The Annals of Statistics
1994, Vol. 22, No. 4, 1635–1673

ANALYSIS OF ADDITIVE DEPENDENCIES AND CONCURVITIES USING SMALLEST ADDITIVE PRINCIPAL COMPONENTS

BY DEBORAH J. DONNELL, ANDREAS BUJA AND WERNER STUETZLE

StatSci, AT&T Bell Laboratories and University of Washington

Additive principal components are a nonlinear generalization of linear principal components. Their distinguishing feature is that linear forms $\Sigma_i a_i X_i$ are replaced with additive functions $\Sigma_i \phi_i(X_i)$. A considerable amount of flexibility for fitting data is gained when linear methods are replaced with additive ones.

Our interest is in the *smallest* principal components, which is somewhat uncommon. Smallest additive principal components amount to data descriptions in terms of approximate implicit equations: $\Sigma_i \phi_i(X_i) \approx 0$. Estimation of such equations is a data-analytic method in its own right, competing in some cases with the more customary regression approaches. It is also a diagnostic tool in additive regression for detection of so-called "concurvity." This term describes degeneracies among predictor variables in additive regression models, similar to collinearity in linear regression models. Concurvity may lead to statistically unstable contributions of variables to additive models. As an example, we show in a reanalysis of the ozone data from Breiman and Friedman that concurvity does indeed exist in this particular data, a fact which may impact the interpretation of the additive fits.

In the second half of this paper, we give some second-order theory, including the description of null situations and eigenexpansions derived from associated eigenproblems. We show how ACE and additive principal components are related, and we outline some analytical methods for theoretical calculations of additive principal components. Lastly we consider methods of estimation and computation.

Additive principal components have had a long tradition in psychometric research and correspondence analysis. They have started receiving attention by statisticians only in recent years.

1. Introduction. The smallest *linear* principal component is a linear function of the data, $\Sigma_i a_i X_i$, with smallest variance, which implies the data lie near the hyperplane defined by the linear constraint, $\Sigma a_i x_i = 0$. The smallest *additive* principal component is an additive function of the data, $\Sigma_i \phi_i(X_i)$, with smallest variance, which implies the data satisfy as nearly as possible the corresponding additive constraint:

$$\sum_{i=1}^{p} \phi_i(X_i^{'}) = 0.$$

Analogous to the linear case, an additive constraint defines an additive manifold of codimension 1, and data nearly satisfying this constraint lie near this additive manifold.

Received May 1991; revised March 1993.

AMS 1991 *subject classifications.* Primary 62H25; secondary 62G07

Key words and phrases. Additive models, nonlinear multivariate analysis, multiple correspondence analysis, scaling, nonlinear transformations, horseshoe effect.

We considered smallest additive principal components (APCs henceforth) first in the context of detecting instability in additive regression models. Nonlinear additive dependencies among predictor variables in additive regression are analogous to collinearity in linear regression. Suppose we were to fit an additive model $Y \approx \Sigma_{i=1}^{p} \psi_i(X_i)$ in the presence of what we shall call an exact *concurvity* between the predictors, that is, $\Sigma \phi_i = 0$ for some $\phi_i = \phi_i(X_i)$. In this situation, the alternative fit

$$Y \approx \sum_{i=1}^{p} (\psi_i + \varepsilon \phi_i)(X_i)$$

is indistinguishable from the initial one. While exact concurvity is unlikely, approximate concurvity may cause harm, too, in that some or all of the estimated ψ_i are likely to be unstable. [The need for diagnostic tools was mentioned in Buja and Kass (1985). The term "concurvity" was introduced in Buja, Donnell and Stuetzle (1986) and used in Buja, Hastie and Tibshirani (1989). See Hastie and Tibshirani (1990) for a broad discussion of additive regression models.]

Besides their role as regression diagnostics, smallest APCs form a data analysis tool in their own right. Smallest APCs estimate constraints, and constraints are a valuable exploratory tool for investigating dependencies of *low codimension* as opposed to factor analysis and largest principal components, both of which search for structure of *low dimension*. Estimation of constraints is more natural than regression when the search for structure is undirected; that is, no variables are designated a priori as predictors of a response of interest. Even if the problem asks for a regression treatment, it may still be worthwhile to analyze the data in a symmetric fashion via APCs to find the variables with strong dependencies. These variables may include the response, in which case regression may be successful, or they may not, in which case the data may be inappropriate for the intended purpose.

The idea of using other than linear functions in the analysis of dependencies has generated a large literature. The most comprehensive treatment of extensions of principal component analysis is the optimal scaling techniques developed by psychometricians [Kruskal and Shepard (1974) and Young, Takane and de Leeuw (1978)], nonlinear multivariate analysis of the Dutch school [Gifi (1990)] and correspondence analysis of the French school [Benzecri (1980), Lebart, Morineau and Warwick (1984) and Greenacre (1984)]. All of these techniques were originally intended for scoring/scaling categorical and ordinal data, to make them optimally suited for linear multivariate methods such as principal components. But extensions for "scoring" (nonlinearly transforming) *continuous* variables were also developed, for example by van Rijckevorsel (1982). Some population theory of "continuous correspondence analysis" was developed by Naouri (1970). Within statistics, related work is canonical analysis of contingency tables [Gilula and Haberman (1988)], while nonlinear transformation of continuous variables in the main thrust of the ACE method of Breiman and Friedman (1985). See Buja (1990) for more references and an examination of scoring/scaling methods in some analytically tractable cases.

The present paper gives a treatment of scoring and nonlinear transformations for principal components in the style of multiple correspondence analysis. The paper is unusual in its emphasis on the low end of the principal components spectrum. The first reference that points out the possibility of using small eigenvalues in a related context is Kettenring (1971). The scaling literature is almost exclusively concerned with dimension reduction and hence the upper end of the spectrum. Smallest and largest eigenvalues differ drastically in use and interpretation. Indeed, largest APCs pose a methodological problem which is nonexistent in linear principal components. A plot of the largest two linear principal components can be interpreted as a data projection covering a maximal amount of variance, but a similar interpretation is not available for a corresponding plot of the two largest APCs $\Sigma_i \phi_i^{(1)}(X_i)$ and $\Sigma_i \phi_i^{(2)}(X_i)$: these are two one-dimensional projections of two sets of *different* data transformations. [This problem does not exist for other scaling approaches, e.g., Gifi (1990).]

Additive principal components should not be confused with principal curves and surfaces [Hastie and Stuetzle (1989)]. These, too, are intended for nonlinear dimension reduction. Principal curves and surfaces attempt to parametrize one- or two-dimensional manifolds, while smallest APCs describe manifolds of low codimension in terms of implicit additive equations.

In what follows, we give first a minimal set of motivations and definitions (Section 2) to enable the reader to follow the data-analytic methodology, exemplified by a reanalysis of the ozone data of Breiman and Friedman (1985) (Section 3). We show the variable selection made by Breiman and Friedman for their ACE regression fit is far from serendipitous. There are two additive constraints among the variables of the full data set which could have caused instability in the ACE fit. Two other examples demonstrate the analysis for artificial data with known structure, one of which exhibits the so-called "horseshoe effect," an artifact well-known among psychometricians.

Section 4 may be of less interest to the casual reader, but is nevertheless necessary for an understanding of APCs: we describe properties of APCs, including sufficient conditions under which population APCs exist; give a characterization of APC null situations; present some eigenexpansions; describe tools for theoretical APC calculations; and dispose of some technicalities related to centering of variables. No asymptotic theory of APC estimation is attempted in this paper [see, however, Dauxois and Pousse (1977)]. Section 5 deals with finite sample estimation of APCs and computational methods. We consider two methods for the computation of APC estimates: (1) reduction to a finite-dimensional eigenvalue problem if conditional expectations are estimated by least squares regressions and smoothers and (2) power iterations based on general (possibly nonlinear) smoothers, a method which is more flexible but less well understood.

2. Motivation and definitions for smallest additive principal components.

2.1. *Motivation.* A natural approach to defining APCs is to extend any one of the well-known characterizations of linear principal components. A linear

combination $\Sigma a_i X_i$ is a smallest linear principal component iff the following equivalent characterizations hold:

- $\Sigma a_i X_i$ has minimal variance among all linear combinations with $\Sigma a_i^2 = 1$.
- $\Sigma a_i x_i = 0$ defines the manifold of codimension 1 minimizing expected squared distance to the data.
- $(a_1, a_2, \ldots, a_p)$ is an eigenvector for the smallest eigenvalue of $\Sigma = \mathbf{var}\,\mathbf{X}$.

A minimum variance definition of smallest APCs would use the additive function $\Sigma\phi_i(X_i)$ minimizing $\mathbf{var}\,\Sigma\phi_i(X_i)$, subject to some normalizing constraint. A geometric definition would find the additive manifold, described by $\Sigma\phi_i(x_i) = 0$, minimizing the expected squared distance to the observations. Unlike linear principal components, the APCs defined by these two definitions will not be the same. We decided to use the minimum variance characterization, which has two useful properties not shared by the geometric approach: (1) minimum variance leads to a characterization of APCs as solutions to an eigenproblem that generalizes the third characterization of linear principal components; and (2) finite sample estimates are easy to compute, since the criterion involves estimation of variance rather than estimation of the Euclidean distance between a nonlinear manifold and the data.

To proceed with a formal definition, we need (1) spaces of functions for the variable transformations and (2) a suitable normalizing constraint that generalizes $\Sigma a_i^2 = 1$ used in linear principal component analysis.

We assume the variable transformations $\phi_i = \phi_i(X_i)$ are in some closed subspace H_i of centered L_2 variables:

$$\phi_i \in H_i \subset \{\phi\colon \mathbf{E}\,\phi = 0,\ \mathbf{var}\,\phi < \infty\}.$$

Closedness is necessary for the existence of orthogonal projections, and centering is a sensible condition to get rid of unidentifiable constants in additive functions. Denote

$$\mathbf{\Phi} \stackrel{\text{def}}{=} (\phi_1, \ldots, \phi_p) \in \mathbf{H} \stackrel{\text{def}}{=} H_1 \times H_2 \times \cdots \times H_p.$$

We capture APC definitions for populations and for finite-sample estimation in a single framework by (1) allowing the joint distribution of the variables $X_1, \ldots, X_p$ to be either a (generally continuous) probability measure or an empirical distribution based on a sample, and (2) selecting the spaces H_i accordingly. In particular, H_i may be the set of all centered L_2 functions of X_i for a population definition of nonparametric APCs, while a choice of H_i as a finite-dimensional space of splines or polynomials may lead to useful nonparametric finite sample estimators of APCs. If H_i is just the set of linear functions $\phi_i = a_i X_i$, one gets back to population or finite sample linear principal components. This specialization can be used as a check that our definitions are proper extensions of linear principal components. In this instance, we assume of course that the raw variables X_i are centered and L_2. Furthermore, we will make the assumption

that they are standardized, $\mathbf{var} X_i = 1$, since we will only attempt to generalize linear principal components based on correlations as opposed to covariances.

These remarks are relevant for our choice of a normalizing constraint for APCs: $\Sigma_i \mathbf{var}\,\phi_i = 1$, which is seen to specialize to $\Sigma_i a_i^2 = 1$ when $\phi_i = a_i X_i$, using the assumption that the raw variables are standardized: $\mathbf{var}\,\phi_i = a_i^2$.

DEFINITION 2.1. The smallest additive principal component vector of the spaces $H_1, H_2, \ldots, H_p$, if it exists, is a random vector $\Phi = (\phi_i)_i$ which solves

$$\mathbf{var} \sum_{i=1}^{p} \phi_i = !\min \quad \text{subject to} \quad \sum_{i=1}^{p} \mathbf{var}\,\phi_i = 1, \qquad \phi_i \in H_i.$$

We can define a sequence of APCs, analogous to the sequence of (uncorrelated) linear principal components. This requires an additional orthogonality constraint. We define orthogonality between $\Phi^{(1)} = (\phi_i^{(1)})_i$ and $\Phi^{(2)} = (\phi_i^{(2)})_i$ by

$$\sum_i \mathbf{cov}\left(\phi_1^{(1)}, \phi_i^{(2)}\right) = 0.$$

This specializes to the proper orthogonality condition for linear principal components: if $\phi_i^{(1)} = a_i^{(1)} X_i$ and $\phi_i^{(2)} = a_i^{(2)} X_i$, we get $\Sigma_i \mathbf{cov}(\phi_i^{(1)}, \phi_i^{(2)}) = \Sigma_i a_i^{(1)} a_i^{(2)} = 0$.

DEFINITION 2.2. The kth smallest additive principal component vector of the spaces H_i, if it exists, is random vector $\Phi^{(k)} = (\phi_i^{(k)})_i$, which solves

$$\mathbf{var} \sum_i \phi_i = !\min \quad \text{subject to} \quad \sum_i \mathbf{cov}\left(\phi_i, \phi_i^{(l)}\right) = 0 \quad \text{for } l = 1, \ldots, k-1,$$
$$\text{and} \sum_i \mathbf{var}\,\phi_i = 1, \qquad \phi_i \in H_i.$$

If the spaces H_i are finite-dimensional, linear algebra grants existence of APCs. For infinite-dimensions, Section 4.3 discusses some of the usual sufficient conditions.

2.2. *Eigenproperties.* Linear principal components are characterized by the eigenequation $\Sigma \mathbf{a} = \lambda \mathbf{a}$. This generalizes to APCs as follows:

$$P_i \sum_j \phi_j = \lambda \phi_i, \tag{1}$$

where P_i denotes the orthogonal projection from the space of all L_2 functions onto the space H_i, and orthogonality is understood w.r.t. the covariance as the

inner product. If the underlying distribution is a (generally continuous) probability measure and each space H_i is chosen as the set of all centered L_2 functions of X_i, then $P_i = E^{X_i}$, the conditional expectation given X_i. If $X_1, \ldots, X_p$ carry the empirical distribution of a sample and $H_1, \ldots, H_p$ are finite-dimensional spaces of splines or polynomials, then $P_1, \ldots, P_p$ are spline or polynomial regressions. If H_i consists of the linear functions $\phi_i = a_i X_i$, we get back to the eigenequation for linear principal components, since P_i then is a simple linear regression onto X_i: $P_i Y = \mathbf{cov}(Y, X_i)X_i$ and hence $P_i \Sigma_j \phi_j = \Sigma_j a_j P_i X_j = \Sigma_j a_j \Sigma_{i,j} X_i = \lambda a_i X_i$, assuming the variables X_i are centered and standardized.

The eigenequation (1) for APCs will be derived in Section 4. The hierarchy of APCs, if it exists, is identical with the hierarchy of normalized eigensolutions, and the variance of an APC equals its eigenvalue: $\mathbf{var}\, \Sigma_i \phi_i = \lambda$. Here are some consequences (for proofs see Section 4):

1. APCs can only capture structure in pairwise marginals: the l.h.s. of (1) depends only on $P_i \phi_j$, that is, the pair H_i, H_j. Structure that depends on higher-order marginals is elusive for this class of techniques, as is common knowledge among psychometricians.
2. Pairwise orthogonality of the spaces H_i defines the "null" analysis for APCs: all eigenvalues are identical +1. For populations and H_i being all centered L_2 functions, this amounts to pairwise (but not joint) independence.
3. The eigenvalues are bounded below by 0 (since eigenvalues are variances), and above by p (the number of variables). The existence of a zero eigenvalue indicates perfect degeneracy/concurvity.

It is natural to define small and large eigenvalues according to whether they are below or above +1. APCs with small eigenvalues ($< +1$) describe degeneracies (small codimension), while those with large eigenvalues ($> +1$) describe factor structure (small dimension).

3. Interpretation and use of smallest APCs. In this section we first give a brief guide to APC analysis, then demonstrate the use of APC analysis in a reanalysis of the ozone data. Finally, we give two examples of artificial data to address the problem of parabolic transforms, known as "horseshoes." In the first example, the parabolas are artifactual, but in the second example they are meaningful.

3.1. *Data analysis with smallest APCs.* The APCs of a data set are characterized by the eigenvalues and the APC functions ϕ_i (synonymous: APC transforms). We begin with the interpretation of relevant APC quantities:

1. Eigenvalues: $\lambda = \mathbf{var}\, \Sigma_i \phi_i$. Eigenvalues of small APCs measure the strength of additive degeneracy. They are nonnegative and, by definition, below 1. An APC with zero eigenvalue corresponds to exact additive degeneracy. If the smallest eigenvalue is 1, the spaces H_i are pairwise orthogonal. In addition, size and spacing of different eigenvalues provide information about stability and uniqueness of APC estimates. If an eigenvalue has multiplicity

greater than 1, the APCs are indeterminate. While in practice this can be ruled out with probability 1 (except for certain purely discrete data), *approximate* APC multiplicity is of practical relevance.

2. APC: $\tilde{\phi} = \Sigma_i \phi_i$. The smallest few APCs, by definition, have minimal variance (under successive orthogonality constraints), hence their interpretation is akin to analysis of residuals. Ideally, they will be distributed symmetrically about 0: departures from symmetry, such as outliers or grouping in the APCs, indicate cases which are unusual with respect to the estimated implicit equation. Plotting APCs against each other will reveal patterns in the residual structure.
3. APC weights: $\mathbf{sd}\,\phi_i$. The standard deviations, or relative *weights*, of the transforms indicate the relative importance of the variables in an APC. In the presence of approximate APC multiplicity, where adjacent eigenvalues are very close, even large weights may not be stable. Interpretation therefore requires caution. Recall that the normalization constraint of APCs is $\Sigma_i\, \mathbf{sd}^2 \phi_i = 1$.
4. APC functions: ϕ_i. The shape of the transform indicates sensitivity to the values of x_i in the approximate degeneracy: a section with steep gradient defines a region of high sensitivity to changing values, while an (approximate) step function indicates sensitivity only to the corresponding levels of the variable. In trying to make sense of the additive equation $\Sigma_i \phi_i(x_i) \approx 0$, it may help to eliminate the transforms with the smallest standard deviations. For strong transforms, the shape of the resulting surface may be described by conditioning on all but two variables: $\cdots + \phi_i(x_i) + \cdots + \phi_j(x_j) + \cdots \approx 0$; the important point is that, conditional on x_k, $k \neq i, j$, high values of ϕ_i go with low values of ϕ_j and vice versa. However, caution should be used in inferring conditional bivariate relations among variable pairs since additional covariations (e.g., those described by other APCs) may prevent the possibility of holding all but two coordinates fixed.

The problem of APC indeterminacy requires one more comment. If, for example, the smallest and second smallest eigenvalues are very close, $\lambda^{(1)} \approx \lambda^{(2)}$, all linear combinations of the two smallest APCs should be considered as APCs as well, with the same approximate eigenvalue. It may therefore be useful to search for the most interpretable choices among trigonometric linear combinations of the form

$$\phi_i^{(\alpha)} = \cos \alpha \cdot \phi_i^{(1)} + \sin \alpha \cdot \phi_i^{(2)},$$

which constitutes a new version of the factor rotation problem. The use of trigonometric coefficients assures that the linear combination is normalized: $\Sigma_i\, \mathbf{var}\, \phi_i^{(\alpha)} = 1$. The combined weight of X_i in two APCs can be calculated as $(\mathbf{sd}^2 \phi_i^{(1)} + \mathbf{sd}^2 \phi_i^{(2)})^{1/2}$.

Since an APC determines a linear dependence in the *transformed* variables, translating this dependence to the original variables can be nontrivial. In our experience, interactive graphics tools are the most direct means of exploring the corresponding variable dependencies [Donnell (1987)].

3.2. *APC analysis of the ozone data.* Breiman and Friedman (1985) use the ACE algorithm on a data set which was collected for studying the relationship between atmospheric ozone concentration and meteorology in the Los Angeles basin. Breiman and Friedman chose ozone concentration as the response variable in their ACE analysis. In our reanalysis, we choose a self-contained APC approach, treating all variables equally and letting the data decide on the strongest dependencies, as opposed to using APCs for a concurvity analysis of the predictor variables.

The data set consists of daily measurements of ozone concentration (maximum one hour average), eight meteorological quantities for 330 days of 1976, and day of year. The variables are:

ozone:	Upland ozone concentration (ppm);
temp:	Sandburg Air Force Base temperature;
ibh:	inversion base height;
dpg:	Dagget pressure gradient (mmhg);
vis:	visibility in miles;
vh:	Vandenburg 500 millibar height;
humidity:	humidity (percent);
ibt:	inversion base temperature;
wind:	wind speed (mph)
doy:	day of year.

Day of year was included in the original analysis to detect seasonal effects in the ozone variable not captured by the other meteorological variables.

The estimates shown in this section are computed using regression splines with two internal knots located at the $\frac{1}{3}$ and $\frac{2}{3}$ quantiles of each variable, by the direct methods of Section 5.1. The three smallest eigenvalues are 0.03, 0.084 and 0.088. Here is a discussion of their APCs:

1. The smallest APC (Figure 1), with an eigenvalue of 0.030, indicates a very strong dependence between inversion base temperature, inversion base height and Sandburg temperature. Holding inversion base height fixed, there is a positive relationship between inversion base temperature and Sandburg temperature, which is not surprising. Since the transforms for inversion base height and inversion base temperature have the same direction of slope, there may exist a negative relation between these two variables, holding Sandburg temperature fixed.
2. The second-smallest APC (Figure 2), with an eigenvalue of 0.084, shows that Sandburg temperature and Vandenburg height tend to increase together. The transforms for these two variables indicate that their (positive) dependence is strongest for higher temperatures due to steep gradients in that area.
3. The third-smallest APC (Figure 3) has an eigenvalue of 0.088. It is a dependence involving ozone, Sandburg temperature, Dagget pressure and day of year most prominently, with lesser contributions from other variables. Note that ozone, the variable of greatest interest, appears for the first time with substantial weight in this third-smallest APC.

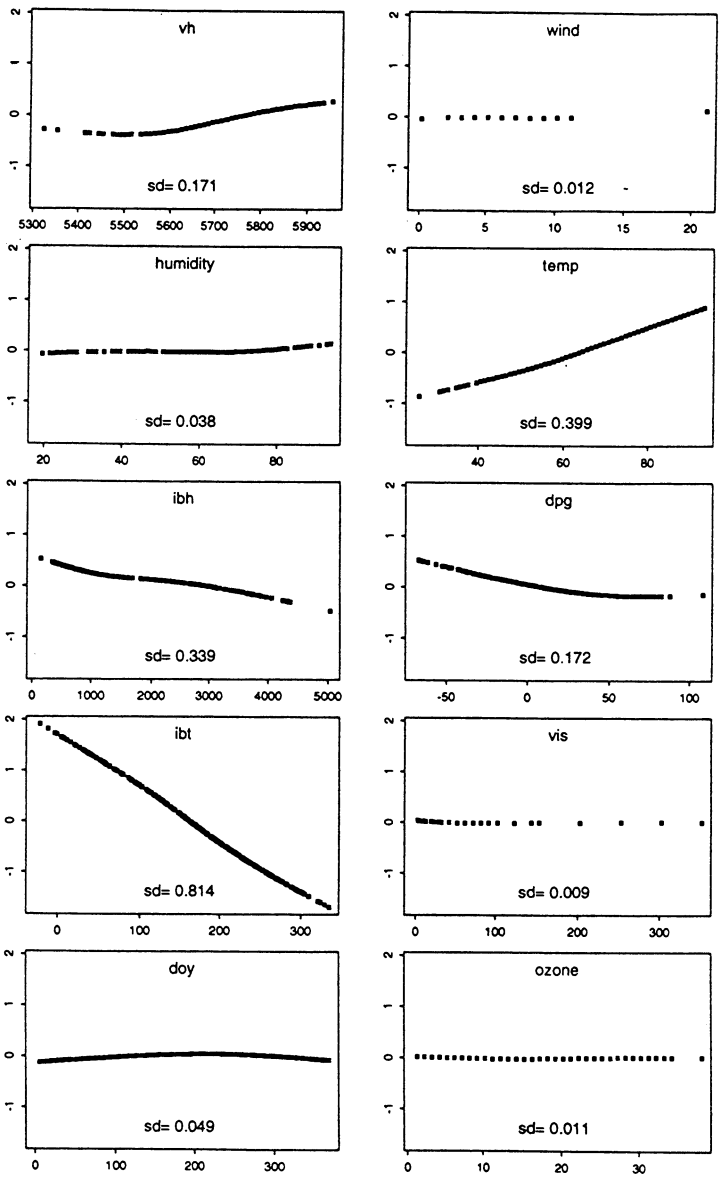

FIG. 1. *The smallest APC functions for the ozone data. The eigenvalue for the APC is* 0.030.

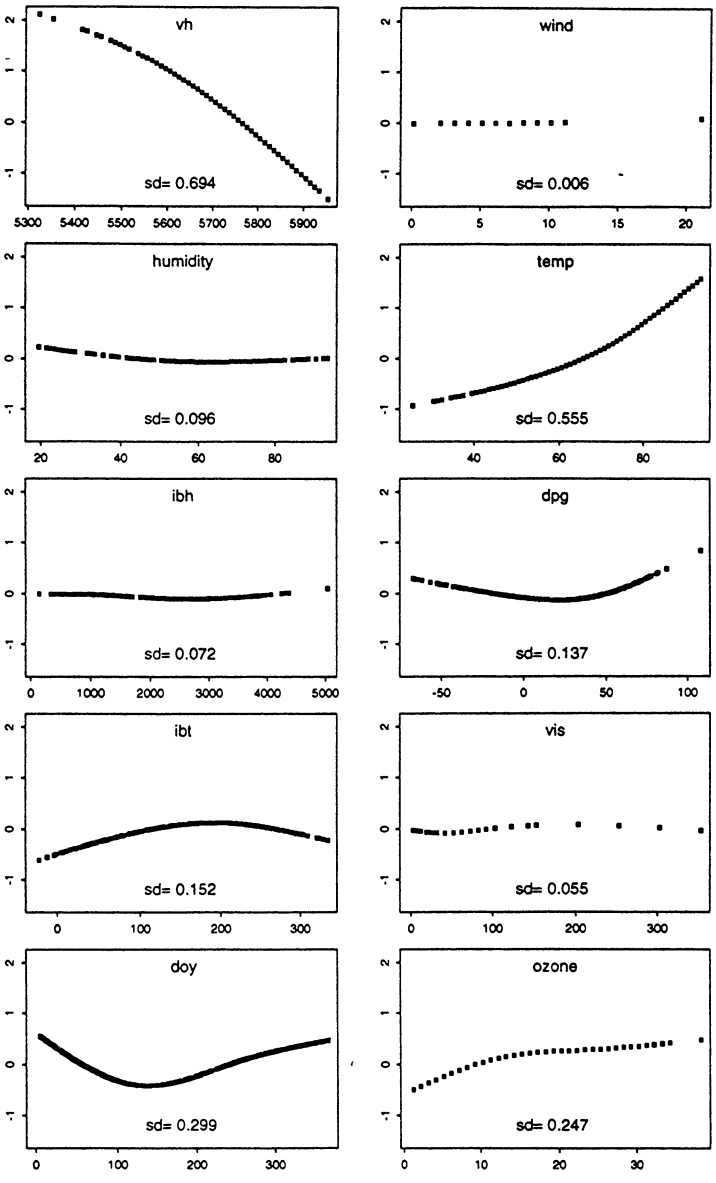

FIG. 2. *The second-smallest APC functions for the ozone data. The eigenvalue for the APC is* 0.084.

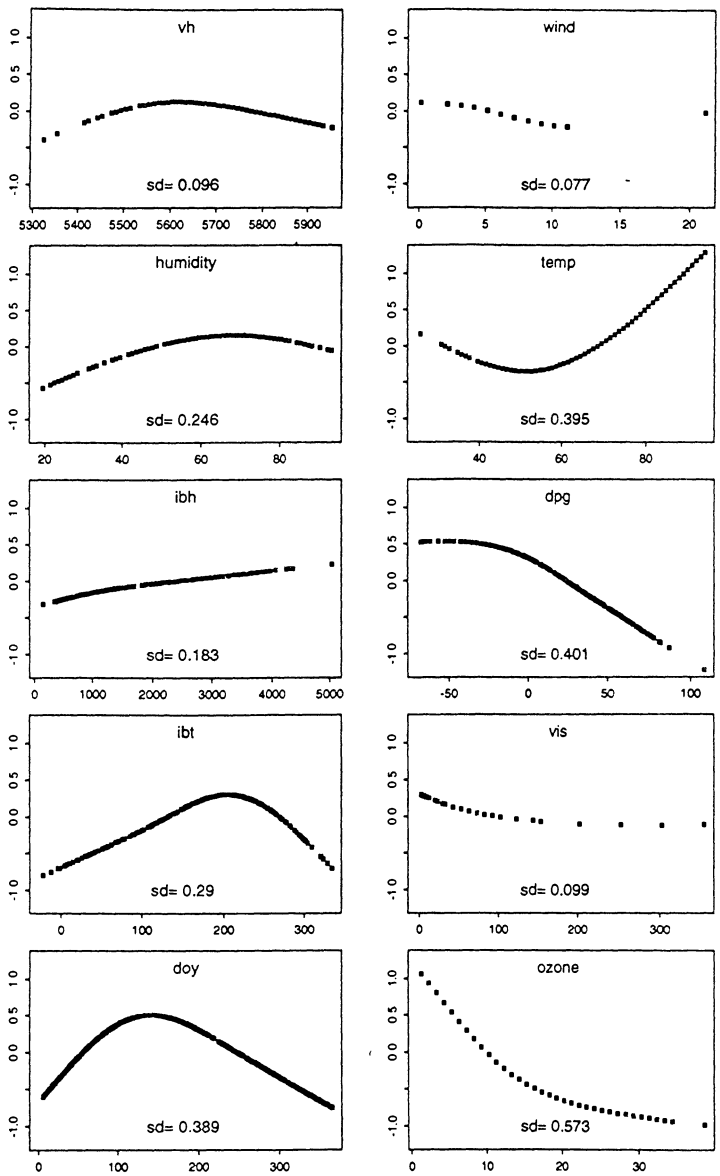

FIG. 3. *The third-smallest APC functions for the ozone data. The eigenvalue for the APC is* 0.088.

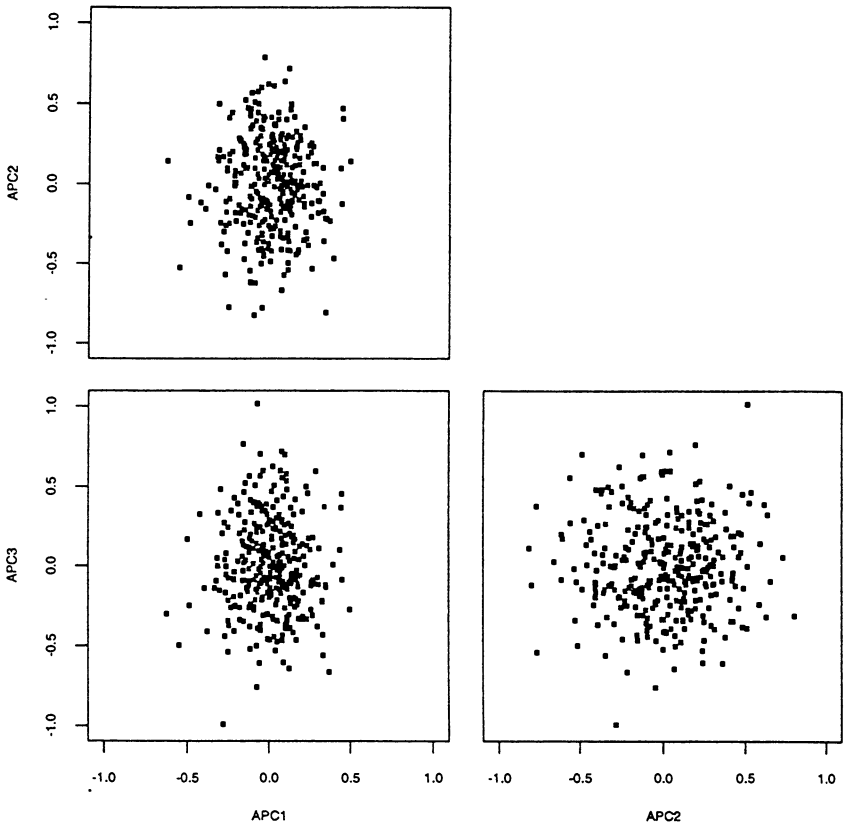

FIG. 4. *Pairwise scatterplots of the three smallest APCs for the ozone data. The eigenvalues for the smallest, second and third smallest are* 0.030, 0.084 *and* 0.088, *respectively.*

For diagnostics purposes, we show the pairwise scatterplots of the smallest three APCs in Figure 4. They do not exhibit any anomalies such as extreme outliers or skewed distributions. The plotting scales chosen are identical for all three APCs so their standard deviations (the square roots of their eigenvalues) are reflected in the elliptic shapes of the point scatters.

3.3. *The ozone dependence: comparing APC and ACE.* The findings of the previous subsection indicate that an additive regression model for ozone based on the complete data set may exhibit instability due to concurvity among the predictors. According to the smallest APC, inversion base temperature (possibly combined with inversion base height) can be traded in for Sandburg temperature to some extent, while the second-smallest APC hints at a similar relation between Vandenburg height and Sandburg temperature. Breiman and Friedman used a forward stepwise approach to variable inclusion in ACE, thus avoiding the concurvity problem ensuing from the strong dependencies between Sand-

burg temperature, inversion base temperature and Vandenburg height: their model includes only the first of these variables. As their forward stepwise approach did not include wind and humidity, we also omit these variables, although humidity is possibly important in the dependence involving ozone. We estimate the two smallest APCs of the variables ozone, Sandburg temperature, day of year, Vandenburg height, visibility and Dagget pressure, using the iterative method (Section 5.2) based on Supersmooth, to facilitate comparison with the ACE fits of Breiman and Friedman.

The smallest APC of the reduced data set (Figure 5) has an eigenvalue of 0.102 and detects the dependence between Sandburg temperature and day of year, with temperatures peaking sharply in summer. The second-smallest APC of the reduced data set (Figure 6) has an eigenvalue of 0.115 and depicts a strong dependence between ozone, day of year, Sandburg temperature and Dagget pressure. The transforms for these variables are for the most part very similar to the transforms found by Breiman and Friedman with ACE, adjusting for the fact that APC yields implicit equations while ACE solves for a dependent variable. (To achieve direct comparability, simply change the sign of the ozone variable.) In comparing APC and ACE transforms, note that we used identical plotting scales for all transforms, in contrast to Breiman and Friedman. Modulo these precautions, our transforms are in very good agreement with Breiman and Friedman's ACE fits, except for the additional tilt in the APC transform of Dagget pressure. Even the mild curvature of the ozone transform is the same if adjusted for the necessary sign change.

For completeness, we also obtained a third-smallest APC. Since its eigenvalue 0.29 is considerably larger than the smallest two eigenvalues, we feel reasonably certain that the smallest two APCs summarize the most important additive dependencies.

The signals given by our reanalysis of the reduced ozone data are ambiguous, as is often the case with real-world data: on the one hand, we obtained an APC which basically recreates the ACE fit and seems to lend credence to its transforms; on the other hand, the presence of a smaller APC, as well as the closeness of the two smallest APC eigenvalues, indicates that some of the ACE transforms should not be taken at face value without considering alternative transforms derived, for instance, as mixtures of the smallest two APCs.

3.4. *Horseshoes: an example with artifactual parabolic transforms.* The following example shows that on occasion APC solutions may not be meaningful. In Figure 7 we plot the three smallest APC transforms of 500 points generated from a Gaussian distribution with correlation matrix

$$\mathbf{R} = \begin{pmatrix} 1.0 & 0.6 & 0.4 & -0.7 \\ 0.6 & 1.0 & 0.5 & -0.3 \\ 0.4 & 0.5 & 1.0 & -0.8 \\ -0.7 & -0.3 & -0.8 & 1.0 \end{pmatrix}.$$

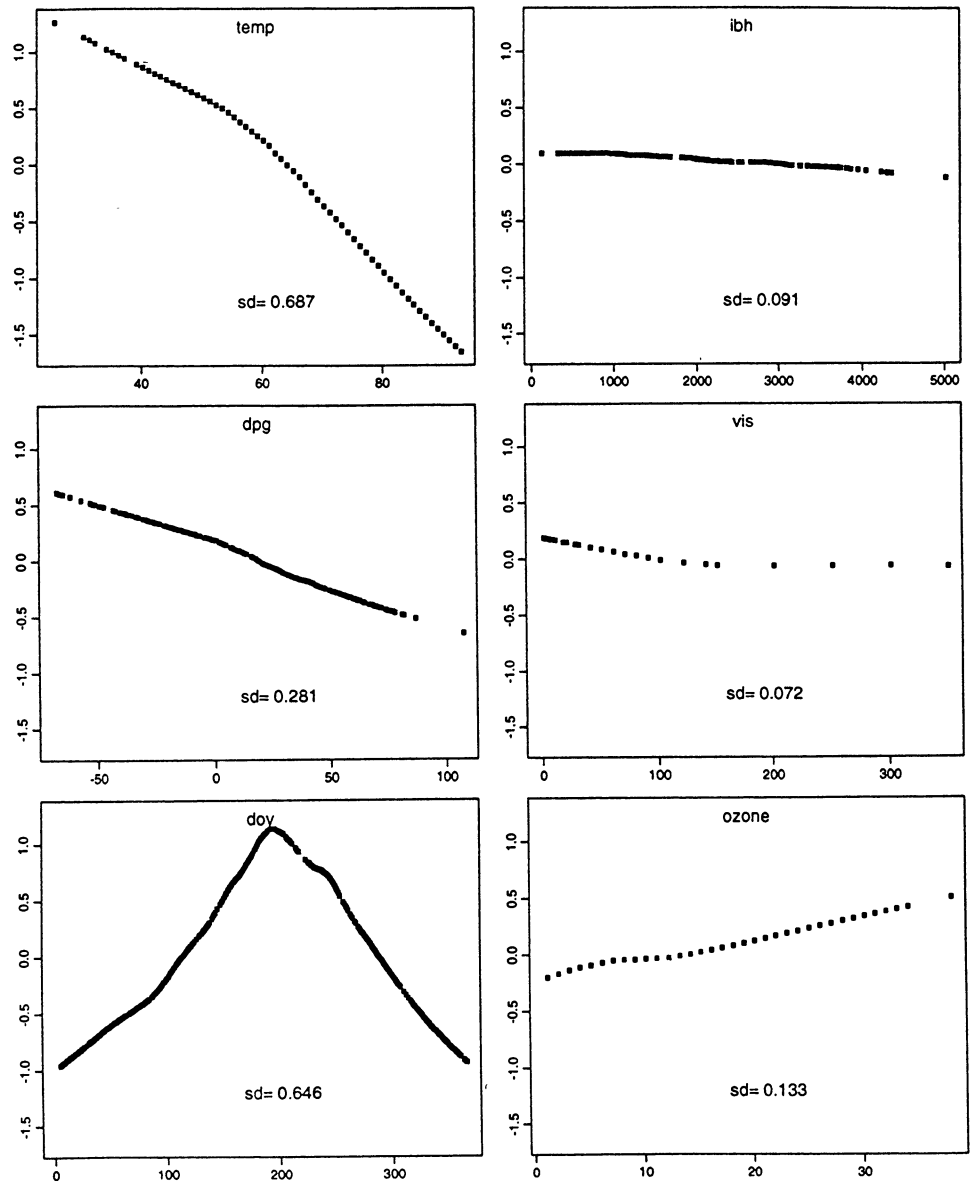

FIG. 5. *The smallest APC of the reduced ozone data:* ozone, temp, doy, vht, vis *and* dpg. *The eigenvalue for the APC is* 0.102.

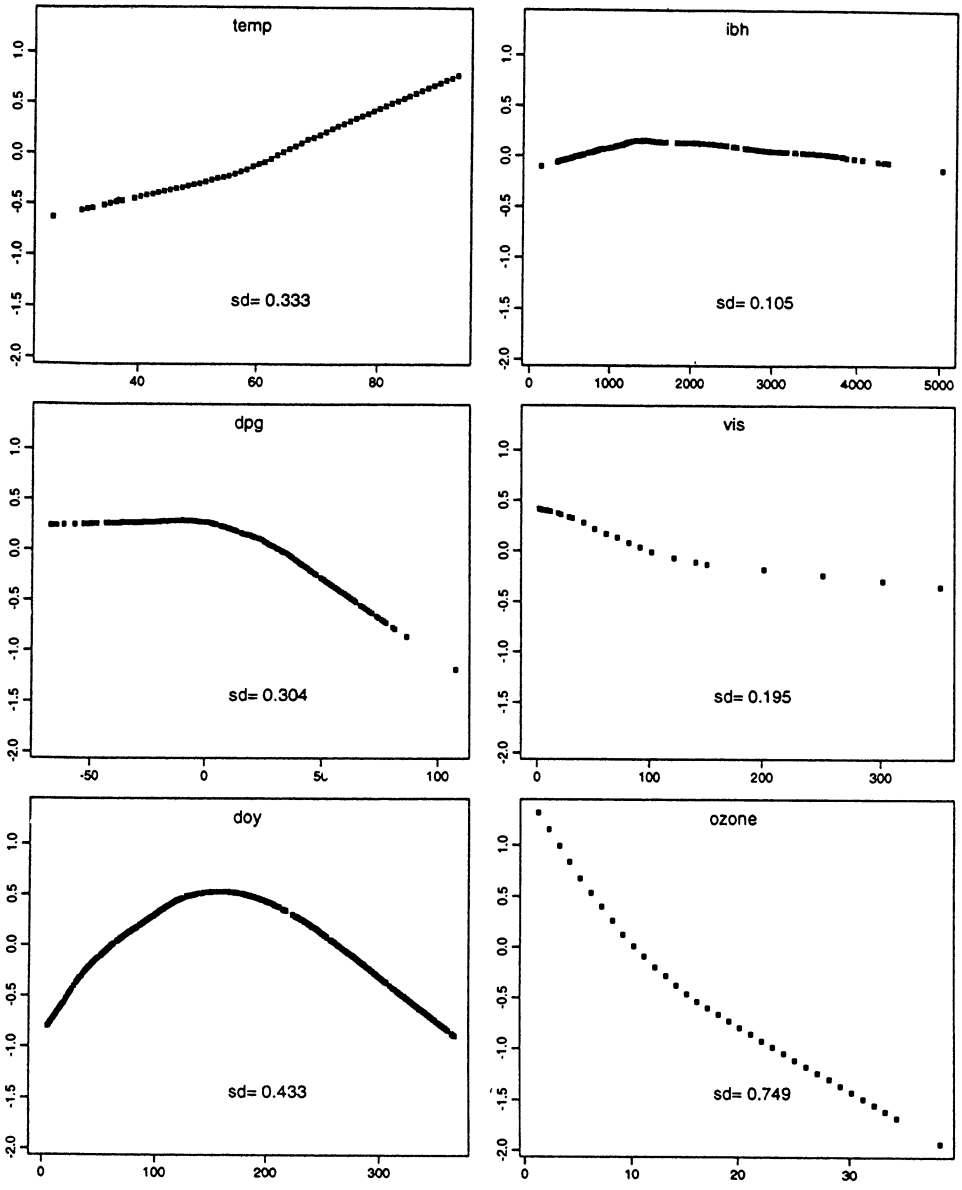

FIG. 6. *The second-smallest APC of the reduced ozone data:* ozone, temp, doy, vht, vis *and* dpg. *The eigenvalue for the APC is* 0.115.

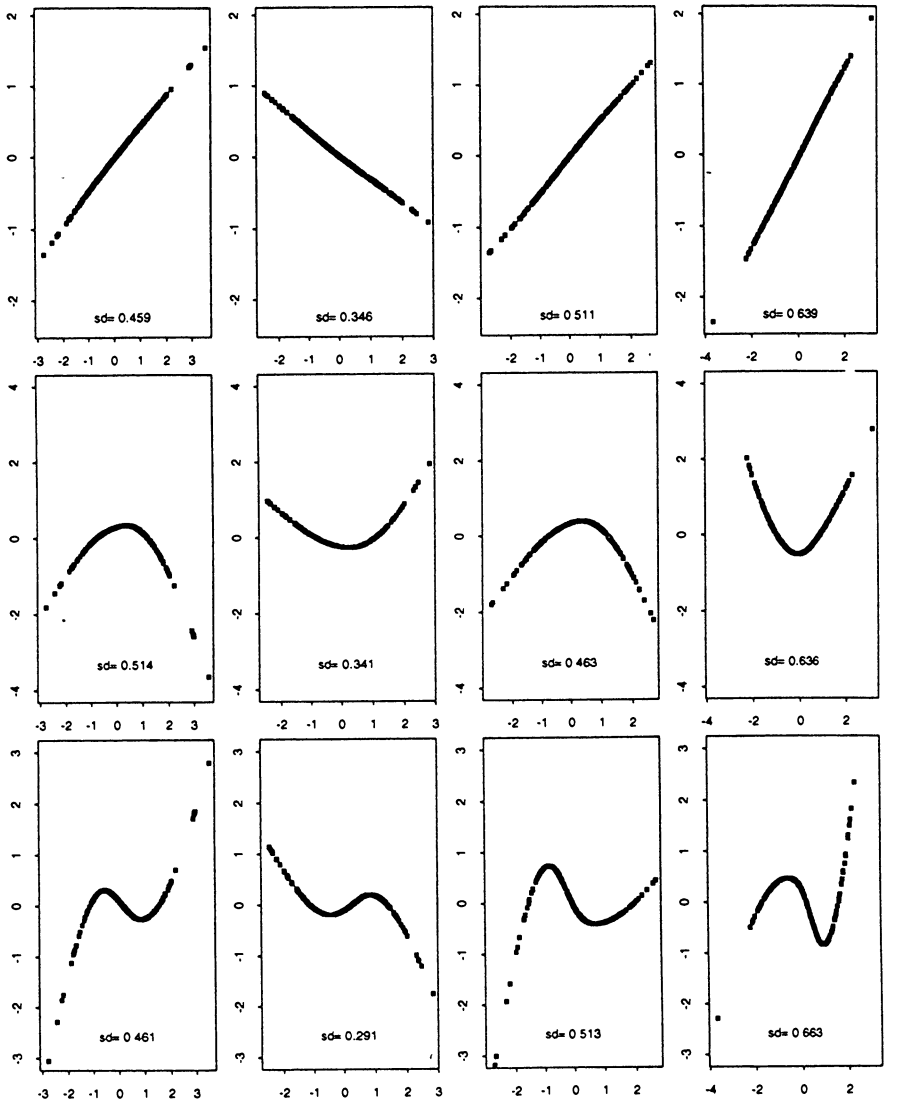

FIG. 7. *The three smallest APCs of* $\mathcal{N}(\mathbf{0}, \mathbf{R})$ *data: variables vary from left to right, smallest APC is the top row, third smallest the bottom row. The second and third APCs exhibit regular horseshoes and "generalized horseshoes," respectively. The estimated eigenvalues are* 0.018, 0.18 *and* 0.437.

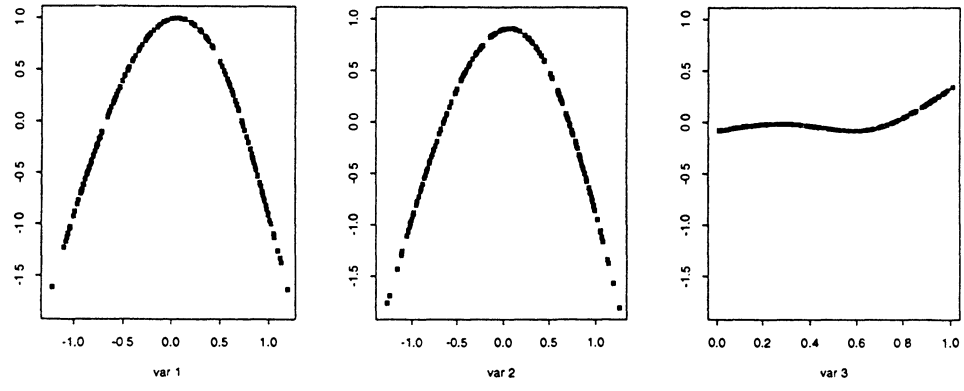

FIG. 8. *Plot of the smallest APC functions for the cylinder data. The eigenvalue for the APC is* 0.189.

The respective eigenvalues are 0.018, 0.18 and 0.437. The smallest is linear, which is comforting, but the second-smallest appears quadratic and the third vaguely cubic. It is intuitive that the second- and third-smallest APCs in this instance are artifacts caused by orthogonality constraints and the Gaussian distribution. Thus, before inferring the presence of peculiar structure, one should scrutinize the data graphically. In the present example, a jump by a factor of 10 from the smallest to the second-smallest eigenvalue should have been a warning. Since artifactual transforms of parabolic shape (as in the middle row of Figure 7) are the most frequent in practice, psychometricians use the term "horseshoe effect." It can be illustrated by the exact APC theory of the Gaussian and many other analytic distributions (Section 4.7).

3.5. *An example with meaningful parabolic transforms.* While it is true that many parabolic transforms in real data are artifactual, there is a rare possibility that these transforms are meaningful. Here is an example. We generated 200 points near the surface of a cylinder in 3-space. The first two variables, X_1, X_2, describe a circular pointscatter, uniform in angle and Gaussian ($\mu = 1, \sigma = .125$) in radial interval. The third variable is pure uniform noise, independent of X_1 and X_2.

From the way the data are generated, we know that the approximate implicit equation $X_1^2 + X_2^2 \approx \textit{constant}$ describes the data well. The smallest APC recovers this structure. Figure 8 shows the estimated transforms based on direct estimates from spline regression with two internal knots for each variable. The variance of the APC is 0.189, with variable weights 0.70, 0.71 and 0.10, respectively.

4. Eigencharacterization of additive principal components. In this section we give some second-order theory for APCs, some of which has been mentioned in Section 2. First, we present some not very deep material on quadratic forms and eigenproblems, which nevertheless is necessary for estimation and

computation (Section 5). Next, we show some relevant propositions regarding null situations, the relation between APCs and ACE, analytical APC calculations, as well as some remarks on the role of centering of APCs.

4.1. *Quadratic forms on products of Hilbert spaces.* We switch to inner product notation: $\langle \phi, \phi' \rangle = \mathbf{E}(\phi\phi')$ and $\|\phi\|^2 = \mathbf{E}(\phi^2)$, where ϕ and ϕ' are centered L_2 random variables. For $\mathbf{\Phi}, \mathbf{\Psi} \in \mathbf{H} = H_1 \times \cdots \times H_p$, the natural inner product and squared norm are

$$\langle \mathbf{\Phi}, \mathbf{\Psi} \rangle_H \stackrel{\text{def}}{=} \sum_i \langle \phi_i, \psi_i \rangle \quad \text{and} \quad \|\mathbf{\Phi}\|_H^2 \stackrel{\text{def}}{=} \sum_i \|\phi_i\|^2.$$

The product space $\mathbf{H}$ thus becomes a Hilbert space for which the natural embeddings of $H_1, \ldots, H_p$ are closed linear subspaces in mutually orthogonal positions. The norm topology of $\mathbf{H}$ trivially coincides with the product topology inherited from the factors H_i. In terms of this norm, the constraint of the APC optimization problem can be cast as a restriction to the unit sphere in $\mathbf{H}$:

$$\sum_i \mathbf{var}\, \phi_i = \|\mathbf{\Phi}\|_H^2 = 1.$$

The APC criterion is the following quadratic form on $\mathbf{H}$:

$$Q(\mathbf{\Phi}) = \mathbf{var} \sum_i \phi_i = \left\| \sum_i \phi_i \right\|^2.$$

This form is bounded with regard to the above norm (see Lemma 4.1 below). The associated bilinear form,

$$B(\mathbf{\Phi}, \mathbf{\Psi}) = \frac{1}{4}\big(Q(\mathbf{\Phi} + \mathbf{\Psi}) - Q(\mathbf{\Phi} - \mathbf{\Psi})\big) = \left\langle \sum_i \phi_i, \sum_j \psi_j \right\rangle,$$

is bounded and symmetric. It follows from elementary theorems in L_2 theory that there exists a bounded, symmetric, linear operator $\mathbf{P}$ on $\mathbf{H}$ such that

$$B(\mathbf{\Phi}, \mathbf{\Psi}) = \langle \mathbf{\Phi}, \mathbf{P}\mathbf{\Psi} \rangle_H.$$

However, we do not need to appeal to existence theorems since $\mathbf{P}$ can be constructed explicitly (see again Lemma 4.1 below).

In the new notation, the optimization problem for the smallest additive principal component can be written as

$$\langle \mathbf{\Phi}, \mathbf{P}\mathbf{\Phi} \rangle_H = \min_{\mathbf{\Phi} \in \mathbf{H}} \quad \text{subject to} \quad \|\mathbf{\Phi}\|_H^2 = 1.$$

It remains to identify $\mathbf{P}$.

LEMMA 4.1. *Define the operator* $\mathbf{P}$: $\mathbf{H} \to \mathbf{H}$ *by the component mappings*

$$[\mathbf{P}\Phi]_i = P_i\left(\sum_j \phi_j\right),$$

where P_i *is the orthogonal projection onto* H_i. *Then* $\mathbf{P}$ *satisfies*

$$Q(\Phi) = \langle \Phi, \mathbf{P}\Phi \rangle_H.$$

It is symmetric, nonnegative definite and bounded above by p. We have $\|\mathbf{P}\Phi\|_H = p$ *iff* $\phi_i = \phi_j$ *a.s.* $\forall i, j$.

PROOF. The operator $\mathbf{P}$ represents B:

$$\langle \Phi, \mathbf{P}\Phi \rangle_H = \sum_i \left\langle \phi_i, P_i \sum_j \phi_j \right\rangle = \sum_i \left\langle \phi_i, \sum_j \phi_j \right\rangle = \left\langle \sum_i \phi_i, \sum_j \phi_j \right\rangle = \left\| \sum_i \phi_i \right\|^2.$$

$\mathbf{P}$ is bounded by p:

$$\|\mathbf{P}\Phi\|_H^2 \overset{\text{def}}{=} \sum_i \left\| P_i \sum_j \phi_j \right\|^2 \le \sum_i \left\| \sum_j \phi_j \right\|^2 = p \left\| \sum_j \phi_j \right\|^2 \le p \left(\sum_j \|\phi_j\| \right)^2.$$

The maximum of $\Sigma_j \|\phi_j\|$ under the constraint $\Sigma_j \|\phi_j\|^2 = 1$ is attained for $\|\phi_j\| = p^{-1/2}$. Hence,

$$\|\mathbf{P}\Phi\|_H^2 \le p\left(\sum \|\phi_j\| \right)^2 \le p^2.$$

The inequality is sharp, with equality occurring iff $\phi_i = \phi_j$ a.s. $\forall i, j$. Symmetry $\langle \Phi, \mathbf{P}\Psi \rangle_H = \langle \mathbf{P}\Phi, \Psi \rangle_H$ follows trivially from the symmetry of $\langle \Phi, \mathbf{P}\Psi \rangle_H = \langle \Sigma_i \phi, \Sigma_j \psi_j \rangle$, and so does nonnegativity: $\langle \Phi, \Phi \rangle_H = \|\Sigma_i \phi_i\|^2 \ge 0$. □

4.2. *The sequence of smallest additive principal components.* The eigencharacterization of the smallest APC now follows from standard results about symmetric operators [e.g., Jorgens (1970), Theorem 6.7, page 125).

PROPOSITION 4.2. *The smallest eigenfunction of the operator* $\mathbf{P}$, *if it exists, is a vector of APC functions for the smallest additive principal component of* $H_1, \ldots, H_p$.

We use the phrase "smallest eigenfunction" as an abbreviation of the more correct description "eigenvector of functions for the smallest eigenvalue." Unlike finite-dimensional matrices, existence of the smallest eigenvalue of $\mathbf{P}$ is not granted a priori. For simplicity, we use the phrase "if it exists" whenever necessary, leaving the discussion of existence conditions to Section 4.3.

The kth additive principal component corresponds to an eigenfunction of $\mathbf{P}$ belonging to the kth smallest eigenvalue (where eigenvalues are repeated according to their multiplicity).

PROPOSITION 4.3. *A vector of APC functions, $\Phi^{(k)}$, defining the kth smallest additive principal component, if it exists, is a kth smallest eigenfunction of the operator* $\mathbf{P}$:

$$\mathbf{P}\Phi^{(k)} = \lambda^{(k)}\Phi^{(k)}.$$

The proof is standard. It relies on invariance of the orthogonal subspaces $\mathbf{H}^{(k)} = \{\Phi \mid \langle \Phi, \Phi^{(i)} \rangle = 0, i = 1, \ldots, k-1\}$ under $\mathbf{P}$. Recursively applying Proposition 4.2 under restriction to $\mathbf{H}^{(k)}$ gives the result. An immediate corollary is as follows.

COROLLARY 4.4. *The variance of the kth smallest APC is* $\lambda^{(k)}$.

PROOF. $\mathbf{var}\, \Sigma\phi_i^{(k)} = \langle \Phi^{(k)}, \mathbf{P}\Phi^{(k)} \rangle_H = \langle \Phi^{(k)}, \lambda^{(k)}\Phi^{(k)} \rangle_H = \lambda^{(k)}$. □

It follows from the eigenproperties that APCs simultaneously diagonalize the quadratic forms $\langle \Phi, \mathbf{P}\Phi \rangle_H$ and $\|\Phi\|_H^2$, implying $\langle \Phi^{(k)}, \mathbf{P}\Phi^{(k')} \rangle_H = 0$ and $\langle \Phi^{(k)}, \Phi^{(k')} \rangle_H = 0$ for $k \neq k'$. The former equality translates to the following corollary.

COROLLARY 4.5. *Additive principal components belonging to two orthogonal eigenfunctions are uncorrelated:* $\langle \Sigma_i \phi_i^{(k)}, \Sigma_i \phi_i^{(k')} \rangle = 0, k \neq k'$.

4.3. *Existence of population APCs.* We now address the issue of existence of eigenvalues when $\mathbf{H}$ is infinite-dimensional. Although the spectrum of $\mathbf{P}$ is bounded, the existence of eigenvalues is complicated by the possibility of $\mathbf{P}$ having a nontrivial continuous spectrum or spectral values that are not eigenvalues. We can rule out these undesirable possibilities by adopting the usual compactness assumptions.

ASSUMPTION. *The restricted projections $P_{i|k}: H_k \to H_i$, $P_{i|k} = P_i|_{H_k}$, are compact operators.*

The definition of a compact operator is that the image of the unit ball, or any norm-bounded set, is a relatively compact set in the norm topology. In the population case, if H_i is the space of *all* centered L_2 functions of X_i, and hence P_i is the conditional expectation given X_i, a sufficient condition for compactness is the Hilbert-Schmidt property, also used by Breiman and Friedman (1985): suppose X_1, X_2 have joint density f_{X_1,X_2}, and marginal densities f_{X_1}, f_{X_2}, then $P_{2|1}$ is Hilbert-Schmidt if

$$\int\int \frac{f_{X_1,X_2}^2(x_1,x_2)}{f_{X_1}(x_1)f_{X_2}(x_2)}\,dx_1\,dx_2 < \infty.$$

Under the assumption of compact $P_{i\,|\,k}$, $\mathbf{P}$ itself is not compact, but we have the following result.

LEMMA 4.6. *The operator* $\mathbf{P} - \mathbf{I}$: $\mathbf{H} \to \mathbf{H}$ *is compact.*

PROOF. We can decompose $\mathbf{P} - \mathbf{I}$ as follows:

$$\mathbf{P} - \mathbf{I} = \sum_i \mathbf{Q}_i,$$

where $\mathbf{Q}_i$: $\mathbf{H} \to \mathbf{H}$ is defined by

$$\mathbf{Q}_i(\Phi) = \left(P_{1\,|\,i}\phi_i, \ldots, P_{i-1\,|\,i}\phi_i, 0, P_{i+1\,|\,i}\phi_i, \ldots, P_{p\,|\,i}\phi_i\right).$$

It is sufficient to show that every summand $\mathbf{Q}_i$ is compact [see, e.g., Jorgens (1970), Theorem 5.10, page 98]. Denoting by $\mathbf{B}$ the unit ball in $\mathbf{H}$, we have to establish that $\mathbf{Q}_i(\mathbf{B})$ is a relatively compact set in $\mathbf{H}$. Let B_i be the unit ball in H_i. Since $\mathbf{B} \subset B_1 \times \cdots \times B_p$, compactness of $\mathbf{Q}_i$ follows if $\mathbf{Q}_i(B_1 \times \cdots \times B_p)$ is shown to be relatively compact. By assumption, $P_{i\,|\,j}(B_j)$ is relatively compact in $H_i \,\forall\, i \neq j$, hence,

$$\mathbf{Q}_i(B_1 \times \cdots \times B_p) = P_{1\,|\,i}(B_i) \times \cdots \times P_{i-1\,|\,i}(B_i) \times \{0\} \times P_{i+1\,|\,i}(B_i) \times \cdots \times P_{p\,|\,i}(B_i)$$

is relatively compact in $\mathbf{H}$, since the norm topology and product topology coincide. □

The assumption of compactness implies that the spectrum of $\mathbf{P} - \mathbf{I}$ has characteristics similar to that of a finite-dimensional symmetric matrix, except for the limiting behavior which is vacuous in finite-dimensions:

1. There exists a sequence $\{l_k\}_1^\infty$ of eigenvalues for which $|l_1| \geq |l_2| \geq \cdots \geq |l_k| \geq \cdots$.
2. $\lim_{k \to \infty} l_k = 0$.
3. The eigenspaces for distinct eigenvalues are orthogonal and the sum of all eigenspaces is dense in the whole space.
4. The nonzero eigenvalues have finite multiplicity.

The spectrum of $\mathbf{P} - \mathbf{I}$ is thus a countable, bounded set with $\{0\}$ as the only possible accumulation point. The eigenvalues $\{l_k\}$ of $\mathbf{P} - \mathbf{I}$ are related to the eigenvalues $\{\lambda_k\}$ of $\mathbf{P}$ through $\lambda_k = l_k + 1$, hence the eigenvalues and eigenspaces of $\mathbf{P}$ inherit all the above properties, with l_k replaced by $\lambda_k - 1$. In particular, we have the following result.

COROLLARY 4.7. *The only accumulation point of the eigenvalues of* $\mathbf{P}$ *is* +1.

4.4. *A null analysis for APCs—comparison of small and large APCs.* A natural question to ask is whether it is possible for either the upper or lower sequence of eigenvalues to be empty. The following proposition establishes that

this occurs only when the spaces H_i are pairwise orthogonal, in which case all eigenvalues are +1. In the population case, when the spaces H_i are all centered L_2 variables of X_i, this is equivalent to pairwise independence. In this situation, there is no structure that could be detected by APCs.

PROPOSITION 4.8. *The following statements about* $\mathbf{P}$ *are equivalent*:

1 *All eigenvalues are greater than or equal to* +1; *that is*, $\mathbf{P} - \mathbf{I}$ *is nonnegative definite.*
2 *All eigenvalues are less than or equal to* +1; *that is*, $-(\mathbf{P} - \mathbf{I})$ *is nonnegative definite.*
3 *The spectrum of* $\mathbf{P}$ *is the singleton* $\{+1\}$; *that is*, $\mathbf{P} = \mathbf{I}$.
4 *The spaces* $H_1, H_2, \ldots, H_p$ *are mutually orthogonal.*
5 $\|\Sigma\phi_i\|^2 = \Sigma\|\phi_i\|^2 \,\forall\, \phi_1 \in H_1, \ldots, \phi_p \in H_p$.

This explains again why +1 is the natural dividing line between small and large APCs.

PROOF. $(3 \Leftrightarrow 4)$ follows from $P_{i\,|\,j} = 0 \Leftrightarrow H_i \perp H_j$ for $i \neq j$. $(4 \Leftrightarrow 5)$ is standard (use $\|\phi_i + \phi_j\|^2 - \|\phi_i\|^2 - \|\phi_j\|^2 = 2\langle\phi_i, \phi_j\rangle$). $(3 \Rightarrow 1, 2)$ is trivial. The proofs of $(1 \Rightarrow 4)$ and $(2 \Rightarrow 4)$ are similar. We therefore show only the first: if $\mathbf{P} - \mathbf{I}$ is nonnegative, then for $\mathbf{\Phi}^{ij} = (0, \phi_i, 0, \ldots, 0, \phi_j, 0, \ldots, 0) \in \mathbf{H}$ we get

$$0 \leq \langle \mathbf{\Phi}, (\mathbf{P} - \mathbf{I})\mathbf{\Phi} \rangle_H = \|\phi_i + \phi_j\|^2 - \left(\|\phi_i\|^2 + \|\phi_j\|^2\right) = 2\langle\phi_i, \phi_j\rangle.$$

Replacing ϕ_j by $-\phi_j$ in the above, we arrive at the conclusion that $\langle\phi_i, \phi_j\rangle = 0 \,\forall\, i \neq j$. It follows that H_i and H_j are orthogonal $\forall\, i \neq j$. □

COROLLARY 4.9. *In any nonnull situation, there exist small and large APCs. Small APCs are variance deficient*:

$$\left\| \sum \phi_i \right\|^2 < \sum \|\phi_i\|^2,$$

while large APCs are variance abundant:

$$\left\| \sum \phi_i \right\|^2 > \sum \|\phi_i\|^2.$$

4.5. *Eigenexpansions associated with APCs.* The eigenanalysis of the operator $\mathbf{P}$ and its quadratic form, $Q = \|\Sigma\phi_j\|^2$, gives rise to some eigenexpansions which illuminate the sense in which APC analysis measures deviation from pairwise orthogonality. As above, we assume for simplicity that $\mathbf{P} - \mathbf{I}$ is compact and hence there exists a complete sequence of eigenvectors $\mathbf{\Phi}^{(k)} = (\phi_i^{(k)})$, $k = 1, 2, \ldots$, and a corresponding sequence of eigenvalues, l_k. With $(\mathbf{P} - \mathbf{I})\mathbf{\Phi}^{(k)} = l_k \cdot \mathbf{\Phi}^{(k)}$, we have the following expansion.

PROPOSITION 4.10. $(\mathbf{P} - \mathbf{I})\mathbf{\Phi} = \Sigma_k l_k \cdot \langle \mathbf{\Phi}, \mathbf{\Phi}^{(k)} \rangle_H \mathbf{\Phi}^{(k)}$.

Convergence of the expansion is in norm for every $\mathbf{\Phi} = (\phi_i) \in \mathbf{H}$. Using the definitions and Lemma 4.1:

$$\langle \mathbf{\Phi}, (\mathbf{P} - \mathbf{I})\mathbf{\Phi} \rangle_H = \langle \mathbf{\Phi}, \mathbf{P}\mathbf{\Phi} \rangle_H - \|\mathbf{\Phi}\|_H^2 = \left\| \sum \phi_i \right\|^2 - \sum \|\phi_i\|^2 = \sum_k l_k \cdot \langle \mathbf{\Phi}, \mathbf{\Phi}^{(k)} \rangle_H^2,$$

we get the following result.

COROLLARY 4.11. $\|\Sigma_i \phi_i\|^2 = \Sigma_i \|\phi_i\|^2 + \Sigma_k l_k \cdot \Sigma_i \langle \phi_i, \phi_i^{(k)} \rangle^2$.

The term $l_k \cdot \Sigma_i \langle \phi_i, \phi_i^{(k)} \rangle^2$ can be interpreted as a correction for deviation from pairwise orthogonality. Under orthogonality, all eigenvalues l_k vanish and the expansion reduces to the Pythagorean identity. For eigenvectors, $\mathbf{\Phi} = \mathbf{\Phi}^{(k)}$, the expansion reduces to a single correction term:

$$\left\| \sum_i \phi_i^{(k)} \right\|^2 = 1 + l_k,$$

assuming that eigenvectors are standardized, $\Sigma_i \|\phi_i^{(k)}\|^2 = 1$.

The expansion of Proposition 4.10 for the operator $\mathbf{P} - \mathbf{I}$ is also interesting. With $\mathbf{\Phi} = (0, \ldots, 0, \phi_j, 0, \ldots, 0)$, where $\phi_j \in H_j$ is the jth component, the expansion specializes to:

COROLLARY 4.12.

$$\sum_k l_k \cdot \langle \phi_j, \phi_j^{(k)} \rangle \phi_j^{(k)} = 0,$$

$$\sum_k l_k \cdot \langle \phi_j, \phi_i^{(k)} \rangle \phi_i^{(k)} = P_i \phi_j \quad \text{for } i \neq j.$$

Thus, the eigensystem can be used to reconstruct the projections between any pair of subspaces H_i and H_j. As such, the eigenvectors are optimally tailored to the collection of spaces $H_1, \ldots, H_p$, but not to any specific pair of spaces. For an expansion which is optimal for a particular pair, a two-variable APC should be used.

4.6. *APC analysis and ACE.* In general, APC and ACE analyses are not identical, although it may be helpful to perform and compare both types of analyses, as we showed in Section 3.2. However, there is a direct link in one simple situation: single-predictor ACE is equivalent to two-variable APC. This special case is known as (continuous) correspondence analysis in the psychometric, Dutch and French literature.

Single-predictor ACE optimizes the correlation or (modulo scale factors) percentage of explained variance:

$$\frac{\mathbf{cov}(\phi_1, \phi_2)}{\mathbf{sd}(\phi_1) \cdot \mathbf{sd}(\phi_2)} \quad \text{or} \quad \frac{\mathbf{E}\left[(\phi_2 - \phi_1)^2\right]}{\mathbf{var}\,\phi_1},$$

whereas an unconstrained criterion for two-variable APC is the Rayleigh quotient

$$\frac{\mathbf{var}(\phi_1 + \phi_2)}{\mathbf{var}\,\phi_1 + \mathbf{var}\,\phi_2}.$$

Comparing the stationary or eigenequations of the two problems,

$$\begin{aligned} P_1\phi_2 &= \lambda_{\text{ACE}} \cdot \phi_1, & P_2\phi_1 &= \lambda_{\text{ACE}} \cdot \phi_2 \quad \text{for ACE,} \\ \phi_1 + P_1\phi_2 &= \lambda_{\text{APC}} \cdot \phi_1, & P_2\phi_1 + \phi_2 &= \lambda_{\text{APC}} \cdot \phi_2 \quad \text{for APC,} \end{aligned}$$

one easily recognizes that the solutions are the same (up to a possible scaling factor), and the eigenvalues are related via

$$\lambda_{\text{APC}} = \lambda_{\text{ACE}} + 1,$$

which is consistent with $0 \leq \lambda_{\text{APC}} \leq 2$ and $-1 \leq \lambda_{\text{ACE}} \leq +1$. Small and large APCs are in a trivial correspondence since any large APC given by (ϕ_1, ϕ_2) for the eigenvalue $\lambda_{\text{APC}} \geq +1$ generates a small APC given by $(\phi_1, -\phi_2)$ for the eigenvalue $2 - \lambda_{\text{APC}}$, as the transforms (ϕ_1, ϕ_2) and $(\phi_1, -\phi_2)$ are ACE eigentransforms for the eigenvalues λ_{ACE} and $-\lambda_{\text{ACE}}$, respectively.

As a consequence, several examples given in Buja (1990) for single-predictor ACE carry over to APCs. For instance, the illustrations of the step function behavior in the presence of bivariate clustering and the bivariate horseshoe examples carry over to APC analysis. Multivariate extensions of the latter can be obtained with the theory of the following subsection.

4.7. *Some horseshoe theory.* This subsection is a theoretical follow-up to Section 3.4, where we illustrated the horseshoe effect, that is, the appearance of artifactual parabolic and other (possibly nonmonotonic) transforms. The reason why "unnatural transforms" must exist lies simply in the nonparametric nature of the APC eigenproblem, where infinitely many eigensolutions are bound to appear if simple variables are represented by infinite-dimensional spaces of transforms. Orthogonality properties of eigensystems lead to qualitative behaviors similar or identical to orthogonal polynomial systems. Thus, coupled with "natural" monotonic or linear transforms, we should expect "artifactual" nonmonotonic transforms that may appear parabolic, cubic, and so forth.

The following theory works for many classes of continuous distributions in unrestricted function spaces (H_i = all centered L_2 functions of X_i, $P_i = E^{X_i}$), the simplest being the multivariate Gaussian. Useful references are de Leeuw (1982) and Gifi (1990), Section 11.3. Here is the basic structure that often leads to invariant subspaces for the APC operator $\mathbf{P}$.

LEMMA 4.13. *If the subspaces $V_i \subset H_i$ project into each other, $P_j V_i \subset V_j$ for all $i, j = 1, \ldots, p$, then the subspace $V_1 \times \cdots \times V_p \subset \mathbf{H}$ is invariant under $\mathbf{P}$.*

PROOF. Trivial; if $\phi_i \in V_i$, then $[\mathbf{P}\Phi]_j = P_j \Sigma_i \phi_i = \Sigma_i P_j \phi_i \in V_j$. □

Invariant product subspaces have some strong orthogonality properties: if $\mathbf{V} = V_1 \times \cdots \times V_p$ and $\mathbf{V}' = V_1' \times \cdots \times V_p'$ are two invariant subspaces with zero intersection, then $V_i \perp V_j'$. For $i = j$ this follows from $\langle , \rangle_H$-orthogonality of $\mathbf{V}$ and $\mathbf{V}'$, and for $i \neq j$, from $P_j V_i \subset V_j \perp V_j'$. See Dauxois and Pousse (1976) and de Leeuw (1982).

LEMMA 4.14. *If the normalized functions $\phi_i \in H_i$ ($\|\phi_i\|^2 = 1$) project onto each other, $P_j \phi_i = \rho_{ij} \phi_j$ for all $i, j = 1, \ldots, p$, then the subspace $\{(a_1\phi_1, a_2\phi_2, \ldots, a_p\phi_p) \mid a_1, \ldots, a_p \in \mathbb{R}\}$ is invariant under $\mathbf{P}$. Each eigenvector $(a_1, \ldots, a_p)^T$ of the matrix $\mathbf{R} = (\rho_{ij})_{ij}$ results in an APC vector $\Phi = (a_i \phi_i)_i$. In other words, a linear principal component analysis of the variables $\phi_1, \ldots, \phi_p$ yields p APCs.*

PROOF. The first part of the lemma is just the previous lemma specialized to $\dim V_j = 1$. The rest is a simple consequence of the fact that the coefficients ρ_{ij} are the correlations between the variables ϕ_i and ϕ_j. □

LEMMA 4.15. *If the direct sum $\Sigma_\nu V_i^{(\nu)}$ of subspaces $V_i^{(\nu)} \subset H_i$ is dense in H_i, $i = 1, \ldots, p$, the direct sum $\Sigma_\nu \mathbf{V}^{(\nu)}$ of products $\mathbf{V}^{(\nu)} = V_1^{(\nu)} \times \cdots \times V_p^{(\nu)}$ is dense in $\mathbf{H}$.*

The proof is obvious. This is a convenient condition to establish when a sequence of invariant eigenspaces is complete.

COROLLARY 4.16. *If the normalized functions $\phi_i^{(\nu)} \in H_i$ ($\|\phi_i^{(\nu)}\|^2 = 1$) form a basis of H_i, $i = 1, \ldots, p$, and if for each ν these functions project onto each other, $P_j \phi_i^{(\nu)} = \rho_{ij}^{(\nu)} \phi_j^{(\nu)}$, then* all *APCs can be found through linear principal component analyses of the variables $\phi_1^{(\nu)}, \ldots, \phi_p^{(\nu)}$. The systems $(\phi_i^{(\nu)})_\nu$ are necessarily orthonormal in H_i.*

This corollary applies to a surprisingly large number of multivariate distributions with $\phi_i^{(\nu)}$ being the normalized orthogonal polynomial of degree ν of X_i. The underlying structure that gives rise to this property is called "polynomial biorthogonality" of the bivariate marginals, a term introduced by Lancaster (1969). See Buja (1990) for a presentation close to our context and many references.

Among the distributions covered by the corollary is the multivariate Gaussian. If $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{R})$, where $\mathbf{R} = (\rho_{ij})$ is a correlation matrix, then normalized Hermite polynomials $\phi_i^{(1)} = h^{(1)}(X_i)$, $\phi_i^{(2)} = h^{(2)}(X_i), \ldots$ ($\|h^{(\nu)}(X_i)\|^2 = 1$) form

bases that satisfy the assumptions of the corollary as follows ($P_j = E^{X_j}$):

$$P_j h^{(\nu)}(X_i) = \rho_{ij}^{\nu} \cdot h^{(\nu)}(X_j). \tag{2}$$

Hence, all Gaussian APCs can be gotten through eigenanalyses of the elementwise (Hadamard or Schur) powers $(\rho_{ij}^{(\nu)})_{ij} = (\rho_{ij}^{\nu})_{ij}$ of the raw correlation matrix $(\rho_{ij})_{ij}$. It is known [Styan (1973)] that

$$\lambda_1^{(\nu)} \leq \lambda_1^{(\nu')} \ (\nu \leq \nu') \quad \text{and} \quad \lambda_p^{(\nu)} \geq \lambda_p^{(\nu')} (\nu \geq \nu');$$

that is, the extreme linear principal components of the raw variables are the extreme APCs of the multivariate Gaussian, a finding first attributed to Kolmogorov [Lancaster (1969)]. However, the second-smallest APC is not necessarily the second-smallest linear principal component of the raw variables; that is, $\lambda_1^{(2)} < \lambda_2^{(1)}$ is possible. If this is the case, the second-smallest APC functions are a set of parabolic transforms of the form $a_i h_i^{(2)}(X_i)$, which illustrates the horseshoe effect.

In the example of Section 3.4, the theoretical linear principal component eigenvalues for the first five Hermite transforms are:

ν	$\lambda_1^{(\nu)}$	$\lambda_2^{(\nu)}$	$\lambda_3^{(\nu)}$	$\lambda_4^{(\nu)}$
1	0.02	0.55	0.77	2.66
2	0.20	0.76	1.01	2.03
3	0.38	0.86	1.07	1.70
4	0.52	0.91	1.06	1.50
5	0.63	0.94	1.05	1.38

The smallest eigenvalues 0.20, 0.38 and 0.52 of the second, third and fourth Hermite transforms are less than the second-smallest eigenvalue 0.55 of the first (linear) Hermite transform. Thus, not only do we get parabolic transforms in the second-smallest APCs, but cubic and quartic ones as third- and fourth-smallest APCs, confirming the empirical findings of Section 3.4. This goes beyond what the term "horseshoe effect" indicates. See Gifi (1990) and de Leeuw (1982) for further discussions of this topic.

Although the Gaussian yields linear transforms at least for the extreme APCs, in real data nothing prevents horseshoes from appearing as the extreme APCs, at least in cases of weak structure. This can be illustrated with certain families of elliptical distributions that differ from the Gaussian [Buja (1990), Section 11].

4.8. *The role of centering.* We conclude this section with a technicality which has some bearing on APC computation. So far we assumed that the spaces H_i have centered variables only, that is, $\mathbf{E}\,\phi_i = 0$, or, equivalently, $\langle \phi_i, 1 \rangle = 0$. This assumption can be weakened, but care is needed if the constants are part

of some or all spaces H_i. For uncentered APC analysis, the constrained optimization problem is in terms of cross-moments rather than covariances:

$$\mathbf{E}\left(\left(\sum \phi_i\right)^2\right) = \min \quad \text{subject to } \sum_i \mathbf{E}(\phi_i^2) = 1,$$

which is formally the same problem in terms of inner products on $\mathbf{H}$.

In practice, the "real" APCs are barely affected by the presence of constants. Some artifactual eigenspaces are generated, but they can be identified and eliminated.

LEMMA 4.17. *Assume that k of the p spaces, $H_1, H_2, \ldots, H_k$ (say), contain the constant functions, while the remaining $p - k$ are spaces of centered functions. The associated operator $\mathbf{P}$ has a trivial eigenvector $(1_1, 1_2, \ldots, 1_k, 0, \ldots, 0)$ with eigenvalue $\lambda = k$ and one trivial eigenspace of dimension $k - 1$ for the eigenvalue $\lambda = 0$, spanned by $k-1$ vectors of the form $(1_1, 0, \ldots, 0, -1_j, 0, \ldots, 0)$ for $2 \le j \le k$. Any other set of transformations orthogonal to these trivial ones consists of real APCs of the equivalent centered problem.*

We use the obvious notation 1_j for the constant 1 function in H_j.

PROOF. It is obvious that $\phi_i' = 1$ for $i \le k$ and $\phi_i' = 0$ for $i > k$ defines an eigenfunction for $\mathbf{P} : P_i \Sigma_j \phi_j' = k \cdot \phi_i'$ since $\Sigma_i \phi_i' = k$ and $P_i 1 = 1$ or 0 depending on whether $i \le k$ or $i > k$, due to the assumptions on the spaces H_i. It is trivial that $(1_1, \ldots, -1_j, \ldots)$ are eigenfunctions for $\lambda = 0$.

Let now $\{\phi_i\}_i$ be some other eigenfunction orthogonal to the above ones. Orthogonality to the $\lambda = k$ eigenfunction means $\Sigma_{i=1}^k \langle \phi_i, 1 \rangle = 0$, that is, $\Sigma_{i=1}^k \mathbf{E}\phi_i = 0$. Orthogonality to the $\lambda = 0$ eigenfunctions implies $\langle \phi_1, 1 \rangle - \langle \phi_j, 1 \rangle = 0$ for $j \le k$, that is, for these j's, $\mathbf{E}\,\phi_j$ are all equal. Together it follows that $k \cdot \mathbf{E}\,\phi_j = 0$ for $j \le k$. Since $\mathbf{E}\phi_j = 0$ for $j > k$ by assumption, the last statement of the proposition follows. □

The lemma implies that meaningful eigenvalues exactly or very near 0 might be indistinguishable from the trivial eigenvalues of the $k-1$-dimensional space. Thus, if the eigenvalue 0 has multiplicity greater than $k - 1$, the trivial eigenfunctions should be removed by centering before attempting interpretation. A similar warning applies of course to the eigenvalue k, but the danger is considerably smaller by comparison.

5. Estimation and computation. We present two approaches to the estimation of APCs. In the first we restrict APC function estimates to finite-dimensional spaces, in which case we can use orthogonal projections, and computations amount to straightforward numerical linear algebra. The second approach is based on more general smoothers which we use as building blocks in an iterative algorithm reminiscent of ACE.

5.1. *Finite-dimensional function spaces.* This method simply solves the APC optimization for empirical variances in finite-dimensional function spaces. Since our notation applies to either populations or empirical distributions, specialization to estimation is trivial: **E** and **var** stand for sample average and sample variance, respectively, and orthogonal projections amount to least squares regressions. Thus, computations of estimates involve standard matrix algebra; see, for example, de Leeuw (1982). Additive scaling methods are usually formulated in terms of matrix algebra, a fact which may have kept statistical audiences away from this literature.

Let the functions f_{ik} for $k = 1, \ldots, d_i$ form an orthonormal basis of the d_i-dimensional space H_i:

$$\langle f_{ik}, f_{ik'} \rangle = \mathbf{E}(f_{ik} f_{ik'}) = \delta_{kk'}, \qquad k, k' = 1, \ldots, d_i.$$

We express any $\phi_i \in H_i$ as $\phi_i = \sum_{k=1}^{d_i} a_{ik} f_{ik}$. In rewriting the APC problem, we distinguish between coefficient vectors that are represented as *column* vectors and function vectors represented as *row* vectors:

$$\begin{aligned}
\sum_i \phi_i &= \sum_i \sum_k^{d_i} a_{ik} f_{ik} \\
&= \sum_i \mathbf{f}_i \mathbf{a}_i && \text{where } \mathbf{f}_i = (f_{i1}, \ldots, f_{id_i}),\ \mathbf{a}_i^t = (a_{i1}, \ldots, a_{id_i}), \\
&= \mathbf{F}\mathbf{a} && \text{where } \mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_p),\ \mathbf{a}^t = (\mathbf{a}_1^t, \ldots, \mathbf{a}_p^t), \\
\left\| \sum_i \phi_i \right\|^2 &= \mathbf{a}^t \mathbf{E}(\mathbf{F}^t \mathbf{F}) \mathbf{a} && \text{where } \mathbf{E}(\mathbf{F}^t \mathbf{F}) \text{ is the cross-moment matrix of the} \\
& && \textstyle\sum d_i\text{-dimensional random vector } \mathbf{F}.
\end{aligned}$$

The normalizing constraint becomes

$$\sum_i \|\phi_i\|^2 = \sum_i \sum_{k=1}^{d_i} \|a_{ik} f_{ik}\|^2 = \sum_i \sum_{k=1}^{d_i} a_{ik}^2 = \mathbf{a}^t \mathbf{a} = 1.$$

PROPOSITION 5.1. *If all spaces H_i are finite-dimensional, their additive principal components are obtained by an eigenanalysis of the cross-moment matrix* $\mathbf{E}(\mathbf{F}^t \mathbf{F})$, *that is, a linear principal component analysis of a collection* $\mathbf{F}$ *of orthonormal bases for the spaces H_i.*

For estimation of APCs from the finite sample cross-moment or variance-covariance matrix of a collection $\mathbf{F}$ of basis functions, it is of course crucial to choose for H_i suitable spaces of low dimensions, commensurate with the sample size. The problem of proper choice of dimensionality is a hard one and has not been tackled to our knowledge. At this point, we do not have more sophisticated rules other than the recommendation that the fitted degrees per APC ($\approx \Sigma_i \dim H_i$ in the fully centered case, adjusted for irrelevant constants otherwise) be controlled such that they do not exceed a fraction such as 1/5

of the sample size. Depending on the strength of the structure present, this may be either conservative or liberal. Some of the nonlinear smoothers used in the iterative methods of the next section perform bandwidth choice on a per-smoother basis. How these individually cross-validating choices interact with each other in additive methods is not understood.

There are several typical choices for the finite-dimensional spaces H_i:

1. If X_i is continuous, it is natural to use a space of splines with a low number of knots, or a space of polynomials of low degree (sometimes as low as 1, i.e., linear). Polynomials have fallen out of favor due to global stiffness and an inability to adapt to local features. Splines fare better with local adaptation if the knots are placed in a data-driven way, for example, at a ladder of quantiles such as the three interior quartiles. To make the above proposition directly applicable, the bases have to be orthonormal with regard to the empirical distribution of the respective variables.
2. If X_i is discrete and takes on d_i responses $c_{i1}, c_{i2}, \ldots, c_{id_i}$, a variable transformation amounts to "scaling" or "scoring" these responses by assigning them real values. The natural basis for the space H_i is given by indicator variables $I_{c_{ij}}(X_i)$ for response categories. The space contains constants, so Lemma 4.17 needs to be applied to weed out the trivial eigenelements. The indicator basis is orthonormal if standardized versions are used: $f_{im} = I_{c_{im}}(X_i)/\sqrt{\pi_{im}}$, where $\pi_{im} = \mathbf{E}(I_{im}(X_i)) = \mathbf{prob}(X_i = c_{im})$. Variables can of course be of mixed type: some continuous, others discrete, but if all variables are discrete and largest rather than smallest APCs are extracted, the technique is called "multiple correspondence analysis." In this case, the matrix $\mathbf{E}(\mathbf{F}^t\mathbf{F})$ contains the joint probabilities (for populations) or relative frequencies (for empirical measures) of occurrence for every pair of categories, weighted by the marginal probabilities:

$$\left[\mathbf{E}(\mathbf{F}^t\mathbf{F})\right]_{im,\,jn} = \frac{\mathbf{prob}(X_i = c_{im},\, X_j = c_{jn})}{\mathbf{prob}(X_i = c_{im})^{1/2}\,\mathbf{prob}(X_j = c_{jn})^{1/2}}.$$

In psychometric work, the unweighted matrix of joint probabilities (relative frequencies) $G_{im,\,jn} = \mathbf{prob}(X_i = c_{im}, X_j = c_{jn})$ is called the Burt table, the contingency table of all pairwise combinations of variables. Note that

$$\mathbf{E}(\mathbf{F}^t\mathbf{F}) = D^{-1/2}GD^{-1/2} \quad \text{where } D = \operatorname{diag}(\pi_{11}, \pi_{12}, \ldots, \pi_{pd_p}).$$

Multiple correspondence analysis is thus the eigenanalysis of the weighted Burt table.

It is one of the strengths of scaling methods such as APC that mixed types of data, that is, both discrete and continuous, can be incorporated in the same analysis by a suitable choice of indicator, linear, spline or polynomial transformations for each variable individually. This has been extensively discussed in the psychometric literature, which also considers ordinal data and monotone

transformations thereof. We omit the latter case since monotone transformations form convex cones rather than linear subspaces.

5.2. *An iterative method.* The methods just introduced are characterized by the use of least squares regressions to estimate conditional expectations $P_i = E^{X_i}$. This is a severe limitation since some of the most flexible and promising fitting methods are not of the least squares type: smoothing or Reinsch splines, running means/lines/parabolas, as well as any nonlinear modification based on global or local cross-validation and/or robustification, such as Supersmooth [Friedman and Stuetzle (1982)], Lowess [Cleveland (1979)] and Turbo [Friedman and Silverman (1989)].

In this subsection, we introduce an iterative method of computation which allows us to use other than least squares and possibly nonlinear fitting techniques to estimate conditional expectations. The algorithm has only a heuristic justification if other than least squares methods are used, similar to ACE [Breiman and Friedman (1985)]. It is a power algorithm for the computation of eigenelements of the operator $\mathbf{P}$, adapted so as to extract the smallest rather than largest APCs.

For most initial $\Phi_o \in \mathbf{H}$, the sequence

$$\Phi^{[k]} = \frac{\mathbf{P}^k \Phi_o}{\|\mathbf{P}^k \Phi_o\|_H}, \qquad k = 1, 2, \ldots,$$

converges in norm to an eigenfunction of $\mathbf{P}$ belonging to the maximal eigenvalue, that is, a largest APC. For finding smallest APCs with eigenvalues smaller than 1, the spectrum needs to be flipped and shifted by replacing $\mathbf{P}$ with an operator of the form $\alpha\mathbf{I} - \mathbf{P}$. Its eigenfunctions are the same as those of $\mathbf{P}$, and its eigenvalues are of the form $\alpha - \lambda$ for eigenvalues λ of $\mathbf{P}$. To ensure that the sequence

$$\frac{(\alpha\mathbf{I} - \mathbf{P})^k \Phi_o}{\|(\alpha\mathbf{I} - \mathbf{P})^k \Phi_o\|_H}$$

converges to a smallest eigenfunction to $\mathbf{P}$, the constant α has to be chosen sufficiently large—yet not so large as to slow down convergence unnecessarily. A reasonably efficient choice is $\alpha = (p + 1)/2$. For this value, the large eigenvalues of $\mathbf{P}$ are mapped to an interval centered at 0:

$$\lambda \in [1,p] \quad \Rightarrow \quad \alpha - \lambda \in \left[-\frac{(p-1)}{2}, \frac{(p-1)}{2}\right],$$

while the small eigenvalues are tacked on at the right end:

$$\lambda \in [0,1] \quad \Rightarrow \quad \alpha - \lambda \in \left[\frac{(p-1)}{2}, \frac{(p-1)}{2} + 1\right].$$

Thus, the sequence of small eigenvalues of $\mathbf{P}$ (in ascending order) have become dominant in $\alpha\mathbf{I}-\mathbf{P}$ (in descending order), yet only enough to achieve reasonable convergence speed. With this choice of α, the essentials of the algorithm are as follows:

ALGORITHM. Choose initial transformations $\phi_1^{[0]}, \phi_2^{[0]}, \ldots, \phi_p^{[0]}$, and set $\alpha = (p+1)/2$.

Repeat for $N = 1, 2, \ldots,$

Do for $i = 1, \ldots, p,$

$$\phi_i \leftarrow \alpha\phi_i^{[N-1]} - P_i \sum_{j=1}^{p} \phi_j^{[N-1]} \quad \text{Update}$$

Standardize with $c = \left(\sum_i \|\phi_i\|^2 \right)^{-1/2}$

$$\left(\phi_1^{[N]}, \phi_2^{[N]}, \ldots, \phi_p^{[N]}\right) \leftarrow (c\phi_1, c\phi_2, \ldots, c\phi_p) \quad \text{Outer iteration}$$

Until var $\sum \phi_i^{[N]}$ converges.

Up to the spectrum shift, the iteration scheme employed here is reminiscent of the ACE algorithm. Note, however, that the innermost update step is parallel, not sequential; that is, for ϕ_i we do not use $\phi_1, \ldots, \phi_{i-1}$, although they have already been computed.

For computation of the second-smallest and other higher-order small APCs, we simply add orthogonality constraints. This is implemented by adding a series of Gram-Schmidt steps following the update step:

$$\phi_i \leftarrow \phi_i - \left(\sum_{j=1}^{p} \langle \phi_j, \phi_j^{(1)} \rangle \right) \phi_i^{(1)}.$$

If, for purposes of estimation, smoothers other than projections are used as building blocks for P_i, no guarantee can be given for convergence. We have rarely experienced nonconvergence of the algorithm although convergence is apt to be slow, particularly in estimation of higher-order APCs.

If the algorithm converges, the resulting functions are defined to be APC function estimates. The outputs may depend on the initializations, which therefore become part of the definition of the estimate. The particular choice of smoothers implicitly dictates the smoothness constraints placed on the function estimates. We have used Supersmooth [Friedman and Stuetzle (1982)], Lowess [Cleveland (1979)] and Turbo [Friedman and Silverman (1989)]. Supersmooth tends to produce smaller eigenvalues than other methods and fairly wiggly transforms, possibly indicating a tendency to overfit. We used Lowess with locally quadratic fits, the approximate algorithm and a 50 percent data

window: the fits were consequently very smooth. Turbo typically gave fits very similar to Lowess, although we observed strange convergence patterns, probably due to the nonlinear behavior caused by the choice of the number of knots. For these estimates, there is no associated extremal problem that the algorithm solves. Some comfort may be derived from asymptotics, as it seems likely that the consistency results of Breiman and Friedman (1985) can be carried over to APCs.

6. Discussion: extensions and open problems. The material given in this paper is far from exhaustive. There are obvious extensions of APC analysis, and there are many open problems.

A simple extension follows from noting that the APC formalism depends only on subspaces $H_1, H_2, \ldots, H_p$, but not on the assumption that these spaces are sets of transforms of variables $X_1, X_2, \ldots, X_p$. Therefore, we can use other specifications of subspaces, an example being subspaces that arise in functional interaction models [see Stone (1994) for recent developments]. In these models, multivariate response surfaces $f(x_1, x_2, \ldots)$ are approximated by additive decompositions into main effects and interactions: $\psi_1(x_1) + \psi_2(x_2) + \psi_{12}(x_1, x_2) + \cdots \in H_1 + H_2 + H_{12} + \cdots$. For identifiability, one assumes that the interactions (e.g., $\psi_{12} \in H_{12}$) are orthogonal to their subordinate main effect and interaction spaces (H_1 and H_2 in this instance). The notion of concurvity associated with interaction models is defined in terms of (approximate) degeneracies of the form $\phi_1(x_1) + \phi_2(x_2) + \phi_{12}(x_1, x_2) + \cdots \approx 0$. Because of their increased flexibility, interaction models are even more at risk from concurvity than additive models. This gives urgency to the task of developing diagnostics: APC analysis of the spaces $H_1, H_2, H_{12}, \ldots$ [Buja (1994)] may prove a valuable candidate.

The most basic open problems in APC analysis concern inference. For example, which eigenvalues should be considered "real," that is, how small is small given the sample size and number of variables? An obvious idea for testing significance of the smallest eigenvalue is to use a permutation test for the null hypothesis of unrelated variables. The more difficult problem is deciding when to stop accepting eigenvalues of higher order: the classical debate on the number of factors problem could be directly transferred to APCs. Related to this problem is the question of when data are appropriate for APC analysis. For example, if the intrinsic dimension of 10-dimensional data is very nearly only 2, it would not make sense to estimate eight implicit equations with APCs. Instead one should use a dimension reduction approach based on largest eigenvalues or principal surfaces. Another problem concerns multiplicity of small eigenvalues: how do we infer it from data; and how do we cope with the resulting indeterminacies in the estimated transforms, if, for example, the smallest two eigenvalues are established as equal? Our comments in Section 3.1 barely touch the issue. It would also be nice to have a formal method for weeding out horseshoe transforms, although they are rarely a problem in practice for the informed user.

Finally, we need a better understanding of the use of modern smoothers as APC building blocks. The algorithm of Section 5.2 provides a gateway for im-

porting any fitting procedure to APCs, but how can we understand what APC analysis does when the fitting procedures are not least squares but penalized least squares regressions, such as smoothing splines? In this instance, we do have an answer, which will appear in a future paper. More difficult is the question of how to incorporate adaptivity, that is, bandwidth selection, in APC. This problem area is wide open.

Acknowledgment. This work was done while the second author was with Bellcore.

REFERENCES

BENZECRI, J. P. (1980). *Pratique de L'Analyse des Données* **1**. Dunod, Paris.

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Amer. Statist. Assoc.* **80** 580–619.

BUJA, A. (1990). Remarks on functional canonical variates, alternating least squares methods and ACE. *Ann. Statist.* **18** 1032–1069.

BUJA, A. (1994). Discussion of "The use of polynomial splines and their tensor products in multivariate function estimation," by C. J. Stone. *Ann. Statist.* **22** 171–177.

BUJA, A., DONNELL, D. and STUETZLE, W. (1986). Additive principal components. Technical Report, Dept. Statistics, Univ. Washington, Seattle.

BUJA, A., HASTIE, T. J. and TIBSHIRANI, R. (1989). Linear smoothers and additive models. *Ann. Statist.* **17** 453–555.

BUJA, A. and KASS, R. (1985). Comment on "Estimating optimal transformations for multiple regression and correlation," by L. Breiman and J. H. Friedman. *J. Amer. Statist. Assoc.* **80** 602–607.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74** 829–836.

DAUXOIS, J. and POUSSE, A. (1976). Les analyses factorielles en calcul de probabilités et en statistique. Ph.D. thesis, Univ. Paul-Sabatier, Toulouse.

DAUXOIS, J. and POUSSE, A. (1977). Some convergence problems in factor analysis. In *Recent Developments in Statistics* (J. R. Barra, B. Van Cutsem, F. Broadeau and G. Romier, eds.) 387–402. North-Holland, Amsterdam.

DE LEEUW, J. (1982). Nonlinear principal component analysis. *COMPSTAT, IASC* 77–86.

DONNELL, D. J. (1987). Additive principal components—a method for estimating equations with small variance from data. Ph.D. thesis, Univ. Washington, Seattle.

FRIEDMAN, J. H. and SILVERMAN, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31** 3–40.

FRIEDMAN, J. H. and STUETZLE, W. (1982). Smoothing of scatterplots. Technical Report ORION 3, Dept. Statistics, Stanford Univ.

GIFI, A. (1990). *Nonlinear Multivariate Analysis*. Wiley, New York.

GILULA, Z. and HABERMAN, S. J. (1988). The analysis of multivariate contingency tables by restricted canonical and restricted association models. *J. Amer. Statist. Assoc.* **83** 760–771.

GREENACRE, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic, London.

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

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, New York.

JORGENS, K. (1970). *Linear Integral Operators*. Pitman Articles LTD, London.

KETTENRING, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika* **58** 433–451.

KRUSKAL, J. B. and SHEPARD, R. N. (1974). A nonmetric variety of linear factor analysis. *Psychometrika* **39** 123–157.

LANCASTER, H. O. (1969). *The Chi-squared Distribution*. Wiley, New York.
LEBART, L., MORINEAU, A. and WARWICK, K. M. (1984). *Multivariate Descriptive Statistical Analysis*. Wiley, New York (translated by E. M. Berry).
NAOURI, J. C. (1970). Analyse factorielle des correspondances continue. *Publ. Inst. Statist. Univ. Paris* **19** 1–100.
STONE, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22** 118–184.
STYAN, G. P. (1973). Hadamard products and multivariate statistical analysis. *Linear Algebra Appl.* **6** 217–240.
VAN RIJCKEVORSEL, J. (1982). Canonical analysis with B-splines. *COMPSTAT, IASC* 393–398.
YOUNG, F. W., TAKANE, Y. and DE LEEUW, J. (1978). The principal components of mixed measurement level multivariate data. *Psychometrika* **43** 279–281.

DEBORAH J. DONNELL
STATSCI, SUITE 500
1700 WESTLAKE AVENUE NORTH
SEATTLE, WASHINGTON 98109

ANDREAS BUJA
AT&T BELL LABORATORIES
ROOM 2C-261
600 MOUNTAIN AVENUE
MURRAY HILL, NEW JERSEY 07974-0636

WERNER STUETZLE
STATISTICS DEPARTMENT, GN-22
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195

DISCUSSION

BERNARD D. FLURY
Indiana University

The past 15 years have shown considerable progress in generalizations and modifications of classical principal component analysis. This includes distribution theory for sample principal components in elliptical families [Muirhead (1982) and references therein], robust estimation and testing [Campbell (1980), Devlin, Gnanadesikan and Kettenring (1981) and Tyler (1981, 1983)], *common principal components* in several groups [Flury (1988)], principal components of patterned covariance matrices [Neuenschwander (1991) and Flury and Neuenschwander (1993)] and last but not least, generalizations to nonlinear situations, which constitute perhaps the thorniest area. Donnell, Buja and Stuetzle (DBS) give a fundamental building block in this field, the previous most significant building block being the *principal curves* of Hastie and Stuetzle (1989).

DBS stress that their method is rooted in the psychometric literature, but there is at least one (to my knowledge) direct predecessor in the statistical literature as well: Gnanadesikan and Wilk's (1969) *Generalized Principal Component Analysis*, described in detail also in Gnanadesikan (1977), Seber (1984) and Jackson (1991), which imitates the successful use of polynomials in regression. Gnanadesikan and Wilk's approach was to introduce powers and products of the