right|\right) < \infty$ for $i \in \{1, ..., p\}$ is a symmetric third order tensor with dimension p. The third cumulant $\mathcal{K}_{3,\mathbf{x}}$ of $\mathbf{x}$ is the third moment of $\mathbf{x} - \boldsymbol{\mu}$, where $\boldsymbol{\mu}$ is the mean of $\mathbf{x}$. The third standardized moment $\mathcal{M}_{3,\mathbf{z}}$ of $\mathbf{x}$ is the third moment of $\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$, where $\boldsymbol{\Sigma}$ is the positive definite covariance matrix of $\mathbf{x}$.

Tensor unfolding is the process which rearranges the tensor's elements into a matrix according to the index which is most meaningful for the problem at hand. Each row of the resulting matrix contains the tensor elements identified by the same value of the unfolding index. Within each row, tensor's elements are arranged beginning with those identified by smallest values of the first other indices. More formally, let $\mathbf{A}_{(u)}$ be the matrix whose $i-$th row contains all elements of the tensor $\mathcal{A}$ with the i-th value of the u-th index, while the elements of $\mathbf{A}_{(u)}$ in the same row are ordered according to the reflected lexicographic ordering of their indices. For example, the third-order tensor $\mathcal{A} = \left\{a_{ijk}\right\} \in \mathbb{R}^3 \times \mathbb{R}^4 \times \mathbb{R}^2$ can be unfolded in three different ways, to obtain the matrices $\mathbf{A}_{(1)}$, $\mathbf{A}_{(2)}$ and $\mathbf{A}_{(3)}$. They are represented below, with the index of the unfolding mode in bold and the other indices in smaller font, to emphasize the different unfoldings:

$$\mathbf{A}_{(1)} = \begin{pmatrix} a_{\mathbf{1}1_1} & a_{\mathbf{1}2_1} & a_{\mathbf{1}3_1} & a_{\mathbf{1}4_1} & a_{\mathbf{1}1_2} & a_{\mathbf{1}2_2} & a_{\mathbf{1}3_2} & a_{\mathbf{1}4_2} \\ a_{\mathbf{2}1_1} & a_{\mathbf{2}2_1} & a_{\mathbf{2}3_1} & a_{\mathbf{2}4_1} & a_{\mathbf{2}1_2} & a_{\mathbf{2}2_2} & a_{\mathbf{2}3_2} & a_{\mathbf{2}4_2} \\ a_{\mathbf{3}1_1} & a_{\mathbf{3}2_1} & a_{\mathbf{3}3_1} & a_{\mathbf{3}4_1} & a_{\mathbf{3}1_2} & a_{\mathbf{3}2_2} & a_{\mathbf{3}3_2} & a_{\mathbf{3}4_2} \end{pmatrix},$$

$$\mathbf{A}_{(2)} = \begin{pmatrix} a_{1\mathbf{1}_1} & a_{2\mathbf{1}_1} & a_{3\mathbf{1}_1} & a_{1\mathbf{1}_2} & a_{2\mathbf{1}_2} & a_{3\mathbf{1}_2} \\ a_{1\mathbf{2}_1} & a_{2\mathbf{2}_1} & a_{3\mathbf{2}_1} & a_{1\mathbf{2}_2} & a_{2\mathbf{2}_2} & a_{3\mathbf{2}_2} \\ a_{1\mathbf{3}_1} & a_{2\mathbf{3}_1} & a_{3\mathbf{3}_1} & a_{1\mathbf{3}_2} & a_{2\mathbf{3}_2} & a_{3\mathbf{3}_2} \\ a_{1\mathbf{4}_1} & a_{2\mathbf{4}_1} & a_{3\mathbf{4}_1} & a_{1\mathbf{4}_2} & a_{2\mathbf{4}_2} & a_{3\mathbf{4}_2} \end{pmatrix}$$

$$\text{and } \mathbf{A}_{(3)} = \begin{pmatrix} a_{11\mathbf{1}_1} & a_{21\mathbf{1}_1} & a_{31\mathbf{1}_1} & a_{12\mathbf{1}_1} & a_{22\mathbf{1}_1} & a_{32\mathbf{1}_1} & a_{13\mathbf{1}_1} & a_{23\mathbf{1}_1} & a_{33\mathbf{1}_1} & a_{14\mathbf{1}_1} & a_{24\mathbf{1}_1} & a_{34\mathbf{1}_1} \\ a_{11\mathbf{2}_1} & a_{21\mathbf{2}_1} & a_{31\mathbf{2}_1} & a_{12\mathbf{2}_1} & a_{22\mathbf{2}_1} & a_{32\mathbf{2}_1} & a_{13\mathbf{2}_1} & a_{23\mathbf{2}_1} & a_{33\mathbf{2}_1} & a_{14\mathbf{2}_1} & a_{24\mathbf{2}_1} & a_{34\mathbf{2}_1} \end{pmatrix}.$$

The unfolding of a symmetric tensor does not depend on the unfolding index. We therefore denote with $\mathbf{A}$ the unfolding of the symmetric tensor $\mathcal{A}$, without men-

tioning the unfolding index. The coskewness of a p -dimensional random vector $\mathbf{x}$ with finite third moments and mean $\boldsymbol{\mu}$ is the unfolding of the third cumulant of $\mathbf{x}$: $\boldsymbol{\Gamma} = \mathrm{E}\left\{(\mathbf{x}-\boldsymbol{\mu}) \otimes (\mathbf{x}-\boldsymbol{\mu})^\top \otimes (\mathbf{x}-\boldsymbol{\mu})^\top\right\} \in \mathbb{R}^p \times \mathbb{R}^{p^2}$, where"$\otimes$" denotes the Kronecker product. Similarly, the standardized coskewness of $\mathbf{x}$ is the unfolding of the third standardized cumulant of $\mathbf{x}$: $\boldsymbol{\Pi} = \mathrm{E}\left(\mathbf{z} \otimes \mathbf{z}^\top \otimes \mathbf{z}^\top\right) \in \mathbb{R}^p \times \mathbb{R}^{p^2}$.

There are other ways to denote and arrange multivariate moments and cumulants (De Luca and Loperfido 2015; Doss et al. 2023; Rao Jammalamadaka et al. 2021; Pereira et al. 2022). In this paper, we favor the coskewness due to its close connection with the eigenpairs of third-order tensors.

Consider now the problem of finding the stationary points of a homogeneous polynomial of degree k in p variables, under the constraint that the squared sum of the variables themselves is one. When k equals 2 the polynomial is a quadratic form and the problem reduces to the derivation of the eigenpairs of the symmetric matrix which characterizes the polynomial itself. Eigenvalues and eigenvectors of symmetric tensors generalize eigenvectors and eigenvalues of symmetric matrices to polynomials of degree greater than 2.

More formally, let $\mathcal{A}$ be a symmetric tensor of order k and dimension p. Also, let $\mathbf{A}$ be the matrix obtained by unfolding $\mathcal{A}$ along one of its modes. A scalar λ and a p -dimensional, nonnull vector $\mathbf{x}$ are an eigenvalue and the corresponding eigenvector of $\mathcal{A}$ if they satisfy $\mathbf{A}\mathbf{x}^{\otimes(k-1)} = \lambda\mathbf{x}$, where $\mathbf{x}^{\otimes(k-1)}$ denotes the product $\mathbf{x} \otimes \cdots \otimes \mathbf{x}$, in which the symbol "$\otimes$" appears $k-1$ times. In particular, if $\mathcal{A}$ is a third-order tensor, λ and $\mathbf{v}$ satisfy $\mathbf{A}\,(\,\mathbf{x} \otimes \mathbf{x}) = \lambda\mathbf{x}$. The eigenvectors of a tensor are the stationary points of the homogeneous polynomial uniquely associated to the tensor itself.

Lim (2005) and Qi (2005) independently introduced tensor eigenvalues and tensor eigenvectors. Sturmfels (2016) thoroughly reviews the topic and states some open problems. Eigenvalues and eigenvectors are defined for any real tensor, including the asymmetric ones. In such cases, however, tensor eigenvalues and eigenvectors depend on the choice of the unfolding index and may not be real. Moreover, such cases are not directly connected to skewness-based projection pursuit and are therefore ignored in the rest of the paper.

The tensor eigenvalue with the greatest norm is the dominant tensor eigenvalue, while the associated tensor eigenvector of unit length is the dominant tensor eigenvector. The constraint on the eigenvector's norm is necessary because if $\mathcal{A}$ is a symmetric tensor of order k and λ is an eigenvalue of $\mathcal{A}$ then λc^{k-2} is the eigenvalue of $\mathcal{A}$ associated with the eigenvector $c\mathbf{x}$, where c is a nonnull scalar. Clearly, this constraint is not necessary for ordinary matrix eigenvectors and for base eigenvectors, that is tensor eigenvectors associated with null eigenvalues. As an example, let $\mathcal{A} = \{a_{ijk}\}$ be a tensor of order 3 and dimension 3 such that a_{ijk} equals one when the indices i, j and k differ from each other, and zero otherwise. Its unfolding is

$$\mathbf{A} = \begin{pmatrix} 0&0&0&0&0&1&0&1&0 \\ 0&0&1&0&0&0&1&0&0 \\ 0&1&0&1&0&0&0&0&0 \end{pmatrix}.$$

As shown in Loperfido (2018), the dominant eigenvector and the dominant eigenvalue of $\mathcal{A}$ are

$$\mathbf{v} = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \text{ and } \lambda = \frac{2}{\sqrt{3}}.$$

Other, nondominant eigenvectors are proportional to one of the following vectors:

$$\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}.$$

Base eigenvectors, that is tensor eigenvectors associated with null tensor eigenvalues, are proportional to one of the following vectors:

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

The connection between dominant tensor eigenpairs and skewness-based projection pursuit becomes apparent when considering the directional skewness of a random vector, that is the maximal skewness achievable by a linear projection of the random vector itself:

$$\gamma_D(\mathbf{x}) = \max_{\mathbf{a} \in \mathbb{S}^{p-1}} \frac{\mathrm{E}\left\{\left(\mathbf{a}^\top \mathbf{x} - \mathbf{a}^\top \boldsymbol{\mu}\right)^3\right\}}{\left(\mathbf{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}\right)^{3/2}},$$

where $\mathbb{S}^{p-1}$ is the p-dimensional unit hypersphere. As shown in Section 3 of Loperfido (2018), the projection achieving maximal skewness is an affine function of $\mathbf{v}^\top \boldsymbol{\Sigma}^{-1/2}\mathbf{x}$, where $\mathbf{v}$ is the dominant eigenvector of the third standardized cumulant $\mathcal{M}_{3,\mathbf{z}}$ of $\mathbf{x}$, while the skewness of $\mathbf{v}^\top \boldsymbol{\Sigma}^{-1/2}\mathbf{x}$ is the dominant tensor eigenvalue λ of $\mathcal{M}_{3,\mathbf{z}}$: $\boldsymbol{\Pi}(\mathbf{v} \otimes \mathbf{v}) = \lambda \mathbf{v}$. On the other hand, the third cumulant of $\mathbf{u}^\top \mathbf{x}$ is zero, if $\mathbf{u}$ is base eigenvector of the third cumulant of $\mathbf{x}$: $\boldsymbol{\Gamma}(\mathbf{u} \otimes \mathbf{u}) = \mathbf{0}_p$, where $\mathbf{0}_p$ is the p-dimensional null vector.

The present paper contributes to the literature on projection pursuit by using tensor concepts to investigate the statistical properties of skewness maximization related to model-based clustering and large sample inference. The results in the paper support a tensor approach to projection pursuit both in the exploratory and the inferential steps of the statistical analysis. The paper is interdisciplinary in nature, since it bridges tensor algebra and projection pursuit. The rest of the paper is organized as follows: Section 2 applies skewness maximization, and therefore dominant tensor eigenvectors, to cluster separation. Section 3 investigates the asymptotic properties of dominant and base eigenvectors of sample third-order cumulants. Section 4 illustrates the results of the previous sections with six well-known data sets. Section 5 contains some concluding remarks and hints for future research. The Appendix contains the proofs.

2 Clustering

Friedman and Tukey (1974) proposed to use projection pursuit to isolate a cluster and then to repeat the procedure on the remaining data. Independently, Hennig (2004) proposed a similar approach, aimed to cluster data where one cluster is homogeneous and well separated from the remaining, possibly more scattered, clusters. The theoretical results in this section support both proposals, when projection pursuit is based on skewness maximization.

The following proposition states that a function of a finitely supported random variable maximizes skewness if it maps every outcome of the random variable itself which has not minimal probability onto the same value, thus obtaining a dichotomous distribution.

Proposition 1 *Let X be a random variable with finite support $X = \{x_1, ..., x_k\}$. Also, let $Y = g(X)$ be a real, nondegenerate function of X: $var(Y) > 0$. Finally, let x_j be the unique element of X occurring with minimal probability: $0 < Pr\left(X = x_j\right) < Pr\left(X = x_i\right) : \ i \neq j$. Then the third standardized cumulant of Y attains its maximum absolute value if and only if Y is dichotomous with $Pr\left\{Y = g\left(x_j\right)\right\} = Pr\left(X = x_j\right)$.*

Proposition 1 is instrumental in proving Theorem 2, but it is also of interest by itself. As seen in the proof of Proposition 1 in the Appendix, the third standardized cumulant of Y is

$$\gamma_1(Y) = \frac{1 - 2p_1}{\sqrt{p_1(1 - p_1)}}, \text{ where } p_1 = \min_i \Pr(X = x_i).$$

Consider now a (not necessarily random) sample X_1,..., X_n, whose mean, variance and skewness are

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \ S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \text{ and } G_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{S}\right)^3.$$

Theorem 1 implies that G_1 achieves its maximum value when all but one observations equal each other, that is when $p_1 = 1/n$ and $G_1 = (n-2)/\sqrt{n-1}$.

Since the third sample standardized cumulant is a continuous function of the observations, it tends to be close to $(n-2)/\sqrt{n-1}$ when one oservation is very different from the remaning ones, while the latter are very close to each other. This reasoning motivates the use of skewness when testing for the presence of outliers, as argued by Ferguson (1961) under the more restrictive normality assumption.

A weakly symmetric distribution is a distribution whose third cumulant is a null matrix (Loperfido 2014). Symmetric distributions with finite third moments are weakly symmetric but the opposite is not necessarily true. Loperfido (2013), Loperfido (2015) and Loperfido (2019), uses skewness-based projection pursuit for cluster detection, when data come from finite mixtures of weakly symmetric distributions. In particular, Loperfido (2013) and Loperfido (2015) dealt with finite weakly symmetric location mixtures, that is finite mixtures of weakly symmetric distributions only differing in

their means. The following theorem shows that, for finite weakly symmetric location mixtures, the component with the smallest weight is best separated from the remaining ones by the projections attaining maximal skewness, when the component's mean is far away from the other components' means. The theorem supports skewness maximization as a tool for the iterative detection and removal of clusters, as suggested in Friedman and Tukey (1974).

Theorem 1 *Let the distribution of the random vector* $\boldsymbol{x}$ *be a finite location mixture of weakly symmetric distributions with linearly independent means. Also, let the mean of the component with the smallest weight have norm* $c > 0$. *Finally, let* $\boldsymbol{u}^\top \boldsymbol{x}$ *and* $\boldsymbol{v}^\top \boldsymbol{x}$ *be the best discriminating projection of* $\boldsymbol{x}$ *and the projection of* $\boldsymbol{x}$ *which maximizes skewness. Then*

$$\lim_{c \to +\infty} \rho^2 \left(\boldsymbol{u}^\top \boldsymbol{x}, \boldsymbol{v}^\top \boldsymbol{x} \right) = 1.$$

Theorem 2 provides the mathematical background for the following sequential clustering procedure. Data are projected onto the direction which maximizes skewness in order to separate a cluster from the others. The detected cluster is then removed from the data and the procedure is repeated until no clusters are left. Theorem 2 might also be used for detecting outliers, which might be regarded as limiting cases of small-sized, well-separated clusters (Hou and Wentzell 2014) and have been modelled by means of finite normal location mixtures (Archimbaud et al. 2018). We illustrate the use of skewness maximization for the iterative detection and removal of clusters with a mixture of three normal distributions with identical covariance matrices. Let C be the random variable representing the cluster memberships. It takes the values 1, 2 and 3 with probabilities 0.1, 0.4 and 0.5: $\mathrm{P}(C = 1) = 0.1$, $\mathrm{P}(C = 2) = 0.4$, $\mathrm{P}(C = 3) = 0.5$. Also, let $\mathbf{x}|C = i \sim N\left(\boldsymbol{\mu}_i, \mathbf{I}_2\right)$ be the distribution of $\mathbf{x}$ in the i-th cluster, where $\mathbf{I}_2$ is the bivariate identity matrix and

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 10 \\ 10 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 5 \\ -5 \end{pmatrix}, \boldsymbol{\mu}_3 = \begin{pmatrix} -4 \\ 4 \end{pmatrix}.$$

The distribution of $\mathbf{x}$ is then a location normal mixture with three components, where the mean of the component with the smallest weight has a norm much greater than the other ones: $\mathbf{x} \sim 0.1 \cdot N(\boldsymbol{\mu}_1, \mathbf{I}_2) + 0.4 \cdot N(\boldsymbol{\mu}_2, \mathbf{I}_2) + 0.5 \cdot N(\boldsymbol{\mu}_3, \mathbf{I}_2)$. The mean, the within-group covariance, the between-group covariance and the total covariance are

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mathbf{W} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mathbf{B} = \begin{pmatrix} 27 & -9 \\ -9 & 27 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 28 & -9 \\ -9 & 28 \end{pmatrix}.$$

The Fisher's discriminating direction is the dominant eigenvector of the matrix

$$\boldsymbol{\Sigma}^{-1}\mathbf{B} = \begin{pmatrix} 28 & -9 \\ -9 & 28 \end{pmatrix}^{-1} \begin{pmatrix} 27 & -9 \\ -9 & 27 \end{pmatrix} = \frac{1}{703} \begin{pmatrix} 675 & -9 \\ -9 & 675 \end{pmatrix} \approx \begin{pmatrix} 0.960 & -0.013 \\ -0.013 & 0.960 \end{pmatrix},$$

which is proportional to the bidimensional vector of ones $\mathbf{1}_2$. The Fisher linear discriminant projection is $\mathbf{1}_2^\top \mathbf{x}$, and it is also the linear projection which best separates Cluster 1 from Cluster 2 and Cluster 3, which are merged together: $\mathbf{1}_2^\top \mathbf{x} \sim$

$0.1 \cdot N(20, 2) + 0.9 \cdot N(0, 2)$. The coskewness of $\mathbf{x}$ and the positive definite square root of the concentration matrix $\boldsymbol{\Sigma}^{-1}$ are

$$cos(\mathbf{x}) = \begin{pmatrix} -29.61 & 6.39 & 6.39 & 42.39 \\ 6.39 & 42.39 & 42.39 & -65.61 \end{pmatrix} \text{ and } \boldsymbol{\Sigma}^{-1/2} = \begin{pmatrix} 0.197 & 0.033 \\ 0.033 & 0.197 \end{pmatrix}.$$

The standardized coskewness of $\mathbf{x}$ is

$$cos(\mathbf{z}) = \boldsymbol{\Sigma}^{-1/2} cos(\mathbf{x}) \left(\boldsymbol{\Sigma}^{-1/2} \otimes \boldsymbol{\Sigma}^{-1/2}\right) = \begin{pmatrix} -0.174 & 0.110 & 0.110 & 0.274 \\ 0.110 & 0.274 & 0.274 & -0.338 \end{pmatrix}.$$

The bidimensional vector of ones is the dominant eigenvector of the third standardized cumulant of $\mathbf{x}$:

$$cos(\mathbf{z})(\mathbf{1}_2 \otimes \mathbf{1}_2) = \begin{pmatrix} -0.174 & 0.110 & 0.110 & 0.274 \\ 0.110 & 0.274 & 0.274 & -0.338 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = 0.32 \cdot \mathbf{1}_2.$$

As remarked in the Introduction, the projection of $\mathbf{x}$ with maximal skewness is $\mathbf{1}_2^\top \boldsymbol{\Sigma}^{-1/2}\mathbf{x}$. Since $\mathbf{1}_2$ is an eigenvector of $\boldsymbol{\Sigma}$, it is also an eigenvector of $\boldsymbol{\Sigma}^{-1/2}$. The projection of $\mathbf{x}$ with maximal skewness is then $\mathbf{1}_2^\top \mathbf{x}$, which coincides with the Fisher linear discriminant function.

In order to separate Cluster 2 from Cluster 3 we assume that we can take out Cluster 1, so we obtain the distribution $\mathbf{x}|C \neq 1 \sim (4/9) \cdot N(\boldsymbol{\mu}_2, \mathbf{I}_2) + (5/9) \cdot N(\boldsymbol{\mu}_3, \mathbf{I}_2)$, which is a mixture with unequal weights of two normal distributions with the same covariance matrices. As shown in Loperfido (2013), the linear projection which maximizes skewness is $(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3)^\top cov(\mathbf{x}|C \neq 1)\mathbf{x} \propto \boldsymbol{\mu}_2^\top \mathbf{x} \propto X_1 - X_2$, where X_1 and X_2 are the first and the second component of $\mathbf{x}$. The projection $X_1 - X_2$ coincides, up to location and scale changes, with the Fisher linear discriminant projection. We used the projection $X_1 + X_2$ to separate the first cluster from the other two, and then the projection $X_1 - X_2$ to separate the second cluster from the third one. The example suggests that Theorem 2 might hold under more general assumptions, since $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ are proportional to each other ($-0.2\boldsymbol{\mu}_2 = 0.25\boldsymbol{\mu}_3$), thus violating the assumptions of Theorem 2.

3 Asymptotics

Let $\mathbf{x}_i^\top$ be the i-th row of $\mathbf{X}$, $i \in \{1, ..., n\}$. The mean, the covariance and the coskewness of $\mathbf{X}$ are

$$\mathbf{m} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i, \mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^\top \text{ and}$$

$$\mathbf{G} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m}) \otimes (\mathbf{x}_i - \mathbf{m})^\top \otimes (\mathbf{x}_i - \mathbf{m})^\top.$$

Let $\mathbf{z}_i^\top$ be the i-th row of the standardized data $\mathbf{Z} = \mathbf{H}_n\mathbf{X}\mathbf{S}^{-1/2}$, $i \in \{1, ..., n\}$, where $\mathbf{H}_n = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^\top/n$ is the $n \times n$ centring matrix, $\mathbf{1}_n$ is the $n-$ dimensional vector of ones, $\mathbf{I}_n$ is the $n-$dimensional identity matrix, and $\mathbf{S}^{-1/2}$ is the symmetric, positive definite square root of the sample concentration matrix $\mathbf{S}^{-1}$. The standardized coskewness of $\mathbf{X}$ is just the coskewnness of $\mathbf{Z}$:

$$\mathbf{Q} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i \otimes \mathbf{z}_i^\top \otimes \mathbf{z}_i^\top.$$

The dominant eigenvalue l_1 of $\mathbf{Q}$ is also the maximal skewness achievable by a linear projection of $\mathbf{X}$. Inferential projection pursuit investigates the connections between l_1 and its population counterpart, that is the dominant tensor eigenvalue of the third standardized moment of the underlying distribution. As mentioned in the Introduction, the first inferential use of moment optimizing projections dates back to Malkovich and Afifi (1973), within a multivariate normality testing framework. Machado (1983) shows that these statistics have an asymptotic distribution, under normality. Baringhaus and Henze (1991) relates the asymptotic distribution of the same statistics to the maximum of a gaussian process, under the assumption of elliptical symmetry. Naito (1997) uses the results in Baringhaus and Henze (1991) and Sun (1993) for approximating the tail probabilities of a generalized moment index which includes the one proposed by Jones and Sibson (1987). Kuriki and Takemura (2008) uses a geometric approach to derive exact formulae for the tail probabilities of Malkovich and Afifi (1973) statistics and other maxima of multilinear forms. Loperfido (2018), supported by both theoretical and empirical arguments, conjectures that the asymptotic distribution of maximal skewness might be conveniently approximated by a skew-normal distribution, under the null hypothesis of normality.

All of the above papers deal with hypothesis testing, and none of them with point estimation. We address the latter inferential issue by showing that the dominant eigenpair of the third sample moment converges almost surely to its population counterpart, under mild assumptions.

Theorem 2 *Let λ and $\boldsymbol{v}$ be the simple, dominant tensor eigenvalue and its tensor eigenvector of the third moment of the p-dimensional random vector $\boldsymbol{x}$. Also, let the n-th elements of the sequences $\{\boldsymbol{X}_n\}$, $\{\mathcal{M}_n\}$, $\{\lambda_n\}$ and $\{\boldsymbol{v}_n\}$, be the $n \times p$ data matrix whose rows are independent outcomes of $\boldsymbol{x}$, the third moment of $\boldsymbol{X}_n$, the dominant tensor eigenvalues of $\mathcal{M}_n$ and the tensor eigenvector of λ_n. Then $\{\lambda_n\}$ and $\{\boldsymbol{v}_n\}$ converge almost surely to λ and $\boldsymbol{v}$ as n tends to infinity: $\lambda_n \xrightarrow{a.s.} \lambda$ and $\boldsymbol{v}_n \xrightarrow{a.s.} \boldsymbol{v}$.*

Skewness-based projection pursuit is also concerned with base tensor eigenvectors, given their close connection with weakly symmetric projections, that is projections whose coskewnesses are null matrices. Weakly symmetric projections may be used before skewness-based projection pursuit as tools for data reduction, following the approach in Jones and Sibson (1987), Hui and Lindsay (2010), Ray (2010), Lindsay and Yao (2012) and Loperfido (2023). Weakly symmetric projections may also be used after skewness-based projection pursuit, to facilitate the search for interesting structures other than skewness, as proposed by Huber (1985) and Daszykowski (2007).

Statistical applications of weakly symmetric projections are not limited to projection pursuit. For example, they may also be useful in multivariate mean testing (Loperfido 2014, 2019).

Weakly symmetric sampled distributions and weakly symmetric data projections are characterized by having null Mardia's skewnesses (Mardia 1970). Let $\mathbf{x}$ and $\mathbf{y}$ two $p-$dimensional, independent and identically distributed random vectors with mean $\boldsymbol{\mu}$, nonsingular variance $\boldsymbol{\Sigma}$ and finite third moments. The Mardia's skewness of $\mathbf{x}$ ($\mathbf{y}$) is

$$\beta_{1,M}^{M}(\mathbf{x}) = \mathrm{E}\left[\left\{(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}^{3}\right].$$

Its sample counterpart is

$$b_{1,M}(\mathbf{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\left[(\mathbf{x}_i-\mathbf{m})^{\top}\mathbf{S}^{-1}(\mathbf{x}_j-\mathbf{m})\right]\right\}^{3}.$$

The Mardia's skewness equals the squared norm of the standardized coskewness, so that the Mardia's skewness equal zero if and only if the coskewness is a null matrix, that is under weak symmetry. In particular, a projection onto the direction of a base eigenvector of the coskewness is weakly symmetric. However, due to sampling variability, the sample coskewness might not have base eigenvectors while the coskewness of the underlying distribution does. In such situations, almost weakly symmetric projections, that is projections having the smallest Mardia's skewness, are intuitively appealing. The following theorem supports this approach.

Theorem 3 *Let the third cumulant of the p-dimensional random vector $\boldsymbol{x}$ have base eigenvectors constituting a linear space of dimension $q < p$. Also, let the elements of the sequences $\{\boldsymbol{X}_n\}$ and $\{\boldsymbol{B}_n\}$ be $n \times p$ data matrices whose rows are independent outcomes of $\boldsymbol{x}$ and $p \times q$ matrices of full rank minimizing the Mardia's skewness of $\boldsymbol{X}_n\boldsymbol{B}_n$. Then each row of $\{\boldsymbol{X}_n\boldsymbol{B}_n\}$ converges almost surely to a weakly symmetric random vector.*

Base matrix eigenvectors constitute a linear space, but the same does not necessarily happens for base tensor eigenvectors. As an example, consider the generalized skew-normal distribution $2\phi(z_1)\phi(z_2)\phi(z_3)\Phi(\theta z_1 z_2 z_3)$, where $\phi(\cdot)$ is the pdf of a standard normal random distribution, $\Phi(\cdot)$ is the cdf of a standard normal random distribution and θ is a nonnull, real value. As shown in Loperfido (2018, 2019), the distribution is standardized and its only nonnull third moment is $\mathrm{E}(Z_1Z_2Z_3) = \gamma = \gamma(\theta)$ (a function of θ), so that its coskewness is

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \gamma & 0 & \gamma & 0 \\ 0 & 0 & \gamma & 0 & 0 & 0 & \gamma & 0 & 0 \\ 0 & \gamma & 0 & \gamma & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Table 1 The first and the second row of the table contain the percentages of correctly classified units by linear discrimination and skewness maximization

	Athletes	Crabs	Breast	Returns	Banknotes	Sparrows
Linear discrimination	92.6%	100%	97.4%	57.3%	78%	65.3%
Skewness maximization	74.8%	80%	77.5%	55.7%	63%	55%

The columns refers to the six datasets

The base eigenvectors of $\boldsymbol{\Gamma}$ are

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \text{ and } \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

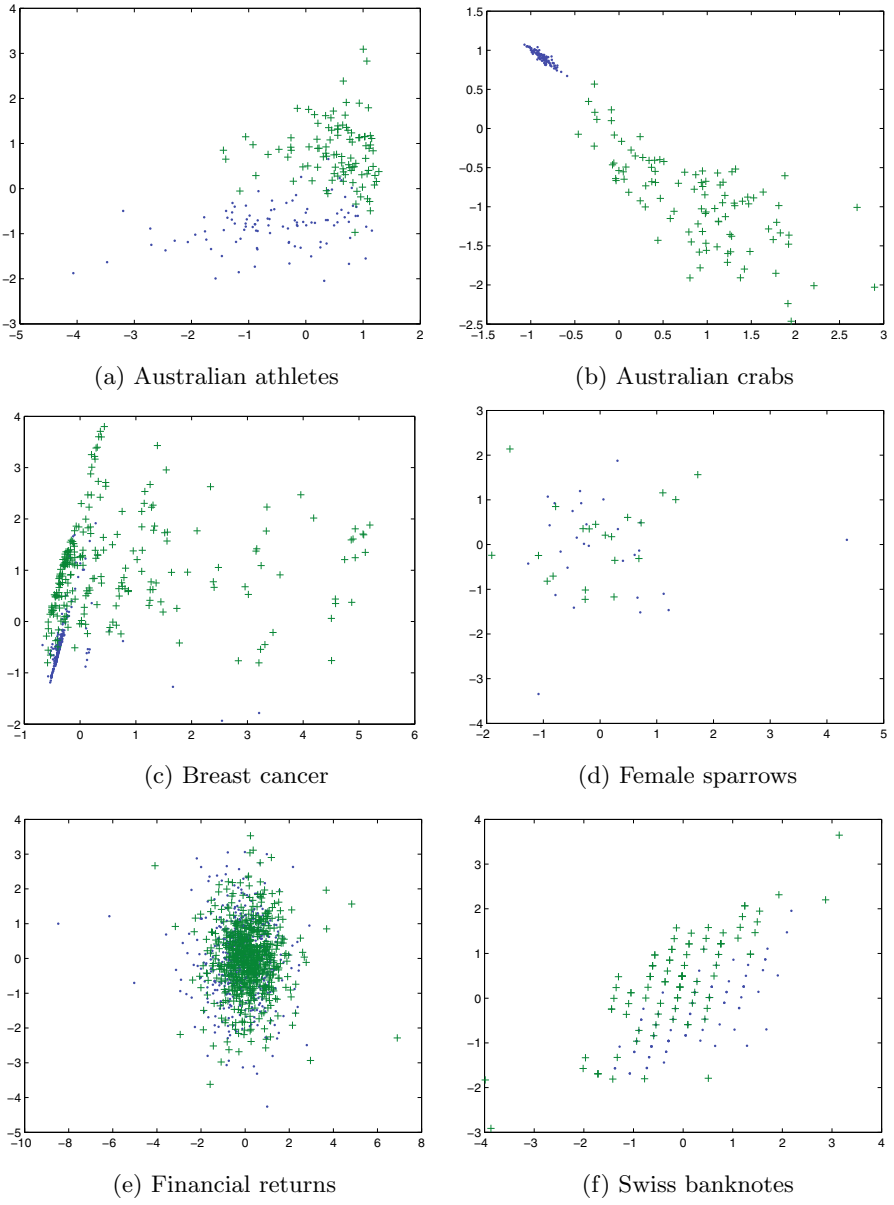
However, no nontrivial linear combination of them is a base eigenvector of $\boldsymbol{\Gamma}$.

4 Examples

In this section we use six well-known data sets to assess the practical usefulness of skewness-based projection pursuit as a clustering method. Each of them is divided into two groups, so the group membership of each sample unit is known. The data sets differ with respect to the skewnesses of their groups and the performance of the linear discriminant fuction in separating the groups themselves. We first classified the observations using the linear discriminant function, which relies on the knowledge of group memberships. Then we classified the same data with skewness-based projection pursuit, which does not rely on the knowledge of group memberships. The classification procedure based on projection pursuit articulates into two steps. First, the data are projected onto the direction which maximizes their skewness. Second, the projected data are classified into two groups using k-means clustering, which is quite efficient when applied to univariate data. As expected, the former method outperforms the latter, since it uses more information. However, the difference is small enough to encourage the use of skewness-based projection pursuit for classifying data when group memberships are unknown. We also visually inspected the data with scatterplots of the two most skewed projections, which revealed further insight into the clustering structure of data. Table 1 summarizes the performances of the two classification methods.

Next, we give a more detailed description of the data and the classification results.

Australian athletes. The Australian Institute of Sports collected several body measurements from 202 elite athletes of both genders competing in different disciplines. Since the seminal paper by Azzalini and Dalla Valle (1996) the data are known to be skewed. We aim at classifying the 100 female athletes and the 102 male athletes by means of their body fat and lean body mass indices. The linear discriminant function correctly classifies 187 athletes, that is about 92.6% of them. Skewness-based projection pursuit correctly classifies 151 athletes, that is about 74.8% of them. The scatterplot of the two most skewed projections (Fig. 1a) shows a clear separation of the two groups, together with their marked non-elliptical shapes.



(a) Australian athletes (b) Australian crabs

(c) Breast cancer (d) Female sparrows

(e) Financial returns (f) Swiss banknotes

Fig. 1 **a** Australian athletes (dots represent female athletes, pluses represent male athletes); **b** Australian crabs (dots represent blue crabs, pluses represent orange crabs); **c** Breast cancer (dots represent benign tumors, pluses represent malignant tumors); **d** Female sparrows (dots represent deceased sparrows, pluses represent survived sparrows); **e** Financial returns (dots represent negative signs, pluses represent positive signs); **f** Swiss banknotes (dots represent forged bills, pluses represent genuine bills) (colour figure online)

Australian crabs. Campbell and Mahon (1974) collected 5 morphological measurements (frontal lobe size, rear width, carapace length, carapace width, and body depth) of the blue and orange species of crabs living in Fremantle, Western Australia. More precisely, there are 100 specimen of blue crabs and 100 specimen of orange crabs. Measurements in both groups are often modelled by normal mixtures with equal or proportional covariances. The linear discriminant function correctly classifies all crabs. Skewness-based projection pursuit correctly classifies 160 crabs, that is exactly 80% of them. The separation between the two groups becomes even more apparent from the scatterplot of the two most skewed projections (Fig. 1b), which also shows a much smaller scatter in the blue crabs group.

Breast cancer. Street et al. (1993) computed ten integer-valued features from digitized images of fine needle aspirates of breast masses belonging to 699 women diagnosed with breast cancer. The features describe characteristics of the cell nuclei present in the image. The tumor was benign for 458 women in the sample, and malignant for 241. We found data in both groups to be significantly skewed. The linear discriminant function correctly classifies 681 women, that is about 97.4% of them. Skewness-based projection pursuit correctly classifies 542 women, that is about 77.5% of them. The difference in performances between the methods might be due to the presence of potential outliers, as hinted by the scatterplot of the two most skewed projections (Fig. 1c).

Female sparrows. Manly and Navarro Alberto (2016) considered total length, alar extent, length of beak and head, length of humerous, and length of keel of sternum of 49 female sparrows. Data were collected after a severe storm, after which 21 of them survived. The sample sizes of both groups are too small to test the symmetry hypothesis. However, an exploratory data analysis (not reported here) hint that skewness may be negligible. The linear discriminant function correctly classifies 32 sparrows, that is about 65.3% of them. Skewness-based projection pursuit correctly classifies 27 sparrows, that is about 55% of them. The poor performance of both methods, and especially of the latter, could have been anticipated by looking at the scatterplot of the two most skewed projections (Fig. 1d), where the groups are not well separated.

Financial returns. Morgan Stanley Capital International Inc. recorded 1291 percentage logarithmic daily returns (simply returns, henceforth) in the financial markets of France, Netherlands and Spain. De Luca and Loperfido (2015) clustered the returns according to the sign of the previous day U.S. return, obtaining two groups with 597 and 694 returns each, which were found to be significantly skewed. The linear discriminant function correctly classifies 740 returns, that is about 57.3% of them. Skewness-based projection pursuit correctly classifies 719 returns, that is about 55.7% of them. As in the previous data set, the two groups are very poorly separated in the scatterplot of the two most skewed projections (Fig. 1e), with the exceptions of a few outliers, which constitute a well-known stylized fact of financial returns.

Swiss banknotes. Flury (1988) reported several measurements from 100 genuine and 100 forged old Swiss 1000 franc bills. Greselin et al. (2011) focused on their width, measured on both sides, and found them to be bivariate normal in the forged group, but not in the genuine one. They also rejected the homoscedastic hypothesis. Other statistical analyses, not shown here, clearly suggest that some skewness is present in the genuine group, but not in the forged one. The linear discriminant function correctly

classifies 156 bills, that is 78% of them. Skewness-based projection pursuit correctly classifies 126 bills, that is about 63% of them. The two groups appear to be even better separated in the scatterplot of the two most skewed projections (Fig. 1f), which also hints the presence of some possible outliers in the genuine group.

5 Conclusions

This paper investigated some connections between third-order tensor eigenvectors and skewness-based projection pursuit. The former concept belongs to multilinear algebra, while the latter concept belongs to multivariate analysis. The theoretical results in the paper support the use of skewness-based projection pursuit both in the exploratory and in the inferential stages of statistical analysis. The practical usefulness of the method is illustrated with six dataset which already appeared in the statistical literature: the Australian Athletes dataset, the Australian Crabs dataset, the Breast Cancer dataset, the Female Sparrows dataset, the Financial Returns dataset and the Swiss Banknotes dataset. They all suggest that skewness-based projection pursuit might be used to recover the linear discriminant function when the group memberships are unknown.

On the other hand, the above examples are limited in several ways. Firstly, they only consider two clusters, while the theorem and the example in Section 2 support the use of skewness-based projection pursuit in the presence of more clusters. Secondly, the optimal discriminant function might not be linear, as it happens when there are two multivariate normal distributions with different means and covariances (see, e.g., Mardia et al. 1979, page 312). Thirdly, the comparison between the performances of the two approaches should not rely on the misclassification rate only, but should include other performance measures, as for example the receiving operating curve (ROC) and the area under the ROC curve (AUC). Space constraints prevented us from investigating these issues in the present paper, but we are planning to address them in the future by means of both real and synthetic data.

Maximally skewed projections of some well-known distributions admit simple and insightful interpretations (Arevalillo and Navarro 2019, 2020). It is then worth asking which widely used multivariate probability distributions have third-order cumulants whose eigenvectors admit a simple tractable analytical form. This would simplify both their computation and their interpretation. It would also give more insight into the asymptotic properties of skewness-based projection pursuit. Similar remarks also hold for kurtosis-based projection pursuit, which relies on kurtosis optimization and is closely related to the eigenvectors of fourth-order symmetric tensors (Loperfido 2017). Moreover, the joint use of skewness and kurtosis optimization might lead to some additional insight into data features (Arevalillo and Navarro 2021a). We are currently investigating these topics.

Acknowledgements The author would like to thank Professor Christian Hennig for the very interesting conversations about model-based clustering and skewness-based projection pursuit. The author would also like to thank two anonymous Reviewers for their useful and detailed comments, which greatly helped to improve the quality of the paper.

Appendix A Proofs

Proof of Proposition 1 Let μ and $\sigma^2 > 0$ be the mean and the variance of Y, and let $Z = (Y - \mu)/\sigma$ be the standardized version of Y. Since Y is a function of X, whose support contains k elements, the support of Z contains at most k elements, denoted as z_1,..., z_h with $h \leq k$. Let us put $\Pr(Z = z_i) = p_i$, for $i = 1$,..., h. Maximizing the third standardized cumulant of Y is equivalent to maximizing $\mathrm{E}\left(Z^3\right)$ under the constraints $\mathrm{E}(Z) = 0$ and $\mathrm{E}\left(Z^2\right) = 1$. We can then write the Lagrangian equation

$$L(z_1, ..., z_h, \lambda, \eta) = \sum_{i=1}^{h} z_i^3 p_i - \lambda \left(\sum_{i=1}^{h} z_i^2 p_i - 1 \right) - \eta \sum_{i=1}^{h} z_i p_i.$$

By differentiating the Lagrangian equation with respect to z_i we obtain

$$\frac{\partial}{\partial z_i} L(z_1, ..., z_k, \lambda, \eta) = 3z_i^2 p_i - 2\lambda z_i p_i - \eta p_i = 0,$$

which can be simplified into $3z_i^2 - 2\lambda z_i - \eta = 0$ by recalling that $p_i > 0$. The second degree equation $3x^2 - 2\lambda x - \eta = 0$ has at most two distinct real roots, which means that Z is a dichotomous random variable. As such, Z may be represented either as

$$p = \Pr\left(Z = -\sqrt{\frac{1-p}{p}} \right) = 1 - \Pr\left(Z = \sqrt{\frac{p}{1-p}} \right) \text{ or as}$$

$$p = \Pr\left(Z = \sqrt{\frac{1-p}{p}} \right) = 1 - \Pr\left(Z = -\sqrt{\frac{p}{1-p}} \right), \text{ where } p \in \{p_1, ..., p_h\}.$$

Let z_1 and z_2 be the outcomes of Z associated with the probabilities p and $1 - p$: $\Pr(Z = z_1) = p$ and $\Pr(Z = z_2) = 1 - p$. The squared third moment of Z is

$$\mathrm{E}^2\left(Z^3\right) = \left\{ \left(\frac{1-p}{p} \right)^{1.5} p - \left(\frac{p}{1-p} \right)^{1.5} (1-p) \right\}^2 = \frac{(1-2p)^2}{p(1-p)}.$$

Without loss of generality we can assume that $p \neq 0.5$: when $p = 0.5$ the squared third moment $\mathrm{E}^2\left(Z^3\right)$ attains its minimum value, that is zero. We first consider the case $p < 0.5$, where $\mathrm{E}^2\left(Z^3\right)$ increases as p decreases. By definition, p is the probability of an outcome of Z and by assumption there is a unique p_i which is smaller than any p_j, with $i \neq j$ and $i, j = 1, ..., h$. Hence the absolute skewness of Z is maximized if the probability of z_1 is the smallest probability associated with an element in the support of X: $\Pr(Z = z_1) = \min_{i=1,...,h} p_i$.

We now consider the case $p > 0.5$, where $\mathrm{E}^2\left(Z^3\right)$ increases as $1-p$ increases. By an argument similar to the one above, the absolute skewness of Z is maximized if

$$\begin{aligned}\Pr\left(Z=z_2\right) &= \max_{i=1,\ldots,h}\left(1-p_i\right)=1-\min_{i=1,\ldots,h}p_i, \text{ so that } \Pr\left(Z=z_1\right)=1\\ &\quad -\Pr\left(Z=z_2\right)=\min_{i=1,\ldots,h}p_i.\end{aligned}$$

Therefore, either when $p < 0.5$ or when $p > 0.5$, the absolute skewness is maximized when there is an outcome of the dicothomous random variable Y which coincides with the outcome of X with minimal probability. □

Proof of Theorem 1 Let $\boldsymbol{\Omega}$, π_i and $\boldsymbol{\mu}_i$ be the components' covariance, the weight of the i -th mixture's component and the mean vector of the i-th mixture's component, for $i = 1,..., g$. Also, let $\mathbf{y}$ be the random vector taking the value $\boldsymbol{\mu}_i$ with probability π_i: $P\left(\mathbf{y}=\boldsymbol{\mu}_i\right)=\pi_i$. Finally, let the mean of the component with the smallest weight be $c\cdot\mathbf{m}$, where $\mathbf{m}$ is a unit norm vector. Without loss of generality we can assume that that the component with the smallest weight is the last one: $\boldsymbol{\mu}_g = c\cdot\mathbf{m}$.

By assumption, the vector means of the components are linearly independent. Without loss of generality, we can also assume that $\mathbf{m}$ is orthogonal to all other mixture's components. If it were not so, there would be a linear transformation of the random vector $\mathbf{x}$, based on the Gram–Schmidt orthogonalization, which would be a location mixture of g weakly symmetric components and where the mean of the g-th component is orthogonal to the remaining ones. Then the projection $\mathbf{m}^\top\mathbf{y}$ is a dichotomous random variable placing the smallest mixture's weight on the nonnull outcome. By Proposition 1, $\mathbf{m}^\top\mathbf{y}$ is the projection of $\mathbf{y}$ maximizing skewness. The covariance of $\mathbf{y}$ is

$$cov\left(\mathbf{y}\right)=\sum_{i=1}^{g}\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}\right)\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}\right)^\top\pi_i, \text{ where } \boldsymbol{\mu}=\mathrm{E}\left(\mathbf{y}\right)=\sum_{i=1}^{g}\boldsymbol{\mu}_i\pi_i.$$

Ordinary properties of covariance decomposition, the identity $\boldsymbol{\mu}_g = c\cdot\mathbf{m}$ and some straightforward, but tedious matrix algebra, imply

$$\begin{aligned}cov\left(\mathbf{y}\right) &= c^2\pi_g\left(1-\pi_g\right)\left(\mathbf{m}-\frac{\boldsymbol{\mu}_-}{c}\right)\left(\mathbf{m}-\frac{\boldsymbol{\mu}_-}{c}\right)^\top\\ &\quad +\sum_{i=1}^{g-1}\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_-\right)\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_-\right)^\top\pi_i,\ \boldsymbol{\mu}_-=\sum_{i=1}^{g-1}\frac{\boldsymbol{\mu}_i\pi_i}{1-\pi_g}.\end{aligned}$$

The ratio of the variance $\sigma^2\left(\mathbf{m}^\top\mathbf{y}\right)$ of $\mathbf{m}^\top\mathbf{y}$ to the variance $\sigma^2\left(\mathbf{m}^\top\mathbf{x}\right)$ of $\mathbf{m}^\top\mathbf{x}$ converges to its maximum value one as c increases:

$$\lim_{c\to+\infty}\frac{\sigma^2\left(\mathbf{m}^\top\mathbf{y}\right)}{\sigma^2\left(\mathbf{m}^\top\mathbf{x}\right)}=\lim_{c\to+\infty}\frac{c^2\mathbf{m}^\top\mathbf{m}}{c^2\mathbf{m}^\top\mathbf{m}+\mathbf{m}^\top\boldsymbol{\Omega}\boldsymbol{m}}=1.$$

As a direct consequence, the best linear discriminant projection $\mathbf{u}^\top\mathbf{x}$ of $\mathbf{x}$ converges to $\mathbf{m}^\top\mathbf{x}$ as $\mathbf{c}$ increases, up to location and scale changes:

$$\lim_{c\to+\infty}\rho^2\left(\mathbf{u}^\top\mathbf{x},\mathbf{m}^\top\mathbf{x}\right)=1.$$

By assumption, the mixture's components are weakly symmetric and have the same covariance matrices. We can then apply Theorem 1 in Loperfido (2019) and show that the third cumulant of $\mathbf{m}^\top\mathbf{x}$ and $\mathbf{m}^\top\mathbf{y}$ coincide. The skewness of $\mathbf{m}^\top\mathbf{x}$ is then

$$\gamma_1\left(\mathbf{m}^\top\mathbf{x}\right)=\frac{\kappa_3\left(\mathbf{m}^\top\mathbf{y}\right)}{\left\{\sigma^2\left(\mathbf{m}^\top\mathbf{y}\right)+\mathbf{m}^\top\boldsymbol{\Omega}\boldsymbol{m}\right\}^{1.5}}=\frac{\gamma_1\left(\mathbf{m}^\top\mathbf{y}\right)}{\left\{1+\left(\mathbf{m}^\top\boldsymbol{\Omega}\boldsymbol{m}\right)/\sigma^2\left(\mathbf{m}^\top\mathbf{y}\right)\right\}^{1.5}},$$

where $\kappa_3\left(\mathbf{m}^\top\mathbf{y}\right)$ and $\gamma_1\left(\mathbf{m}^\top\mathbf{y}\right)$ are the third cumulant and the third standardized cumulant (i.e. the skewness) of $\mathbf{m}^\top\mathbf{y}$. As c increases, the covariance of the components' means, that is the covariance of $\mathbf{y}$, increases, while the mean of the covariances' components remains unchanged, so that we have

$$\lim_{c\to+\infty}\frac{\mathbf{m}^\top\boldsymbol{\Omega}\mathbf{m}}{\sigma^2\left(\mathbf{m}^\top\mathbf{y}\right)}=0 \text{ and } \lim_{c\to+\infty}\gamma_1\left(\mathbf{m}^\top\mathbf{x}\right)=\gamma_1\left(\mathbf{m}^\top\mathbf{y}\right).$$

Therefore, as c tends to infinity, $\mathbf{m}^\top\mathbf{x}$ becomes the projection of $\mathbf{x}$ achieving maximal skewness. We conclude that, as c tends to infinity, the best linear discriminant projection and the skewness-maximizing projection converges to each other, up to location and scale changes. □

Proof of Theorem 2 Let $\mathbf{M}_n$ and $\mathbf{M}$ be the unfoldings of $\mathcal{M}_n$ and $\mathcal{M}$. By ordinary properties of sample moments, the sequence $\{\mathbf{M}_n\}$ converges almost surely to $\mathbf{M}$: $\mathbf{M}_n \xrightarrow{a.s.} \mathbf{M}$. The cubic form $\mathbf{a}^\top\mathbf{M}_n\,(\,\mathbf{a}\otimes\mathbf{a})$, where $\mathbf{a}$ is any vector of the same dimension of $\mathbf{v}$ and $\mathbf{v}_n$, is a continuous function of the third-order tensor $\mathbf{M}_n$ and therefore converges almost surely to the cubic form $\mathbf{a}^\top\mathbf{M}\,(\mathbf{a}\otimes\mathbf{a})$:

$$\mathbf{v}^\top\mathbf{M}_n\,(\mathbf{v}\otimes\mathbf{v})\xrightarrow{a.s.}\mathbf{v}^\top\mathbf{M}\,(\mathbf{v}\otimes\mathbf{v}) \text{ and } \mathbf{v}_n^\top\mathbf{M}_n\,(\mathbf{v}_n\otimes\mathbf{v}_n)\xrightarrow{a.s.}\mathbf{v}_n^\top\mathbf{M}\,(\mathbf{v}_n\otimes\mathbf{v}_n).$$

Taking into account that $\mathbf{v}$ is the dominant eigenvector of $\mathcal{M}$ we can put

$$\Pr\left\{\lambda=\mathbf{v}^\top\mathbf{M}\,(\mathbf{v}\otimes\mathbf{v})\geq\lim_{n\longrightarrow\infty}\mathbf{v}_n^\top\mathbf{M}\,(\mathbf{v}_n\otimes\mathbf{v}_n)=\lim_{n\longrightarrow\infty}\mathbf{v}_n^\top\mathbf{M}\,(\mathbf{v}_n\otimes\mathbf{v}_n)\right\}=1.$$

Taking into account that $\mathbf{v}_n$ is the dominant eigenvector of $\mathcal{M}_n$ we can put

$$\Pr\left\{\lim_{n\longrightarrow\infty}\mathbf{v}_n^\top\mathbf{M}_n\,(\mathbf{v}_n\otimes\mathbf{v}_n)\geq\lim_{n\longrightarrow\infty}\mathbf{v}^\top\mathbf{M}_n\,(\mathbf{v}\otimes\mathbf{v})=\mathbf{v}^\top\mathbf{M}\,(\mathbf{v}\otimes\mathbf{v})\right\}=1.$$

Taking into account that λ_n and λ are the dominant eigenvalues of $\mathcal{M}_n$ and $\mathcal{M}$, the above probability inequalities may be restated as

$$\Pr\left\{\lim_{n\longrightarrow\infty}\lambda_n \leq \lambda\right\} = 1 \text{ and } \Pr\left\{\lim_{n\longrightarrow\infty}\boldsymbol{\lambda}_n \geq \lambda\right\} = 1,$$

which are mutually consistent if and only if $\{\lambda_n\}$ converges almost surely to λ: $\lambda_n \xrightarrow{a.s.} \lambda$. We recall again that λ_n is a tensor eigenvalue of $\mathbf{M}_n$ associated to the tensor eigenvector $\mathbf{v}_n$: $\mathbf{M}_n\left(\mathbf{v}_n^\top \otimes \mathbf{v}_n^\top\right) = \lambda_n \mathbf{v}_n$. We also recall again that the sequences $\{\mathbf{M}_n\}$ and $\{\lambda_n\}$ converges almost surely to $\mathbf{M}$ and λ: $\mathbf{M}\left(\mathbf{v}_n^\top \otimes \mathbf{v}_n^\top\right) \xrightarrow{a.s.} \lambda \mathbf{v}_n$. The sequence $\{\mathbf{v}_n\}$ therefore converges almost surely to a tensor eigenvector of $\mathbf{M}$ associated to the tensor eigenvalue λ, which is simple by assumption. As a direct consequence, the sequence $\{\mathbf{v}_n\}$ converges almost surely to $\mathbf{v}$: $\mathbf{v}_n \xrightarrow{a.s.} \mathbf{v}$. □

Proof of Theorem 3 Let $\mathbf{C}_{h,k}$ be the $hk \times hk$ commutation matrix (Magnus and Neudecker 1979), that is the matrix rearranging the elements of the vectorized $h \times k$ matrix $\mathbf{M}$ into its vectorized transpose: $\mathbf{C}_{h,k} vec\,(\mathbf{M}) = vec\left(\mathbf{M}^\top\right)$. As a special case, the commutation matrix $\mathbf{C}_{p,p}$ rearranges the elements of the tensor product $\mathbf{v}_1 \otimes \mathbf{v}_2$ into the tensor product $\mathbf{v}_2 \otimes \mathbf{v}_1$, where $\mathbf{v}_1$ and $\mathbf{v}_2$ are p-dimensional real vectors: $\mathbf{C}_{p,p}\,(\mathbf{v}_1 \otimes \mathbf{v}_2) = \mathbf{v}_2 \otimes \mathbf{v}_1$: $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^p$. By definition, any tensor eigenvector of the third cumulant of $\mathbf{x}$ is a nonnull p -dimensional vector satisfying

$$\mathbf{K}_{3,\mathbf{x}}\,(\mathbf{v} \otimes \mathbf{v}) = \lambda \mathbf{v},\ \lambda \in \mathbb{C},\ \mathbf{v} \in \mathbb{C}_0^p,\ \mathbf{K}_{3,\mathbf{x}} = \mathrm{E}\left\{(\mathbf{x} - \boldsymbol{\mu}) \otimes (\mathbf{x} - \boldsymbol{\mu})^\top \otimes (\mathbf{x} - \boldsymbol{\mu})^\top\right\}$$
$$\text{and } \mathrm{E}\,(\mathbf{x}) = \boldsymbol{\mu},$$

where $\mathbb{C}$ is the set of complex numbers and $\mathbb{C}_0^p$ is the set of non-null p-dimensional complex vectors. As shown in Loperfido (2015a), the $p \times p^2$ matrix $\mathbf{K}_{3,\mathbf{x}}$, that is the coskewness of $\mathbf{x}$, is invariant to multiplication by a symmetric commutation matrix: $\mathbf{K}_{3,\mathbf{x}} = \mathbf{K}_{3,\mathbf{x}}\mathbf{C}_{p,p}$ and therefore $\mathbf{K}_{3,\mathbf{x}}\,(\mathbf{v}_2 \otimes \mathbf{v}_1) = \mathbf{K}_{3,\mathbf{x}}\,(\mathbf{v}_1 \otimes \mathbf{v}_2)$. By assumption, the third cumulant of the p-dimensional random vector $\mathbf{x}$ has base eigenvectors constituting a linear space of dimension $q < p$. Let $\mathbf{A}$ be a full rank $q \times p$ matrix whose rows span the linear space $\mathbb{A}$ of the base eigenvectors of $\mathbf{K}_{3,\mathbf{x}}$:

$$\mathbf{A} \in \mathbb{R}^q \times \mathbb{R}^p,\ rank\,(\mathbf{A}) = q,\ \mathbf{A} = \begin{pmatrix} \mathbf{a}_1^\top \\ \ldots \\ \mathbf{a}_q^\top \end{pmatrix},\ \mathbf{K}_{3,\mathbf{x}}\,(\mathbf{a}_i \otimes \mathbf{a}_i) = \mathbf{0}_p,\ span\left(\mathbf{a}_1, \ldots \mathbf{a}_q\right) = \mathbb{A}.$$

Since $\mathbb{A}$ is a linear space, any nonnull linear combination of two base eigenvectors of $\mathbf{K}_{3,\mathbf{x}}$ is a base eigenvector, too: $\mathbf{K}_{3,\mathbf{x}}\left\{\left(c_i\mathbf{a}_i + c_j\mathbf{a}_j\right) \otimes \left(c_i\mathbf{a}_i + c_j\mathbf{a}_j\right)\right\} = \mathbf{0}_p$, with $c_i c_j \neq 0$.The assumption of $\mathbf{a}_i$ and $\mathbf{a}_j$ being base eigenvectors of $\mathbf{K}_{3,\mathbf{x}}$, together with the above mentioned identity $\mathbf{K}_{3,\mathbf{x}} = \mathbf{K}_{3,\mathbf{x}}\mathbf{C}_{p,p}$, leads to

$$\begin{aligned}
\mathbf{0}_p &= \mathbf{K}_{3,\mathbf{x}}\left\{\left(c_i\mathbf{a}_i + c_j\mathbf{a}_j\right) \otimes \left(c_i\mathbf{a}_i + c_j\mathbf{a}_j\right)\right\} \\
&= \mathbf{K}_{3,\mathbf{x}}\left(c_i^2\mathbf{a}_i \otimes \mathbf{a}_i + c_i c_j \mathbf{a}_j \otimes\ \mathbf{a}_i + c_i c_j \mathbf{a}_i \otimes \mathbf{a}_j + c_j^2 \mathbf{a}_j \otimes \mathbf{a}_j\right) \\
&= \mathbf{K}_{3,\mathbf{x}}\left(c_i c_j \mathbf{a}_j \otimes \mathbf{a}_i + c_i c_j \mathbf{a}_i \otimes \mathbf{a}_j\right) = c_i c_j \mathbf{K}_{3,\mathbf{x}}\left(\mathbf{a}_j \otimes \mathbf{a}_i\right) + c_i c_j \mathbf{K}_{3,\mathbf{x}}\mathbf{C}_{p,p}\left(\mathbf{a}_i \otimes \mathbf{a}_j\right) \\
&= 2c_i c_j \mathbf{K}_{3,\mathbf{x}}\left(\mathbf{a}_i \otimes \mathbf{a}_j\right).
\end{aligned}$$

The coskewness of $\mathbf{Ax}$ may be derived using multilinear properties of third cumulants (see, e.g., Loperfido 2015a): $\mathbf{K}_{3,\mathbf{Ax}} = \mathbf{A}\mathbf{K}_{3,\mathbf{x}}\left(\mathbf{A}^\top \otimes \mathbf{A}^\top\right) = \left\{\mathbf{a}_i^\top \mathbf{K}_{3,\mathbf{x}}\left(\mathbf{a}_j \otimes \mathbf{a}_k\right)\right\}$, $i, j, k \in \{1, ..., q\}$. The identities $\mathbf{K}_{3,\mathbf{x}}\left(\mathbf{a}_i \otimes \mathbf{a}_j\right) = \mathbf{0}_p$ imply that the coskewness of $\mathbf{Ax}$ is a $q \times q^2$ null matrix, which in turn implies that the Mardia's skewness of $\mathbf{Ax}$ equals zero: $\beta_1(\mathbf{Ax}) = 0$.

We prove the theorem by contradiction, assuming that the sequence $\{b_1(\mathbf{X}_n\mathbf{B}_n)\}$ of Mardia's skewnesses of $\{\mathbf{X}_n\mathbf{B}_n\}$ does not converge almost surely to zero. Let $b_1\left(\mathbf{X}_n\mathbf{A}^\top\right)$ be the Mardia's skewness of $\mathbf{X}_n\mathbf{A}^\top$. By ordinary properties of sample cumulants, $b_1\left(\mathbf{X}_n\mathbf{A}^\top\right)$ converges almost surely to zero: $b_1\left(\mathbf{X}_n\mathbf{A}^\top\right) \xrightarrow{a.s.} 0$. Since $\{b_1(\mathbf{X}_n\mathbf{B}_n)\}$ does not converge almost surely to zero there is, almost surely, a number of sample sizes for which $b_1(\mathbf{X}_n\mathbf{B}_n)$ is greater than any preassigned positive value, and therefore some sample sizes for which $b_1(\mathbf{X}_n\mathbf{B}_n)$ is greater than $b_1\left(\mathbf{X}_n\mathbf{A}^\top\right)$: $P\left(\bigcup_{n=1}^{\infty}\{b_1(X_nA^T) < b_1(X_nB_n)\}\right) = 1$. On the other hand, $\mathbf{X}_n\mathbf{B}_n$ minimizes Mardia's skewness among all q-dimensional projections of $\mathbf{X}_n$: $b_1(\mathbf{X}_n\mathbf{B}_n) \leq b_1\left(\mathbf{X}_n\mathbf{A}^\top\right)$. The two inequalities above are mutually inconsistent, unless the sequence of skewnesses $\{b_1(\mathbf{X}_n\mathbf{B}_n)\}$ converges almost surely to zero. Since Mardia's skewness attains its minimum value, that is zero, only if all third-order cumulants equal zero, the sequence $\{\mathbf{X}_n\mathbf{B}_n\}$ converges to a random vector with null third-order cumulants, that is a weakly symmetric random vector. □

References

Archimbaud A, Nordhausen K, Ruiz-Gazen A (2018) ICS for multivariate outlier detection with application to quality control. Comp Statist Data Anal 128:184–199

Arevalillo JM, Navarro H (2015) A note on the direction maximizing skewness in multivariate skew−t vectors. Stat Probab Lett 96:328–332

Arevalillo JM, Navarro H (2019) A stochastic ordering based on the canonical transformation of skew-normal vectors. TEST 28:475–498

Arevalillo JM, Navarro H (2020) Data projections by skewness maximization under scale mixtures of skew-normal vectors. Adv Data Anal Classif 14:435–461

Arevalillo JM, Navarro H (2021) Skewness-kurtosis model-based projection pursuit with application to summarizing gene expression data. Mathematics 9:954

Arevalillo JM, Navarro H (2021) Skewness model based projection pursuit as an eigenvector problem. Symmetry 13:1056

Azzalini A, Dalla Valle A (1996) The multivariate skew-normal distribution. Biometrika 83:715–726

Balakrishnan N, Scarpa B (2012) Multivariate measures of skewness for the skew-normal distribution. J Multivar Anal 104:73–87

Baringhaus L, Henze N (1991) Limit distributions for measures of multivariate skewness and kurtosis based on projections. J Multivar Anal 38:51–69

Campbell N, Mahon R (1974) A multivariate study of variation in two species of rock crab of genus leptograpsus. Aust J Zool 22:417–425

Daszykowski M (2007) From projection pursuit to other unsupervised chemometric techniques. J Chemom 21:270–279

De Luca G, Loperfido N (2015) Modelling multivariate skewness in financial returns: a SGARCH approach. Eur J Financ 21:1113–1131

Doss N, Wu Y, Yang P et al (2023) Optimal estimation of high-dimensional location gaussian mixtures. Ann Stat 51:62–95

Ferguson TS (1961) On the rejection of outliers. In: of the University of California SL (ed) Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, pp 253–287

Flury B (1988) Common principal components and related multivariate models. Wiley, New York
Friedman J, Tukey J (1974) A projection pursuit algorithm for exploratory data analysis. IEEE Trans Comput C–23:881–889
Greselin F, Ingrassia S, Punzo A (2011) Assessing the pattern of covariance matrices via an augmentation multiple testing procedure. Stat Methods Appl 20:141–170
Hennig C (2004) Asymmetric linear dimension reduction for classification. J Comput Graph Stat 13:930–945
Hou S, Wentzell P (2014) Re-centered kurtosis as a projection pursuit index for multivariate data analysis. J Chemomet 370–384
Huber P (1985) Projection pursuit (with discussion). Ann Stat 13:435–475
Hui G, Lindsay B (2010) Projection pursuit via white noise matrices. Sankhya B 72:123–153
Jones MC, Sibson R (1987) What is projection pursuit? (with discussion). J Roy Stat Soc A 150:1–37
Kim H, Kim C (2017) Moments of scale mixtures of skew-normal distributions and their quadratic forms. Commun Stat Theory Methods 46:1117–1126
Kuriki S, Takemura A (2008) The tube method for the moment index in projection pursuit. J Statist Plann Inf 138:2749–2762
Lim LH (2005) Singular values and eigenvalues of tensors: a variational approach. In: First international workshop on computational advances in multi-sensor adaptive processing
Lindsay B, Yao W (2012) Fisher information matrix: a tool for dimension reduction, projection pursuit, independent component analysis, and more. Can J Stat 40:712–730
Loperfido N (2010) Canonical transformations of skew-normal variates. TEST 19:146–165
Loperfido N (2013) Skewness and the linear discriminant function. Stat Probab Lett 83:93–99
Loperfido N (2014) Linear transformations to symmetry. J Multivar Anal 129:186–192
Loperfido N (2015) Singular value decomposition of the third multivariate moment. Linear Algebra Appl 473:202–216
Loperfido N (2015) Vector-valued skewness for model-based clustering. Stat Probab Lett 99:230–237
Loperfido N (2017) A new kurtosis matrix, with statistical applications. Linear Algebra Appl 512:1–17
Loperfido N (2018) Skewness-based projection pursuit: a computational approach. Comput Stat Data Anal 120:42–57
Loperfido N (2019) Finite mixtures, projection pursuit and tensor rank: a triangulation. Adv Data Anal Classif 13:145–173
Loperfido N (2023) Kurtosis removal for data pre-processing. Adv Data Anal Classif 17:239–267
Machado S (1983) Two statistics for testing for multivariate normality. Biometrika 70:713–718
Magnus J, Neudecker H (1979) The commutation matrix: some properties and applications. Ann Stat 7:381–394
Malkovich J, Afifi A (1973) On tests for multivariate normality. J Am Stat Assoc 68:176–179
Manly B, Navarro Alberto J (2016) Multivariate statistical methods: a primer. Chapman & Hall/CRC, New York
Mardia K (1970) Measures of multivariate skewness and kurtosis with applications. Biometrika 57:519–530
Mardia K, Kent J, Bibby J (1979) Multivariate analysis. Academic Press, London
Naito K (1997) A generalized projection pursuit procedure and its significance level. Hiroshima Math J 27:513–554
Pereira JM, Kileel J, Kolda TG (2022) Tensor moments of gaussian mixture models: theory and applications. arXiv:2202.06930
Qi L (2005) Eigenvalues of a real supersymmetric tensor. J Symb Comput 40:1302–1324
Rao Jammalamadaka S, Taufer E, Terdik G (2021) Asymptotic theory for statistics based on cumulant vectors with applications. Scand J Stat 48:708–728
Ray S (2010) Discussion of Projection pursuit via white noise matrices, by G. Hui and B. Lindsay. Sankhya B 72:147–151
Street W, Wolberg W, Mangasarian O (1993) Nuclear feature extraction for breast tumor diagnosis. In: Proceedings SPIE 1905, biomedical image processing and biomedical visualization 1905, pp 861–870
Sturmfels B (2016) Tensors and their eigenvectors. Not Am Math Soc 63:604–606
Sun J (1993) Tail probabilities of the maxima of gaussian random fields. Ann Probab 21:34–71
Tarpey T, Loperfido N (2015) Self-consistency and a generalized principal subspace theorem. J Multivar Anal 133:27–37