Statistical Science
1996, Vol. 11, No. 3, 229–243

Self-Consistency: A Fundamental Concept in Statistics

Thaddeus Tarpey and Bernard Flury

Abstract. The term "self-consistency" was introduced in 1989 by Hastie and Stuetzle to describe the property that each point on a smooth curve or surface is the mean of all points that project orthogonally onto it. We generalize this concept to self-consistent random vectors: a random vector $\mathbf{Y}$ is self-consistent for $\mathbf{X}$ if $\mathscr{E}[\mathbf{X}|\mathbf{Y}] = \mathbf{Y}$ almost surely. This allows us to construct a unified theoretical basis for principal components, principal curves and surfaces, principal points, principal variables, principal modes of variation and other statistical methods. We provide some general results on self-consistent random variables, give examples, show relationships between the various methods, discuss a related notion of self-consistent estimators and suggest directions for future research.

Key words and phrases: Elliptical distribution; EM algorithm; k-means algorithm; mean squared error; principal components; principal curves; principal modes of variation; principal points; principal variables; regression; self-organizing maps; spherical distribution; Voronoi region.

1. INTRODUCTION

One of the fundamental objectives of statistics is to summarize a distribution while retaining as much information as possible. Many statistical techniques designed to summarize or simplify data have been labeled "principal": principal components (Pearson, 1901); principal curves (Hastie and Stuetzle, 1989); principal points and self-consistent points (Flury, 1990, 1993); principal variables (McCabe, 1984); principal modes of variation for curves (Castro, Lawton and Sylvestre, 1986). All of these techniques may be based on the unifying property of self-consistency: a random vector $\mathbf{Y}$ is self-consistent for $\mathbf{X}$ if each point in the support of $\mathbf{Y}$ is the conditional mean of $\mathbf{X}$, given that $\mathbf{X}$ projects onto that point. The term "self-consistency" was inspired by Hastie and Stuetzle (1989), who defined self-consistent curves and principal curves. An earlier definition of self-consistency of estimators is due to Efron (1967); we will illustrate its relationship to our notion of self-consistency in Section 8.

In Section 2 we give a formal definition and elementary properties of self-consistent random variables, as well as technical preliminaries. In Sections 3–6 we show that several statistical techniques may be based on the property of self-consistency: regression and principal variables (Section 3), principal components (Section 4), principal modes of variation (Section 5), and self-consistent points and curves (Section 6). The orthogonal complement of a self-consistent random vector is discussed in Section 7. Section 8 reviews the concept of self-consistency in maximum likelihood estimation with incomplete data and relates it to our definition of self-consistency. Section 9 offers some discussion and points out similarities between the k-means algorithm and the EM algorithm.

2. SELF-CONSISTENT RANDOM VECTORS

Suppose we want to represent or approximate the distribution of a random vector $\mathbf{X}$ by a random vector $\mathbf{Y}$ whose structure is less complex. One measure of how well $\mathbf{Y}$ approximates $\mathbf{X}$ is the mean squared error $\mathscr{E}\|\mathbf{X}-\mathbf{Y}\|^2$. In terms of mean squared error, the approximation of $\mathbf{X}$ by $\mathbf{Y}$ can always be improved using $\mathscr{E}[\mathbf{X}|\mathbf{Y}]$ since, for any function g, $\mathscr{E}\|\mathbf{X}-\mathscr{E}[\mathbf{X}|\mathbf{Y}]\|^2 \le \mathscr{E}\|\mathbf{X}-g(\mathbf{Y})\|^2$. Taking g to be the

Thaddeus Tarpey is Assistant Professor, Department of Mathematics and Statistics, Wright State University, Dayton, Ohio 45435 (e-mail: ttarpey@discover.wright.edu). Bernard Flury is Professor, Department of Mathematics, Indiana University, Bloomington, Indiana 47405.

identity gives

$$\mathscr{E}\|\mathbf{X} - \mathscr{E}[\mathbf{X}|\mathbf{Y}]\|^2 \le \mathscr{E}\|\mathbf{X} - \mathbf{Y}\|^2$$

(Bickel and Doksum, 1977, page 36). Thus the random vector $\mathbf{Y}$ is locally optimal for approximating $\mathbf{X}$ if $\mathbf{Y} = \mathscr{E}[\mathbf{X}|\mathbf{Y}]$, in which case we call $\mathbf{Y}$ self-consistent for $\mathbf{X}$.

DEFINITION 2.1. For two jointly distributed random vectors $\mathbf{X}$ and $\mathbf{Y}$, we say that $\mathbf{Y}$ is *self-consistent* for $\mathbf{X}$ if $\mathscr{E}(\mathbf{X}|\mathbf{Y}) = \mathbf{Y}$ almost surely.

We will assume implicitly that moments exist as required. The notion of self-consistency is not vacuous, as the two extreme cases demonstrate. The random vector $\mathbf{X}$ is self-consistent for $\mathbf{X}$ and represents no loss of information. $\mathbf{Y} = \mathscr{E}[\mathbf{X}]$ is also self-consistent for $\mathbf{X}$ and represents a total loss of information except for the location of the distribution. Interesting self-consistent distributions range in between these two extremes. Many relevant cases of self-consistency are obtained by taking conditional means over subsets of the sample space of $\mathbf{X}$.

Another simple example of self-consistency is the following:

EXAMPLE 2.1. *Partial sums.* Let $\{X_n\}$ denote a sequence of independent, mean-zero random variables, and let $S_n = \sum_{i=1}^n X_i$. Then

$$\begin{aligned}\mathscr{E}[S_{n+k}|S_n] &= S_n + \mathscr{E}[X_{n+1} + \cdots + X_{n+k}|S_n] \\ &= S_n + \mathscr{E}[X_{n+1} + \cdots + X_{n+k}] \\ &= S_n.\end{aligned}$$

Thus, S_n is self-consistent for S_{n+k}, $k \ge 1$. The same property holds more generally if $\{S_n\}_{n\ge1}$ represents a martingale process.

For a given $\mathbf{X}$, a self-consistent approximation $\mathbf{Y}$ can be generated by partitioning the sample space of $\mathbf{X}$ and defining $\mathbf{Y}$ as a random variable taking as values the conditional means of subsets in the partition. This is illustrated by our next example, in which the support of $\mathbf{X}$ is partitioned into two half-planes.

EXAMPLE 2.2. *Two principal points.* Let $\mathbf{X} = (X_1, X_2)' \sim N_2(\mathbf{0}, \mathbf{I}_2)$. Note that $\mathscr{E}[X_1|X_1 \ge 0] = \sqrt{2/\pi}$. Let $\mathbf{Y} = (-\sqrt{2/\pi}, 0)'$ if $X_1 < 0$ and $\mathbf{Y} = (\sqrt{2/\pi}, 0)'$ if $X_1 \ge 0$. Then $\mathbf{Y}$ is self-consistent for $\mathbf{X}$. See Section 6 for a definition of principal points, and see Figure 7 for a generalization of this example.

The preceding example illustrates the purpose of self-consistency quite well. It is actually an application of our first lemma.

LEMMA 2.1. *For a p-variate random vector $\mathbf{X}$, suppose $\mathscr{S} \subset \mathbb{R}^p$ is a measurable set such that $\forall\ \mathbf{y} \in \mathscr{S}$, $\mathbf{y} = \mathscr{E}[\mathbf{X}|\mathbf{X} \in \mathscr{D}_\mathbf{y}]$, where $\mathscr{D}_\mathbf{y}$ is the domain of attraction of $\mathbf{y}$, that is, $\mathscr{D}_\mathbf{y} = \{\mathbf{x} \in \mathbb{R}^p\colon \|\mathbf{x} - \mathbf{y}\| < \|\mathbf{x} - \mathbf{y}^*\|,\ \forall\ \mathbf{y}^* \in \mathscr{S}\}$. Define $\mathbf{Y} = \mathbf{y}$ if $\mathbf{X} \in \mathscr{D}_\mathbf{y}$. Then $\mathbf{Y}$ is self-consistent for $\mathbf{X}$.*

PROOF. $\mathscr{E}[\mathbf{X}|\mathbf{Y} = \mathbf{y}] = \mathscr{E}[\mathbf{X}|\mathbf{X} \in \mathscr{D}_\mathbf{y}] = \mathbf{y}$. □

In Example 2.2, $\mathscr{S}$ consists of only two points, and the associated domains of attraction are the half-planes given by $x_1 < 0$ and $x_1 > 0$.

The following three lemmas give elementary properties of self-consistent random vectors.

LEMMA 2.2. *If $\mathbf{Y}$ is self-consistent for $\mathbf{X}$, then $\mathscr{E}[\mathbf{Y}] = \mathscr{E}[\mathbf{X}]$.*

PROOF. The lemma follows from $\mathscr{E}[\mathscr{E}[\mathbf{X}|\mathbf{Y}]] = \mathscr{E}[\mathbf{X}]$. □

We now introduce notation for the *mean squared error* (MSE) of a random vector $\mathbf{Y}$ for $\mathbf{X}$,

$$\mathrm{MSE}(\mathbf{Y};\mathbf{X}) = \mathscr{E}\|\mathbf{X} - \mathbf{Y}\|^2.$$

The next lemma relates the MSE of a self-consistent $\mathbf{Y}$ for $\mathbf{X}$ in terms of their respective covariance matrices. Here, $\mathbf{\Psi}_\mathbf{X}$ and $\mathbf{\Psi}_\mathbf{Y}$ denote the covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively.

LEMMA 2.3. *If $\mathbf{Y}$ is self-consistent for $\mathbf{X}$, then the following hold:*

(i) $\mathbf{\Psi}_\mathbf{X} \ge \mathbf{\Psi}_\mathbf{Y}$, *that is, $\mathbf{\Psi}_\mathbf{X} - \mathbf{\Psi}_\mathbf{Y}$ is positive semi-definite;*

(ii) $\mathrm{MSE}(\mathbf{Y};\mathbf{X}) = \mathrm{tr}(\mathbf{\Psi}_\mathbf{X}) - \mathrm{tr}(\mathbf{\Psi}_\mathbf{Y})$.

See the Appendix for a proof.

It follows from Lemma 2.3 that $\mathrm{Cov}[\mathbf{Y}] = \mathrm{Cov}[\mathbf{X}]$ exactly if $\mathrm{Cov}[\mathbf{X}|\mathbf{Y}] = \mathbf{0}$ a.s., that is, if $\mathbf{Y} = \mathbf{X}$ a.s. For one-dimensional random variables X and Y, if Y is self-consistent for X, then $\mathrm{var}[Y] \le \mathrm{var}[X]$, with equality exactly if $Y = X$ a.s.

There is a similarity between the two preceding lemmas and the Rao–Blackwell theorem (Casella and Berger, 1990, page 316), which in a simplified version states the following. If X is an unbiased estimator of a parameter θ, and if Y is a sufficient statistic for θ, then $\mathscr{E}[X|Y]$ is an unbiased estimator of θ, and $\mathrm{var}[\mathscr{E}[X|Y]] \le \mathrm{var}[X]$. If $\mathscr{E}[X|Y] = Y$,

then Lemma 2.2 gives $\mathscr{E}[Y] = \mathscr{E}[X]$, and part (i) of Lemma 2.3 gives $\mathrm{var}[Y] \le \mathrm{var}[X]$.

The next lemma demonstrates a dimensionality-reducing property of self-consistent random variables. Here, $\mathscr{S}(\mathbf{Y})$ denotes the support of $\mathbf{Y}$.

LEMMA 2.4. *Suppose* $\mathbf{Y}$ *is self-consistent for a* p*-variate random vector* $\mathbf{X}$ *with* $\mathscr{E}[\mathbf{X}] = \mathbf{0}$*, and* $\mathscr{S}(\mathbf{Y})$ *is contained in a linear subspace spanned by* q *orthonormal column vectors in the* $p \times q$ *matrix* $\mathbf{A}$*. Let* $\mathbf{P} = \mathbf{A}\mathbf{A}'$ *denote the associated projection matrix. Then* $\mathbf{Y}$ *and* $\mathbf{A}'\mathbf{Y}$ *are self-consistent for* $\mathbf{P}\mathbf{X}$ *and* $\mathbf{A}'\mathbf{X}$*, respectively.*

See the Appendix for a proof.

Lemma 2.4 means that the marginal distribution of a self-consistent $\mathbf{Y}$ in the linear subspace spanned by its support is self-consistent for the marginal distribution of $\mathbf{X}$ in the same subspace. For example, a self-consistent distribution for $\mathbf{X}$ whose support consists of a circle (see Section 6) is determined by the bivariate marginal distribution of $\mathbf{X}$ in the subspace containing the circle. In Example 2.2, the linear subspace spanned by the support of $\mathbf{Y}$ is the x_1-axis, the marginal distribution of $\mathbf{X}$ in this subspace is standard normal, and the random variable $Y_1 = \mathrm{sgn}(X_1)\sqrt{2/\pi}$ is self-consistent for X_1.

We conclude this section with a general method of finding self-consistent random variables.

LEMMA 2.5. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *denote two jointly distributed random vectors, not necessarily of the same dimension. Then* $\mathscr{E}[\mathbf{X}|\mathbf{Y}]$ *is self-consistent for* $\mathbf{X}$*.*

PROOF. Let $\mathbf{Z} = \mathscr{E}[\mathbf{X}|\mathbf{Y}]$. Then $\mathscr{E}[\mathbf{X}|\mathbf{Z}] = \mathscr{E}[\mathscr{E}[\mathbf{X}|\mathbf{Y}]|\mathbf{Z}] = \mathscr{E}[\mathbf{Z}|\mathbf{Z}] = \mathbf{Z}$. □

In particular, setting $\mathbf{Y} = \mathbf{X}$ in Lemma 2.5 gives again self-consistency of $\mathbf{X}$ for itself. If $\mathbf{Y}$ is independent of $\mathbf{X}$, then it follows that $\mathscr{E}[\mathbf{X}]$ is self-consistent for $\mathbf{X}$.

3. REGRESSION AND PRINCIPAL VARIABLES

For jointly distributed random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$, the conditional expectation $\mathscr{E}[\mathbf{X}_2|\mathbf{X}_1]$ is called the regression of $\mathbf{X}_2$ on $\mathbf{X}_1$. Not surprisingly, there are close connections to self-consistency.

In a classical regression setup, let $\mathbf{X}_1$ denote an m-variate random vector, $f(\cdot)$ a function from $\mathbb{R}^m$ to $\mathbb{R}^k$, and define

$$\mathbf{X}_2 = f(\mathbf{X}_1) + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is a k-variate random vector, independent of $\mathbf{X}_1$, with $\mathscr{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$. Then $\mathscr{E}[\mathbf{X}_2|f(\mathbf{X}_1)] = f(\mathbf{X}_1)$; that is, $f(\mathbf{X}_1)$ is self-consistent for $\mathbf{X}_2$. The mean $f(\mathbf{X}_1) = \mathscr{E}[\mathbf{X}_2]$ is a special case. However, in this section we will be interested in the problem of approximating a p-variate random vector $\mathbf{X}$ by a self-consistent $\mathbf{Y}$, where some q of the variables are replaced by their conditional means, as illustrated by our first theorem.

THEOREM 3.1. *Suppose the* p*-variate random vector* $\mathbf{X}$ *is partitioned into* q *and* $p - q$ *components as* $\mathbf{X} = \binom{\mathbf{X}_1}{\mathbf{X}_2}$*. Then the random vector*

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathscr{E}[\mathbf{X}_2|\mathbf{X}_1] \end{bmatrix}$$

is self-consistent for $\mathbf{X}$*.*

PROOF. Write

$$\mathscr{E}[\mathbf{X}|\mathbf{Y}] = \begin{pmatrix} \mathscr{E}[\mathbf{X}_1|\mathbf{Y}] \\ \mathscr{E}[\mathbf{X}_2|\mathbf{Y}] \end{pmatrix}.$$

Then $\mathscr{E}[\mathbf{X}_1|\mathbf{Y}] = \mathscr{E}[\mathbf{X}_1|\mathbf{X}_1, \mathscr{E}[\mathbf{X}_2|\mathbf{X}_1]] = \mathscr{E}[\mathbf{X}_1|\mathbf{X}_1] = \mathbf{X}_1 = \mathbf{Y}_1$, and $\mathscr{E}[\mathbf{X}_2|\mathbf{Y}] = \mathscr{E}[\mathbf{X}_2|\mathbf{X}_1] = \mathbf{Y}_2$. Hence $\mathbf{Y}$ is self-consistent for $\mathbf{X}$. □

Theorem 3.1 has an important interpretation in view of the aspect of distributions being "summarized by simpler ones," according to the criterion of self-consistency. It states that the q regressor variables $\mathbf{X}_1$, along with the regression of $\mathbf{X}_2$ on $\mathbf{X}_1$, are self-consistent for $\mathbf{X}$.

EXAMPLE 3.1. Suppose $\mathbf{X}$ is bivariate normal with mean $\mathbf{0}$ and covariance matrix $\binom{1 \quad \rho}{\rho \quad 1}$. Then

$$\mathbf{Y} = \begin{pmatrix} X_1 \\ \mathscr{E}[X_2|X_1] \end{pmatrix} = \begin{pmatrix} X_1 \\ \rho X_1 \end{pmatrix}$$

is self-consistent for $\mathbf{X}$. This is a bivariate normal but singular random vector, with $\mathrm{MSE}[\mathbf{Y}; \mathbf{X}] = 1 - \rho^2$. See also Example 4.2 and Figure 2.

In regression, the partition of $\mathbf{X}$ into "independent" variables $\mathbf{X}_1$ and "dependent" variables $\mathbf{X}_2$ is usually given by the setup of the analysis. However, for given (fixed) q, $1 \le q \le p-1$, one may ask for the subset of variables which, in some sense to be defined, gives the best summary of the p-variate distribution. This problem has been studied by McCabe (1984), who called the "best" subset of q variables the *principal variables* of $\mathbf{X}$. Suppose all conditional means of a subset of variables, given the remaining variables, are linear, as in the case of elliptical distributions. Let $\boldsymbol{\Psi} := \mathrm{Cov}[\mathbf{X}]$, and denote by $\mathbf{P}$ a

permutation matrix of dimension $p \times p$. Set

$$\mathbf{X}^* = \mathbf{P}\mathbf{X} = \begin{pmatrix} \mathbf{X}_1^* \\ \mathbf{X}_2^* \end{pmatrix},$$

where $\mathbf{X}^*$ has q components and $\mathbf{X}_2^*$ has $p-q$ components. Partition the mean vector and the covariance matrix of $\mathbf{X}^*$ analogously as

$$\mathscr{E}[\mathbf{X}^*] = \begin{bmatrix} \boldsymbol{\mu}_1^* \\ \boldsymbol{\mu}_2^* \end{bmatrix}, \quad \mathrm{Cov}[\mathbf{X}^*] = \begin{pmatrix} \boldsymbol{\Psi}_{11}^* & \boldsymbol{\Psi}_{12}^* \\ \boldsymbol{\Psi}_{21}^* & \boldsymbol{\Psi}_{22}^* \end{pmatrix}.$$

Then, assuming nonsingularity of $\boldsymbol{\Psi}$,

$$\mathscr{E}[\mathbf{X}_2^*|\mathbf{X}_1^*] = \boldsymbol{\mu}_2^* + (\boldsymbol{\Psi}_{21}^*)(\boldsymbol{\Psi}_{11}^*)^{-1}(\mathbf{X}_1^* - \boldsymbol{\mu}_1^*),$$

and the conditional variance formula (see the proof of Lemma 2.3) gives

$$\mathscr{E}[\mathrm{Cov}(\mathbf{X}_2^*|\mathbf{X}_1^*)] = \boldsymbol{\Psi}_{22}^* - \boldsymbol{\Psi}_{21}^*(\boldsymbol{\Psi}_{11}^*)^{-1}\boldsymbol{\Psi}_{12}^* =: \boldsymbol{\Psi}_{22\cdot 1}^*.$$

An intuitively reasonable optimality criterion is to choose $\mathbf{P}$ such that $\mathrm{tr}(\boldsymbol{\Psi}_{22\cdot 1}^*)$ is as small as possible. This can be motivated as follows. If we set

$$\mathbf{Y}^* = \begin{bmatrix} \mathbf{Y}_1^* \\ \mathbf{Y}_2^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^* \\ \mathscr{E}[\mathbf{X}_2^*|\mathbf{X}_1^*] \end{bmatrix},$$

then $\mathbf{Y}^*$ is self-consistent for $\mathbf{X}^*$ and can be regarded as a good approximation to $\mathbf{X}^*$ if $\mathrm{MSE}(\mathbf{Y}^*;\mathbf{X}^*)$ is as small as possible. Assuming linearity of the conditional mean of $\mathbf{X}_2^*$, given $\mathbf{X}_1^*$, and setting $\mathscr{E}[\mathbf{X}^*] = \mathbf{0}$ without loss of generality, we obtain

$$\begin{aligned} &\mathscr{E}\|\mathbf{X}^* - \mathbf{Y}^*\|^2 \\ &\quad = \mathrm{tr}(\boldsymbol{\Psi}^*) - \mathrm{tr}(\mathrm{Cov}(\mathbf{Y}^*)) \quad \text{(by Lemma 2.3)} \\ &\quad = \mathrm{tr}(\boldsymbol{\Psi}_{22}^*) - \mathrm{tr}(\boldsymbol{\Psi}_{21}^*(\boldsymbol{\Psi}_{11}^*)^{-1}\boldsymbol{\Psi}_{12}^*) \\ &\quad = \mathrm{tr}(\boldsymbol{\Psi}_{22\cdot 1}^*). \end{aligned}$$

Hence, for given q, principal variables identify an optimal subset of q variables $\mathbf{X}_1^*$, which (along with the regression of $\mathbf{X}_2^*$ on $\mathbf{X}_1^*$) defines a self-consistent approximation

$$\mathbf{Y}^* = \begin{pmatrix} \mathbf{X}_1^* \\ \mathscr{E}[\mathbf{X}_2^*|\mathbf{X}_1^*] \end{pmatrix}$$

to $\mathbf{X}^*$. Returning to the original order of variables, $\mathbf{Y} = \mathbf{P}'\mathbf{Y}^*$ is then self-consistent for $\mathbf{X}$.

Finding principal variables is computationally intensive because, for a p-dimensional random vector $\mathbf{X}$ and q principal variables, there are $\binom{p}{q}$ ways to select q candidates. If the assumption of linearity of the conditional means is dropped, one may of course still search for the "best" partition of $\mathbf{X}$ into q of the original variables and $p-q$ conditional means, according to the criterion of minimizing $\mathrm{MSE}(\mathbf{Y}^*;\mathbf{X}^*)$, but the problem becomes intractable without making further assumptions.

4. THE PRINCIPAL SUBSPACE THEOREM AND LINEAR PRINCIPAL COMPONENTS

For high-dimensional random variables it is often desirable to find a low-dimensional approximation, or more precisely, an approximation whose support is a low-dimensional manifold. The main result of this section, Theorem 4.1, is called the *principal subspace theorem*. It appeared originally in Tarpey, Li and Flury (1995) for the special case of self-consistent approximations whose support consists of k distinct points. We show the theorem for random vectors $\mathbf{X}$ such that, for any orthogonal matrix $\mathbf{A}$, the conditional mean of any subset of variables in $\mathbf{A}'\mathbf{X}$, given the remaining variables, is linear. This is the case, for instance, for all elliptical distributions.

THEOREM 4.1. *Suppose* $\mathbf{X}$ *is a* p*-variate random vector with mean* $\mathbf{0}$ *and positive definite covariance matrix* $\boldsymbol{\Psi}$. *Assume linearity of conditional means as explained above. Suppose* $\mathbf{Y}$ *is self-consistent for* $\mathbf{X}$, *and the support of* $\mathbf{Y}$ *spans a linear subspace* $\mathscr{A}$ *of dimension* $q < p$. *Suppose, furthermore, that* $\mathbf{Y}$ *is measurable with respect to the orthogonal projection of* $\mathbf{X}$ *on* $\mathscr{A}$. *Then* $\mathscr{A}$ *is spanned by* q *eigenvectors of* $\boldsymbol{\Psi}$.

See the Appendix for a proof.

Theorem 4.1 is of considerable theoretical appeal because it says that for certain types of self-consistent approximations attention may be restricted to subspaces spanned by eigenvectors of the covariance matrix. Principal components are a particular case, as we shall see later in this section. Another important case is self-consistent points, as illustrated in Example 4.1 and later in Section 6. Theorem 4.1 also provides justification for restricting estimators of principal points, principal curves or other self-consistent distributions of high-dimensional data to lie in a lower-dimensional linear subspace. These subspace restrictions can improve the estimation of principal points (Flury, 1993).

EXAMPLE 4.1. Sets of four self-consistent points of an elliptical distribution with mean zero and positive definite covariance matrix $\boldsymbol{\Psi}$. Suppose we want to approximate the distribution of a mean-zero, trivariate elliptical random vector $\mathbf{X}$ by four points. Let $\mathbf{Y}$ be a self-consistent random vector for $\mathbf{X}$ whose support consists of four points $\mathbf{y}_1, \ldots, \mathbf{y}_4 \in \mathbb{R}^3$ which span a subspace of dimension 2. If $\mathbf{Y}$ is chosen so that each $\mathbf{y}_j$ is the conditional mean of $\mathbf{X}$, given that $\mathbf{X}$ is closer to $\mathbf{y}_j$ than to all other $\mathbf{y}_h$, then the points $\mathbf{y}_1, \ldots, \mathbf{y}_4$ are called

self-consistent points of $\mathbf{X}$ (Flury, 1993). Since $\mathbf{Y}$ satisfies the conditions of Theorem 4.1, only subspaces of dimension 2 spanned by two eigenvectors of $\boldsymbol{\Psi}$ are candidates for the plane that contains the four points. Assuming that all eigenvalues of $\boldsymbol{\Psi}$ are distinct, this implies that only three two-dimensional subspaces need to be considered. Figure 1 illustrates this where four points are chosen to form a rectangular pattern.

Next we show how principal components can be based on the notion of self-consistency.

Suppose $\mathbf{A}_1$ is a $p \times q$ matrix ($q < p$) such that all columns of $\mathbf{A}_1$ have unit length and are mutually orthogonal. Let $\mathbf{P} = \mathbf{A}_1\mathbf{A}_1'$ denote the projection matrix associated with the orthogonal projection from $\mathbb{R}^p$ into the subspace spanned by the columns of $\mathbf{A}_1$. For a p-dimensional random vector $\mathbf{X}$ and some fixed $\mathbf{b} \in \mathbb{R}^p$, consider the transformation $\mathbf{Y} = \mathbf{b} + \mathbf{PX}$. If $\mathbf{Y}$ is to be self-consistent for $\mathbf{X}$, then Lemma 2.2 implies $\mathscr{E}[\mathbf{Y}] = \mathscr{E}[\mathbf{X}]$, that is, $\mathbf{b} = (\mathbf{I}_p - \mathbf{P})\boldsymbol{\mu}$. Thus, if $\mathbf{Y}$ is a self-consistent projection of $\mathbf{X}$ into a linear manifold, then

$$\mathbf{Y} = (\mathbf{I}_p - \mathbf{P})\boldsymbol{\mu} + \mathbf{PX}.$$

Assume, without loss of generality, that $\boldsymbol{\mu} = \mathbf{0}$. Let $\mathbf{A}_2$ denote a $(p-q) \times p$ matrix such that $[\mathbf{A}_1 : \mathbf{A}_2]$ is orthogonal and let $\mathbf{Q} = \mathbf{I}_p - \mathbf{P}$. Then $\mathbf{Y} = \mathbf{PX}$, and self-consistency of $\mathbf{Y}$ for $\mathbf{X}$ implies $\mathscr{E}[\mathbf{QX}|\mathbf{PX}] = \mathbf{0}$ a.s. However, for any random variables U and V, $\mathscr{E}[U|V] = 0$ implies $\operatorname{cov}(U, V) = 0$. Thus $\mathbf{A}_2'\boldsymbol{\Psi}\mathbf{A}_1 = \mathbf{0}$. Using the same argument as in the proof of Theorem 4.1 we thus have the following theorem.

THEOREM 4.2. *If* $\mathbf{Y} = (\mathbf{I}_p - \mathbf{P})\boldsymbol{\mu} + \mathbf{PX}$ *is self-consistent for* $\mathbf{X}$, *where* $\boldsymbol{\mu} = \mathscr{E}[\mathbf{X}]$ *and* $\mathbf{P}$ *is the projection matrix associated with the orthogonal projection from* $\mathbb{R}^p$ *into a linear subspace* $\mathscr{M}$ *of dimension* $q < p$, *then* $\mathscr{M}$ *is spanned by some* q *eigenvectors of* $\boldsymbol{\Psi} = \operatorname{Cov}(\mathbf{X})$.

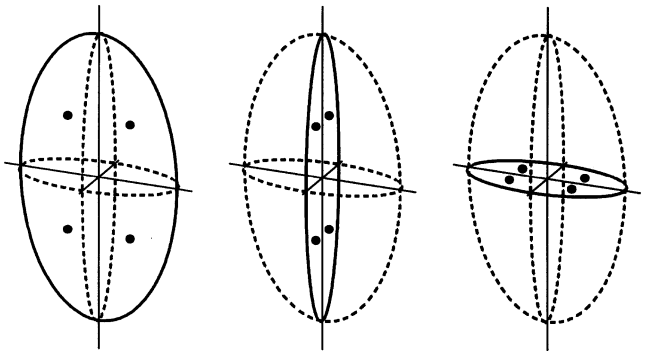

FIG. 1. *Approximation of an elliptical random vector* $\mathbf{X}$ *by a self-consistent* $\mathbf{Y}$ *whose support consists of four points in a plane. The plane containing the support of* $\mathbf{Y}$ *is spanned by two eigenvectors of the covariance matrix of* $\mathbf{X}$.

EXAMPLE 4.2 (Continuation of Example 3.1). With the same setup as in Example 3.1, assume $\rho \neq 0$. Then $\boldsymbol{\Psi} = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$ has two normalized eigenvectors

$$\boldsymbol{\beta}_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} -1 \\ 1 \end{pmatrix},$$

with associated eigenvalues $1 + \rho$ and $1 - \rho$. Let

$$\mathbf{P} = \boldsymbol{\beta}_1\boldsymbol{\beta}_1' = \tfrac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$

then

$$\mathbf{Y} = \tfrac{1}{2}\begin{pmatrix} X_1 + X_2 \\ X_1 + X_2 \end{pmatrix}$$

is self-consistent for $\mathbf{X}$, with $\operatorname{MSE}[\mathbf{Y}; \mathbf{X}] = 1 - \rho$. For $\rho > 0$, the support of $\mathbf{Y}$ is the first principal component axis. Note that the mean squared error $1-\rho$ for $\mathbf{Y}$ is smaller than $1-\rho^2$, which is the mean squared error of the self-consistent distribution whose support is along the regression line of X_2 on X_1 in Example 3.1. See Figure 2. The same Example 3.1 is a case where the support of a self-consistent random vector is a one-dimensional linear subspace which is not spanned by an eigenvector of the covariance matrix. The reason Theorem 4.1 does not apply in this case is that the $\mathbf{Y}$ of Example 3.1 is not measurable with respect to the orthogonal projection of $\mathbf{X}$ into the subspace spanned by the support of $\mathbf{Y}$.

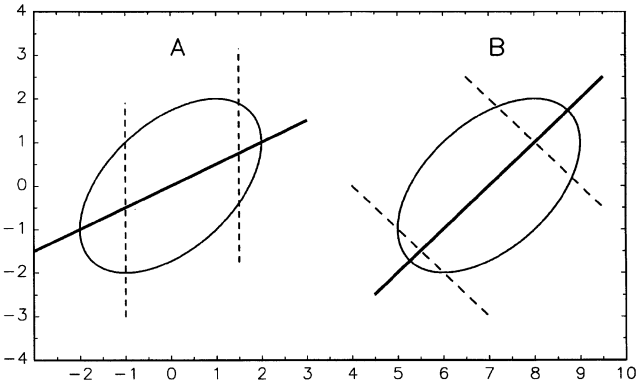

FIG. 2. *Two self-consistent approximations of the bivariate normal distribution with mean* $\mathbf{0}$ *and covariance matrix* $\left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$, *with* $\rho = 0.5$. *In both graphs the support of* $\mathbf{Y}$ *is indicated as a solid line, and broken lines represent sets of points in* $\mathbb{R}^2$ *that are projected onto the same value of* $\mathbf{Y}$. (a) *The approximation from Example* 3.1, *with*

$$\mathbf{Y} = \begin{pmatrix} X_1 \\ \rho X_1 \end{pmatrix};$$

(b) *the approximation from Example* 4.2, *with*

$$\mathbf{Y} = \tfrac{1}{2}\begin{pmatrix} X_1 + X_2 \\ X_1 + X_2 \end{pmatrix}.$$

Principal components are traditionally introduced in terms of a stepwise maximization procedure which does not depend on any distributional assumptions beyond the existence of second moments. With $\mathbf{\Psi}$ denoting the covariance matrix of $\mathbf{X}$, the first principal component is defined as $U_1 = \boldsymbol{\beta}_1'\mathbf{X}$, where $\boldsymbol{\beta}_1 \in \mathbb{R}^p$ is such that

$$\operatorname{var}[\boldsymbol{\beta}'\mathbf{X}] = \max_{\substack{\mathbf{a}\in\mathbb{R}^p \\ \mathbf{a}'\mathbf{a}=1}} \operatorname{var}[\mathbf{a}'\mathbf{X}].$$

The coefficients of the subsequent principal components $U_j = \boldsymbol{\beta}_j'\mathbf{X}$, $j = 2, \ldots, p$, are obtained from the same maximization problem, subject to the additional constraints $\operatorname{cov}[\mathbf{a}'\mathbf{X}, U_h] = 0$ for $h = 1, \ldots, j-1$. In the traditional definition of principal components of a p-variate random vector $\mathbf{X}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Psi}$, any linear combination $U = \boldsymbol{\beta}'(\mathbf{X} - \boldsymbol{\mu})$, where $\boldsymbol{\beta}$ is a normalized eigenvector of $\mathbf{\Psi}$, is called a principal component of $\mathbf{X}$. In view of the fact that the projection associated with a given eigenvector may or may not be self-consistent, we suggest adding the word "linear" to this definiton: $U = \boldsymbol{\beta}'(\mathbf{X} - \boldsymbol{\mu})$ is called a *linear* principal component of $\mathbf{X}$. This parallels the terminology used in regression, and distinguishes the classical method better from nonlinear generalizations (Hastie and Stuetzle, 1989). Ideally, a linear principal component with coefficient vector $\boldsymbol{\beta}$ is associated with a self-consistent projection $\mathbf{Y} = (\mathbf{I}_p - \mathbf{P})\boldsymbol{\mu} + \mathbf{PX}$, as is the case for multivariate elliptical distributions. In other cases, none or only a few linear principal components may correspond to self-consistent projections.

It is not difficult to construct examples of p-variate random vectors with any number of k self-consistent orthogonal projections, as we show in the following examples.

EXAMPLE 4.3. Suppose $\mathbf{X}$ is elliptical with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Psi}$, where all eigenvalues of $\mathbf{\Psi}$ are distinct. Then there are exactly 2^p different orthogonal projections that are self-consistent, including the projection matrices $\mathbf{P} = \mathbf{I}_p$ and $\mathbf{P} = \mathbf{0}$.

Tarpey (1995) showed that orthogonal projections into subspaces spanned by sets of eigenvectors of the covariance matrix are self-consistent for $\mathbf{X}$ in a large class of symmetric multivariate distributions, of which elliptical distributions are a special case. Ellipticity is therefore not a necessary condition for linear principal component approximations to be self-consistent. A simple example is the bivariate uniform distribution in a rectangle with unequal side lengths.

EXAMPLE 4.4. Let $\mathbf{X}$ denote a bivariate discrete random vector which puts probability $1/k$ on each of $k \geq 2$ equally spaced points on a circle centered at the origin. Then there exist exactly k self-consistent projections into one-dimensional subspaces. The same construction can be applied to the uniform distribution inside the regular polygon spanned by the k points.

EXAMPLE 4.5. Suppose $\mathbf{X}$ is uniformly distributed in the set $x_1^2 + x_2^2 \leq 1$, $x_2 \geq 0$. Then there is a single self-consistent projection into a linear subpace of dimension one, namely, the x_2 axis. See Figure 3.

For observed data given in the coordinate system of the eigenvectors of the covariance matrix, the question naturally arises whether a given coordinate direction corresponds to a self-consistent projection or not. This is illustrated in the next example.

EXAMPLE 4.6. Figure 4 shows a scatterplot of the first (U_1) versus the second (U_2) linear principal component computed for a sample of 24 female turtles (Jolicoeur and Mosimann, 1960), using variables $X_1 = \log(\text{shell length})$ and $X_2 = \log(\text{shell width})$. Self-consistency of the projection on the first principal component axis is desirable here because of the interpretation of the coefficients of the first principal component vector of log-transformed variables as constants of allometric growth. If self-consistency of the projection on the first principal component axis holds, then we would expect the local average of U_2 to be approximately zero over the whole range of U_1. The data in Figure 4 contradict the assumption of self-consistency because at

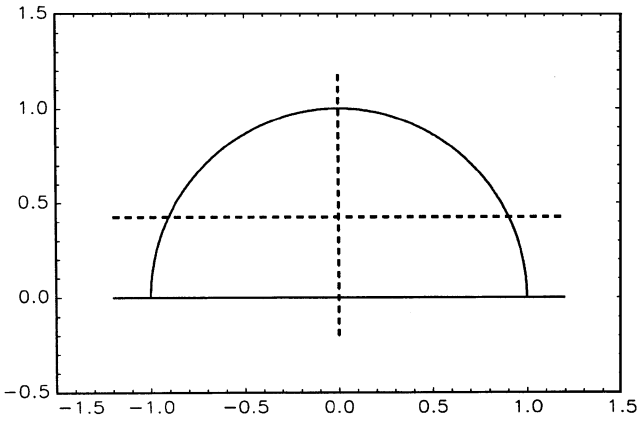

FIG. 3. *Principal component axes in Example* 4.5; $\mathbf{X}$ *is uniform in the half-circle. The projection of* $\mathbf{X}$ *on the vertical axis is self-consistent, but the projection on the horizontal axis is not.*

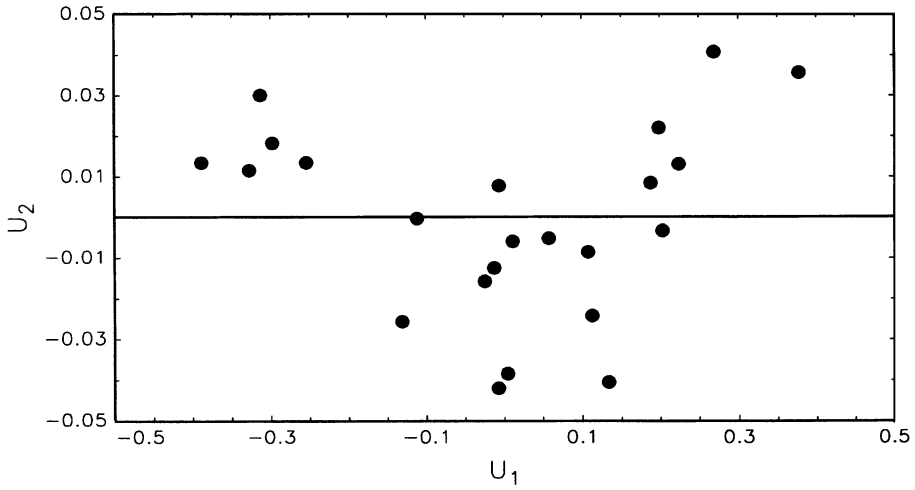

FIG. 4. *Scatterplot of centered linear principal components in Example* 4.6.

both ends of the range of U_1 we find only positive values of U_2, while mostly negative values of U_2 are found in the middle.

To our knowledge, no formal testing procedures for self-consistency of a given projection have been developed so far, although tests for subspaces spanned by eigenvectors of covariance matrices do exist (Kshirsagar, 1961; Mallows, 1961; Anderson, 1963; Jolicoeur, 1968; Tyler, 1983; Schott, 1991). This leaves a variety of research questions; see Section 9.

The last example in this section goes beyond principal component analysis, by combining self-consistent projections. Although it is probably of no practical importance, it is nevertheless interesting in view of the variety of ways to construct self-consistent random variables for a given distribution.

EXAMPLE 4.7. Suppose $\mathbf{X} = \binom{X_1}{X_2}$ is bivariate elliptical with mean $\mathbf{0}$ and covariance matrix $\operatorname{diag}(\sigma_1^2, \sigma_2^2)$. Let $\mathbf{Y} = \binom{Y_1}{Y_2}$, where

$$Y_1 = \begin{cases} X_1, & \text{if } |X_1| \geq |X_2|, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$Y_2 = \begin{cases} 0, & \text{if } |X_1| \geq |X_2|, \\ X_2, & \text{otherwise.} \end{cases}$$

Then $\mathbf{Y}$ is self-consistent for $\mathbf{X}$. This provides a "summary" of $\mathbf{X}$ where all probability mass is concentrated on the two coordinate axes. The MSE of $\mathbf{Y}$ when $\mathbf{X} \sim N_2(\mathbf{0}, \mathbf{I}_2)$ for this example is $1 - 2/\pi$.

5. PRINCIPAL MODES OF VARIATION

Principal components have been used for data reduction when the observed response is a continuous curve rather than a vector variable. Let $x(t)$, $0 \leq t \leq T$, denote a random process with continuous sample paths and continuous covariance function $C(s, t) = \operatorname{cov}(x(s), x(t))$, $0 \leq s, t \leq T$. Observing the process at p distinct points $t_1, \ldots, t_p$ yields observed random vectors, and principal components techniques can be used to study the x-process. In fact, for a vector process $\mathbf{X} = (x_1, \ldots, x_p)'$, an expression of the form

$$\mathbf{Z} = \boldsymbol{\mu} + \sum_{i=1}^{k} \boldsymbol{\xi}_i U_i,$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ denotes the vector of means of the process, the $\boldsymbol{\xi}_i$ are fixed, normalized p-vectors, and the U_i are scalar variates dependent on $\mathbf{X}$, is called a k-dimensional linear model for $\mathbf{X}$. The random vector $\mathbf{Z}$ which minimizes the MSE $\mathscr{E}\|\mathbf{X} - \mathbf{Z}\|^2$ over all possible choices of normalized vectors $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k$ and all choices of scalar variates $U_1, \ldots, U_k$ is called a *best k-dimensional linear model* for $\mathbf{X}$. Such a best k-dimensional linear model is given by choosing the $\boldsymbol{\xi}_i$ as the normalized eigenvector corresponding to the ith largest eigenvalue of the covariance matrix of $\mathbf{X}$.

For a random process $x(t)$, $0 \leq t \leq T$, let $C(s, t)$ denote the covariance function and let $\mu(t) = \mathscr{E}[x(t)]$. Then a k-dimensional linear model for the process $x(t)$ is a linear combination

$$z(t) = \mu(t) + \sum_{i=1}^{k} \alpha_i f_i(t)$$

of k linearly independent functions $f_1, \ldots, f_k$ and k scalar variates $\alpha_1, \ldots, \alpha_k$ that depend on $x(t)$.

The *best* k-dimensional model for $x(t)$ minimizes the mean squared error

$$\mathscr{E}\left\{\int_0^T |x(t) - z(t)|^2 dt\right\}$$

over all choices of k linearly independent nonrandom functions $f_1, \ldots, f_k$ and all real coefficients $\alpha_1, \ldots, \alpha_k$ which may be functions of the sample paths $x(t)$.

A solution to this problem is given by $f_i = \phi_i$, $i = 1, \ldots, k$, where ϕ_i is the normalized eigenfunction corresponding to the ith largest eigenvalue of the covariance kernel $C(s, t)$ (Castro, Lawton and Sylvestre, 1986). That is, since $C(s, t)$ is symmetric and nonnegative definite, there are real numbers $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq \cdots \geq 0$, called eigenvalues, and functions $\phi_1, \phi_2, \ldots$, satisfying

$$\int_0^T C(s, t)\phi_i(t)\, dt = \lambda_i \phi_i(s), \quad 0 \leq s \leq T,$$

and

$$\int_0^T \phi_i(t)\phi_j(t)\, dt = \delta_{i,j}.$$

Castro, Lawton and Sylvestre (1986) call ϕ_1 the *first principal mode of variation* in $x(t)$ and ϕ_2 the second principal mode of variation and so on.

The following theorem relates principal modes of variation to the notion of self-consistency.

THEOREM 5.1. *Consider a random process with continuous sample paths $x(t)$, $0 \leq t \leq T$, with a covariance function $C(s, t)$ which is continuous in the pair (s, t). Let f denote a measurable function such that $\int_0^T C(s, t)f(t)dt$ exists and is finite for all $s \in \mathbb{R}$. Suppose that $\alpha f(t)$ is self-consistent for the process $x(t)$, where $\alpha = \int_0^T (x(t) - \mu(t))f(t)dt$. Then $f(t)$ is a principal mode of variation for the process $x(t)$; that is, $f(t)$ is an eigenfunction of $C(s, t)$.*

See the Appendix for a proof.

6. SELF-CONSISTENT CURVES AND POINTS

Suppose $\mathbf{Y}$ is self-consistent for $\mathbf{X}$ where the support $\mathscr{S}(\mathbf{Y})$ is a smooth (C^∞) curve parameterized over a closed interval of $\mathbb{R}$ that has finite length inside any finite ball in $\mathbb{R}^p$. If each point in $\mathscr{S}(\mathbf{Y})$ is equal to the conditional mean of $\mathbf{X}$ given that $\mathbf{X}$ projects onto that point, then the curve $\mathscr{S}(\mathbf{Y})$ is called a *self-consistent curve* (Hastie and Stuetzle, 1989). The theory of self-consistent curves and surfaces has found several applications in addition to those mentioned in Hastie and Stuetzle (1989); Banfield and Raftery (1992); Tibshirani (1992); LeBlanc and Tibshirani (1994).

Hastie and Stuetzle's definition of a self-consistent curve has the added constraint that the curve cannot intersect itself. However, our definition of self-consistency does not exclude self-consistent distributions whose support consists of intersecting curves, unions of intersecting curves, combinations of curves and points and so on.

Next we define self-consistent points and principal points for a distribution.

DEFINITION 6.1. The points in the set $\{\mathbf{y}_1, \ldots, \mathbf{y}_k\}$ are called k *self-consistent points* of $\mathbf{X}$ if

$$\mathscr{E}[\mathbf{X}|\mathbf{X} \in D_j] = \mathbf{y}_j, \quad j = 1, \ldots, k,$$

where $D_j = \{\mathbf{x} \in \mathbb{R}^p\colon \|\mathbf{x} - \mathbf{y}_j\| < \|\mathbf{x} - \mathbf{y}_l\|,\ l \neq j\}$ is the *domain of attraction* of point $\mathbf{y}_j$.

Setting

$$\mathbf{Y} = \sum_{j=1}^{k} \mathbf{y}_j I\{\mathbf{X} \in D_j\},$$

it follows that $\mathscr{E}[\mathbf{X}|\mathbf{Y}] = \mathbf{Y}$ and $\mathbf{Y}$ is self-consistent for $\mathbf{X}$. More generally, one can obtain a self-consistent discrete random vector $\mathbf{Y}$ for a given random vector $\mathbf{X}$ by choosing an arbitrary (finite or countably infinite) partition $\{\mathscr{A}_i\}$ of the support of $\mathbf{X}$ and defining a joint distribution of $\mathbf{X}$ and $\mathbf{Y}$ by $\mathbf{Y} = \mathscr{E}[\mathbf{X}|\mathbf{X} \in \mathscr{A}_i]$ if $\mathbf{X} \in \mathscr{A}_i$. However, such examples are usually not interesting unless some rule is imposed on the way the partition is created, as in our Definition 6.1. In this section it will therefore be assumed implicitly that the partition of the support of $\mathbf{X}$ is in terms of domains of attraction D_j.

Suppose $\mathbf{Z}$ is a random vector measurable with respect to $\mathbf{X}$ and the support of $\mathbf{Z}$ consists of k points $\mathscr{S}(\mathbf{Z}) = \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k\}$. If

$$\mathrm{MSE}(\mathbf{Z}; \mathbf{X}) \leq \mathrm{MSE}(\mathbf{Y}; \mathbf{X})$$

for all k-point distributions $\mathbf{Y}$ which are measurable with respect to $\mathbf{X}$, then $\mathbf{Z}$ is self-consistent for $\mathbf{X}$ and the points $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k$ are called k *principal points* of $\mathbf{X}$ (Flury, 1993).

EXAMPLE 6.1. *Principal points of the normal distribution.* Figure 5 shows $k = 2$ to $k = 5$ principal points of the standard normal distribution. For $k > 2$, the principal points have to be found by numerical methods; see, for example, Lloyd (1982), Zoppè (1995) and Rowe (1996). Figure 6 shows an example of $k = 5$ self-consistent points of the bivariate normal distribution considered in Examples 3.1 and 4.2, with correlation coefficient $\rho = 0.5$, along with the partition of $\mathbb{R}^2$ by domains of attraction

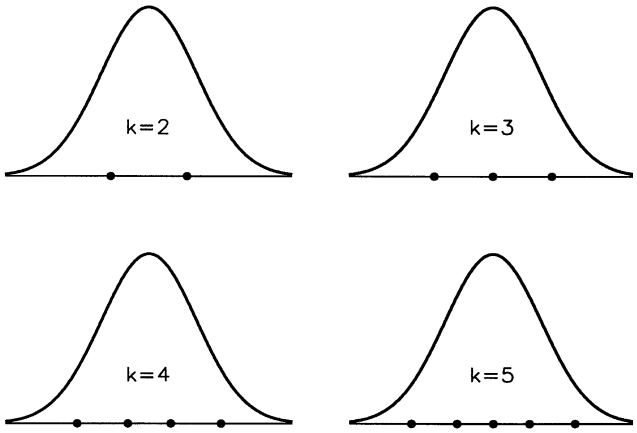

FIG. 5. *$k = 2$ to $k = 5$ principal points of the standard normal distribution,*

k	Principal Points	MSE
2	$\pm\sqrt{2/\pi}$	0.3634
3	$0, \pm 1.227$	0.1900
4	$\pm 1.507, \pm 0.451$	0.1170
5	$0, \pm 0.754, \pm 1.707$	0.0800

of the self-consistent points. The pattern shown in Figure 6 has been found numerically; see Tarpey (1996).

Principal points and self-consistent points have found applications in optimal grouping (Cox, 1957), optimal stratification (Dalenius, 1950), determining optimal representatives of a population for fitting masks (Flury, 1990, 1993), standardizing clothing (Fang and He, 1982) and curve selection (Flury and Tarpey, 1993). Eubank (1988) noted that determining principal points is equivalent to finding the best

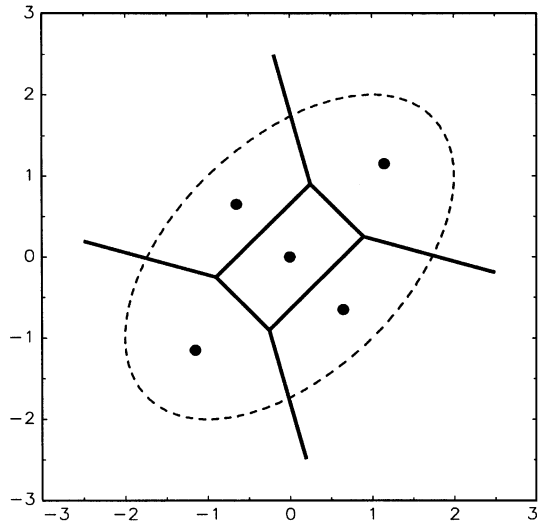

FIG. 6. *$k = 5$ self-consistent points of the bivariate normal distribution from Figure 2, along with the partition of $\mathbb{R}^2$ by domains of attraction of the five points.*

set of knots for approximating the quantile function of the distribution by a piecewise constant function. Iyengar and Solomon (1983) point out that the mathematical problems that arise in the theory of representing a distribution by a few optimally choosen points, such as principal points, are often the same as those in quantization for signal processing. The March 1982 *IEEE Transactions on Information Theory* (IEEE, 1982) is a special issue devoted to the subject of quantization—the mapping of vectors from an information source into k distinct values of the signal for transmission over a digital channel.

Noting that k principal points partition the space into k Voronoi regions according to minimal distance, principal points are intimately connected to the problem of clustering. The k-means algorithm (Hartigan, 1975, Chapter 4) converges by construction to a set of k self-consistent points of the empirical distribution given by the sample. The cluster means of the k-means algorithm are strongly consistent estimators of k principal points provided the k principal points are unique (Pollard, 1981). Convergence results of cluster means to principal points in more general settings have been given by Cuesta and Matran (1988) and by Pärna (1990).

We are now going to illustrate a relationship between self-consistent curves and self-consistent points for spherical distributions.

Let $\mathbf{X} = (X_1, X_2)'$ denote a bivariate spherical random vector with distribution function F. For each positive integer k, there exists a self-consistent $\mathbf{Y}$ whose support consists of k points $\mathbf{y}_1, \ldots, \mathbf{y}_k$, evenly spaced on a circle centered at the origin. In order to determine the radius r_k of this circle, orientate the k points such that $\mathbf{y}_1$, the first point, lies on the positive x_1-axis: $\mathbf{y}_1 = (r_k, 0)'$. Then the domain of attraction D_1 of $\mathbf{y}_1$ is a pie-slice-shaped region given by the linear inequalities $x_1 > 0$ and $|x_2| < mx_1$, where $m = \tan(\pi/k)$. Because the k points are equally spaced on the circle and $\mathbf{X}$ is spherical, $\mathscr{P}(\mathbf{X} \in D_1) = 1/k$. Self-consistency requires

$$r_k = \mathscr{E}[X_1 | \mathbf{X} \in D_1] = \frac{\int_0^\infty \int_{-mx_1}^{mx_1} x_1 \, dF(x_1, x_2)}{1/k}.$$

For the spherical normal, $r_k = k \sin(\pi/k)/\sqrt{2\pi}$ with $\text{MSE} = 2 - k^2 \sin^2(\pi/k)/(2\pi)$. For $k = 2$ to $k = 4$, the evenly spaced points on the circle are also principal points of $\mathbf{X}$, but for $k > 4$ patterns of principal points become more complicated (Tarpey, 1996). As k increases, the radii of the circles containing the k self-consistent points increase, and in

the limit we have a self-consistent circle with radius $r = \lim_{k\to\infty} r_k$. For the bivariate normal distribution, the self-consistent circle has radius

$$r = \lim_{k\to\infty} \frac{k\sin(\pi/k)}{\sqrt{2\pi}} = \sqrt{\frac{\pi}{2}} = E\|\mathbf{X}\|.$$

This is illustrated in Figure 7. More generally, for any p-variate spherical random vector $\mathbf{X}$, the uniform distribution on the sphere with radius $r = \mathscr{E}\|\mathbf{X}\|$ centered at the origin is self-consistent for $\mathbf{X}$. To see this, let $\mathbf{X} = R\mathbf{S}$ denote the stochastic representation of $\mathbf{X}$, where $R = \|\mathbf{X}\| > 0$ is independent of $\mathbf{S} = \mathbf{X}/R$ and $\mathbf{S}$ is uniformly distributed on the unit sphere in $\mathbb{R}^p$ (Fang, Kotz and Ng, 1990, page 30). Let $r = \mathscr{E}\|\mathbf{X}\| = \mathscr{E}[R]$. Then $\mathscr{E}[\mathbf{X}|r\mathbf{S} = \mathbf{a}] = \mathscr{E}[R\mathbf{S}|r\mathbf{S} = \mathbf{a}] = \mathbf{a}/r\mathscr{E}[R|\mathbf{S} = \mathbf{a}/r] = \mathbf{a}$ since R is independent of $\mathbf{S}$. Thus $r\mathbf{S}$ is self-consistent for $\mathbf{X}$. If each component of $\mathbf{X}$ has unit variance, then, by Lemma 2.3, the MSE of the self-consistent sphere is $p - r^2$.

Suppose $\mathbf{X} = (X_1, X_2, X_3)'$ is trivariate standard normal. Then the MSE of the self-consistent uniform distribution on a sphere is $3 - 8/\pi \approx 0.4535$, which is less than the MSE of the self-consistent marginal distribution $(X_1, X_2, 0)'$. In other words, the self-consistent sphere (which is a two-dimensional manifold) is a better approximation to the distribution of $\mathbf{X}$ than the marginal distribution in a two-dimensional linear subspace in terms of mean squared error.

Next we relate sets of self-consistent points to self-consistent distributions whose support consists of concentric spheres for spherical distributions. For a spherically distributed random vector $\mathbf{X}$, consider once again its stochastic representation $\mathbf{X} = R\mathbf{S}$. Let $r_1 < \cdots < r_k$ denote a set of k self-consistent points of R, and let $D_j = \{x \in \mathbb{R}: (r_{j-1} + r_j)/2 < x < (r_j + r_{j+1})/2\}$. Then $\mathscr{E}[R|R \in D_j] = r_j$. Let $\mathbf{Y} = \sum_{j=1}^{k} r_j\mathbf{S}I(R \in D_j)$. Then $\mathbf{Y}$ is self-consistent for $\mathbf{X}$, and $\mathscr{S}(\mathbf{Y})$ consists of k concentric spheres with radii $r_1, \ldots, r_k$. To see this, if $\mathbf{Y} = \mathbf{t}$, $\mathbf{t} \in \mathbb{R}^p$, then $\mathbf{t} = r_j\mathbf{s}$ for some j and a unit vector $\mathbf{s}$. Thus $\mathscr{E}[\mathbf{X}|\mathbf{Y} = \mathbf{t}] = \mathscr{E}[R\mathbf{S}|R \in D_j,\ \mathbf{S} = \mathbf{s}] = \mathbf{s}\mathscr{E}[R|R \in D_j] = r_j\mathbf{s} = \mathbf{t}$. Therefore, $\mathbf{Y}$ is self-consistent for $\mathbf{X}$.

For spherically symmetric random vectors there exist also self-consistent distributions whose support consists of any number of concentric spheres along with a point at the origin. For instance, for a bivariate spherical distribution, there exists a set of k self-consistent points where one of the points is at the origin and the remaining $k - 1$ points are equally spaced on a circle centered at the origin. As $k \to \infty$, the distribution of these points converges to a self-consistent distribution whose support consists of a circle along with a point at the origin. For the bivariate spherical normal distribution, $k = 5$ principal points are such that one point lies at the origin and the remaining four points lie on a circle forming a square pattern. For $k = 6$ principal points, one point lies at the origin and the remaining five points lie on a circle forming a pentagonal pattern (Tarpey, 1996).

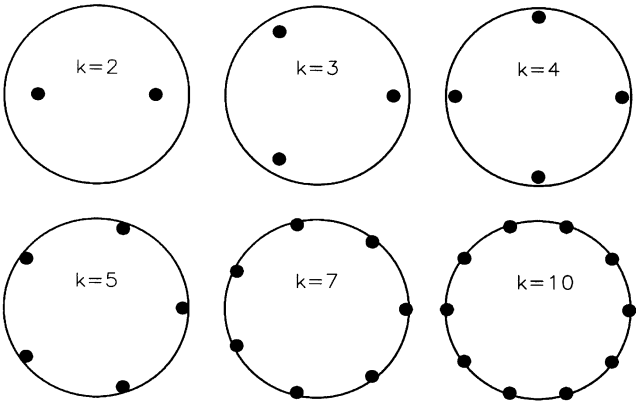

FIG. 7. *Sets of k evenly spaced points on a circle, indicating the support of a self-consistent approximation to the bivariate standard normal distribution, for $k = 2, 3, 4, 5, 7$ and 10, along with the limiting self-consistent circle as $k \to \infty$.*

The next example illustrates a self-consistent distribution whose support consists of a principal component axis and two points.

EXAMPLE 6.2. Consider the bivariate normal random vector $\mathbf{X} = \binom{X_1}{X_2}$ with mean $\mathbf{0}$ and covariance matrix $\mathrm{diag}(\sigma^2, 1)$. There exists a self-consistent random vector $\mathbf{Y}(\sigma)$ whose support consists of the x_1-axis along with the points $(0, \pm d)$. More precisely, with $\mathbf{y}_1 = (0, d)'$ and $\mathbf{y}_2 = (0, -d)'$,

$$\mathbf{Y}(\sigma) = \begin{cases} \mathbf{y}_1, & \text{if } \|\mathbf{X} - \mathbf{y}_1\| < |X_2|, \\ \mathbf{y}_2, & \text{if } \|\mathbf{X} - \mathbf{y}_2\| < |X_2|, \\ (X_1, 0)', & \text{else.} \end{cases}$$

For $\sigma = 1$, numerical computations indicate that $d \approx 1.43$ and the MSE of $\mathbf{Y}(1)$ is approximately 0.55. For $\sigma = 1.5$, we have $d \approx 1.49$ with MSE ≈ 0.83.

Table 1 gives the MSE for a few self-consistent approximations to the $N_2(\mathbf{0}, \mathbf{I}_2)$ distribution.

7. ORTHOGONAL COMPLEMENTS OF SELF-CONSISTENT DISTRIBUTIONS

Let $\mathscr{L}^2$ be the Hilbert space of square-integrable random vectors on a probability space $(\Omega, \mathscr{F}, \mathscr{P})$. Suppose $\mathbf{Y}$ is self-consistent for $\mathbf{X}$. The operator $P_{\mathbf{Y}}$ defined as $P_{\mathbf{Y}}(\mathbf{U}) = \mathscr{E}[\mathbf{U}|\mathbf{Y}]$ is a projection operator onto the closed subspace $M_{\mathbf{Y}} = P_{\mathbf{Y}}(\mathscr{L}^2)$. Each

TABLE 1
Self-consistent approximations $\mathbf{Y}$ *for the* $N_2(\mathbf{0}, \mathbf{I}_2)$ *distribution*

Support $\mathscr{S}(\mathbf{Y})$	MSE
Mean $(0, 0)$	2
Two principal points $\pm\begin{pmatrix}\sqrt{2/\pi}\\0\end{pmatrix}$	$2-\frac{2}{\pi}\approx 1.36$
x_1-axis	1.0
Four principal points $\begin{pmatrix}\pm 2/\sqrt{\pi}\\0\end{pmatrix}, \begin{pmatrix}0\\\pm 2/\sqrt{\pi}\end{pmatrix}$	$2-\frac{4}{\pi}\approx 0.73$
x_1-axis and points $\begin{pmatrix}0\\\pm 1.43\end{pmatrix}$ (see Example 6.1)	0.55
Circle $(r=\sqrt{\pi/2})$	0.43
x_1-axis and x_2-axis (see Example 4.4)	$1-\frac{2}{\pi}\approx 0.36$
Circle $(r=1.50)$ and origin $(0, 0)$	0.30
Two concentric circles $(r_1=0.83,\ r_2=1.92)$	0.15

element $\mathbf{U}\in\mathscr{L}^2$ has a unique decomposition $\mathbf{U}=\mathscr{E}[\mathbf{U}|\mathbf{Y}]+\mathbf{Z}$, where $\mathbf{Z}\in M_{\mathbf{Y}_\perp}=\{\mathbf{W}\in\mathscr{L}^2\colon \mathscr{E}[\mathbf{W}'\mathbf{V}]=0,\ \forall\,\mathbf{V}\in M_{\mathbf{Y}}\}$ (e.g., see Friedman, 1982, page 205). Therefore, $P_{\mathbf{Y}}(\mathbf{X})=\mathbf{Y}$ and we shall define the *orthogonal complement* of a self-consistent $\mathbf{Y}$ as $\mathbf{Y}_\perp=\mathbf{X}-\mathbf{Y}$.

If $\mathbf{Y}$ is self-consistent for $\mathbf{X}$, then $\mathbf{Y}_\perp$ may or may not be self-consistent for $\mathbf{X}$. For instance, $\mathbf{X}$ is self-consistent for $\mathbf{X}$, and the orthogonal complement of $\mathbf{X}$ is $\mathbf{0}$, which is not self-consistent unless $\mathscr{E}[\mathbf{X}]=\mathbf{0}$. The following examples illustrate nontrivial cases where $\mathbf{Y}_\perp$ is self-consistent (Example 7.1) and not self-consistent (Example 7.2).

EXAMPLE 7.1. Suppose X is uniformly distributed on the interval $[-1, 1]$. Then

$$Y=\begin{cases}-1/2, & \text{if } X<0,\\ 1/2, & \text{if } X\ge 0,\end{cases}$$

has support consisting of two principal points of X, and Y is self-consistent for X. The orthogonal complement $Y_\perp$ is uniformly distributed on $[-1/2, 1/2]$ and is self-consistent for X.

EXAMPLE 7.2. Let $X\sim N(0, 1)$. Then

$$Y=\begin{cases}-\sqrt{2/\pi}, & \text{if } X<0,\\ \sqrt{2/\pi}, & \text{if } X\ge 0,\end{cases}$$

is self-consistent for X. However, $Y_\perp=X-Y$ is not self-consistent. If $0<|t|<\sqrt{2/\pi}$, then

$$\mathscr{E}[X|Y_\perp=t]=t+\sqrt{\frac{2}{\pi}}\frac{\phi(t+\sqrt{2/\pi})-\phi(t-\sqrt{2/\pi})}{\phi(t-\sqrt{2/\pi})+\phi(t+\sqrt{2/\pi})}\neq t.$$

As Example 4.5 demonstrates, the orthogonal complement of a self-consistent projection onto a principal component axis may not be self-consistent. Suppose $\mathbf{Y}=\mathbf{P}\mathbf{X}$ is self-consistent for $\mathbf{X}$, where $\mathbf{P}$ is an orthogonal projection matrix onto a subspace spanned by eigenvectors of the covariance matrix of $\mathbf{X}$. If the conditional expectation is linear as in the case of elliptical distributions, then the orthogonal projection $\mathbf{Y}_\perp=(\mathbf{I}-\mathbf{P})\mathbf{X}$ is also self-consistent for $\mathbf{X}$.

8. SELF-CONSISTENCY AND THE EM ALGORITHM

The term self-consistency was, to our knowledge, first used by Efron (1967) to describe a class of estimators of a distribution function $F(t)$ in the presence of censored data. If $x_1,\ldots,x_N$ are observed data from a distribution F, the nonparametric maximum likelihood estimate of F is $\hat{F}(t)=\sum_{i=1}^N I[x_i\le t]/N$, where $I[\cdot]$ is the indicator function. For all censored observations the function $I[x_i\le t]$ cannot be evaluated. If $\mathbf{y}$ denotes the observed data, including censoring times for the censored observations, and F^* denotes a distribution function, then

$$\mathscr{P}(x_i\le t|\mathbf{y}, F^*)=E(I[x_i\le t]|\mathbf{y}, F^*)$$

may be used to estimate $I[x_i\le t]$ for all censored observations. A distribution function $F^*(t)$ is called a self-consistent estimate of the unknown distribution function $F(t)$ if

$$F^*(t)=\frac{1}{N}\sum_{i=1}^N \mathscr{P}(x_i\le t|\mathbf{y}, F^*)$$

for all t (Efron, 1967; Laird, 1988). That is, if we substitute the estimate F^* in the calculation of the expected values, we obtain the same estimate F^*. In other words, the estimate F^* "confirms itself." In an iterative algorithm F^* corresponds to a fixed point. This parallels the interpretation of self-consistent points in a k-means algorithm; see Section 9.

The expectation–maximization (EM) algorithm (Dempster, Laird and Rubin, 1967; Little and Rubin, 1987) is an iterative procedure for maximizing the log-likelihood in the presence of missing data. Suppose we have a model for complete data $\mathbf{X}$, with density $f(\mathbf{x}, \theta)$ indexed by an unknown parameter θ. Write $\mathbf{X}=(\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}})$, where $\mathbf{X}_{\text{obs}}$ represents the

observed part of the data, and $\mathbf{X}_{\text{mis}}$ the missing part. Let $l(\theta|\mathbf{X})$ denote the complete data log-likelihood, and let $\theta^{(t)}$ denote the current value of the parameter estimate in iteration t of the EM algorithm. Each iteration of the EM algorithm consists of an E (expectation) step and an M (maximization) step. The E step corresponds to taking the expectation of the complete data log-likelihood, given the observed data $\mathbf{X}_{\text{obs}}$, and using the current value $\theta^{(t)}$ of the parameter estimate. That is, the E step computes

$$Q(\theta, \theta^{(t)}) = E[l(\theta;\mathbf{X})|\mathbf{X}_{\text{obs}}, \theta^{(t)}].$$

The M step then finds $\theta^{(t+1)}$ which maximizes $Q(\theta, \theta^{(t)})$ over all θ in the parameter space. Convergence is reached if $\theta^{(t+1)} = \theta^{(t)}$. Thus the final estimate, denoted by $\hat{\theta}$, is again a fixed point of the algorithm, and the estimate $\hat{\theta}$ "confirms itself" in any further iteration of the algorithm.

The EM algorithm has been shown to converge under general conditions to a maximum of the likelihood function based on the observed data $\mathbf{X}_{\text{obs}}$. Since an iteration of the EM algorithm can never decrease the log-likelihood, Cox and Oakes (1984, page 171) define the self-consistency condition for the maximum likelihood estimator $\hat{\theta}$ as

$$Q(\theta, \hat{\theta}) \leq Q(\hat{\theta}, \hat{\theta})$$

for all θ in the parameter space.

If the density of the complete data $\mathbf{X}$ is from the exponential family, we can establish a direct connection between our notion of self-consistency and the notion of a self-consistent estimator just explained. Suppose X has a density of the form

$$f(\mathbf{X};\theta) = b(\mathbf{X}) \exp[\theta' s(\mathbf{X})]/a(\theta),$$

where $\theta \in \mathbb{R}^d$ is a parameter vector, $s(\mathbf{X})$ is a d-vector of complete-data sufficient statistics, and a and b are functions of θ and $\mathbf{X}$, respectively. Then the E step simplifies to

$$s^{(t)} = E[s(\mathbf{X})|\mathbf{X}_{\text{obs}}, \theta^{(t)}].$$

By Lemma 2.5, $s^{(t)}$ is self-consistent for $s(\mathbf{X})$, that is, $\mathscr{E}[s(\mathbf{X})|s^{(t)}] = s^{(t)}$. The M step determines the updated estimate $\theta^{(t+1)}$ as the solution of the equation

$$E[s(\mathbf{X});\theta] = s^{(t)},$$

based on which the next conditional expectation is taken. Convergence is reached when the sequence $\{s^{(t)}\}_{t\geq 1}$ of self-consistent random variables has stabilized, that is, $s^{(t+1)} = s^{(t)}$. Thus the EM algorithm generates a sequence of self-consistent random variables for a sufficient statistic $s(\mathbf{X})$, and the maximum likelihood estimator, which corresponds to a stationary point in the sequence, satisfies the self-consistency condition as defined in Cox and Oakes (1984).

9. DISCUSSION

The notion of self-consistency treated in this article gives a unified theoretical basis to principal components and curves, principal points, principal variables and other statistical techniques. Self-consistency also provides a framework for combining these techniques as shown in examples where aspects of principal components are linked with self-consistent points and where self-consistent curves are obtained as limiting cases of sets of self-consistent points. Self-consistency appears occasionally in the statistical literature, without being explicitly named. For instance, Bandeen-Roche (1994, page 1450) applies it to additive mixtures. Another intriguing example is as follows. If $\bar{X}$ and S^2 denote the mean and variance of a sample from a Poisson distribution, then $\bar{X}$ is self-consistent for S^2 (Casella and Berger, 1990, page 339), which by Lemma 2.3 implies $\text{var}[\bar{X}] < \text{var}[S^2]$.

Many research questions remain open. For instance, for nonspherical elliptically symmetric distributions we do not know if there exist self-consistent distributions whose support is a nonlinear curve. More important, the area of estimation of self-consistent "objects" has many open problems. Cluster means obtained from a k-means clustering algorithm (Hartigan, 1975) are nonparametric estimators of self-consistent points because they are self-consistent points of the empirical distribution. Estimation of self-consistent curves as proposed in Hastie and Stuetzle (1989) is quite similar. Starting with a set A_0 which consists of a line spanned by the first eigenvector of the covariance matrix, the conditional mean of each $y \in A_0$ over its domain of attraction is computed, using a smoothing algorithm, and A_1 is defined to be the set of these conditional means. If $A_1 = A_0$, then A_0 is self-consistent and the process stops. Otherwise, continue by letting A_2 denote the set of conditional means of the elements in A_1 over their respective domains of attraction, and so on, until convergence is reached. Similar ideas are used as well in the computation of semiparametric estimators of principal points (Flury, 1993), which are based on the k-means algorithm but restricted to follow certain patterns of principal points as suggested by the theory of principal points for elliptical distributions (Tarpey, Li and Flury, 1995).

The notion of self-consistency shows also a remarkable similarity between the EM algorithm

and the k-means algorithm. Suppose a k-means algorithm is applied to a random vector $\mathbf{X}$. For an initial set of points $\{\mathbf{y}_1(1), \ldots, \mathbf{y}_k(1)\}$, let $\mathbf{Y}_1 = \sum_{j=1}^{k} \mathbf{y}_j(1) I\{\mathbf{X} \in D_j(1)\}$, where $D_j(1)$ is the domain of attraction of point $\mathbf{y}_j(1)$. Then setting $\mathbf{Y}_2 = \mathscr{E}[\mathbf{X}|\mathbf{Y}_1]$ may be viewed as the E step of the k-means algorithm, and, by Lemma 2.5, $\mathbf{Y}_2$ is self-consistent for $\mathbf{X}$. The analog to the maximization step in the EM algorithm is then to update the domains of attraction $D_j(2)$ for the new $\mathbf{y}_j(2)$ and define $\mathbf{Y}_3 = \sum_{j=1}^{k} \mathbf{y}_j(2) I\{\mathbf{X} \in D_j(2)\}$. This may actually be called a minimization step because each point in the support of $\mathbf{X}$ is allocated to the nearest representative among the $\mathbf{y}_j(2)$, thus minimizing the within-group variability. The algorithm continues by iterating between these two steps. Once the algorithm converges so that $\mathbf{Y}(t) = \mathbf{Y}(t+1) = \mathbf{Y}^*$, then $\mathbf{Y}^*$ is self-consistent for $\mathbf{X}$, and the k points in the support of $\mathbf{Y}^*$ correspond to conditional means of Voronoi regions; that is, the support of $\mathbf{Y}^*$ corresponds to k self-consistent points of $\mathbf{X}$ in the sense of Definition 6.1.

Therefore, both the EM algorithm and the k-means algorithm have an expectation step which produces a self-consistent random vector. The final product after convergence is a self-consistent random variable that corresponds to a local maximum of the log-likelihood function for the EM algorithm, and to a set of self-consistent points for the k-means algorithm.

Closely related to the k-means algorithm as well as to principal curves and surfaces is the *self-organizing map* (SOM) (Kohonen, 1995) from the literature on neural networks. Like the k-means algorithm, the self-organizing map begins with a set of k initial points $\{\mathbf{y}_1(1), \ldots, \mathbf{y}_k(1)\}$ or "reference" vectors. Associated with each $\mathbf{y}_j(1)$ is a "neuron," a point in a two-dimensional array. This array is typically arranged as a hexagonal or rectangular lattice. The input to the algorithm consists of observations $\mathbf{x}_t$, $t = 1, 2, \ldots$, from some distribution F. The SOM algorithm updates the reference vectors based on the formula

$$\mathbf{y}_j(t+1) = \mathbf{y}_j(t) + h_{cj}[\mathbf{x}_t - \mathbf{y}_i(t)], \quad t = 1, 2, \ldots .$$

The function h_{cj} is a "neighborhood" function. The subscript c refers to the reference vector which is closest to $\mathbf{x}_t$, that is, $\|\mathbf{x}_t - \mathbf{y}_c(t)\| = \min_i\{\|\mathbf{x}_t - \mathbf{y}_i(t)\|\}$. Thus the neighborhood function allows the closest reference vector $\mathbf{y}_c(t)$ to be updated by $\mathbf{x}_t$ as well as reference vectors that correspond to a "neighborhood" of $\mathbf{y}_c(t)$. This parallels the use of a smoother in estimation of principal curves (Hastie and Stuetzle, 1989), where each sample point influences not only the particular point on the curve on which it is projected, but all points in an interval around the projection. Thus the self-organizing maps may be viewed as a discrete analog of principal curves and surfaces.

A special case of the SOM was given by MacQueen (1967),

$$\mathbf{y}_j(t+1) = \mathbf{y}_j(t) + \frac{1}{w_j(t)+1}(\mathbf{x}_t - \mathbf{y}_j(t)),$$

where the weights $w_j(t)$ are defined by $w_j(1) = 1$ and $w_j(t+1) = w_j(t) + 1$ if $j = c$ and $w_j(t+1) = w_j(t)$ if $j \neq c$. Thus the neighborhood function updates only the reference vector which is closest to the input $\mathbf{x}_t$. If the algorithm converges, then it must converge to a set of k self-consistent points (MacQueen uses the term *unbiased* for a set of self-consistent points) of the distribution F (e.g., see Kohonen, 1995, page 105).

The problem of self-consistency of the orthogonal projection associated with linear principal components opens some questions as well. To our knowledge, all existing tests for principal components or principal component subspaces are based on the fact that principal component subspaces are spanned by eigenvectors of the covariance matrix. The one-dimensional subspace spanned by an eigenvector may or may not be the support of a self-consistent random variable. It would be useful to have a criterion for deciding whether or not, for given data, a principal component in the traditional sense satisfies the criterion of self-consistency, without making parametric assumptions.

APPENDIX: PROOFS OF SELECTED RESULTS

PROOF OF LEMMA 2.3. Without loss of generality assume $\mathscr{E}[\mathbf{X}] = \mathbf{0}$. For part (i), by self-consistency of $\mathbf{Y}$ for $\mathbf{X}$ and using the conditional variance formula $\mathrm{Cov}[\mathbf{X}] = \mathrm{Cov}[\mathscr{E}[\mathbf{X}|\mathbf{Y}]] + \mathscr{E}[\mathrm{Cov}[\mathbf{X}|\mathbf{Y}]]$, we have

$$\mathrm{Cov}[\mathbf{X}] = \mathrm{Cov}[\mathbf{Y}] + \mathscr{E}[\mathrm{Cov}[\mathbf{X}|\mathbf{Y}]].$$

But $\mathrm{Cov}[\mathbf{X}|\mathbf{Y}]$ is positive semidefinite almost surely, and hence (i) follows.

For part (ii) we have

$$\begin{aligned}
\mathscr{E}\|\mathbf{X} - \mathbf{Y}\|^2 &= \mathscr{E}[\mathbf{X}'\mathbf{X}] - 2\mathscr{E}[\mathbf{Y}'\mathbf{X}] + \mathscr{E}[\mathbf{Y}'\mathbf{Y}] \\
&= \mathrm{tr}(\mathbf{\Psi}_{\mathbf{X}}) - 2\mathscr{E}[\mathscr{E}[\mathbf{Y}'\mathbf{X}|\mathbf{Y}]] + \mathrm{tr}(\mathbf{\Psi}_{\mathbf{Y}}) \\
&= \mathrm{tr}(\mathbf{\Psi}_{\mathbf{X}}) - 2\mathscr{E}[\mathbf{Y}'\mathscr{E}[\mathbf{X}|\mathbf{Y}]] + \mathrm{tr}(\mathbf{\Psi}_{\mathbf{Y}}) \\
&= \mathrm{tr}(\mathbf{\Psi}_{\mathbf{X}}) - 2\mathscr{E}[\mathbf{Y}'\mathbf{Y}] + \mathrm{tr}(\mathbf{\Psi}_{\mathbf{Y}}) \\
&= \mathrm{tr}(\mathbf{\Psi}_{\mathbf{X}}) - \mathrm{tr}(\mathbf{\Psi}_{\mathbf{Y}}). \qquad \square
\end{aligned}$$

PROOF OF LEMMA 2.4. Since $\mathbf{Y}$ is self-consistent for $\mathbf{X}$, $\mathscr{E}[\mathbf{PX}|\mathbf{Y}] = \mathbf{P}\mathscr{E}[\mathbf{X}|\mathbf{Y}] = \mathbf{PY} = \mathbf{Y}$ a.s. For a given $\mathbf{y} \in \mathbb{R}^p$, let $\mathbf{w} = \mathbf{A}_1'\mathbf{y}$. Then $\{\mathbf{Y} = \mathbf{y}\} = \{\mathbf{A}_1'\mathbf{Y} = \mathbf{w}\}$. Multiplying both sides of the equation $\mathscr{E}[\mathbf{X}|\mathbf{Y} = \mathbf{y}] = \mathbf{y}$ on the left by $\mathbf{A}_1'$ gives $\mathscr{E}[\mathbf{A}_1'\mathbf{X}|\mathbf{A}_1'\mathbf{Y} = \mathbf{w}] = \mathbf{w}$. □

PROOF OF THEOREM 4.1. Let $\mathbf{A} = [\mathbf{A}_1 : \mathbf{A}_2]$ denote an orthogonal $p \times p$ matrix, partitioned into q columns $\mathbf{A}_1$ and $p - q$ columns $\mathbf{A}_2$ such that the columns of $\mathbf{A}_1$ span $\mathscr{A}$. Then $\mathbf{A}_2'\mathbf{Y} = \mathbf{0}$ a.s.

The covariance matrix of $\mathbf{A}'\mathbf{X} = (\mathbf{A}_1'\mathbf{X}, \mathbf{A}_2'\mathbf{X})'$ is

$$\begin{pmatrix} \mathbf{A}_1'\boldsymbol{\Psi}\mathbf{A}_1 & \mathbf{A}_1'\boldsymbol{\Psi}\mathbf{A}_2 \\ \mathbf{A}_2'\boldsymbol{\Psi}\mathbf{A}_1 & \mathbf{A}_2'\boldsymbol{\Psi}\mathbf{A}_2 \end{pmatrix}.$$

By the linearity of the conditional expectation, $\mathscr{E}[\mathbf{A}_2'\mathbf{X}|\mathbf{A}_1'\mathbf{X}] = \mathbf{A}_2'\boldsymbol{\Psi}\mathbf{A}_1(\mathbf{A}_1'\boldsymbol{\Psi}\mathbf{A}_1)^{-1}\mathbf{A}_1'\mathbf{X}$. Since $\mathbf{Y}$ is measurable with respect to $\mathbf{A}_1'\mathbf{X}$,

$$\begin{aligned}
\mathscr{E}[\mathbf{A}_2'\mathbf{X}|\mathbf{Y}] &= \mathscr{E}[\mathscr{E}[\mathbf{A}_2'\mathbf{X}|\mathbf{A}_1'\mathbf{X}]|\mathbf{Y}] \\
&= \mathscr{E}[\mathbf{A}_2'\boldsymbol{\Psi}\mathbf{A}_1(\mathbf{A}_1'\boldsymbol{\Psi}\mathbf{A}_1)^{-1}\mathbf{A}_1'\mathbf{X}|\mathbf{Y}] \\
&= \mathbf{A}_2'\boldsymbol{\Psi}\mathbf{A}_1(\mathbf{A}_1'\boldsymbol{\Psi}\mathbf{A}_1)^{-1}\mathbf{A}_1'\mathscr{E}[\mathbf{X}|\mathbf{Y}] \\
&= \mathbf{A}_2'\boldsymbol{\Psi}\mathbf{A}_1(\mathbf{A}_1'\boldsymbol{\Psi}\mathbf{A}_1)^{-1}\mathbf{A}_1'\mathbf{Y} \\
&\qquad \text{(by self-consistency of } \mathbf{Y} \text{ for } \mathbf{X}).
\end{aligned}$$

Also by self-consistency, $\mathscr{E}[\mathbf{A}_2'\mathbf{X}|\mathbf{Y}] = \mathbf{A}_2'\mathbf{Y} = \mathbf{0}$. Therefore, $\mathbf{A}_2'\boldsymbol{\Psi}\mathbf{A}_1(\mathbf{A}_1'\boldsymbol{\Psi}\mathbf{A}_1)^{-1}\mathbf{A}_1'\mathbf{Y} = \mathbf{0}$, which implies $\mathbf{A}_2'\boldsymbol{\Psi}\mathbf{A}_1 = \mathbf{0}$. That is, the columns of $\mathbf{A}_2$ are orthogonal to the columns of $\boldsymbol{\Psi}\mathbf{A}_1$, or $\boldsymbol{\Psi}\mathbf{A}_1 = \mathbf{A}_1\mathbf{H}$ for some orthogonal matrix $\mathbf{H}$ of dimension $q \times q$. This means in turn that the columns of $\mathbf{A}_1$ span the same q-dimensional subspace as some q eigenvectors of $\boldsymbol{\Psi}$. □

PROOF OF THEOREM 5.1. Assume without loss of generality that $\mu(t) = 0$. Using an argument similar to that given by Hastie and Stuetzle (1989, page 505), we have

$$\begin{aligned}
\int_0^T C(s,t)f(t)\,dt &= \int_0^T \mathscr{E}[x(s)x(t)]f(t)\,dt \\
&= \mathscr{E}\left[x(s)\int_0^T x(t)f(t)\,dt\right] \\
&= \mathscr{E}[x(s)\alpha] \\
&= \mathscr{E}[\mathscr{E}[x(s)\alpha|\alpha f(s)]] \\
&= \mathscr{E}[\alpha\mathscr{E}[x(s)|\alpha f(s)]] \\
&= \mathscr{E}[\alpha^2]f(s) \quad \text{(by self-consistency)}.
\end{aligned}$$

Thus, $f(t)$ is an eigenfunction of the covariance function of $x(t)$. □

ACKNOWLEDGMENTS

The authors would like to thank the Editor, a referee, Nicola Loperfido and Ann Mitchell for constructive comments on earlier versions of this article.

REFERENCES

ANDERSON, T. W. (1963). Asymptotic theory for principal component analysis. *Ann. Math. Statist.* **34** 122–148.

BANDEEN-ROCHE, K. (1994). Resolution of additive mixtures into source components and contributions: a compositional approach. *J. Amer. Statist. Assoc.* **89** 1450–1458.

BANFIELD, J. and RAFTERY, A. (1992). Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *J. Amer. Statist. Assoc.* **87** 7–16.

BICKEL, P. J. and DOKSUM, K. A. (1977). *Mathematical Statistics*. Holden-Day, San Francisco.

CASELLA, G. and BERGER, R. L. (1990). *Statistical Inference*. Duxbury Press, Belmont, CA.

CASTRO, P. E., LAWTON, W. H. and SYLVESTRE, E. A. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics* **28** 329–337.

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

COX, D. R. and OAKES, D. (1984). *Analysis of Survival Data*. Chapman and Hall, New York.

CUESTA, J. A. and MATRAN, C. (1988). The strong law of large numbers for k-means and best possible nets of Banach valued random variables. *Probab. Theory Related Fields* **78** 523–534.

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

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

EFRON, B. (1967). The two sample problem with censored data. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **4** 831–853. Univ. California Press, Berkeley.

EUBANK, R. L. (1988). Optimal grouping, spacing, stratification, and piecewise constant approximation. *SIAM Rev.* **30** 404–420.

FANG, K. and HE, S. (1982). The problem of selecting a given number of representative points in a normal population and a generalized Mill's ratio. Technical report, Dept. Statistics, Stanford Univ.

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 collection of curves: a case for principal points. *Amer. Statist.* **47** 304–306.

FRIEDMAN, A. (1982). *Foundations of Modern Analysis*. Dover, New York.

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

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

IEEE (1982). *IEEE Trans. Inform. Theory* **28**. (Special issue on quantization.)

IYENGAR, S. and SOLOMON, H. (1983). Selecting representative points in normal populations. In *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on his 60th Birthday* (M. H. Rizvi, J. Rustagi and D. Siegmund, eds.) 579–591. Academic Press, New York.

JOLICOEUR, P. (1968). Interval estimation of the slope of the major axis of a bivariate normal distribution in the case of a small sample. *Biometrics* **24** 679–682.

JOLICOEUR, P. and MOSIMANN, J. E. (1960). Size and shape variation in the painted turtle; a principal component analysis. *Growth* **24** 339–354.

KOHONEN, T. (1995). *Self-Organizing Maps*. Springer, Berlin.

KSHIRSAGAR, A. M. (1961). The goodness of fit of a single (non-isotropic) hypothetical principal component. *Biometrika* **48** 397–407.

LAIRD, N. (1988). Self-Consistency. In *Encyclopedia of Statistical Sciences*, **8** 347–351. Wiley, New York.

LEBLANC, M. and TIBSHIRANI, R. (1994). Adaptive principal surfaces. *J. Amer. Statist. Assoc.* **89** 53–64.

LITTLE, J. and RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

LLOYD, S. (1982). Least squares quantization in PCM. *IEEE Trans. Inform. Theory* **28** 129–149.

MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **3** 281–297. Univ. California Press, Berkeley.

MALLOWS, C. L. (1961). Latent vectors of random symmetric matrices. *Biometrika* **48** 133–149.

MCCABE, G. P. (1984). Principal variables. *Technometrics* **26** 137–144.

PÄRNA, K. (1990). On the existence and weak convergence of k-centres in Banach spaces. *Acta et Commentationes Universitatis Tartuensis* **893** 17–28.

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. (1996). An algorithm for computing principal points with respect to a loss function in the unidimensional case. *Statistics and Computing* **6** 187–190.

SCHOTT, J. R. (1991). A test for a specific principal component of a correlation matrix. *J. Amer. Statist. Assoc.* **86** 747–751.

TARPEY, T. (1995). Principal points and self-consistent points of symmetric multivariate distributions. *J. Multivariate Anal.* **53** 39–51.

TARPEY, T. (1996). Self-consistent patterns for symmetric, multivariate distributions. Unpublished manuscript.

TARPEY, T., LI, L. and FLURY, B. (1995). Principal points and self-consistent points of elliptical distributions. *Ann. Statist.* **23** 103–112.

TIBSHIRANI, R. (1992). Principal curves revisited. *Statistics and Computing* **2** 183–190.

TYLER, D. (1983). A class of asymptotic tests for principal component vectors. *Ann. Statist.* **11** 1243–1250.

ZOPPÈ, A. (1995). Principal points of univariate continuous distributions. *Statistics and Computing* **5** 127–132.