ADDITIVE PRINCIPAL COMPONENTS

by

Andreas Buja,
Deborah J. Donnell
and
Werner Stuetzle

TECHNICAL REPORT No. 76
January 1986

Department of Statistics, GN-22
University of Washington
Seattle, Washington 98195 USA
Additive Principal Components

Andreas Buja
Deborah Donnell
Werner Stuetzle

Department of Statistics, GN-22, University of Washington
Seattle, Washington 98195, USA

ABSTRACT

Additive principal components are proposed as a generalisation of linear principal components, with sums of arbitrary transformations replacing linear combinations of variables. Cocurvity — the presence of additive principal components with small variances — indicates the concentration of the observations around a possibly nonlinear manifold. Cocurvity among predictor variables of an additive regression model leads to similar problems as collinearity among predictors in a linear model.

Additive principal components are shown to be the solution of an eigenproblem in an appropriate function space. An iterative algorithm is given for the sequential computation of the smallest additive principal components, and it is shown to converge to the correct minimising solution.

This work was supported by the grants: DOE DE-FG06-85ER25006, NSF DMS-8504359
Additive Principal Components

Andreas Buja
Deborah J. Donnell
Werner Stuetzle

January 17, 1986
Contents

1 Introduction .................................................. 3

2 The Smallest Additive Principal Component .......... 6

3 An Iterative Algorithm ................................... 8
   3.1 A Naive Algorithm .................................. 8
   3.2 Analysis of Convergence for Linear Transformations ... 9
   3.3 The Hilbert Space Setting .......................... 11
   3.4 The Eigenfunction Characterization ................ 13
   3.5 The Final Algorithm ................................ 17
   3.6 A Null Situation for Additive Principal Components ... 18

4 Further Additive Principal Components ................. 21
   4.1 Definition and Algorithm .......................... 21
   4.2 Use and Interpretation ............................. 24
Additive Principal Components

Andreas Buja
Deborah Donnell
Werner Stuetzle

Department of Statistics, GN-22, University of Washington
Seattle, Washington 98195, USA

ABSTRACT
Additive principal components are proposed as a generalisation of linear principal components, with sums of arbitrary transformations replacing linear combinations of variables. Cocurvity — the presence of additive principal components with small variances — indicates the concentration of the observations around a possibly nonlinear manifold. Cocurvity among predictor variables of an additive regression model leads to similar problems as collinearity among predictors in a linear model.

Additive principal components are shown to be the solution of an eigenproblem in an appropriate function space. An iterative algorithm is given for the sequential computation of the smallest additive principal components, and it is shown to converge to the correct minimising solution.

This work was supported by the grants: DOE DE-FG06-85ER25006, NSF DMS-8504359
1 Introduction

Computers have made it possible to collect and analyze ever larger data sets. This development has created a need for new statistical methods. Small sample size necessarily limits the complexity of models that can be fitted and of structure that can be reliably detected; thus one is restricted to classical parametric methods like linear regression and linear principal components. On the other hand, large sample size allows the detection of complicated structure, and the fitting of complex models. This means that nonparametric methods for description and inference making fewer assumptions about the underlying situation are called for.

These considerations are the reasons why in recent years there has been a surge of interest in methods for nonparametric multiple regression, in particular the models of additive regression [3] and ALS [2] or ACE regression [1]. The former models the response as an additive function of the predictors:

\[ Y \sim \sum_{j=1}^{p} \phi_j(X_j). \]

ACE regression finds transformations \( \phi_1 \ldots \phi_p \) of the predictors as well as a transformation \( \theta \) of the response:

\[ \theta(Y) \sim \sum_{j=1}^{p} \phi_j(X_j). \]

This report is part of an attempt to further the development of methodology suitable for detecting complex structure in data. We present here a method for finding transformations \( \phi_1 \ldots \phi_p \) of the data which satisfy as nearly as possible the constraint

\[ \sum_{j=1}^{p} \phi_j(X_j) \approx 0. \]

Such a constraint indicates a nonlinear association among \( X_1, X_2, \ldots X_p \).

The importance of recognizing nonlinear dependencies among the predictor variables when building additive regression models is analogous to
the importance of detecting collinearity patterns when building linear models. In a linear model collinearity between carriers results in inflated variance of the estimated regression coefficients. The resulting instability of the model is harmful to the use of regression for hypothesis testing, prediction and estimation. In particular it is difficult to infer the separate influence of the collinear explanatory variables on the response variable. In the additive case, similar difficulties arise. Suppose we fit an additive model $Y \approx \sum_{j=1}^{p} \phi_j(X_j)$ to the data. We naturally want to make both qualitative and quantitative statements about the contributions of each $X_i$ in the model, based on the estimated $\phi_i$. Consider the analogy to the extreme case of exact collinearity, where there exist functions of the variables such that $\sum \theta_i(X_i) = 0$. In this situation, the alternative fit

$$Y \sim \sum_{j=1}^{p} (\phi_j + \theta_j)(X_j)$$

is indistinguishable from the initial one. If the data come close to satisfying this constraint, some or all of the estimated $\phi_j$ will not be stable. We are clearly in no position to interpret the component functions of the fitted model when this is the case. A method which enables us to examine how close the data come to satisfying an additive constraint would thus be a diagnostic check for global stability of the transforms in additive or ACE regression.

In addition to detecting dependencies among carrier variables in the regression setting, the ability to find additive dependencies is an important contribution to multivariate analysis. Additive constraints describe a geometric structure underlying the data that is of interest for its own sake. Consider the linear case. A linear constraint $l(x) = a \cdot x = 0$ defines a linear manifold of co-dimension 1. If the data nearly satisfy such a constraint, they lie close to this linear manifold. Analogously, an additive constraint $\sum \theta_i(X_i) = 0$ defines an additive manifold of co-dimension 1, and data nearly satisfying this constraint lie near the additive manifold. Although the additive form of the equation places some restrictions on the surfaces that can be modelled, they nevertheless are considerably more general than linear manifolds. Hence if they can be reliably estimated, and properly displayed and interpreted, additive manifolds have the potential
to be an important tool for better understanding the multivariate nature of data.

In this report we define additive principal components of a random vector $\mathbf{X} = (X_1, X_2 \ldots X_p)$ and suggest a method for their calculation. Only the abstract version of the algorithm, operating on distributions, is described. Details of the algorithm for the sample case as well as examples will be the subject of a forthcoming report.
2 The Smallest Additive Principal Component

For the remainder of this report we will assume the random variables $X_1, X_2, \ldots X_p$ to have $E(X_i) = 0$ and $\text{var}(X_i) = 1$.

We wish to determine whether $X_1, X_2, \ldots X_p$ come close to satisfying an additive constraint $\sum \phi_i(X_i) = 0$, where the transformations $\phi_i$ of the constraint are unknown. Hence we want to estimate from the sample the additive constraint that the data come nearest to satisfying, in some sense yet to be made precise.

It is instructive to first consider the simpler case where we restrict our attention to linear functions of the variables. Then we wish to find the linear combination of the variables that is closest to zero in some sense. We might search for the vector $a$ minimizing the variance of the sum, $\text{var} \sum a_i X_i = \text{var}(X \cdot a)$. To avoid the trivial solution, $a = 0$, we impose the constraint $\sum a_i^2 = \sum \text{var} a_i X_i = 1$. Then the solution vector is known to be an eigenvector for the smallest eigenvalue of $\text{cov}(X) = \Sigma$. The random variable $\sum a_i X_i$ is called a smallest principal component of $X$. The corresponding linear function $l_1(x) = a \cdot x$ defines a linear manifold of co-dimension 1 in $p$-space through $l_1(x) = 0$. It can be shown that this manifold minimizes the expected squared distance from the observations to any linear manifold of codimension 1. This second characterization is an intrinsically geometric property of the smallest principal component.

We now wish to make the transition to additive functions. Suppose there exist non trivial functions $\phi_i$ which make the sum of the transformed variables “closest” to zero. The function $f(x) = \sum \phi_i(x_i)$ describes through $f(x) = 0$ an additive manifold which lies close to the data. We need to decide on a criterion for uniquely defining the set of transformations $\Phi = (\phi_1, \phi_2, \ldots , \phi_p)$ of the additive constraint. It is natural to consider extending the characterizations of the smallest principal component of $X$. We could define $\Phi$ as the vector of transformations of the variables minimizing $\text{var} \sum \phi_i(X_i)$ subject to $\sum \text{var} \phi_i(X_i) = 1$. Alternatively, we could determine the additive manifold described by $\sum \phi_i(X_i) = 0$ which minimizes the expected squared distance from the observations to any additive manifold of co-dimension 1. In the additive form, these characterizations are not
equivalent.

We chose to investigate here the minimum variance approach, as it is computationally more tractable:

Definition

The smallest additive principal component of \( X = (X_1, \ldots X_p) \) is the random variable \( Y = \sum \phi_i(X_i) \) minimizing \( \text{var} \sum \phi_i(X_i) \) subject to \( \sum \text{var} \phi_i(X_i) = 1 \).

At this point we would like to introduce a relaxation in our terminology. The additive principal component is unambiguously defined by the vector of transformations \( \Phi = (\phi_1, \ldots, \phi_p) \). It corresponds to the score vector \( a \) defining the linear principal component. Since we will be primarily discussing these functions directly, rather than make somewhat cumbersome reference to “the vector of transformations defining the additive component” we will call \( \Phi \) itself “the additive principal component”. While this usage is not technically correct, it should not prove to be confusing.

The first purpose of this report is to investigate estimation of the smallest additive principal component of a set of variables.

It might not be enough to estimate only the smallest additive component of the data, as there may be other additive dependencies of importance. In classical principal components searching for other linear dependencies would correspond to examining the second smallest principal component.

The second purpose of this report is to define second and further smallest additive principal components, and then to suggest methods for their estimation.
3 An Iterative Algorithm

3.1 A Naive Algorithm

We wish to find functions $\Phi = (\phi_1, \phi_2 \ldots \phi_p)$ that minimize $\text{var} \sum \phi_i(X_i)$ subject to $\sum \text{var} \phi_i(X_i) = 1$. Rewriting the variance of the sum,

$$\text{var} \sum_i \phi_i(X_i) = \text{var} \left( -\sum_{i \neq 1} \phi_i(X_i) - \phi_1(X_1) \right)$$

suggests a straightforward componentwise minimization scheme, in the spirit of ACE [1]. Let us ignore the constraint $\sum \text{var} \phi_i(X_i) = 1$ for the moment. If we assume $\phi_2 \ldots \phi_p$ to be known, then the minimizing transformation of $X_1$ is given by

$$\hat{\phi}_1(X_1) = E^{X_1}( -\sum_{i \neq 1} \phi_i(X_i)).$$

(Here, $E^{X_1} \equiv E(\cdot | X_1).$) This suggests the following algorithm.

Naive algorithm

Choose initial estimate $\Phi^{(0)}$.

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

For $i = 1 \ldots p$

$$\phi_i := E^{X_i}( -\sum_{j \neq i} \phi_j^{(k-1)}).$$

Set

$$\Phi^{(k)} \equiv (\phi_1^{(k)}, \ldots, \phi_p^{(k)})$$

$$:= (\sum_i \text{var} \phi_i)^{-1/2}(\phi_1, \phi_2 \ldots \phi_p)$$

Until $\text{var} \sum \phi_i^{(k)}$ converges
Notice that the iteration scheme employed here in the inner loop is different to that used in the ACE algorithm of Breiman and Friedman [1]. They replace each $\phi_i$ by its new estimate as the inner loop proceeds, whereas we obtain the new $p$-tuple using *only the previous $p$-tuple throughout the entire inner loop*! This provides us with a natural way of restandardizing in the outer loop and will allow a transparent analysis of the convergence of the algorithm. Breiman and Friedman comment in their work [1] that the successive replacement algorithm converges more quickly. They are unable to prove convergence for their inner loop procedure, and yet they report empirical evidence that convergence to the correct solution usually occurs. We may explore a faster algorithm based on this modification in a future report.

Before we attempt to establish whether the naive algorithm above is a method of calculating the smallest additive component of $X$, it is instructive to analyze the procedure for the special case where the transformations are restricted to be linear.

### 3.2 Analysis of Convergence for Linear Transformations

The problem in the linear case is to find the vector $a$ minimizing the variance of the corresponding linear combination of $X$, i.e., to minimize $\text{var} \sum l_i(X_i) = \text{var} \sum a_i X_i$ subject to $\sum a_i X_i = \sum a_i^2 = 1$. It is well known that the solution is an eigenvector for the smallest eigenvalue of $\Sigma$. When the transformations are restricted to be linear, analysis of the convergence of our algorithm is straightforward. Let us consider the first step of an inner loop. Following the previous section, we initially ignore the side condition and assume $a_2, \ldots, a_p$ to be known. We then find $a_1$ minimizing $\text{var} (\sum_{j \neq 1} a_j X_j - a_1 X_1)$. The coefficient $a_1$ can be estimated directly by
a simple linear regression of $-$ $\sum_{j \neq 1} a_j X_j$ on $X_1$:

$$a_1 = \frac{\mathbb{E} \left( -\sum_{j \neq 1} a_j X_j X_1 \right)}{\mathbb{E} X_1^2}$$

Assuming that this step is part of the inner loop of an algorithm in which we compute $a^{(\text{new})}_1$ from $(a^{(\text{old})}_1, a^{(\text{old})}_2, \ldots, a^{(\text{old})}_p)$, and also making use of the assumptions $\mathbb{E} X_i = 0$ and $\text{var} X_i = 1$, we can write:

$$a^{(\text{new})}_1 = -\sum_{j \neq 1} a^{(\text{old})}_j \text{cov}(X_j, X_1)$$

$$= a^{(\text{old})}_1 - \sum_{j=1}^p a^{(\text{old})}_j \text{cov}(X_j, X_1)$$

Using this equation, the full inner loop iteration over all variables can be written in vector notation:

$$a^{(\text{new})} = a^{(\text{old})} - \mathbb{E} (X'X) a^{(\text{old})}$$

$$= (I - \Sigma) a^{(\text{old})}$$

We can then impose the constraint $\sum a_i^2 = 1$ by rescaling by the factor $(\sum (a_i^{(\text{new})})^2)^{1/2}$ after the full inner loop is completed. At the kth iteration we have

$$a^{(k)} = \frac{(I - \Sigma) a^{(k-1)}}{\| (I - \Sigma) a^{(k-1)} \|} = \frac{(I - \Sigma)^{k-1} a^{(0)}}{\| (I - \Sigma)^{k-1} a^{(0)} \|}$$

This algorithm is seen to be a power method which converges to the eigenvector corresponding to the eigenvalue of $I - \Sigma$ with the largest absolute value. This may not be the vector we are seeking, which is rather the smallest eigenvector of $\Sigma$. Suppose, for the sake of illustration, that the largest eigenvalue of $\Sigma$ is 3, and the smallest 0.5. Then the eigenvalue of $I - \Sigma$ with the largest absolute value is $|1 - 3| \geq |1 - 0.5|$ and the algorithm will converge to the eigenvector for the largest eigenvalue of $\Sigma$ instead of the smallest. How the algorithm can be modified to ensure convergence to the desired eigenvector is made clear by the following remark, together with the fact that the eigenvalues of $\Sigma$ lie between 0 and $p$. 

10
Remark 3.1

If $A_{n \times p}$ has eigenvalues $0 \leq \lambda_p \ldots \leq \lambda_1 \leq p$ with corresponding eigenvectors $\nu_p \ldots \nu_1$, then $pI - A$ has eigenvalues $0 \leq p - \lambda_p \ldots \leq p - \lambda_1 \leq p$ with the same eigenvectors in reverse order $\nu_1, \ldots \nu_p$.

An eigenvector for the largest (absolute) eigenvalue of $pI - \Sigma$ is an eigenvector for the smallest eigenvalue of $\Sigma$. To ensure convergence to the smallest eigenvector of $\Sigma$, we apply the matrix $pI - \Sigma$ at each step, instead of the naive $I - \Sigma$. Then the new estimate of $a_i$ in the inner loop is

$$a_i^{(\text{new})} := pa_i^{(\text{old})} - \sum_{j=1}^{p} a_j^{(\text{old})} \text{cov}(X_j, X_i)$$

To summarize, the modified algorithm, powering up $pI - \Sigma$ instead of $I - \Sigma$, will converge to the correct solution of our problem in the case where we restrict ourselves to linear functions of the variables. Now we return to considering the additive models of Section 2, and show that the minimizing set of transformations corresponds to the smallest eigenfunction of an operator in an appropriate function space. The next section presents the additive principal component in this more formal setting.

3.3 The Hilbert Space Setting

The analysis of the algorithm in the general case is made tractable by formulating the problem in a Hilbert space setting. Since the additive principal component is defined by a vector of functions $\Phi = (\phi_1 \ldots \phi_p)$, the product space of $L_2$ functions is a natural space to consider. For $i = 1 \ldots p$, define the function spaces $H(X_i)$ as the sets of measurable functions $\phi_i(X_i)$, satisfying

$$\mathbb{E} \phi_i(X_i) = 0, \quad \mathbb{E} \phi_i^2(X_i) < \infty$$
Each of these forms a Hilbert space with inner product 
\[ \langle \phi_i, \phi'_i \rangle = \mathbb{E} \left( \phi_i(X_i) \phi'_i(X_i) \right) \]
and corresponding norm \[ \| \phi_i \|^2 = \mathbb{E} \phi_i^2(X_i). \]
Define the cartesian product space \( H = H(X_1) \times H(X_2) \times \cdots \times H(X_p) \).
The natural inner product on \( H \) for \( \Phi = (\phi_1, \ldots, \phi_p) \in H \) is
\[ \langle \Phi, \Phi' \rangle = \sum_i \langle \phi_i, \phi'_i \rangle \]
\[ = \sum_i \mathbb{E} (\phi_i \phi'_i) \]
with corresponding norm
\[ \| \Phi \|^2 = \sum_i \mathbb{E} (\phi_i^2). \]
In [1] it is established that \( H \) is a Hilbert space for which the natural embeddings of \( H(X_i), i = 1, 2, \ldots, p \) are all closed linear subspaces. Also, the norm topology of \( H \) coincides with the product topology inherited from the factors \( H(X_j) \).
We now return to the naive algorithm for finding the additive principal component stated in section 3.1. In the inner loop we obtain a new estimate of each \( \phi_i \) by:
\[ \phi_i^{(\text{new})} := \mathbb{E} X_i \left( - \sum_{j \neq i} \phi_j^{(\text{old})} \right) \]
\[ = \mathbb{E} X_i (\phi_i^{(\text{old})}(X_i) - \sum_{j=1}^p \phi_j^{(\text{old})}(X_j)). \]
In Hilbert space terminology, the conditional expectation \( \mathbb{E} X_i \) restricted to centered square integrable r.v.'s is an orthogonal projection onto the space \( H(X_i) \), denoted \( P_i \) hereafter.
The equations
\[ \phi_i^{(\text{new})} = \phi_i^{(\text{old})}(X_i) - \sum_{j=1}^p P_i(\phi_j^{(\text{old})}(X_j)), \quad i = 1 \ldots p \]
12
can be written in an operator matrix notation:

\[
\Phi^{(new)}(X) = \left( I - \begin{pmatrix} I & P_1 & \ldots & P_1 \\ P_2 & I & \ldots & P_2 \\ \vdots & \vdots & \ddots & \vdots \\ P_p & P_p & \ldots & I \end{pmatrix} \right) \begin{pmatrix} \Phi_1^{(old)} \\ \Phi_2^{(old)} \\ \vdots \\ \Phi_p^{(old)} \end{pmatrix} = (I - P) \Phi^{(old)}
\]

where \( P = \begin{pmatrix} I & P_1 & \ldots & P_1 \\ P_2 & I & \ldots & P_2 \\ \vdots & \vdots & \ddots & \vdots \\ P_p & P_p & \ldots & I \end{pmatrix} \)

\[ I \equiv \text{the identity mapping in } H. \]

After \( k \) iterations of the outer loop,

\[
\Phi^{(k)} = \frac{(I - P) \Phi^{(k-1)}}{\| (I - P) \Phi^{(k-1)} \|} = \frac{(I - P)^{k-1} \Phi^{(0)}}{\| (I - P)^{k-1} \Phi^{(0)} \|}.
\]

The naive algorithm is seen to be the power method applied to the operator \( I - P \), so it will converge to the eigenfunction for the largest absolute eigenvalue of \( I - P \) (if it exists).

### 3.4 The Eigenfunction Characterization

We now show that under suitable assumptions the smallest eigenfunction of the operator \( P \) exists and is the smallest additive principal component of \( X \). The simple identity below is the crux of the argument.

**Lemma 3.1**

\[ \langle \Phi, P \Phi \rangle = \text{var} \sum \phi_i \]
Proof:
\[
\langle \Phi, P\Phi \rangle = \sum_i \langle \phi_i, P_i \sum_j \phi_j \rangle \\
= \sum_i \langle \phi_i, \sum_j \phi_j \rangle \\
= \langle \sum_i \phi_i, \sum_j \phi_j \rangle \\
= \text{var} \sum_i \phi_i. \quad \blacksquare
\]

Noting that \( \sum \text{var} \phi_i = \| \Phi \|^2 \), we can rephrase the definition of the smallest additive principal component as an extremum problem in \( H \):

A function vector \( \Phi \in H \) minimizes \( \langle \Phi, P\Phi \rangle \) under the constraint \( \| \Phi \|^2 = 1 \) exactly if the set of transformations \( \phi_1, \phi_2, \ldots \phi_p \) minimizes \( \text{var} \sum \phi_i(X_i) \) under \( \sum \text{var} \phi_i(X_i) = 1 \).

It is a well known argument from the theory of self-adjoint operators, [5] (Th 6.7 p.125) that \( \Phi \in H \) minimizing \( \langle \Phi, P\Phi \rangle \) subject to \( \| \Phi \| = 1 \), or equivalently minimizing the Rayleigh quotient

\[
\frac{\langle \Phi, P\Phi \rangle}{\| \Phi \|^2},
\]

is an eigenfunction for the smallest eigenvalue of \( P \) (where it exists). Hence once we have established that \( P \) is selfadjoint, then if there is a smallest eigenfunction of the operator \( P \) it is a smallest additive principal component of \( X \).

At this point then, we turn our attention to establishing the properties of the operator \( P \).

Lemma 3.2

\( P \) is a bounded, selfadjoint, non-negative operator in \( H \).
Proof:

\( \mathbf{P} \) is bounded:

\[
\| \mathbf{P} \phi \|^2 = \Sigma_i \| P_i \Sigma_j \phi_j \|^2 \\
\leq \Sigma_i \| \Sigma_j \phi_j \|^2 \\
= p \| \Sigma_j \phi_j \|^2 \\
\leq p (\sum_j \| \phi_j \|^2)^2.
\]

The maximum of \( \sum_j \| \phi_j \|^2 \) under the constraint \( \sum \| \phi_j \|^2 = 1 \) is attained at \( \| \phi_j \| = p^{-1/2} \). Hence

\[
\| \mathbf{P} \phi \|^2 \leq p (\sum_j \| \phi_j \|^2)^2 \\
\leq p^2.
\]

The inequality is sharp, with equality occurring when \( X_i = X_j \quad \forall i, j \), hence the eigenvalues of \( \| \mathbf{P} \| \) are bounded above by \( p \).

\( \mathbf{P} \) is self adjoint:

\[
\langle \Phi, \mathbf{P} \Psi \rangle = \sum_i \langle \phi_i, P_i \Sigma_j \psi_j \rangle \\
= \sum_i \langle \phi_i, \Sigma_j \psi_j \rangle \\
= \langle \sum_i \phi_i, \Sigma_j \psi_j \rangle.
\]

From the symmetry of this expression it follows that \( \langle \Phi, \mathbf{P} \Psi \rangle = \langle \mathbf{P} \Phi, \Psi \rangle \).

\( \mathbf{P} \) is non-negative: from Lemma 3.1 it can be deduced that

\[
\langle \Phi, \mathbf{P} \Phi \rangle = \text{var} \sum \phi_i \geq 0.
\]

Finding linear principal components is a simple finite dimensional problem, with \( \Sigma \) having at most \( p \) distinct eigenvalues. Finding additive principal components, where we are estimating a set of \( L_2 \) functions, is generally not a finite dimensional problem — for continuous variables the operator \( \mathbf{P} \) will most often have an infinite number of eigenvalues. Given that the
eigenvalues are bounded, around which value or values do they accumulate? Will there arise situations in which the smallest eigenvalue is not distinct, and how might this be reflected in the estimated smallest additive principal component? From the theory of bounded self-adjoint operators we see that there are potential problems: P may have a non-trivial continuous spectrum or may have spectral values that are not eigenvalues. We can rule out these undesirable possibilities by adopting suitable compactness assumptions in the spirit of [1].

Assumption

The restricted projection operator $P_{i/k} : H(X_k) \rightarrow H(X_i)$ is compact for $k \neq i, i = 1 \ldots p$.

This means that the image of the unit ball is relatively compact. Even under this assumption, P itself is not compact since, for example, where $X_1$ is independent of $X_2, \ldots X_p$, the bounded set $\{\Phi : \Phi = (\phi_1(X_1), 0, \ldots, 0)', \|\Phi\| \leq 1\}$ is preserved under P but not relatively compact in H. However we can show:

Lemma 3.3

The operator $P - I : H \rightarrow H$ is compact.

Proof: Let $B$ denote the unit ball in $H$, $B_i$ the unit ball in $H(X_i)$.

$$P - I = \sum_i Q_i$$

where $Q_i : H \rightarrow H$ is defined by

$$Q_i(\Phi) = (P_i\phi_1, \ldots P_i\phi_{i-1}, 0, P_i\phi_{i+1}, \ldots P_i\phi_p)'$$

It is enough to show that every $Q_i$ is compact by [5] (Th 5.10 p.98). Since $B \subset B_1 \times \ldots \times B_p$, compactness of $Q_i$ is established if $Q_i(B_1 \times \ldots \times B_p)$ is shown to be relatively compact.

By assumption $P_i(B_j)$ is relatively compact in $H(X_i)$ $\forall i \neq j$, hence

$$Q_i(B_1 \times \ldots \times B_p) = P_i(B_1) \times \ldots P_i(B_{i-1}) \times 0 \times P_i(B_{i+1}) \ldots \times P_i(B_p)$$

is relatively compact in $H$.

The spectrum of a compact operator in an infinite dimensional Hilbert space has the following properties:
• There exists a sequence \( \{\lambda_i\}_{i=1}^\infty \) of distinct nonzero eigenvalues for which
  \[
  |\lambda_1| \geq |\lambda_2| \geq \ldots |\lambda_k| \ldots
  \]
• \( \lim_{k \to \infty} \lambda_k = 0 \)
• The eigenspaces for distinct eigenvalues are orthogonal and the sum of all the eigenspaces is dense in the whole space.
• The nonzero eigenvalues have finite multiplicity.

The spectrum of \( P - I \) is thus a bounded set with 0 as the only possible accumulation point. Since the eigenvalues \( \{\nu_k\} \) of \( P - I \) are related to the eigenvalues \( \{\lambda_k\} \) of \( P \) through \( \lambda_k = \nu_k + 1 \), the eigenvalues of \( P \) are also bounded, with corresponding accumulation point 1.

3.5 The Final Algorithm

Having shown in the previous section that the smallest additive principal component is a smallest eigenfunction of \( P \), we see that the naive algorithm for the additive case has the same flaw as it had in the linear case. The one-to-one correspondence between eigenvalues of \( P \) and \( I - P \) of Remark 3.1 also holds for operators. Since the naive algorithm converges to the eigenfunction of \( I - P \) for the eigenvalue with the largest absolute value, it will converge to the largest eigenfunction of \( P \) if any eigenvalue of \( P \) exceeds 2. As in the linear case, this can be rectified by applying the modified operator \( pI - P \), since \( \|P\| \) is bounded by \( p \). Thus the final algorithm is:
Algorithm

Repeat for $k = 1, 2\ldots$

For $i = 1, 2, \ldots p$

\[ \phi_i := p\phi_i^{(k-1)}(X_i) - \mathbb{E}X_i(\sum_{j=1}^{p} \phi_j^{(k-1)}(X_j)) \]

Set

\[ \Phi^{(k)} \equiv (\phi_1^{(k)}, \ldots \phi_p^{(k)}) \]

\[ = (\sum \text{var} \phi_i(X_i))^{-1/2}(\phi_1 \ldots \phi_p) \]

Until $\text{var} \sum \phi_i^{(k)}(X_i)$ converges.

3.6 A Null Situation for Additive Principal Components

Commonsense indicates that only additive principal components for which $\text{var} \sum \phi_i < 1$ are worth considering: since $\sum \text{var} \phi_i = 1$, only then have we detected interesting dependencies between the variables. This amounts to restricting our attention to eigenfunctions corresponding to eigenvalues of $\mathbf{P}$ less than one, and it is then natural to explore the possibility that $\mathbf{P}$ has no eigenvalues less than one. The following theorem provides a characterization of this null situation, where no interesting equation of the form $\text{var} \sum \phi_i = \epsilon (< 1)$ exists.

Theorem

The following are equivalent:

1. All the spectral values of $\mathbf{P}$ are $\geq 1$.  

18
2. \( P - I \) is non-negative definite.

3. \( P = I \), i.e., all spectral values are 1.

4. The variables \( X_1, X_2, \ldots X_p \) are pairwise independent.

5. The spaces \( H(X_1), H(X_2), \ldots H(X_p) \) are orthogonal.

Proof:

(1 \( \Rightarrow \) 2) If all the spectral values of \( P \) are \( \geq 1 \), then all the spectral values of \( P - I \) are non-negative, or equivalently \( P - I \) is non-negative.

(2 \( \Rightarrow \) 4) If \( P - I \) is non-negative, then for \( \Phi^{ij} = (0, \phi_i, \ldots 0, \phi_j \ldots 0) \in H \)

\[
\langle \Phi^{ij}, (P - I)\Phi^{ij} \rangle = \langle \Phi^{ij}, P\Phi^{ij} \rangle - \|\Phi^{ij}\|^2 \geq 0
\]

\[\Rightarrow 0 \leq \text{var}(\phi_i + \phi_j) - 1 \quad \text{by lemma 3.1 and } \|\Phi\|^2 = 1\]

\[= \text{var} \phi_i + \text{var} \phi_j + 2\text{cov}(\phi_i, \phi_j) - 1\]

\[= 2 \text{cov} (\phi_i, \phi_j) \quad \text{since } \|\Phi^{ij}\| = \text{var} \phi_i + \text{var} \phi_j = 1.\]

Replacing \( \phi_j \) by \( -\phi_j \) in the above we arrive at the conclusion that \( \text{cov}(\phi_i, \phi_j) = 0 \quad \forall \phi_i, \phi_j \), hence \( X_i \) and \( X_j \) are independent \( \forall i \neq j \).

(4 \( \Rightarrow \) 5) Clear, since \( \text{cov}(\phi_i, \phi_j) = \langle \phi_i, \phi_j \rangle \).

(5 \( \Rightarrow \) 3) \( H(X_i) \perp H(X_j) \Rightarrow P_{ij} = 0 \forall i \neq j \). Hence \( P = I \).

(3 \( \Rightarrow \) 1) Trivially.

Note that orthogonality of spaces \( H(X_1), H(X_2), \ldots H(X_p) \) is equivalent to pairwise independence only, and full independence of \( X_1, \ldots X_p \) does not follow: from \( H(X_1) \perp H(X_2) \) and \( H(X_1) \perp H(X_3) \) we only have \( H(X_1) \perp H(X_2) \oplus H(X_3) \) whereas independence of \( X_1 \) from \( (X_2, X_3) \) is equivalent to \( H(X_1) \perp H(X_2, X_3) \), the latter denoting the centered \( L_2 \) functions which depend only on \( X_2 \) and \( X_3 \).

It is also important that the smallest eigenvalue be clearly separate from 1, since only then will we be able to reliably estimate the functions. This is a consequence of the clustering of the eigenvalues at 1 and the discontinuous behaviour of the smallest eigenfunction as a function of \( P \). If the operator \( P \) has two single eigenvalues close to each other, estimates of the corresponding eigenfunctions will be inherently ill-determined since finite
sample perturbations will be enough to cause dramatic fluctuation between the orthogonal vectors of eigenfunctions. When the smallest eigenvalue is close to 1, there are many near optimal solutions so the estimates will be unstable.

To summarize, only when the smallest eigenvalue is clearly below and separated from 1 are we likely to have an additive principal component of practical interest. Note that we can easily obtain the eigenvalue associated with the smallest principal component since

$$\text{var} \sum \phi_i = \langle \Phi, P\Phi \rangle = \langle \Phi, \lambda \Phi \rangle = \lambda.$$
4 Further Additive Principal Components

4.1 Definition and Algorithm

We have established the characterization of the smallest additive principal component as a smallest eigenfunction of $\mathbf{P}$. This suggests exploring the possible use and interpretation of eigenfunctions associated with other eigenvalues of $\mathbf{P}$. We start with the linear case, since that is well understood.

In section 3.2 we pointed out that the smallest principal component $\mathbf{a}_1$ minimizes $\text{var}(\mathbf{X} \cdot \mathbf{a})$, or equivalently defines through $l(x) = \mathbf{a}_1 \cdot x = 0$ a linear manifold $L^{(1)}$ minimizing the expected squared distance from the observations to any linear manifold of co-dimension 1.

The second smallest principal component $\mathbf{a}_2$ is defined as the unit vector minimizing $\text{var}(\mathbf{X} \cdot \mathbf{a})$ subject to $\mathbf{a}_1 \perp \mathbf{a}_2$, or equivalently subject to $\sum \text{cov}(a_{1i}X_i, a_{2i}X_i) = 0$. The vector $\mathbf{a}_2$ defines a linear function $l_2$ and a corresponding manifold $L^{(2)}$ of co-dimension 1 through $l_2(x) = 0$. Together the two functions $l_1$ and $l_2$ define a linear manifold $L^{(12)}$ of co-dimension 2, which is the intersection of $L^{(1)}$ and $L^{(2)}$. This is the manifold that, among all manifolds of co-dimension 2, has the smallest expected squared distance from the observations.

We define the second smallest additive principal component by extending the variance criterion of the linear definition. We find the vector of transformations minimizing the variance of the corresponding sum of transformed variables, under the constraint that it is orthogonal to the smallest additive principal component. Orthogonality of the vectors of functions is defined with respect to the inner product on the Hilbert space $H$. 

21
Definition

The second smallest additive principal component of $X$ is the random variable $Y = \sum \phi_i (X_i)$ minimizing over $\Phi' \in H$ the criterion

$$\text{var} \sum_j \phi'_j (X_j) = \langle \Phi', P \Phi' \rangle$$

subject to

$$\langle \Phi', \Phi(1) \rangle = \sum_i \langle \phi_i', \phi(1)_i \rangle = 0$$

and

$$||\Phi'||^2 = \sum \text{var} \phi'_j (X_j) = 1$$

Linear principal components are uncorrelated, because the score vectors are orthogonal eigenvectors of $\Sigma$, i.e., the dispersion matrix of the principal components, $y = Ax$, is diagonal:

$$\text{var} y = \text{var} Ax = A' \Sigma A = \text{diag}(\lambda_1, \ldots, \lambda_p)$$

where $A'A = I$.

For additive principal components the same result holds true: the additive principal components simultaneously diagonalize the quadratic forms $\langle \Phi, P \Phi \rangle$ and $||\Phi||^2$, consequently additive principal components belonging to two different eigenvalues are uncorrelated. Note that we then have

$$\text{var} (\sum \phi_{1i} + \sum \phi_{2i}) = \text{var} (\sum \phi_{1i}) + \text{var} (\sum \phi_{2i})$$

a familiar property of linear principal components.

In exact analogy with the smallest additive principal component, the second smallest additive principal component is an eigenfunction corresponding to the next smallest eigenvalue of $P$. This follows from the argument of section 3.4, since the second additive principal component $\Phi^{(2)} \in H$ minimizes the Rayleigh quotient among all $\Phi \perp \Phi^{(1)}$.

It is easy to generalize this idea and define a sequence of additive principal components, each orthogonal to all the preceding ones. The $i^{th}$ additive principal component corresponds to an eigenfunction of $P$ belonging to the $i^{th}$ largest eigenvalue (where we list each eigenvalue according to its multiplicity).

We can investigate the geometric structure implied by further additive principal components by analogy to the linear case. If the covariance matrix

\[22\]
of \( X \) has two vanishing eigenvalues, the observations lie in a linear manifold of co-dimension 2, i.e., there exist \( a_1 \) and \( a_2 \) with \( a_1^T a_2 = 0 \) such that \( \sum a_{1i} X_i = 0 \) and \( \sum a_{2i} X_i = 0 \). If there are two eigenvalues "close" to zero, then the observations have a strong geometric configuration described by the linear manifold.

If the two smallest additive principal components of the variables have small variances, then there are two sets of functions for which

\[
\sum \phi_{1i}(X_i) \approx 0 \text{ and } \sum \phi_{2i}(X_i) \approx 0, \text{ with } \sum \langle \phi_{1i}, \phi_{2i} \rangle = 0,
\]

The observations lie close to the additive manifold defined by the corresponding pair of implicit equations: \( \sum \phi_{1i}(x_i) = 0 \) and \( \sum \phi_{2i}(x_i) = 0 \) define an additive manifold of co-dimension 2 in \( p \)-space. Enabling detection of this high dimensional structure in data is an important aspect of additive principal components, particularly since such structure is impossible to detect visually once more than three or four variables are involved.

An obvious method for finding the \( i^{th} \) component is to use the same algorithm as before, simply orthogonalizing the starting vector, \( \Phi' \), with respect to all the previous principal components.

**Algorithm**

1. **Initialize** \( \Phi' = \Phi' - \langle \Phi', \Phi(1) \rangle \Phi(1) - \ldots - \langle \Phi', \Phi(i-1) \rangle \Phi(i-1) \)
2. **Repeat** for \( k = 1, 2 \ldots \)
   1. **For** \( j = 1, 2 \ldots p \)
      \[
      \phi_j^* = p \phi_j^{k-1} - E X_j \left( \sum_{i=1}^p \phi_i^{k-1} \right)
      \]
   2. **Set** \( \Phi^k := (\sum m \text{ var } \phi_m^*)^{-1/2}(\phi_1^*, \ldots, \phi_p^*) \).
   3. **Until** \( \text{var } \sum \Phi^k \text{ converges.} \)

The algorithm will converge to the \( i^{th} \) largest eigenfunction of \( pI - P \) or equivalently the \( i^{th} \) smallest eigenfunction of \( P \). Since this eigenfunction minimizes \( \text{var } \sum \phi_i(X_i) \) subject to \( \sum \text{var } \phi_i(X_i) = 1 \) and \( \langle \Phi, \Phi(1) \rangle = 0, \)
\( j = 1 \ldots i - 1 \), it is the \( i^{th} \) smallest additive principal component. The associated eigenvalue is given by \( \lambda_i = \text{var} \sum \phi_{ij} \).

In an actual implementation of the algorithm for computing the \( i^{th} \) additive principal component, it will not be sufficient to orthogonalize only once with respect to the previous components. In the finite sample version, conditional expectations are estimated by a scatterplot smoother. Typically the smoother is not a projection operator — it may not even satisfy linearity or symmetry — in which case exact orthogonality is not preserved in the inner loop. Even if the smoother was a projection, rounding error would still reintroduce components in the orthogonal space. To ensure convergence to an orthogonal solution, the new estimate will need to be reorthogonalised in every pass through the outer loop.

A suitably modified version of the above algorithm has been implemented, using the supersmoker [4] to estimate the conditional expectation operators. Our experience with implementation of the algorithm with examples from various situations of interest will be the subject of a forthcoming report.

### 4.2 Use and Interpretation

There are several ways in which further additive components enable a heuristic assessment of the significance of the structure detected in the data.

Suppose we have estimated the first and second smallest additive components of \( \mathbf{X} \). If the two eigenvalues are well separated, the estimated eigenfunctions are both unique and stable. If, on the other hand, the eigenvalues are similar, we are close to the case where the smallest eigenvalue has multiplicity greater than one. This has a bearing on the uniqueness of the eigenfunctions.

Consider the linear case when the eigenspace of the smallest eigenvalue is two dimensional. Then a minimal eigenvector is no longer uniquely defined, since any vector lying in the eigenplane will suffice — only the plane spanned by two orthogonal eigenvectors is unique. Clearly, the eigenvectors cannot separately provide information about specific dependencies among the variables, since their realization is arbitrary. Analogously, when the variables exactly satisfy two additive constraints, so that the small-
est eigenspace is two dimensional, the eigenfunctions are not unique. An arbitrary pair of orthogonal eigenfunctions are clearly not separately informative, although jointly they uniquely determine the associated eigenspace. In a finite sample setting, we come near to this situation of non-uniqueness if the two smallest eigenvalues are close. Slight changes in the data sample can result in completely different realizations of the smallest additive component, since the eigenfunctions are intrinsically unstable. The estimates will depend on starting values for the algorithm and the ordering of the variables, among other things.

In this situation, then, it is not reasonable to try to infer the form of dependencies between the variables based on only the smallest additive component. It may be fruitful to consider the subspace spanned by the two smallest principal components, and while it would be appealing to be able to choose the simplest versions of the equations, or those most readily interpreted, it is not clear how one might do this algorithmically.

Another use for the second smallest and further additive components is to provide comparison values for the relative size of the smallest eigenvalue. In the same spirit as examining the profile of eigenvalues in a classical principal component analysis, comparing the estimated eigenvalues will enable evaluation of whether the observations lie close to an additive manifold of co-dimension 1 or 2 or neither.

The smallest additive principal component of $X$ is exactly the smallest linear principal component of the transformed and restandardized variables

$\Phi^*_1(X) = \left( \frac{\phi_1(X_1)}{\|\phi_1(X_1)\|}, \frac{\phi_2(X_2)}{\|\phi_2(X_2)\|}, \ldots, \frac{\phi_p(X_p)}{\|\phi_p(X_p)\|} \right)$.

This smallest eigenvector has the associated vector of scores $a = (\|\phi_1(X_1)\|, \ldots, \|\phi_p(X_p)\|)$. Once we consider more than one additive principal component, this duality no longer exists, as the transformations of the variables are different for each additive principal component. Moreover, the $i^{th}$ additive principal component does not correspond to the smallest linear principal component of $\Phi^*_i(X)$, since the side conditions are not enforced. This is true even in the case where the smallest additive principal component is linear.

An alternative approach, where it is still meaningful to transform then

25
use a linear analysis, is to find a set of transforms for which the transformed variables are closest to a linear manifold of co-dimension \( k \geq 1 \), instead of co-dimension 1. A modification of a criterion considered by Koyak[6] gives a suitable formulation of the problem: minimize the sum of the first \( k \) eigenvalues of the correlation matrix of the transformed variables. Since this is equivalent to maximizing the first \( p-k \) eigenvalues of this correlation matrix, which is exactly the problem that Koyak considers, interested readers are referred to [6].
References


