The Annals of Statistics
1995, Vol. 23, No. 1, 103–112

PRINCIPAL POINTS AND SELF-CONSISTENT POINTS OF ELLIPTICAL DISTRIBUTIONS

BY THADDEUS TARPEY,[1] LUNING LI AND BERNARD D. FLURY

National Institute of Standards and Technology, Indiana University and Indiana University

The k principal points of a p-variate random vector $\mathbf{X}$ are those points $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k \in \mathbb{R}^p$ which approximate the distribution of $\mathbf{X}$ by minimizing the expected squared distance of $\mathbf{X}$ from the nearest of the $\boldsymbol{\xi}_j$. Any set of k points $\mathbf{y}_1, \ldots, \mathbf{y}_k$ partitions $\mathbb{R}^p$ into "domains of attraction" $D_1, \ldots, D_k$ according to minimal distance; following Hastie and Stuetzle we call $\mathbf{y}_1, \ldots, \mathbf{y}_k$ self-consistent if $E[\mathbf{X}|\mathbf{X} \in D_j] = \mathbf{y}_j$ for $j = 1, \ldots, k$. Principal points are a special case of self-consistent points. In this paper we study principal points and self-consistent points of p-variate elliptical distributions. The main results are the following: (1) If k self-consistent points of $\mathbf{X}$ span a subspace of dimension $q < p$, then this subspace is also spanned by q principal components, that is, self-consistent points of elliptical distributions exist only in principal component subspaces. (2) The subspace spanned by k principal points of $\mathbf{X}$ is identical with the subspace spanned by the principal components associated with the largest roots. This proves a conjecture of Flury. We also discuss implications of our results for the computation and estimation of principal points.

1. Introduction. The k principal points of a p-variate random vector $\mathbf{X}$ are those points $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k \in \mathbb{R}^p$ that minimize the expected squared distance of $\mathbf{X}$ from the nearest of the $\boldsymbol{\xi}_j$ [Flury (1990)]. The term "principal points" was introduced to stress their similarity with the least-squares definition of principal components by Pearson (1901), and to distinguish them from cluster means. Methods of cluster analysis are most often understood and presented in terms of finite samples [Hartigan (1975)], while principal points deal with similar questions of partitioning and optimal representation for theoretical distributions.

Principal points have ancestors in the theory of stratified sampling; see Dalenius (1950), Dalenius and Gurney (1951) and Cox (1957). However, these authors studied only the univariate case. Zoppè (1992) gives an extensive review of the history of principal points, including their connections with stratification and clustering.

Finding k principal points of a continuous distribution usually requires iterative computations even in the univariate case [Rowe (1995), Zoppè

Received November 1992; revised March 1994.

[1]The work of Thaddeus Tarpey is a contribution of the National Institute of Standards and Technology and is not subject to copyright in the United States.

AMS 1991 *subject classifications*. Primary 62H30; secondary 62H05, 62H25.

Key words and phrases. k-means cluster analysis, normal distribution, principal components, uniform distribution.

(1992)]. However, the computations become prohibitive for multivariate distributions, and hence it is desirable to establish theoretical results that allow us to reduce the dimensionality. The current paper reflects the progress in the theory of principal points since Flury (1990) published some initial results and conjectures. Further results are given in Tarpey (1992).

Principal points have some theoretical appeal: they pose challenging mathematical problems with no "standard" way to solve. They are also practically useful because they lead to new statistical procedures: Flury (1993) discusses the application of principal points to a problem of finding optimal sizes and shapes of protection masks. Flury and Tarpey (1993) use principal points to define "representative curves" from a large collection of curves, similar to the technique of Jones and Rice (1992).

Principal points are special cases of self-consistent points, a notion that was inspired by the self-consistent curves of Hastie and Stuetzle (1989). Self-consistent points (see Section 2 for an exact definition) are conditional means over subsets in a partition of the support of a random variable. In the univariate case such partitions are intervals, the endpoints of the intervals being the midpoints between two self-consistent points. This fact can be used for the efficient calculation of principal points of univariate distributions, as in Dalenius (1950), Dalenius and Gurney (1951), Cox (1957), Zoppè (1992) and Rowe (1995). Self-consistent points are an important concept because the k-means algorithm [Hartigan (1975), Hartigan and Wong (1979)] converges by definition to a set of self-consistent points of a sample, but not necessarily to the set which minimizes the average squared minimal distance. It also turns out that many results are just as easy or difficult to obtain for self-consistent points as for principal points, as we shall see shortly.

This article is organized as follows. In Section 2 we give formal definitions, review the relevant theory and give some preliminary results. In Section 3 we state and prove the "principal subspace theorem," which says that self-consistent points of elliptical distributions exist only in principal component subspaces. In Section 4 we treat the special case of principal points and prove that the subspace spanned by k principal points of an elliptical distribution is the subspace of the *first* principal components. In Section 5 we give some numerical results on the multivariate normal distribution and the uniform distribution inside an ellipsoid to illustrate the theory. Finally, Section 6 offers a discussion and outlook on unresolved problems.

2. Preliminaries. Throughout this paper, $\mathbf{X}$ will denote a p-variate random vector, and $F(\cdot)$ its distribution function. Whenever needed, it will implicitly be assumed that all first or second moments are finite. The Euclidean norm of $\mathbf{x} \in \mathbb{R}^p$ will be denoted by $\|\mathbf{x}\| = (\mathbf{x}'\mathbf{x})^{1/2}$.

The following is a summary of relevant definitions and preliminary results from Flury (1990, 1993). For a set of k points $W = \{\mathbf{y}_1, \ldots, \mathbf{y}_k\}$, all $\mathbf{y}_j \in \mathbb{R}^p$, the *minimal distance* of $\mathbf{x} \in \mathbb{R}^p$ to W is denoted by $d(\mathbf{x}|\mathbf{y}_1, \ldots, \mathbf{y}_k) = \min_{1 \le j \le k} \|\mathbf{x} - \mathbf{y}_j\|$. The vectors $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k \in \mathbb{R}^p$ are called *k principal points* of $\mathbf{X}$ if they minimize the expected squared minimal distance over all sets of k

points in $\mathbb{R}^p$. We will write $P_{\mathbf{X}}(k) := E[d^2(\mathbf{X}|\xi_1, \ldots, \xi_k)]$ for the minimum, which may be considered as the loss in approximating the distribution of $\mathbf{X}$ by its k principal points. For a set $W = \{\mathbf{y}_1, \ldots, \mathbf{y}_k\}$, the *domain of attraction* D_j of $\mathbf{y}_j$ consists of all $\mathbf{x} \in \mathbb{R}^p$ which have $\mathbf{y}_j$ as their nearest point in W; the boundary of a domain of attraction is also known as the Voronoi or Dirichlet polygon. Since we consider only continuous distributions in this article, we will not worry about the boundaries between different domains of attraction, because they have probability zero. A set $W = \{\mathbf{y}_1, \ldots, \mathbf{y}_k\}$ is called *self-consistent* for the random vector $\mathbf{X}$ if $E[\mathbf{X}|\mathbf{X} \in D_j] = \mathbf{y}_j$ for all $j = 1, \ldots, k$. Thus a set of k points is self-consistent if each of the points is a conditional mean, given that $\mathbf{X}$ is in the respective domain of attraction. For simplicity we will often refer to the points $\mathbf{y}_j$ as self-consistent, meaning that the set is self-consistent. Principal points are self-consistent [Flury (1993)], but the converse is not necessarily true. If $\mathbf{X}$ has m different sets of k self-consistent points, then those sets which minimize the expected squared minimal distance are sets of principal points.

Our first lemma generalizes a result of Flury [(1990), page 38] from principal points to self-consistent points.

LEMMA 2.1. *If $\mathbf{X}$ is a p-variate random vector with a self-consistent set of points $\{\mathbf{y}_1, \ldots, \mathbf{y}_k\}$, then $E[\mathbf{X}]$ is in the convex hull of $\mathbf{y}_1, \ldots, \mathbf{y}_k$.*

PROOF. Let $\bigcup_{j=1}^k D_j$ denote a partition of $\mathbb{R}^p$, then $E[\mathbf{X}] = \sum_{j=1}^k \int_{D_j} \mathbf{x}\, dF(\mathbf{x}) = \sum_{j=1}^k \pi_j E[\mathbf{X}|\mathbf{X} \in D_j]$, where $\pi_j = \Pr[\mathbf{X} \in D_j]$. If the D_j are domains of attraction associated with k self-consistent points $\mathbf{y}_1, \ldots, \mathbf{y}_k$, then $E[\mathbf{X}|\mathbf{X} \in D_j] = \mathbf{y}_j$, and $E[\mathbf{X}] = \sum_{j=1}^k \pi_j \mathbf{y}_j$. □

Thus the linear manifold spanned by k self-consistent points has dimension at most $k - 1$.

LEMMA 2.2. *Let $\mathbf{X}_1$ denote a p-variate random vector, and let $\mathbf{X}_2 = \boldsymbol{\delta} + \rho\mathbf{H}\mathbf{X}_1$ for some $\boldsymbol{\delta} \in \mathbb{R}^p$, $\rho \in \mathbb{R}$ and some orthogonal matrix $\mathbf{H}$ of dimension $p \times p$.*

(a) *If $\{\mathbf{y}_1, \ldots, \mathbf{y}_k\}$ is a set of k self-consistent points of $\mathbf{X}_1$, then $\boldsymbol{\delta} + \rho\mathbf{H}\mathbf{y}_j$, $j = 1, \ldots, k$, form a set of k self-consistent points of $\mathbf{X}_2$.*

(b) *If $\xi_1, \ldots, \xi_k$ are principal points of $\mathbf{X}_1$, then $\boldsymbol{\delta} + \rho\mathbf{H}\xi_j$, $j = 1, \ldots, k$, are principal points of $\mathbf{X}_2$, and $P_{\mathbf{X}_2}(k) = \rho^2 P_{\mathbf{X}_1}(k)$.*

The proof is omitted [see Tarpey (1992), page 64]. The lemma allows us to assume, without loss of generality, that $E[\mathbf{X}] = \mathbf{0}$ and that the covariance matrix of $\mathbf{X}$ is diagonal. Henceforth we shall always assume $E[\mathbf{X}] = \mathbf{0}$, which means that the linear manifold spanned by k self-consistent points is a subspace of dimension at most $k - 1$.

LEMMA 2.3 [Tarpey (1992)]. *Let $\mathbf{X}$ denote a p-variate random vector with mean $\mathbf{0}$. Suppose $\mathbf{y}_1, \ldots, \mathbf{y}_k$ are k self-consistent points of $\mathbf{X}$, and $\mathbf{y}_1, \ldots, \mathbf{y}_k$*

span a subspace of dimension $q < p$. Let $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q \in \mathbb{R}^p$ denote an orthonormal basis of this subspace, and set $\mathbf{A}_1 := [\boldsymbol{\alpha}_1 : \cdots : \boldsymbol{\alpha}_q]$. Then the random vector $\mathbf{A}_1'\mathbf{X}$ has a set of k self-consistent points $\mathbf{A}_1'\mathbf{y}_1, \ldots, \mathbf{A}_1'\mathbf{y}_k$.

PROOF. Let $D_j \subset \mathbb{R}^p$ denote the domain of attraction of $\mathbf{y}_j$, and let $D_j^* \subset \mathbb{R}^q$ denote the domain of attraction of $\mathbf{A}_1'\mathbf{y}_j$, $j = 1, \ldots, k$. Then $\mathbf{X} \in D_j$ is equivalent to $\mathbf{A}_1'\mathbf{X} \in D_j^*$, and therefore $E[\mathbf{A}_1'\mathbf{X} | \mathbf{A}_1'\mathbf{X} \in D_j^*] = \mathbf{A}_1'\mathbf{y}_j$. □

The converse of Lemma 2.3 is in general not true, as the following example shows. Let $\mathbf{X}$ denote a $\mathcal{N}_2\left(\mathbf{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$ random vector, $k = 2$, $q = 1$ and $\mathbf{A}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Then $\pm(2/\pi)^{1/2}$ are two self-consistent points of $X_1 = \mathbf{A}_1'\mathbf{X}$ [see Flury (1990)], but $\pm(2\pi)^{1/2} \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ are not self-consistent points of $\mathbf{X}$ unless $\rho = 0$.

It is interesting (and also simplifies the proofs of the main results) to study discrete distributions defined on sets of self-consistent points or principal points. Here, $\mathbf{Y}$ will denote a p-variate discrete random vector, distributed jointly with $\mathbf{X}$, and $S(\mathbf{Y})$ will denote the support of $\mathbf{Y}$.

DEFINITION 2.1. The random vector $\mathbf{Y}$ is a *best k-point approximation to* $\mathbf{X}$ if $S(\mathbf{Y})$ contains exactly k distinct points $\mathbf{y}_1, \ldots, \mathbf{y}_k$, and $E[\|\mathbf{X} - \mathbf{Y}\|^2] \leq E[\|\mathbf{X} - \mathbf{Z}\|^2]$ for all $\mathbf{Z}$ whose support has at most k points.

LEMMA 2.4. *If $\mathbf{Y}$ is a best k-point approximation to $\mathbf{X}$, then the following two conditions hold:*

(i) $\|\mathbf{X} - \mathbf{Y}\| \leq \|\mathbf{X} - \mathbf{y}_j\|$ *a.s. for all* $\mathbf{y}_j \in S(\mathbf{Y})$.
(ii) $E[\mathbf{X}|\mathbf{Y}] = \mathbf{Y}$ *a.s.*

PROOF. If (i) does not hold, define a discrete random vector $\mathbf{Z}$ by $\mathbf{Z} = \mathbf{y}_j$ if $\mathbf{X} \in D_j$, where D_j is the domain of attraction of $\mathbf{y}_j$. Then $E[\|\mathbf{X} - \mathbf{Y}\|^2] > E[\|\mathbf{X} - \mathbf{Z}\|^2]$, which contradicts the assumption of the lemma. Condition (ii) follows from self-consistency of the $\mathbf{y}_j$. □

Hence each set $W = \{\mathbf{y}_1, \ldots, \mathbf{y}_k\}$ of self-consistent points defines a k-point random variable $\mathbf{Y}$, up to a set of probability zero, according to $\Pr[\mathbf{Y} = \mathbf{y}_j] = \Pr[\mathbf{X} \in D_j]$, satisfying (i) and (ii). Using this setup, Lemma 2.1 follows from $E[\mathbf{X}] = E[E\{\mathbf{X}|\mathbf{Y}\}]$; $\mathbf{Y}$ is a best k-point approximation exactly if W is a set of k principal points.

3. Self-consistent points of elliptical distributions. Suppose the p-variate random vector $\mathbf{X}$ is partitioned into q and $p - q$ components as $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$, with mean vector $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$ and covariance matrix $\boldsymbol{\Psi} = \begin{pmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{pmatrix}$ partitioned analogously. If $\mathbf{X}$ follows an elliptical distribution with finite

second moments, then the following two properites hold:

(a) $E[\mathbf{X}_2|\mathbf{X}_1] = \boldsymbol{\mu}_2 + \Psi_{21}\Psi_{11}^{-1}(\mathbf{X}_1 - \boldsymbol{\mu}_1)$ (provided $\boldsymbol{\Psi}_{11}$ is positive definite);

(b) for a matrix $\mathbf{A}$ of dimension $p \times m$, $\mathbf{A}'\mathbf{X}$ is elliptical. [See Fang, Kotz and Ng (1990) and Muirhead (1982), Chapter 1.5.]

The following theorem was first proved by Tarpey (1992) and called the principal subspace theorem.

THEOREM 3.1. *Suppose* $\mathbf{X}$ *is p-variate elliptical with* $E[\mathbf{X}] = 0$ *and* $\mathrm{Cov}(\mathbf{X}) = \boldsymbol{\Psi}$. *If* $\mathscr{V}$ *is the subspace spanned by a self-consistent set of points* $\{\mathbf{y}_1, \ldots, \mathbf{y}_k\}$ *of* $\mathbf{X}$, *then* $\mathscr{V}$ *is spanned by a set of eigenvectors of* $\boldsymbol{\Psi}$.

PROOF. Define a k-point random vector $\mathbf{Y}$ with support $S(\mathbf{Y}) = \{\mathbf{y}_1, \ldots, \mathbf{y}_k\}$ as at the end of Section 2, and let $q = \dim \mathscr{V}$. Let $\mathbf{A} := (\mathbf{A}_1 : \mathbf{A}_2)$ denote an orthogonal $p \times p$ matrix such that the q columns of $\mathbf{A}_1$ span $\mathscr{V}$. By self-consistency, we have $E[\mathbf{A}'_2\mathbf{X}|\mathbf{Y}] = \mathbf{A}'_2\mathbf{Y} = \mathbf{0}$ a.s. Using properties (a) and (b) of elliptical distributions,

$$\begin{aligned} E[\mathbf{A}'_2\mathbf{X}|\mathbf{Y}] &= E\big[\, E\{\mathbf{A}'_2\mathbf{X}|\mathbf{A}'_1\mathbf{X}\}|\mathbf{Y}\big] \\ &= E\Big[\mathbf{A}'_2\boldsymbol{\Psi}\mathbf{A}_1(\mathbf{A}'_1\boldsymbol{\Psi}\mathbf{A}_1)^{-1}\mathbf{A}'_1\mathbf{X}|\mathbf{Y}\Big] \\ &= \mathbf{A}'_2\boldsymbol{\Psi}\mathbf{A}_1(\mathbf{A}'_1\boldsymbol{\Psi}\mathbf{A}_1)^{-1}\mathbf{A}'_1\mathbf{Y} \quad \text{a.s.} \end{aligned}$$

Since $S(\mathbf{Y})$ spans $\mathscr{V}$, $S(\mathbf{A}'_1\mathbf{Y})$ spans $\mathbb{R}^q$, and therefore $\mathbf{A}'_2\boldsymbol{\Psi}\mathbf{A}_1 = \mathbf{0}$. Writing $\mathbf{P} = \mathbf{A}_1\mathbf{A}'_1$ for the projection matrix associated with $\mathscr{V}$, this implies

$$\Psi = \mathbf{A}\begin{pmatrix} \mathbf{A}'_1\boldsymbol{\Psi}\mathbf{A}_1 & 0 \\ 0 & \mathbf{A}'_2\boldsymbol{\Psi}\mathbf{A}_2 \end{pmatrix}\mathbf{A}' = \mathbf{P}\boldsymbol{\Psi}\mathbf{P} + \mathbf{A}_2\mathbf{A}'_2\boldsymbol{\Psi}\mathbf{A}_2\mathbf{A}'_2$$

and $\boldsymbol{\Psi}\mathbf{A}_1 = \mathbf{P}\boldsymbol{\Psi}\mathbf{A}_1$, that is, the columns of $\mathbf{A}_1$ are spanned by q eigenvectors of Ψ. □

A cautionary remark is perhaps in order. In the foregoing proof it is assumed that $\mathbf{A}'_1\boldsymbol{\Psi}\mathbf{A}_1$ is nonsingular, although $\boldsymbol{\Psi}$ itself may be singular. However, it is tacitly assumed that no self-consistent points are allowed to be outside $S(\mathbf{X})$ (such points would have domains of attraction with associated probability zero). Thus the subspace spanned by the columns of $\mathbf{A}_1$ will always be such that $\mathrm{Cov}(\mathbf{A}'_1\mathbf{X})$ is nonsingular. Note also that multiple eigenvalues do not affect the proof.

Before discussing the consequences of Theorem 3.1, we present another useful result.

THEOREM 3.2. *Let* $\mathbf{X}$ *denote an elliptically distributed random vector with mean* $\mathbf{0}$ *and covariance matrix* $\boldsymbol{\psi} = \mathbf{B}\boldsymbol{\Lambda}\mathbf{B}'$, *where* $\mathbf{B} = [\boldsymbol{\beta}_1 : \cdots : \boldsymbol{\beta}_p]$ *is orthogonal and* $\boldsymbol{\Lambda}$ *is diagonal. Suppose the k vectors* $\mathbf{y}_1, \ldots, \mathbf{y}_k$ *span the same subspace as the q eigenvectors* $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q$ (*which are not necessarily ordered*). *Let* $\mathbf{B}_1 := [\boldsymbol{\beta}_1 : \cdots : \boldsymbol{\beta}_q]$, *and* $\mathbf{X}^* = \mathbf{B}'_1\mathbf{X}$. *If the* $\mathbf{z}_j = \mathbf{B}'_1\mathbf{y}_j$, $j = 1, \ldots, k$, *are self-consistent points of* $\mathbf{X}^*$, *then* $\mathbf{y}_1, \ldots, \mathbf{y}_k$ *are self-consistent points of* $\mathbf{X}$.

PROOF. The proof follows standard arguments; see also Tarpey [(1992), page 47], who proves the result for a larger class of symmetric distributions. □

Theorems 3.1 and 3.2 have important consequences for the computation of self-consistent points. Suppose for instance that $\mathbf{X}$ is p-variate elliptical, $p \geq 4$, and we wish to find sets of $k = 4$ self-consistent points. By Lemma 2.1, we need to look at subspaces of dimension at most 3. By Theorem 3.1, it suffices to look at all $\binom{p}{3}$ subspaces spanned by three different eigenvectors to find three-dimensional patterns, all $\binom{p}{2}$ subspaces of dimension 2 to find two-dimensional patterns, and finally, all one-dimensional sets can be found by computing four self-consistent points of a single principal component. By Theorem 3.2, each such pattern found determines a set of four self-consistent points of the p-variate distribution. Although this may still be a considerable amount of work, it is typically much less work than finding sets of four self-consistent points in dimension p. We will give an illustrative example in Section 5.

Figure 1 gives an illustration for subspaces of dimension at most 2 spanned by four self-consistent points of a multivariate normal. Each of the

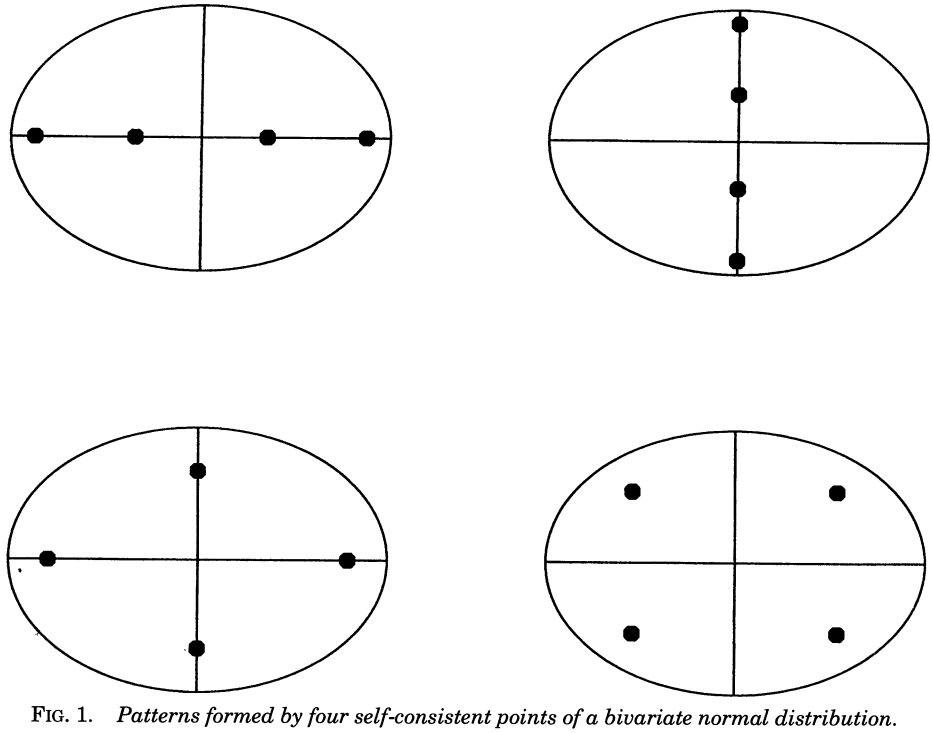

FIG. 1. *Patterns formed by four self-consistent points of a bivariate normal distribution.*

two two-dimensional patterns in the lower half of Figure 1, applied to a two-dimensional principal component subspace of a multivariate normal variable $\mathbf{X}$, generates a set of four self-consistent points, as Theorems 3.1 and 3.2 show. We currently have no proof (besides numerical evidence) that patterns other than the four shown in Figure 1 can occur for the bivariate normal.

4. Principal points of elliptical distributions. In this section we state and prove a theorem conjectured by Flury [(1990), pages 40–41].

THEOREM 4.1. *Suppose* $\mathbf{X}$ *is* p*-variate elliptical with* $E[\mathbf{X}] = \mathbf{0}$ *and* $\mathrm{Cov}(\mathbf{X}) = \Psi$. *If a set of* k *principal points of* $\mathbf{X}$ *spans a subspace* $\mathscr{V}$ *of dimension* q, *then* Ψ *has a set of eigenvectors* $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ *with associated ordered eigenvalues* $\lambda_1 \geq \cdots \geq \lambda_p$ *such that* $\mathscr{V}$ *is spanned by* $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q$.

PROOF. Without loss of generality (see Lemma 2.2) assume that $\Psi = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, where the λ_i are not necessarily ordered. Assume also that all $\lambda_i > 0$ (otherwise see the remark following Theorem 3.1, and reduce the dimension to the rank of Λ). By Theorem 3.1, $\mathscr{V}$ is spanned by q standard basis vectors of $\mathbb{R}^p$, and without loss of generality we can assume that these are the *first* q basis vectors. Thus we must show that $\lambda_j \geq \lambda_m$ for all $j \leq q$ and all $m > q$. Assume for the moment that $\mathbf{X} = (X_1, \ldots, X_p)$ has a unique set of k principal points, and let $\mathbf{Y} = (Y_1, \ldots, Y_p)'$ denote a best k-point approximation to $\mathbf{X}$ (which is uniquely determined by $\mathbf{X}$ up to a set of probability zero). Let $\mathbf{W} = (W_1, \ldots, W_p)' = \Lambda^{-1/2}\mathbf{X}$, and $\mathbf{Y}^* = (Y_1^*, \ldots, Y_p^*)' = \Lambda^{-1/2}\mathbf{Y}$. Setting $\alpha_i = E[(W_i - Y_i^*)^2]$, $i = 1, \ldots, p$, we have $E[\|\mathbf{X} - \mathbf{Y}\|^2] = \sum_{i=1}^p \lambda_i \alpha_i$. If $i > q$, then $Y_i = 0$ a.s., and therefore $\alpha_i = 1$ for all $i > q$. For $i \leq q$, Lemma 2.4(ii) shows that $E[Y_i(X_i - Y_i)] = 0$, and therefore $E(X_i^2) = E[(X_i - Y_i)^2] + E[Y_i^2]$, which implies $E[Y_i^2] = \lambda_i(1 - \alpha_i)$. If $\alpha_i = 1$, then $Y_i = 0$ a.s., which contradicts the assumption that $\dim \mathscr{V} = q$. Hence $\alpha_i < 1$ for $i = 1, \ldots, q$.

Consider now a permutation π of $\{1, \ldots, p\}$ transposing j and m. Let $\mathbf{W}_\pi = (W_{\pi(1)}, \ldots, W_{\pi(p)})'$, and note that $\mathbf{W}_\pi$ has the same spherically symmetric distribution as $\mathbf{W}$. Define $Z_i = (\lambda_i/\lambda_{\pi(i)})^{1/2} Y_{\pi(i)}$, $i = 1, \ldots, p$. Then $\mathbf{Z} = (Z_1, \ldots, Z_p)'$ is a k-point random vector, and

$$
\begin{aligned}
E\left[\|\mathbf{X} - \mathbf{Z}\|^2\right] &= \sum_{i=1}^{p} \lambda_i E\left[\left(W_i - Y^*_{\pi(i)}\right)^2\right] \\
&= \sum_{i=1}^{p} \lambda_i E\left[\left(W_{\pi(i)} - Y^*_{\pi(i)}\right)^2\right] \\
&= \sum_{i=1}^{p} \lambda_i \alpha_{\pi(i)}.
\end{aligned}
$$

Thus $E[\|\mathbf{X} - \mathbf{Z}\|^2] - E[\|\mathbf{X} - \mathbf{Y}\|^2] = (1 - \alpha_j)(\lambda_j - \lambda_m)$. If $\lambda_m > \lambda_j$, then this expression would be negative, which contradicts the assumption that $\mathbf{Y}$ is a best k-point approximation. Hence $\lambda_j \geq \lambda_m$.

If $\mathbf{X}$ has more than one set of principal points, the same proof can be applied to each such set, by defining a best k-point approximation $\mathbf{Y}$ according to the particular choice of the principal points. □

Ellipticity of $\mathbf{X}$ is a sufficient condition, but not a necessary one, for the conclusion of the theorem to hold. It is currently not known what the necessary conditions are. It is easy, though, to find (discrete or continuous) examples of nonelliptical distributions such that the subspace spanned by k principal points is not a subspace spanned by eigenvectors of the covariance matrix.

Besides its theoretical appeal, Theorem 4.1 is useful because it allows us to restrict further the search for principal points. By Lemma 2.1, k principal points span a subspace of dimension at most $k - 1$, and hence at most $k - 1$ subspaces with dimensions $k - 1$, $k - 2$ and so on must be searched. (Multiple eigenvalues may lead to unpleasant computational difficulties, which we do not elaborate upon in this article). For $k < p + 1$, one may start out by searching for all sets of k self-consistent points that span the same subspace as the first $k - 1$ eigenvectors of $\mathrm{Cov}(\mathbf{X})$, and then reduce the dimension step by step, eliminating one eigenvector at a time. If in any given step the expected squared minimum distance increases, the procedure can stop.

5. Numerical example. Omitting details of the algorithms used, we report some results found by numerical calculations in which the respective continuous distributions were approximated by discrete distributions on approximately 500,000 equispaced gridpoints inside an ellipsoid. The two distributions chosen were the trivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Lambda} = \mathrm{diag}(4, 2, 1)$, and the trivariate uniform distribution inside the ellipsoid $x_1^2/4 + x_2^2/2 + x_3^2 = 5$, which has the same covariance matrix $\mathbf{\Lambda}$. All calculations were done for $k = 4$ points. For both distributions, *no* sets of four self-consistent points spanning $\mathbb{R}^3$ were found. In the subspace of the first two variables, both the normal and the uniform examples yielded a "cross"-pattern of the form $(\pm a, 0, 0)$ and $(0, \pm b, 0)$ and a "rectangle"-pattern of the form $(\pm c, \pm d)$, see the bottom part of Figure 1. The numerical calculations yielded $a = 2.613$ and $b = 1.306$, for the normal, and $a = 2.522$ and $b = 1.511$, for the uniform. The expected squared minimum distance was 2.987 for the normal and 2.642 for the uniform. In both cases the rectangle-pattern yielded a larger expected squared distance, as did the one-dimensional solutions with four points along the first axis. Thus the numerical evidence suggests that in both cases the sets of $k = 4$ principal points form a cross-pattern with coordinates as indicated.

It is interesting that no three-dimensional patterns of four self-consistent points could be found. So far we have not been able to prove that such a pattern does not always exist, but it is clear that it exists in special cases. For instance, if $\mathbf{\Lambda}$ is proportional to $\mathbf{I}_3$, then any four points in $\mathbb{R}^3$ spanning a tetrahedron centered at the origin define a partition of $\mathbb{R}^3$ into four domains D_j according to minimal distance. All D_j have equal shape, and the four points $\mathbf{y}_j := E[\mathbf{X}|\mathbf{X} \in D_j]$ are self-consistent. Numerical calculations indicate

that the tetrahedron spanned by the $\mathbf{y}_j$ has side length approximately 1.94 for $\mathbf{X}$ normal with covariance matrix $\mathbf{I}_3$. Evidently, any rotation of the four self-consistent points in $\mathbb{R}^3$ will yield another set of self-consistent points in this case.

6. Discussion and outlook. Principal points originated in a problem of determining optimal sizes and shapes of gas masks, as described in Flury (1993). In the same article, methods of estimation of principal points were defined using our Theorem 4.1, which at that time was a conjecture. Thus the current article gives a late theoretical justification for earlier work.

Principal points and self-consistent points have much in common with k-cluster means [Hartigan (1975)]; k-cluster means of a sample are strongly consistent estimators of k principal points of a distribution [Pollard (1981)]. Methods of cluster analysis are typically viewed as purely data-oriented, with no statistical model in the background, with the pragmatic purpose of finding optimal partitions of observed data. Principal points, on the other hand, find optimal partitions of theoretical distributions. To our knowledge, only homogeneous theoretical models have been studied so far, with the exception of some univariate examples [Flury (1990), Zoppè (1992)]. It would be interesting to study principal points of theoretical distributions that reflect group structure, such as finite mixtures, for which cluster analysis is meant to work. Future developments in the theory of principal points may help to understand cluster analysis better. Alternatively, principal points may be used to define *best k-point approximations* to continuous distributions, as in Definition 2.1. This approach is interesting in itself and does not need any justification in terms of clustering or stratification.

Many challenging problems remain open. For instance, despite strong numerical evidence it is not known whether the two-dimensional patterns in Figure 1, formed by four self-consistent points of elliptical distributions, are the only possible ones, and whether the cross-pattern is always better than the rectangle-pattern. There appears to be no easy way to prove uniqueness or existence of high-dimensional patterns. Only in the univariate case are sufficient conditions for uniqueness of principal points known [Tarpey (1994)]. Other challenging problems arise in estimation of principal points: all results obtained so far refer to univariate distributions and $k = 2$ points [Tarpey (1992), Chapter 6]. Again, no standard methodology seems to be available to generalize these results.

Acknowledgments. The authors wish to thank an Associate Editor and two referees for their careful reviews which led to a much improved version of this article. In particular, the Associate Editor's suggestion of introducing the notion of *best k-point approximations* (Definition 2.1) helped significantly to simplify and shorten the proofs of Theorem 3.1 and 4.1.

REFERENCES

Cox, D. R. (1957). Note on grouping. *J. Amer. Statist. Assoc.* **52** 543–547.

Dalenius, T. (1950). The problem of optimum stratification. *Skandinavisk Aktuarietidskrift* **33** 203–213.

DALENIUS, T. and GURNEY, M. (1951). The problem of optimum stratification II. *Skandinavisk Aktuarietidskrift* **34** 203–213.

FANG, K., KOTZ, S. and NG, K. (1990). *Symmetric Multivariate and Related Distributions*. Chapman and Hall, New York.

FLURY, B. (1990). Principal points. *Biometrika* **77** 33–41.

FLURY, B. (1993). Estimation of principal points. *J. Roy. Statist. Soc. Ser. C* **42** 139–151.

FLURY, B. and TARPEY, T. (1993). Representing a large number of curves: a case for principal points. *Amer. Statist.* **47** 304–306.

HARTIGAN, J. A. (1975). *Clustering Algorithms*. Wiley, New York.

HARTIGAN, J. A. and WONG, M. A. (1979). Algorithm AS 136: a K-means clustering algorithm. *J. Roy. Statist. Soc. Ser. C* **28** 100–108.

HASTIE, T. and STUETZLE, W. (1989). Principal curves. *J. Amer. Statist. Assoc.* **84** 502–516.

JONES, M. C. and RICE, J. A. (1992). Displaying the important features of large collections of similar curves. *Amer. Statist.* **46** 140–145.

MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York.

PEARSON, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2** 559–572.

POLLARD, D. (1981). Strong consistency of k-means clustering. *Ann. Statist.* **9** 135–140.

ROWE, S. A. (1995). An algorithm for computing principal points with respect to a loss function in the unidimensional case. *Statistics and Computing*. To appear.

TARPEY, T. (1992). Principal points. Ph.D. dissertation, Dept. Mathematics, Indiana Univ.

TARPEY, T. (1994). Two principal points of symmetric, strongly unimodal distributions. *Statist. Probab. Lett.* **20** 253–258.

ZOPPÈ, A. (1992). I punti principali di distribuzioni univariate. Ph.D. dissertation, Dept. Statistics and Operations Research, Univ. Trento (Italy).

THADDEUS TARPEY
DEPARTMENT OF MATHEMATICS
AND STATISTICS
WRIGHT STATE UNIVERSITY
DAYTON, OHIO 45435

LUNING LI
ADVANCED STUDY PROGRAM
NATIONAL CENTER FOR
ATMOSPHERIC RESEARCH
BOULDER, COLORADO 80307

BERNARD D. FLURY
DEPARTMENT OF MATHEMATICS
INDIANA UNIVERSITY
BLOOMINGTON, INDIANA 47405