Self-consistency and a generalized principal subspace theorem

Thaddeus Tarpey a,⁎, Nicola Loperfido b

a Department of Mathematics and Statistics, Wright State University, 120 MM Building, Dayton, OH, 45435, USA
b Dipartimento di Economia, Società e Politica, Università degli Studi di Urbino, “Carlo Bo”, Via Saffi 2, 61029 Urbino (PU), Italy

HIGHLIGHTS

• A principal subspace theorem for multivariate mixture distributions is proved based on the notion of self-consistency.
• The results are used to characterize principal points for multivariate skew-normal distributions.
• The results are used in projection pursuit to find projections of multivariate data that deviate from normality.

ARTICLE INFO

Article history:
Received 21 October 2013
Available online 19 September 2014

AMS subject classifications:
primary 62H30
secondary 62H25

Keywords:
Location mixtures
Principal points
Projection pursuit
Skewness
Skew-normal distribution
Spherical distribution

ABSTRACT

Principal subspace theorems deal with the problem of finding subspaces supporting optimal approximations of multivariate distributions. The optimality criterion considered in this paper is the minimization of the mean squared distance between the given distribution and an approximating distribution, subject to some constraints. Statistical applications include, but are not limited to, cluster analysis, principal components analysis and projection pursuit. Most principal subspace theorems deal with elliptical distributions or with mixtures of spherical distributions. We generalize these results using the notion of self-consistency. We also show their connections with the skew-normal distribution and projection pursuit techniques. We also discuss their implications, with special focus on principal points and self-consistent points. Finally, we access the practical relevance of the theoretical results by means of several simulation studies.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

One of the overarching goals of statistics is to obtain a useful summary of data by means of well-chosen statistics. In multivariate data analysis, this goal is often achieved via dimension reduction, using tools such as principal components and projection pursuit. Just as the mean is frequently used to summarize an entire distribution by a single point (i.e. a measure of location), cluster analysis generalizes the mean from one to several points. The cluster means serve as prototypical points that represent the heterogeneity within a probability distribution. Cluster analysis is related to the problem of optimally stratifying a probability distribution whereby the cluster means determine a partition or stratification of the population. These well-known statistical methods, as well as other methods, are related to the notion of self-consistency [45]. In the context of this paper, given a random vector \( \mathbf{X} \) of interest, the notion of self-consistency basically refers to a random vector, say \( \mathbf{Y} \), that provides a summarization of \( \mathbf{X} \) (e.g. via dimension reduction or an optimal partitioning of the support of \( \mathbf{X} \)).
First, we begin with some definitions, starting with the notion of self-consistency. A very broad definition is to say a random vector \( Y \) is self-consistent for a random vector \( X \) if \( E[X|Y] = Y \) almost surely. The following definition is more restrictive but encompasses most applications of self-consistency:

**Definition 1.1.** For two jointly distributed \( p \)-variate random vectors \( X \) and \( Y \), we say that \( Y \) is self-consistent for \( X \) if \( E[X|Y] = Y \) almost surely where the support of \( Y \) spans a linear subspace, say \( \delta \), of dimension \( q \leq p \) and that \( Y \) is measurable with respect to the projection of \( X \) on \( \delta \).

This notion of self-consistency includes not only cluster means and principal components, but also principal curves [24], principal variables [38], and others.

Applications of self-consistency that we shall focus on primarily are principal points [14] and self-consistent points [15]:

**Definition 1.2.** Let \( X \) denote a \( p \)-variate random vector. Consider a set of \( k \) distinct points \( \{y_1, \ldots, y_k\} \) in \( \mathbb{R}^p \). Define \( Y = y_j \) if \( \|X - y_j\| < \|X - y_h\|, \ h \neq j \). If \( Y \) is self-consistent for \( X \), then the points \( y_1, \ldots, y_k \) are called \( k \) self-consistent points for \( X \).

**Definition 1.3.** Given a set of \( k \) distinct points \( \xi_1, \ldots, \xi_k \), define \( Y = \xi_j \) if \( \|X - \xi_j\| < \|X - \xi_h\|, \ h \neq j \). Then the points \( \xi_1, \ldots, \xi_k \) are called \( k \) principal points for \( X \) if \( E\|X - Y\|^2 \leq E\|X - Y^*\|^2 \) where \( Y^* \) is any other random vector with support consisting of at most \( k \) points.

For a given value of \( k \), a distribution can have several different sets of \( k \) self-consistent points [e.g. see [42]] and it is also possible to have more than one set of \( k \) principal points. [47] proved that the mean of a distribution must lie in the convex hull of any set of \( k \) self-consistent points. Without loss of generality, we shall assume throughout that the mean of the distribution under consideration is zero. We shall also assume that the underlying distribution \( X \) is continuous and has finite second moments. Thus, the inequality sign in Definition 1.2 can be changed to less-than-or-equal without effecting the definition. [15] showed that principal points must be essentially self-consistent points. Basically, principal points are cluster means for theoretical distributions. The cluster means from the standard \( k \)-means algorithm [e.g. [22,23,33]] are self-consistent points for the empirical distribution and represent nonparametric estimators of the principal points of the underlying distribution.

Typically, analytical solutions for the principal points (and other self-consistent objects) are not available, particularly for multivariate distributions. The search for self-consistent objects, such as principal points, becomes much easier if the search can be confined to a smaller dimensional subspace. Principal subspace theorems stipulate that the support of a self-consistent approximation lies in some particular subspace, and they are the primary theoretical motivation for this paper.

The rest of the paper is organized as follows. Section 2 provides reviews of the main results on principal subspaces, with special emphasis on principal points. Section 3 contains the main results, dealing with self-consistency and with mixtures spherical distributions. Section 4 shows their connections with the skew-normal distribution and skewness-based projection pursuit. The simulation study in Section 5 assesses the usefulness of the latter method for estimating principal subspaces. Section 6 provides a simulation example that illustrates the connection between a principal component axis and self-consistency using the theoretical results in Section 3. Section 7 discusses the paper’s results, their limitations and gives some hints for future research.

### 2. Prior results

[14] was the first to conjecture a principal subspace theorem that provided a connection between principal points and principal component subspaces. In particular, it was conjectured that for elliptical distributions, if \( k \) principal points span a space of dimension \( q < p \), then this linear space coincides with the space spanned by the \( q \) eigenvectors of the covariance matrix associated with the \( q \) largest eigenvalues. [14] proved the conjecture for \( k = 2 \) principal points, and the conjecture was proved for any value of \( k \) by [47] who were the first to use the term principal subspace theorem. In particular, they proved that if \( k \) self-consistent points of an elliptical distribution span a space of dimension \( q < p \), then this space must be spanned by \( q \) eigenvectors of the covariance matrix (i.e. be a principal component space) and for principal points, this space must be spanned by the \( q \) eigenvectors of the covariance matrix associated with the largest eigenvalues. The principal subspace theorem was extended to the infinite dimensional realm by [46], who proved the theorem for Gaussian random functions.

A general form of the principal subspace theorem was proved by [45, Theorem 4.1] for distributions with a linear conditional expectation (which includes elliptical distributions) that states that if \( Y \) is self-consistent for \( X \) and the support of \( Y \) spans a linear subspace \( \delta \) of dimension \( q < p \), then \( \delta \) is spanned by \( q \) eigenvectors of the covariance matrix of \( X \).

Since these publications, other principal subspace theorems have appeared in the literature for non-elliptical distributions, primarily mixture distributions. First, [49] proved a principal subspace result for \( k = 2 \) principal points for multivariate location mixtures of spherically symmetric distributions. This result for \( k = 2 \) was extended to a broader class of mixture distributions by [28], [35] provided another extension for \( k = 2 \) for general location mixtures of spherically symmetric distributions. Following up on this, in [36], a principal subspace result was proved for an arbitrary number of principal points for mixtures of spherically symmetric distributions defined as

\[
X = \mu + V + U,
\]  
(1)

where \( V \) is a spherically symmetric \( p \)-dimensional random vector with covariance matrix \( \sigma^2 I \), and \( U \) is an arbitrary \( p \)-dimensional random vector with mean zero and finite second moments, independent of \( V \). They assume the support
of \( \mathbf{U} \) lies in the span of a \( p \times r \) matrix \( \mathbf{M} \) whose rank is \( r \) in which case \( \mathbf{U} = \mathbf{M} \mathbf{u} \) for a \( q \times 1 \) random vector \( \mathbf{u} \), with \( \mathbf{u} \neq \mathbf{0} \). This class contains the family of location mixtures of spherically symmetric distributions. Their proof follows from their Theorem 3 which says that given any set of \( k \) points spanning a subspace of dimension \( q \leq r \), there will exist another set of points spanning a subspace of dimension \( q \) in the span of \( \mathbf{M} \) that have a lower mean squared error (mse). Here, mean squared error is the expected squared distance between \( \mathbf{X} \) and the nearest of the \( k \) support points. A similar result, proved by [41] for elliptically symmetric random vectors, states that for any set of arbitrary points lying on a line, the mse decreases if the points are rotated onto the first principal component axis — this result was later extended from a finite set of points to all points on a line [43]. More recently, [37] proved that if \( k \) principal points are constrained to lie in an \( m \) dimensional subspace, then this \( m \)-dimensional subspace must coincide with the subspace spanned by the first \( m \) eigenvectors of the covariance matrix for elliptical distributions.

3. Main results

In this section, we prove a generalized principal subspace theorem using the notion of self-consistency which provides a generalization of the main result of [36] when the support of the self-consistent approximation is one-dimensional.

Without loss of generality, we assume the distributions under consideration are centered at zero and we will consider random vectors \( \mathbf{X} \) such that

\[
\mathbf{X} = \mathbf{V} + \mathbf{U},
\]

(2)

with \( E[\mathbf{V}] = E[\mathbf{U}] = \mathbf{0} \) and \( \text{supp}(\mathbf{U}) \subseteq \text{span}(\mathbf{M}) \), where \( \mathbf{M} \) is a \( p \times q \) matrix with orthonormal columns, and \( \mathbf{V} \) has a \( p \)-dimensional spherical distribution with covariance matrix \( \sigma^2 \mathbf{I} \). Then we can write \( \mathbf{U} = \mathbf{M} \mathbf{u} \) for a \( q \times 1 \) non-degenerate random vector \( \mathbf{u} \), and (2) becomes

\[
\mathbf{X} = \mathbf{V} + \mathbf{M} \mathbf{u}.
\]

(3)

Let \( \mathbf{M}_1 \) denote a \( p \times (p - q) \) matrix so that \([ \mathbf{M} : \mathbf{M}_1 ]\) is orthogonal.

Suppose \( \mathbf{Y} \) is self-consistent for \( \mathbf{X} \) (using Definition 1.1) with support of dimension \( q^* < p \). Then, by definition, \( \mathbf{Y} \) is measurable with respect to the projection of \( \mathbf{X} \) in this subspace. Let \( \mathbf{A}_1 \) denote a \( p \times q^* \) matrix with orthonormal columns that span the support of \( \mathbf{Y} \). Then, by self-consistency, \( \mathbf{Y} \) is measurable with respect to \( \mathbf{A}_1 \mathbf{X} \) in which case the \( \sigma \)-field generated by \( \mathbf{Y} \) is \( \sigma(\mathbf{Y}) \subseteq \sigma(\mathbf{A}_1 \mathbf{X}) \). In the derivations below, almost sure (a.s.) equalities will simply be denoted with an equal sign for notational convenience. Additionally, we assume that a self-consistent \( \mathbf{Y} \) is non-trivial, i.e. \( \mathbf{Y} \neq \mathbf{0} \).

The main results presented below for random vectors \( \mathbf{X} \) in the class (2). Specifically, if a random vector is self-consistent for the projection of \( \mathbf{X} \) onto \( \mathbf{M} \) or \( \mathbf{M}_1 \), then the random vector is self-consistent for \( \mathbf{X} \). The main result is that if \( q = 1 \) in (2), then the only \( 1 \)-dimensional self-consistent approximations to \( \mathbf{X} \) must have support \( \mathbf{M} \) or the support lies in the span of \( \mathbf{M}_1 \).

Before proving the main result, we first give a couple of lemmas. The first lemma generalizes a result that was proved for elliptical distributions in [47, Theorem 3.2, page 107]. Since \( \mathbf{V} \) is spherical, without loss of generality, we can partition \( \mathbf{X} \) as

\[
\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 + \mathbf{U}_2 \end{pmatrix},
\]

(4)

where \( \mathbf{V}_1 \) has dimension \( p - q \) and \( \mathbf{U} = (\mathbf{0}', \mathbf{U}_1')' \).

**Lemma 1.** If \( \mathbf{Y}_2 \) is self-consistent for \( \mathbf{V}_2 + \mathbf{U}_2 \) in (4), then \( (\mathbf{0}', \mathbf{Y}_1')' \) is self-consistent for \( \mathbf{X} \) defined by (2). Also, if \( \mathbf{Y}_1 \) is self-consistent for \( \mathbf{V}_1 \), then \( (\mathbf{Y}_1', \mathbf{0}')' \) is self-consistent for \( \mathbf{X} \).

**Proof.** Since \( \mathbf{Y}_2 \) is self-consistent for \( \mathbf{V}_2 + \mathbf{U}_2 \), then \( \mathbf{Y}_2 \) is measurable with respect to \( \mathbf{V}_2 + \mathbf{U}_2 \) and

\[
E[\mathbf{V}_1 | \mathbf{Y}_2] = E[E[\mathbf{V}_1 | (\mathbf{V}_2, \mathbf{U}_2)] | \mathbf{Y}_2] = E[E[\mathbf{V}_1 | \mathbf{V}_2] | \mathbf{Y}_2] = \mathbf{0}.
\]

From this we have \( E[\mathbf{X} | \mathbf{Y}_2] = (0', \mathbf{Y}_1')' \) which proves the first part.

Similarly, if \( \mathbf{Y}_1 \) is a random vector that is self-consistent for \( \mathbf{V}_1 \), then \( \mathbf{Y}_1 \) is measurable with respect to \( \mathbf{V}_1 \) implying that \( \mathbf{Y}_1 \) is independent of \( \mathbf{U}_2 \) which in turn gives \( E[\mathbf{U}_2 | \mathbf{Y}_1] = \mathbf{0} \). Also, \( E[\mathbf{V}_2 | \mathbf{Y}_1] = E[E[\mathbf{V}_2 | \mathbf{V}_1] | \mathbf{Y}_1] = \mathbf{0} \). Thus \( E[\mathbf{X}_2 | \mathbf{Y}_1] = E[\mathbf{V}_2 + \mathbf{U}_2 | \mathbf{Y}_1] = \mathbf{0} \).

At this point, we can also note that if \( \mathbf{Y} \) is a measurable function of \( \mathbf{U} \) and self-consistent for \( \mathbf{U} \) with support of dimension \( q \), then \( \mathbf{Y} \) is independent of \( \mathbf{V} \). It follows that

\[
E[\mathbf{X} | \mathbf{Y}] = E[\mathbf{V} + \mathbf{U} | \mathbf{Y}] = E[\mathbf{V} | \mathbf{Y}] + E[\mathbf{U} | \mathbf{Y}] = \mathbf{0} + \mathbf{Y} = \mathbf{Y},
\]

by self-consistency and therefore \( \mathbf{Y} \) is self-consistent for \( \mathbf{X} \) too.

Next we give a technical lemma needed in the proof of our main result. The following are regularity conditions for the lemma as well as the theorem that follows. First assume that the spherical distribution in the definition (3) is non-degenerate
with a density function. If a random vector has a spherical distribution with a density, then by sphericity, the density has the form $h(\| \cdot \|^2)$ for some nonnegative function $h$ of a scalar variable and $h$ is called the density-generator for the spherical random vector. The marginal densities of a spherical distribution are also spherical and we assume that the marginal density generators of $V$ in (3) have bounded continuous derivatives.

Lemma 2. Suppose $X = V + Mu$ has a representation given by (3) with $q < p$ that satisfy the regularity conditions above. Let $A_i$ denote a unit vector that is not orthogonal to the columns of $M$. Set $u_1 = A_iMu$ and $v_1 = A_iV$. Then $E[u_1|v_1 + u_1]$ is non-degenerate.

Proof. Let $g_1$ be the density for $u_1$. Because $V$ is spherically symmetric, each of its marginal distributions are also spherically symmetric and the density for $v_1$ is a function of $v_1^2$ and can be expressed as $h_1(v_1^2)$ where $h_1(\cdot)$ is the density generator for $v_1$. Furthermore, it follows that the derivative $h'_1(x)$ is negative for $x > 0$ (see for example formula 2.24 on page 37 of [13]). By independence, the joint density for $u_1$ and $v_1$ is the product of $g_1$ and $h_1(\| \cdot \|^2)$. Let $w = v_1 + u_1$ in which case $w$ is a convolution and the joint density of $(u_1, w)$ is $g_1(u_1)h_1((w - u_1)^2)$. Thus,

$$E[u_1|v_1 + u_1 = c] = E[u_1|w = c] = \int u_1f(u_1|w = c)du_1 = \frac{H(c)}{f_w(c)},$$

where $f_w(\cdot)$ is the marginal density of $w$ and

$$H(c) = \int u_1g_1(u_1)h_1((c - u_1)^2)du_1$$

denote the numerator of this conditional expectation. Differentiating $H$ with respect to $c$, we get

$$\frac{\partial}{\partial c}H(c) = \int 2(c - u_1)u_1g_1(u_1)h_1'((c - u_1)^2)du_1,$$

where the interchange of differentiation and integration follows from the boundedness of $h_1'$ and applying Lebesgue’s dominated convergence theorem. Setting $c = 0$ gives

$$\frac{\partial}{\partial c}H(c)|_{c=0} = -\int 2u_1^2g_1(u_1)h_1'(u_1^2)du_1. \tag{5}$$

This last expression is the expected value of the non-negative quantity $-2u_1^2h_1'(u_1^2)$. Since $u$ is non-degenerate and $A_i'M \neq 0$, $u_1$ is non-degenerate, and the expected value in (5) is positive. Therefore, the derivative of $H$ at $c = 0$ is positive. Since $H$ has a positive derivative at zero, $H$ is not identically equal to zero and $E[u_1|v_1 + u_1 = c] = H(c)/f_w(c)$ is non-degenerate. 

Note that the condition that the marginal density generator be differentiable and bounded is not that restrictive because all marginal density generators of dimension $p - 2$ or less are guaranteed to be differentiable and take their maximum value at zero [page 37 [13]].

Now we present our main result, a generalized principal subspace theorem:

Theorem 3. Let $X$ be a $p$-variate random vector with representation (3) with $q = 1$ and satisfying the conditions of Lemma 2. If $Y$ is self-consistent for $X$ with support spanned by a unit vector $A_1 \in \mathbb{R}^p$, then either $A_1 = M$ or $A_1$ is orthogonal to $M$.

Proof. If $A_1 \perp M$, then use representation (4) and $Y = (Y', 0')'$. Since

$$Y = E[X|Y] = \begin{pmatrix} E[V_1|Y_1] \\ E[V_2 + U_2|Y_1] \end{pmatrix},$$

if $Y_1$ is self-consistent for $V_1$ then $Y_1$ is a measurable function of $V_1$ and hence independent of $U_2$ in which case $E[U_2|Y_1] = 0$. Also, from Lemma 1, $E[V_2|Y_1] = 0$. This gives

$$E[X|Y] = \begin{pmatrix} E[V_1|Y_1] \\ E[V_2 + U_2|Y_1] \end{pmatrix} = \begin{pmatrix} E[V_1|Y_1] \\ 0 \end{pmatrix} = \begin{pmatrix} Y_1 \\ 0 \end{pmatrix} = Y,$$

showing $Y$ is self-consistent for $X$ when the support of $Y$ is orthogonal to the support of $U$.

Now suppose that $A_1$ is not orthogonal to the column span of $M$. Define $A_2$ so that $A = [A_1 : A_2]$ is a $p \times p$ orthogonal matrix. Since $A_2'Y = 0$, by self-consistency we have $Y = E[X|Y]$ and thus

$$AY = \begin{pmatrix} A_1'Y \\ A_2'Y \end{pmatrix} = \begin{pmatrix} E[A_1'X|Y] \\ E[A_2'X|Y] \end{pmatrix},$$

which implies

$$E[A_2'(V + Mu)|Y] = 0.$$
Since the $\sigma$-field $\sigma(Y) \subset \sigma(A_1, V, A_1, U)$ we have
\[ 0 = E[A_1'(V + Mu)|Y] = E[E[A_1'V|A_1', V, A_1, U]|Y] + E[E[A_1'Mu|A_1', V, A_1, U]|Y]. \]
Independence of $U$ and $V$ [e.g. 29, Proposition 6.1.7, page 366] implies that
\[ E[A_2'V|A_1', V, A_1, U] = E[A_2'V|A_1']. \]
and since $V$ is spherical, it follows that $E[A_2'V|A_1'] = 0$ and
\[ A_2'M(u|Y) = 0. \]
By Lemma 2, $E[A_1'Mu|Y]$ is a non-degenerate random variable and if $A_1'M \neq 0$, it follows that $E[u|Y] \neq 0$. Therefore $A_1'M = 0$ and $A_1 = M$. ■

Lemma 1 indicates where the support space for self-consistent approximations, such as lines, planes, and self-consistent curves [24], may lie for distributions that satisfy (2). Theorem 3 provides a generalized version of the principal subspace theorem proved in [36] for one-dimensional self-consistent approximations. In particular, the principal subspace theorem in [36] is a result for sets of principal points only. In fact, the result of [36] follows as a corollary to Theorem 3 (see below) when a set of principal points do lie on a line. The principal subspace Theorem 3 presented here applies not only to principal points, but to any self-consistent approximation on the support line. This is highlighted with the example presented in Section 6. The proof presented here is based on a geometrical approach that uses the notion of self-consistency which is a different approach than the proof given by [36].

The following example illustrates that when a self-consistent projection has dimension greater than one, it is possible that its support may not reside exclusively in the span of $M$ from (3) nor its complement. Thus, in order to generalize Theorem 3 to higher dimensions, additional constraints may be needed.

Example. Consider a random vector $X$ that has a distribution of the form (2) for $p = 3$ where $U$ has equally weighted support points
\[ (c, 0, 0), (-c, 0, 0), (0, c, 0), \text{ and } (0, -c, 0) \]
for some $c > 0$ in the $(x, y, z)$ coordinate system; let $V$ be an arbitrary 3-dimensional spherical distribution. Then the support of $U$ is the $x$-$y$ plane. Consider the projection of $X$ onto the plane spanned by the $z$-axis (with support vector $\alpha_1 = (0, 0, 1)$) and the $45^\circ$ diagonal line across the $x$-$y$ plane with support vector $\alpha_2 = (1, 1, 0)/\sqrt{2}$. Let $P$ be the orthogonal projection onto the space spanned by $\alpha_1$ and $\alpha_2$. Then it follows that $Y = PX$ is self-consistent for $X$ although the support of $Y$ does not coincide with $M$, the $x$-$y$ plane from (3), nor $M_\perp$.

Corollary 4. For models of the form (2) with $q = 1$ that satisfy the regularity conditions of Theorem 3, if $k$ principal points lie on a line, they lie on the line spanned by $M$.

Proof. By the above mentioned theorem, the $k$ principal points, being self-consistent points, must either lie on the line spanned by $M$ or lie on a line in the plane perpendicular to $M$. Suppose the $k$ principal points lay on a line spanned by a unit vector $\alpha_1 \perp M$. We will show that this leads to a contradiction. If it holds, then the principal points can be expressed as $y_j\alpha_1$, for some scalars $y_j$, $j = 1, \ldots, k$. Let $Y$ be the self-consistent $k$-principal point approximation to $X$, i.e. $Y = y_1\alpha_1$ if $X$ is closer to $y_1\alpha_1$ than any of the other principal points. Then $Y = y_1\alpha_1$ where $y = y_j$ if $X$ is closest to the $j$th principal point.

The mean squared error for the $k$ principal point approximation to $X$ is
\[ E\|(X - Y)^2\| = E[(V + Mu - y_1\alpha_1)'\|V + Mu - y_1\alpha_1\|] = E\|(V - y_1\alpha_1)'\|V - y_1\alpha_1\| + (Mu)'\|Mu + (Mu)'V\| = E\|(V - y_1\alpha_1\| + E\|u\|^2M'M = E\|V - y_1\alpha_1\|^2 + \text{var}(u)^2, \]

since $\alpha_1$ is orthogonal to $M$ and because $u$ and $V$ are independent of each other, both with mean zero. Because $V$ is spherically symmetric, $E\|V - y_1\alpha_1\|^2 = E\|V - yM\|^2$. Now let $y_j', j = 1, \ldots, k$, denote the $k$ principal points of the projection of $X$ onto $M$ and let $Y' = y_j'M$ be the self-consistent approximation to $X$ with support points $y_j'M$. Then
\[ E\|X - Y^\ast\|^2 = E\|V + Mu - Y^\ast\|^2 \leq E\|V + Mu - yM\|^2 \quad \text{(by definition of principal points)} \]
\[ < E\|V - Y\|^2 + \text{var}(u)^2 \quad \text{(by the triangle inequality)} \]
\[ = E\|X - Y\|^2 \]
which is a contradiction because $Y$ is supposed to be the best $k$-point approximation to $X$. ■

This next corollary generalizes a result of [44]:
Corollary 5. For models satisfying (2), the projection of $X$ onto the space spanned by the columns of $M$ is self-consistent for $X$. If $q = 1$ in (2), then the projection onto the first principal component axis is self-consistent for $X$.

Proof. Let $P = MM'$ denote the projection matrix onto the column space of $M$ and let $P_\perp$ be the projection matrix onto the orthogonal complement of this space. Then


since

$$E[P_\perp V|P(V + Mu)] = E[E[P_\perp V|PV, PMu]|P(V + Mu)] = 0. \blacksquare$$

4. Multivariate skewness

Self-consistency is theoretically motivated under mean squared error, which only considers first and second-order moments. Also, main results for principal points were first developed for elliptical distributions, which are multivariate symmetric. However, location mixtures of spherical distributions can cause deviations from symmetry, thus leading to skewed data (see, for example, [32]). A natural question to ask then is whether skewness and principal subspaces of these distributions are somehow related. This section addresses the issue in the parametric as well as in the nonparametric setting, by means of the multivariate skew-normal distribution and projection pursuit techniques, respectively.

[28,35,36] provided fundamental results regarding the principal subspaces of possibly infinite location mixtures of spherical distributions. Now we address the question as to whether or not their results apply to some well-known parametric families of multivariate distributions. This section addresses the problem by considering the multivariate skew-normal distribution (SN, hereafter) introduced in [4]. The pdf of a SN random vector $X$ with location parameter $\xi$, scale parameter $\Omega$ and shape parameter $\alpha$, denoted by $X \sim SN_p (\xi, \Omega, \alpha)$, is $f(x; \xi, \Omega, \alpha) = 2 \phi_p (x - \xi; \Omega) \cdot \Phi [\alpha' (x - \xi)]$ where $x, \xi, \alpha \in \mathbb{R}^p, \Phi (\cdot)$ is the cdf of a standardized normal variable and $\phi_p (x - \xi; \Omega)$ is the pdf of $N_p (\xi, \Omega)$. The main properties of the SN distribution and its generalizations can be found in [2,27,3,1] as well as in the book [19].

The increasing popularity of the SN in both theoretical and applied statistics is partly due to its tractability. For example, it is closed under affine transformations and includes the normal distribution as a special case. The popularity of the SN is also due to its flexibility. For example, it has been used as an approximating distribution, both in the univariate case [6,9], and in the multivariate case [21,10].

The SN has already appeared in model-based clustering, as a component of finite mixtures [5,7,8,18,30], as well as in infinite mixtures [48]. The following theorem shows that it also plays a role in parametric $k$-means clustering, by modeling the principal subspace.

Proposition 6. Let $X$ be a $p$-dimensional skew-normal random vector with location parameter $0$, scale parameter $I_p + \lambda\lambda'$, with $\lambda \in \mathbb{R}^p$, and shape parameter $\lambda/(1 + \lambda'\lambda)^{1/2}$. Then if $k$ principal points lie on a line, the $k$ principal points of $X$ lie on the line spanned by $\lambda$.

Proof. Let $V$ have a multivariate normal distribution $N_p (\xi, \Psi)$ and let $U$ be independently distributed as $N(0, 1)$. The vector $X = V + \lambda U$ has the distribution $SN_p (\xi + \lambda\lambda', \Psi^{-1} \lambda/(1 + \lambda' \Psi^{-1} \lambda)$ [1]. It follows that the distribution of $X$ is $SN_p (0, I_d + \lambda\lambda', \lambda/(1 + \lambda' \lambda)$ when $\xi = 0$ and $\Psi = I_d$. As a direct consequence, $X$ can be regarded as an infinite location mixture of multivariate standard normal distributions, with the support of the mixing distribution belonging to the linear subspace spanned by the shape parameter. The proof is completed by noticing that $X$ satisfies the assumptions of Corollary 4. $\blacksquare$

The main advantage of the skew-normal distribution is that it allows for maximum likelihood estimates of principal points. When parametric models are inappropriate, alternative inferential methods are called for. We propose projection pursuit [17] as a possible choice. It is a multivariate statistical technique aimed at finding interesting features of the data by means of linear projections. Since most linear projections are approximatively normal [12], projection pursuit is very often based on the maximization of some nonnormal feature of the data, as for example skewness, kurtosis or entropy [25,26]. [40,11] review previous literature on this topic.

In particular, skewness-based projection pursuit aims at finding the direction identified by the $p$-dimensional real vector $c$ which maximizes the squared third standardized cumulant of a linear projection of a $p$-dimensional random vector $X$ with
mean $\mu$, variance $\Sigma$ and finite third-order moments:

$$
\beta_1^D(X) = \max_{c \in \mathcal{M}_0^p} \frac{E^2[(c'X - c'\mu)^3]}{(c'\Sigma c)^3},
$$

(6)

where $\mathcal{M}_0^p$ is the set of all $p$-dimensional, nonnull real vectors. The scalar $\beta_1^D(X)$ was proposed by [34] as a measure of multivariate skewness. [25] showed that skewness maximization provides a valid projection pursuit criterion. In recent years, skewness based projection pursuit has been applied to estimation [31], cluster analysis [32] and normality testing [16]. The following result shows that it is also helpful in finding principal subspaces under semiparametric assumptions.

**Proposition 7.** Let $V$ be a $p$-dimensional, spherical random variable with finite third moments. Also, let $u$ be a continuous random variable centered at the origin with nonnull third moment and independent of $V$. Finally, let $X = V + Mu$, where $M$ is a $p$-dimensional real vector. Then, if $k$ principal points lie on a line, then the $k$ principal points of $X$ lie in the linear subspace which maximizes the skewness of a projection of $X$.

**Proof.** Let $\kappa_3(W)$ and $\kappa_2(W)$ denote the third and second cumulant of a random variable $W$, respectively. Also, let $c'X$ denote a projection of $X$ onto the direction of the $p$-dimensional, nonnull vector $c$. Elementary properties of cumulants, together with the assumed independence between $u$ and $V$ imply both $\kappa_3(c'X) = \kappa_3(c'V) + (c'M)^3 \kappa_3(u)$ and $\kappa_2(c'X) = \kappa_2(c'V) + (c'M)^2 \kappa_2(u)$. By assumption, $V$ is a spherical random vector with finite third-order moments, so that $c'V$ is a symmetric random variable with finite third moment, which in turn implies both $\kappa_3(c'V) = 0$ and $\kappa_3(c'X) = (c'M)^3 \kappa_3(u)$. The same assumption implies that the variance of $V$ is $\omega I_p$, where $I_p$ is the $p \times p$ identity matrix and $\omega$ is a nonnegative scalar which only depends on the distribution of $V$ and hence not on $c$. As a direct consequence, the variance of $c'V$ is $\kappa_2(c'V) = \omega c'c$ and $\kappa_2(c'X) = \omega c'c + (c'M)^2 \kappa_2(u)$. Without loss of generality we can assume that $\|c\| = 1$ and that the variance of $u$ is one, so that $\kappa_2(c'X) = \omega + (c'M)^2$. The squared skewness of a random variable $W$ might be represented as the ratio $\kappa_3^2(W) / \kappa_2^2(W)$, so that the squared skewness of $c'X$ is $\kappa_3(c'M)^3 \kappa_2(u) / [\omega + (c'M)^2]^3$, that is an increasing function of $\|c'M\|$ which attains its maximum when $c$ is proportional to $M$. The proof is completed by noticing that $X$ satisfies the assumptions of Corollary 4. ■

5. Simulation results

In this section we shall use simulations to evaluate the performance of skewness-based projection pursuit as a method for estimating the principal subspace. We simulated 10,000 samples of size $n = 50, 100, 150, 200$ and dimension $p = 2, 4, 6$ from the mixture $\pi_1N_p(0_p, I_p) + (1 - \pi_1)N_p(mI_p, I_p)$, where $\pi_1 = 0.1, 0.2, 0.3, 0.4$ and $m = 1.5, 10$. The symbols $0_p$, $I_p$ and $I_d$ denote the $p$-dimensional null vector, the $p$-dimensional vector of ones and the $p$-dimensional identity matrix, respectively. For each simulated sample, we computed the absolute value of the correlation between data projected onto the direction of the population principal subspace (identified by $1_p$) and the data projected onto the direction of its estimate, obtained by skewness maximization. The simulations in this section were performed using Matlab and the code is available from the authors.

The simulation results are illustrated in Figs. 1–3. The figures clearly show that the proposed method performs better when the data are more skewed (that is, when the mixing weights are very different from each other), the groups are well separated (that is, when the groups’ means are very different from each other), as the sample size increases (by consistency of the sample skewness matrix) and when the dimension of the data is low (so that the number of parameters to be estimated is small). These causes are arranged in decreasing order of perceived magnitude: for example, the difference between mixing weights appears to be more influential than groups’ separation. More precisely, the proposed method achieves virtually perfect accuracy when a mixing weight is either 0.1 or 0.2 and the constant $m$ is at least 5. On the contrary, the proposed method performs very poorly when the mixing weights differ by less than 0.2, and the constant $m$ is about one.

We conclude that skewness based estimate of the principal direction is appropriate whenever the data are very skewed or when well separated groups are present. Otherwise, it could be used as an exploratory tool or as an initial estimate in some other estimating procedure.

6. Self-consistent axis example

Intuitively, a principal component subspace is a good lower-dimensional approximation to a multivariate distribution when it is self-consistent [44]. Self-consistency of a principal component axis occurs when each point on the axis equals the mean of all the points in the distribution that project orthogonally onto that point. In this vein, a self-consistent line represents a generalization of the mean of a distribution from a zero-dimensional object to a one-dimensional object with an uncountable support set that provides another representation of the “center” of the distribution. Projections of elliptically symmetric random vectors into principal component subspaces are self-consistent. Conversely, [45] showed that if a random vector $Y$ is a projection of an elliptically symmetric random vector $X$ onto a $q$-dimensional hyper-plane and $Y$ is self-consistent for $X$, then this hyper-plane is spanned by a set of $q$ eigenvectors of the covariance matrix.
Fig. 1. Simulation results for the skewness-based projection pursuit: the absolute value of the correlation between data projection onto the direction of the population principal subspace versus the dimension $p$ of the data.

This section provides an example demonstrating the self-consistency of the first principal component axis for a mixture distribution of the form (2). Consider a mixture of bivariate spherical distributions with identity covariance matrices with mixture component means $\mu_j = (0, j)'$ for $j = -2, -1, 0, 1, 2$ and equal probability at each of these mixture components. Thus, in the notation above, the support of $U$ is the $y$-axis in this bivariate illustration and $V \sim N(0, I_2)$. Now we consider an initial support set for a self-consistency algorithm [43] of all points on a line making an angle $\pi/20$ from the $x$-axis. For each point $s$ on this initial line, the self-consistency algorithm computes the mean of all points on the line perpendicular to the initial line at the point $s$. The results of this algorithm are shown in Fig. 4. The R software [39] was used to perform this simulation. The initial set is the line through the mean (in this case the origin) forming the angle $\pi/20$ from the $x$-axis. The first iteration produces a curve that deviates slightly from being completely linear that has moved closer to the $y$-axis. If the projection of $X$ onto the initial line was self-consistent, then the iteration of the self-consistency algorithm would produce the exact same line, i.e. the algorithm would not change the support set and the self-consistent line would be a stationary
Fig. 3. Simulation results for the skewness-based projection pursuit: the absolute value of the correlation between data projection onto the direction of the population principal subspace versus the dimension $p$ of the data.

Fig. 4. Steps of the self-consistency algorithm when the support of $Y$ is a line through the origin.

point of the algorithm. However, the projection of $X$ onto the initial line cannot be self-consistent by Theorem 3. Thus, the algorithm continues to iterate except the algorithm was modified whereby at each subsequent iteration shown in Fig. 4, the support set was taken to be a line forming an angle with the $x$-axis equal to the angle the curve from the previous iteration makes with the $x$-axis. Thus, Fig. 4 does not show iterations of an actual self-consistency algorithm because at each step, the support set is re-calibrated to be a line through the origin. As the iterations of this modified self-consistency algorithm show, the resulting curve begins to wiggle more with the first few iterations but by the 4th iteration, the resulting curve straightens out and is essentially equal to the $y$-axis; by Corollary 5, the projection of $X$ onto the $y$-axis is self-consistent and the algorithm has converged.

Each iteration of the algorithm in this example produces a nonlinear curve. By contrast, if the underlying distribution where elliptically symmetric, then each iteration of the algorithm would produce straight line [43].
7. Discussion

The present paper contributes to the principal subspace literature in two different ways. In the first place, it generalizes previous results, using a proof based on the notion of self-consistency. It would be interesting to extend the main result to the case where \( V \) in the theorem is elliptically distributed. Secondly, connections between principal subspaces with the skew–normal distribution and projection pursuit are illuminated.

Limitations of the present paper mostly deal with the skewness-based approach to principal subspaces. In the first place, it allows for parametric inference when the \( K \) principal points are assumed to lie on a line. In the second place, skewness-based projection pursuit does not appear to lead to reliable estimates in the presence of mild group-related skewness. The first problem might be dealt with more general hidden truncation models, such as the closed skew-normal distribution [20]. This name refers to the fact that it is closed with respect to conditioning, convolution and affine transformations (see, for example, [1]). The second problem might be dealt with using other projection pursuit indices, as for example kurtosis and entropy. Also, it might be worth investigating the relationships between principal subspaces and projections onto subspaces maximizing other measures of multivariate skewness. Both lines of research are currently being investigated. We conclude that the skewness-based approach to principal subspaces is a very promising avenue to discover sources of heterogeneity in a distribution, but more research is needed before it provides both general results and reliable algorithms.

Acknowledgments

The authors would like to thank the editors and a reviewer for their helpful comments which improved this paper considerably.

References


Acknowledgments

The authors would like to thank the editors and a reviewer for their helpful comments which improved this paper considerably.
[38] G. McCabe, Principal variables, Technometrics 826 (1984) 137–144.