*Journal of Applied Statistics*
*Vol. 32, No. 9, 969–987, November 2005*

# Interpretable Dimension Reduction

HUGH A. CHIPMAN* & HONG GU**

*Department of Mathematics and Statistics, Acadia University, Wolfville, Canada,
**Department of Mathematics and Statistics, Dalhousie University, Halifax, Canada

ABSTRACT *The analysis of high-dimensional data often begins with the identification of lower dimensional subspaces. Principal component analysis is a dimension reduction technique that identifies linear combinations of variables along which most variation occurs or which best "reconstruct" the original variables. For example, many temperature readings may be taken in a production process when in fact there are just a few underlying variables driving the process. A problem with principal components is that the linear combinations can seem quite arbitrary. To make them more interpretable, we introduce two classes of constraints. In the first, coefficients are constrained to equal a small number of values (homogeneity constraint). The second constraint attempts to set as many coefficients to zero as possible (sparsity constraint). The resultant interpretable directions are either calculated to be close to the original principal component directions, or calculated in a stepwise manner that may make the components more orthogonal. A small dataset on characteristics of cars is used to introduce the techniques. A more substantial data mining application is also given, illustrating the ability of the procedure to scale to a very large number of variables.*

KEY WORDS: Principal component, interpretable, homogeneity, sparsity, stepwise algorithm, dimension reduction, data mining

## Introduction

Principal component analysis can be an effective tool for reducing dimensionality in problems where many variables are measured. This is especially true when there are strong linear relationships among the variables. By identifying linear combinations of the original variables that capture the most variation, the data are reduced. Quite often these new variables (the principal components) can be interpreted. For example, in problems where many physical dimensions of objects are measured, the first direction may be quite close to a sum (or equivalently an average) of the variables, and consequently correspond to overall size. Subsequent directions may identify contrasts between variables in which there is substantial variation.

To interpret the principal components, one must filter through the coefficients (or loadings) of the linear combinations and identify patterns. This can be quite challenging in problems with many variables, which is precisely when principal components may be most helpful. This paper introduces several methods to simplify the linear combinations, making them more interpretable.

*Correspondence Address:* Hugh Chipman, Department of Mathematics and Statistics, Acadia University, Wolfville, NS, B4P 2R6, Canada. Email: hugh.chipman@acadiau.ca

0266-4763 Print/1360-0532 Online/05/090969–19 © 2005 Taylor & Francis
DOI: 10.1080/02664760500168648

To illustrate the approach, consider data on 91 cars (Lock, 1993), with 17 variables such as price, fuel economy, weight, engine size, etc. Variable names are given in Table 1, and a more detailed description of the data and some preliminary processing is given later. The coefficients of the first five and last principal components based on the correlation matrix are given in Table 1. The coefficients in the first component are quite similar in absolute magnitude, ranging from 0.141 to 0.295, with most values in the range (0.20, 0.28). This similarity motivates the *homogeneity* constraint, in which each coefficient is proportional to either $\pm 1$ or 0. In the third section we show that for these data, the best first homogeneous direction has no loadings equal to zero, giving an interpretation corresponding to vehicle size. In some cases we may want to constrain the weights of a homogeneous component to sum to 0, yielding *contrast* constraints.

The second principal component has several coefficients near zero, while others take a wide range of values. To aid interpretation, we may want to make this direction *sparse*, by setting several of the coefficients to zero. If few enough of the coefficients are non-zero, they may still be interpretable without homogeneity restrictions mentioned above.

Other papers have considered methods for simplifying principal components. Hausman (1982) restricts the possible loading values to be proportional to 0 or $\pm 1$, while Vines (2000) constrains loadings to be proportional to integers. Both approaches are similar to our homogeneity and contrast constraints. Groups of principal components can also be rotated, explaining the same variation, but being more interpretable. See for example discussion and references in Jolliffe (1989). Jolliffe & Uddin (2000) optimize a penalized variance function, shrinking loadings towards zero. Jolliffe *et al.* (2003) use a variation of the LASSO (Tibshirani, 1996) to shrink loadings, possibly generating some zero values. The sparsity constraints introduced in the third section of this paper are similar to these approaches. Additional details of these methods and a comparison with the proposed approach are given in the sixth section.

**Table 1.** Loadings of the first five and last principal components, based on correlation matrix of the cars data

| Variable | PC 1 | PC 2 | PC 3 | PC 4 | PC 5 | Last PC |
|---|---|---|---|---|---|---|
| Min price | 0.230 | −0.376 | −0.118 | −0.154 | −0.211 | 0.363 |
| Price | 0.220 | −0.421 | −0.131 | −0.114 | −0.243 | −0.808 |
| Max price | 0.203 | −0.439 | −0.136 | −0.077 | −0.258 | 0.465 |
| City MPG | −0.265 | 0.002 | −0.103 | −0.450 | 0.089 | 0.001 |
| Highway MPG | −0.247 | 0.013 | −0.005 | −0.611 | 0.108 | 0 |
| Engine size | 0.282 | 0.050 | 0.184 | −0.202 | −0.036 | −0.003 |
| Horsepower | 0.243 | −0.289 | 0.190 | −0.005 | 0.343 | 0.001 |
| RPM | −0.141 | −0.411 | −0.149 | 0.140 | 0.754 | −0.001 |
| Rev./mile | −0.241 | −0.135 | −0.344 | 0.126 | −0.013 | −0.001 |
| Fuel tank capacity | 0.273 | 0.004 | −0.064 | 0.214 | 0.113 | 0.001 |
| Passengers | 0.192 | 0.321 | −0.461 | 0.231 | 0.037 | −0.001 |
| Length | 0.263 | 0.073 | 0.058 | −0.295 | 0.153 | 0 |
| Wheelbase | 0.275 | 0.108 | −0.172 | −0.130 | 0.094 | 0 |
| Width | 0.271 | 0.163 | 0.189 | −0.105 | 0.152 | −0.001 |
| Turn circle | 0.247 | 0.175 | 0.196 | −0.117 | 0.189 | 0.001 |
| Rear seat room | 0.178 | 0.195 | −0.637 | −0.260 | 0.108 | 0.001 |
| Weight | 0.295 | 0.011 | 0.017 | 0.097 | 0.081 | 0 |

To set the stage for our approach, we comment briefly on a few important issues here. In all the articles mentioned above, the criterion used to obtain modified components or compare performance with the original principal components is the variance captured by the new components. This criterion can be misleading in some applications. The proportion of variance explained by the first few principal components can be made arbitrarily large by the inclusion of highly correlated variables, while information may be lost in other directions. This can be a problem when regression follows. We remedy this by also considering the "reconstruction error" in estimating the original data with some of the principal components as predictors.

Principal components have other desirable properties in addition to maximization of variance explained and minimization of reconstruction error. We propose quantities that measure these properties, and the extent to which they are retained by interpretable components.

Another important issue is computation. Our approach scales well to very large datasets, a problem for some of the previous approaches.

In the next section, we first address four important properties of principal components and their quantification. Three different constrained principal components and methods for their calculation are introduced and illustrated using the cars data in the section after. The fourth section discusses an alternative stepwise method to calculate the constrained components. In the fifth section we illustrate the method using a data mining application. Comparisons with related work and a discussion of other interesting problems are given in the sixth section.

## Quantifying Optimal Properties of Principal Components

Interpretable components are not principal components, but they can often be quite similar. This section reviews four optimal properties of principal components, and develops methods to quantify these properties for interpretable components.

### *Properties of Principal Components*

Given an $n \times p$ data matrix $X$ consisting of $p$ continuous measurements on each of $n$ objects, we seek a $p \times q$ projection matrix $\Gamma$ with columns $\gamma_i$, $i = 1, \ldots, q$. This matrix is such that $X\Gamma$, the $q$-dimensional projection of $X$ onto $\Gamma$ captures most of the variance in the original $p$ dimensional space. One solution to this problem is principal components, which identifies a solution $\Gamma$ with the following properties:

(a) The columns of $\Gamma$ are orthogonal.
(b) The projected data $X\Gamma$ are uncorrelated.
(c) The variance of $X\gamma_i$ is maximized subject to (a). That is, $\gamma_1$ is chosen to maximize $\text{Var}(X\gamma_1)$, $\gamma_2$ is chosen to maximize $\text{Var}(X\gamma_2)$ among all vectors orthogonal to $\gamma_1$, etc.
(d) Suppose that the original data are reconstructed using a projection of the data onto a $k$-dimensional subspace. The subspace defined by $\gamma_1, \ldots, \gamma_k$ will minimize the reconstruction error. That is, the variance of orthogonal distances from the original data to the first $k$ principal components is minimized over all possible $k$-dimensional subspaces.

The issue of scaling the columns of $X$ is important. We assume that a suitable scaling has been found. In the examples, the variables are scaled to have unit variance, which is equivalent to using the correlation matrix of $X$ to identify the principal components. Throughout this paper, we use $\Gamma = (\gamma_1, \ldots, \gamma_p)$ to represent the principal component directions, and $\Sigma$ to represent the covariance matrix of the appropriately scaled $X$. The variance of the $i$th principal component $X\gamma_i$ is $\lambda_i$, the $i$th eigenvalue of $\Sigma$.

The principal components $\{X\gamma_i\}_{i=1}^{q}$ are then a smaller group of variables representing most of the information in the original data. Since the $\gamma_i'$s are optimal for properties (a)–(d), they are usually real-valued and difficult to interpret. In the next section we explore methods that will find directions $\alpha_i$ that are close to the original $\gamma_i$, but more interpretable. Before giving details of the interpretable components, we introduce methods to quantify properties (a)–(d).

### *Quantifying Closeness to the Principal Components*

In (a), the linear combinations are orthogonal; we calculate the angle between our interpretable directions. For any two directions $\alpha_i$ and $\alpha_j$ this angle is $\arccos(\alpha_i'\alpha_j)$. In (b) the projected data are uncorrelated; thus we calculate the correlation of the data projected onto the interpretable directions. If only a few directions are to be retained, it may be of interest to look at the correlations of the data only in this projected space.

Property (c) requires that the variance of the projected data be maximal, subject to orthogonality of the $\gamma_i'$. Orthogonality of $\Gamma$ ensures that the variances of the projections of the data will sum to the total variance of the original data. Since the $\alpha_i'$s are not orthogonal, the variance of $X\alpha_i$ is not comparable directly with that of $X\gamma_i$. We suggest two ways to measure the variation explained, based on the expression

$$\alpha_i = a_{i1}\gamma_1 + a_{i2}\gamma_2 + \cdots + a_{ip}\gamma_p$$

where $a_{ij}$ is the projection of $\alpha_i$ on $\gamma_j$. Since $\mathrm{Var}(X\alpha_i) = a_{i1}^2\lambda_1 + \cdots + a_{ip}^2\lambda_p$, $a_{ii}^2\lambda_i$ provides a measure of the variance captured by $\alpha_i$ in the direction of $\gamma_i$. We thus compare $\mathrm{Var}(X\gamma_i) = \lambda_i$ with $\mathrm{Var}(Xa_{ii}\gamma_i) = a_{ii}^2\lambda_i$. Since the interpretable component $\alpha_i$ will often be close to $\gamma_i$, $a_{ii}$ can be close to 1.

The variation associated with $a_{ii}\gamma_i$ underestimates the variation associated with $\gamma_i$. However, each sparse component actually captures the variance in every direction of $\Gamma$. Instead of using the variance of $\alpha_i$ in the direction of $\gamma_i$ ($a_{ii}^2\lambda_i$), the total variance of the first $k$ interpretable components $\alpha_i, \ldots, \alpha_k$ in the direction of $\gamma_i$ could be calculated. That is, use $\lambda_i \sum_{l=1}^{k} a_{li}^2$ to summarize variance captured by $\alpha_1, \ldots, \alpha_k$ in the direction $\gamma_i$. This sum of variances could be compared to $\mathrm{Var}(X\gamma_i) = \lambda_i$, A potential problem with this approach is that it can over-estimate the variance associated with $\gamma_i$ because of the non-orthogonality of the interpretable components.

For property (d), the principal components will give minimal reconstruction error (see for example, Rao 1965). This reconstruction error can be quantified as follows. Let $\tilde{X}$ be the $n \times p$ data matrix with column means subtracted, and $A$ a $p \times k$ projection matrix with columns representing linear combinations. The columns of $A$ could be from principal components or interpretable components. The projection of $\tilde{X}$ onto $A$ is $T = \tilde{X}A$. To reconstruct $\tilde{X}$ using the projected data $T$, we calculate

$$\hat{X} = T(T'T)^{-1}T'\tilde{X} = T\hat{P} \tag{1}$$

which may be thought of as the regression of the original data $\tilde{X}$ on the projected data $T$. The (matrix) difference between original and reconstructed data,

$$F = \tilde{X} - \hat{X} = \tilde{X} - T\hat{P} \tag{2}$$

gives the reconstruction errors. The trace of the variance of this difference,

$$\text{trace}(\text{Var}(F)) = \text{trace}(F'F/n) \tag{3}$$

gives the unexplained variance when reconstructing the original data using the $k$ directions in $A$. We shall use equation (3) to quantify property (d). If $A$ contains the first $k$ principal component directions, then (3) simplifies to the total variance minus the variance explained by the first $k$ components,

$$\text{trace}(\text{Var}(F)) = \text{trace}(\Sigma) - \sum_{i=1}^{k} \lambda_i = \sum_{i=k+1}^{p} \lambda_i$$

The principal component directions $\Gamma$ can be obtained by minimizing equation (3) subject to the condition that the columns of $T$ are orthogonal. By increasing the interpretability of $X\gamma_i$, the absolute orthogonality of columns of $T$ is sacrificed. In many cases, $T$ may not be far from orthogonal, and the reconstruction error may be close to that of the same number of principal components.

It is interesting to note that in equation (2) the principal component directions can be calculated one at a time, each time removing the reconstruction of the components already identified. This stepwise method will be applied to interpretable components in the fourth section.

## Constrained Principal Components

This section introduces methods for identifying interpretable directions $\alpha_i$, $i = 1, \ldots, p$. By interpretable we mean either many coefficients will be zero, eliminating many variables from a component, or the coefficients in the components only take a few distinct values. Both cases, either a complex combination of a few variables or a simple combination of possibly more variables, can be easier to interpret than the original component.

### *Homogeneity Constraints*

The $i$th direction $\alpha_i$ could be made interpretable if its elements took very few distinct values, say 0 or $\pm c$ for $c$ such that $\alpha_i'\alpha_i = 1$. This *homogeneity* constraint corresponds to a direction that is an average of some of the variables, or their negative values.

There are $3^p$ possible values for $\alpha_i$. To find the best $\alpha_i$, we minimize $\arccos(\gamma_i'\alpha_i)$, the angle to the $i$th principal component direction, or equivalently maximize the inner product $\gamma_i'\alpha_i$. Note that the maximization is over $-c$, 0, $c$ values, not all real values. The search algorithm is straightforward. Among all $\alpha_i$ with $k$ non-zero elements, identify the $k$ elements of $\gamma_i$ with largest absolute values. Set the corresponding elements of $\alpha_i$ to $\pm 1/\sqrt{k}$, matching signs with the $k$ elements of $\gamma_i$. All other elements of $\alpha_i$ are 0, and $\alpha_i'\alpha_i = 1$. The $\alpha_i$ closest to $\gamma_i$ is then identified by repeating this procedure for $k = 1, 2, \ldots, p$. For example if $\gamma_1 = (0.41, -0.03, -0.42, 0.81)$ the vectors

$$(0, 0, 0, 1), \quad (0, 0, -1, 1)/\sqrt{2}, \quad (1, 0, -1, 1)/\sqrt{3} \quad \text{and} \quad (1, -1, -1, 1)/2$$

would be the closest to $\gamma_1$ among $\alpha$ with $k = 1, \ldots, 4$ non-zero elements. In this case, $(1, 0, -1, 1)/\sqrt{3}$ is closest to $\gamma_1$ with an angle of 18.8 degrees. As $k$ increases, the angle will not necessarily decrease.

An alternative constraint would set $\alpha_{ij} \in \{-c_1, 0, c_2\}$ such that $\sum_{j=1}^{p} \alpha_{ij} = 0$. That is, the linear combination is a difference of the average of one set of variables and the average of another set of variables, called a *contrast*. Constants $c_1$ and $c_2$ are chosen so that $\alpha_i' \alpha_i = 1$. For example, in the cars data a minimum, typical, and maximum price are given. The typical price is just the average of the other two. If the three prices are scaled by the same factor, the last principal component is quite close to

$$0.41(\texttt{Min price}) - 0.82(\texttt{Price}) + 0.41(\texttt{Max price}).$$

This corresponds to a difference between the `Price` and the average of `Min price` and `Max price`. Since the variance of the last principal component is almost zero (non-zero due to rounding error), this indicates that `Price` is simply the average of the other two variables. In this case, the principal component needs little simplification since all other elements are at least an order of magnitude smaller. In situations where noise levels are higher, such contrast constraints may focus attention on simplified directions.

The algorithm for identifying contrast directions is very similar to that for homogeneous directions. One minor difference is that the coefficient vector must contain at least one element of each sign. In the modified algorithm, the loadings corresponding to the largest positive and negative coefficients of the original linear combination are never set to zero, and other components are selected by absolute magnitude.

### *Homogeneity Constraints Applied to the Cars Data*

The original data contained 27 variables, including manufacturer, model, and other categorical variables. All categorical variables were removed, along with the variable `luggage room`, which has 11 missing values. Two cars with missing values of `rear seat room` were also removed, leaving 17 variables and 91 observations. The variables were scaled to have unit variance, and the principal components calculated. The first five homogeneous and contrast directions are given in Table 2, with the angles to the corresponding principal component directions.

The first principal component is better summarized by a homogeneous direction than by a contrast, since it has a much smaller angle. Other components are less clear. Before interpreting the components, we examine properties (a)–(d). There are $\binom{17}{2} = 136$ angles between the 17 interpretable directions for both the homogeneous and contrast components. The distribution of these angles is summarized in Figure 1. Most angles are between 70 and 110 degrees, indicating relatively orthogonal directions.

Criterion (b), the correlation between projected variables, is given by boxplots in Figure 2, along with correlations in the original dataset. The original data has many more correlations above 0.50. Although the data projected onto the interpretable directions are not uncorrelated (as the principal components projections would be), the correlations in the data are greatly reduced.

For criterion (c), we compared the variance of the data projected on the homogeneous and contrast directions with that of principal components. Results were quite similar to later comparisons made for the sparse components, and are not shown.

Criterion (d), the unexplained reconstruction variance, equation (3), is given in Figure 3. Both constrained methods have reconstruction error only slightly larger than the principal components. For comparison, we also include the average reconstruction error when a random orthogonal transformation is used (based on 50 random simulations of uniform angular rotations of the original basis). In this case, the highly correlated nature of the

**Table 2.** Cars data: interpretable directions, and angles with corresponding principal component directions. The normalizing constant for the homogeneous components is omitted

| | Homogeneous | | | | | Contrast | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Min price | 1 | −1 | 0 | 0 | −1 | 0.13 | −0.30 | 0 | −0.23 | −0.37 |
| Price | 1 | −1 | −1 | 0 | −1 | 0.13 | −0.30 | 0 | 0 | −0.37 |
| Max price | 1 | −1 | −1 | 0 | −1 | 0.13 | −0.30 | 0 | 0 | −0.37 |
| City MPG | −1 | 0 | 0 | −1 | 0 | −0.44 | 0 | 0 | −0.23 | 0 |
| Highway MPG | −1 | 0 | 0 | −1 | 0 | −0.44 | 0 | 0 | −0.23 | 0 |
| Engine size | 1 | 0 | 1 | −1 | 0 | 0.13 | 0.26 | 0.33 | −0.23 | 0 |
| Horsepower | 1 | −1 | 1 | 0 | 1 | 0.13 | −0.30 | 0.33 | 0 | 0.55 |
| RPM | −1 | −1 | −1 | 0 | 1 | −0.44 | −0.30 | 0 | 0.40 | 0.55 |
| Rev/mile | −1 | 0 | −1 | 0 | 0 | −0.44 | −0.30 | −0.44 | 0.40 | 0 |
| Fuel tank capacity | 1 | 0 | 0 | 1 | 0 | 0.13 | 0 | 0 | 0.40 | 0 |
| Passengers | 1 | 1 | −1 | 1 | 0 | 0.13 | 0.26 | −0.44 | 0.40 | 0 |
| Length | 1 | 0 | 0 | −1 | 0 | 0.13 | 0.26 | 0 | −0.23 | 0 |
| Wheelbase | 1 | 0 | −1 | 0 | 0 | 0.13 | 0.26 | 0 | −0.23 | 0 |
| Width | 1 | 1 | 1 | 0 | 0 | 0.13 | 0.26 | 0.33 | 0 | 0 |
| Turn circle | 1 | 1 | 1 | 0 | 1 | 0.13 | 0.26 | 0.33 | 0 | 0 |
| Rear seat room | 1 | 1 | −1 | −1 | 0 | 0.13 | 0.26 | −0.44 | −0.23 | 0 |
| Weight | 1 | 0 | 0 | 0 | 0 | 0.13 | 0 | 0 | 0 | 0 |
| Angle (deg.) | 10 | 22 | 33 | 31 | 35 | 35 | 26 | 29 | 40 | 31 |

data means that even if directions are selected at random, the data can be reconstructed with reasonable accuracy.

In this problem, contrast and especially homogeneous components perform well. We now give some possible interpretations of the simplified components. The first component can be interpreted as the size of the car, with all the positive weights positively

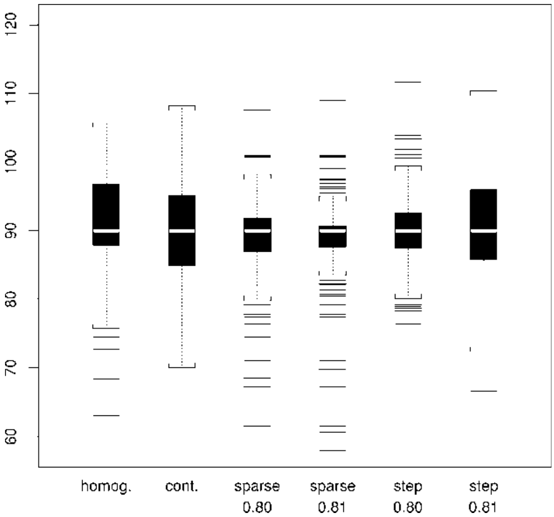

**Figure 1.** Boxplots of angles between components, cars example. Each boxplot represents angles for components from a specific method

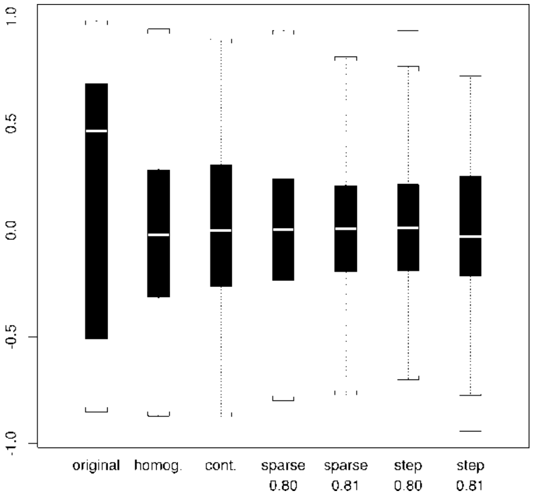

**Figure 2.** Correlations of projected data, Cars example. The first boxplot represents the original data, and the others correspond to various methods described in Sections 3 and 4

related to the size of the car and negative weights negatively related to the size of the car. The second homogeneous component has a negative relation with the price variables, horsepower, RPM and a positive relation with passengers, width, turn circle and rear seat room. It can be interpreted as a contrast between cheap, weak and large cars (e.g. minivans) versus the expensive, powerful and smaller cars (e.g. sports cars). Interpretation of other homogeneous components and contrast components is possible, but not given here.

### *Sparsity Constraints*

Very often a principal component direction may have some near-zero coefficients. Setting them to zero may not cause a large change in variance explained by this component or the total variance explained by the first several components. Such a component with many zero elements will be called a *sparse* component.

As in the homogeneous and contrast cases, to find sparse components that nearly obey the four properties of principal components, we minimize the angle between the sparse component direction and the corresponding principal component direction. The angle is minimized when no elements are zero, so we add a penalty term, such as the number of nonzero elements. Thus we minimize

$$C1 = \theta/(\pi/2) + \eta k/p, \tag{4}$$

over $\alpha_i$ and $k$, the number of non-zero elements in $\alpha_i$. In equation (4), $\theta$ is the angle between sparse component $\alpha_i$ and principal component $\gamma_i$, and $\eta$ is a tuning parameter. As $\eta$ increases, the component minimizing $C1$ becomes more sparse. Criterion $C1$ may be thought of as the percent angular error (since the angle to the principal component will not exceed $\pi/2$) plus a constant times the percentage of non-zero loadings.

Other criteria are possible. For example, one could maximize $C2 = (p-k)(\cos\theta)^{\eta}$. The $p-k$ term is large when most coefficients are zero. The $\cos\theta$ term is large when the interpretable direction is close to the principal component direction. Larger $\eta$ values

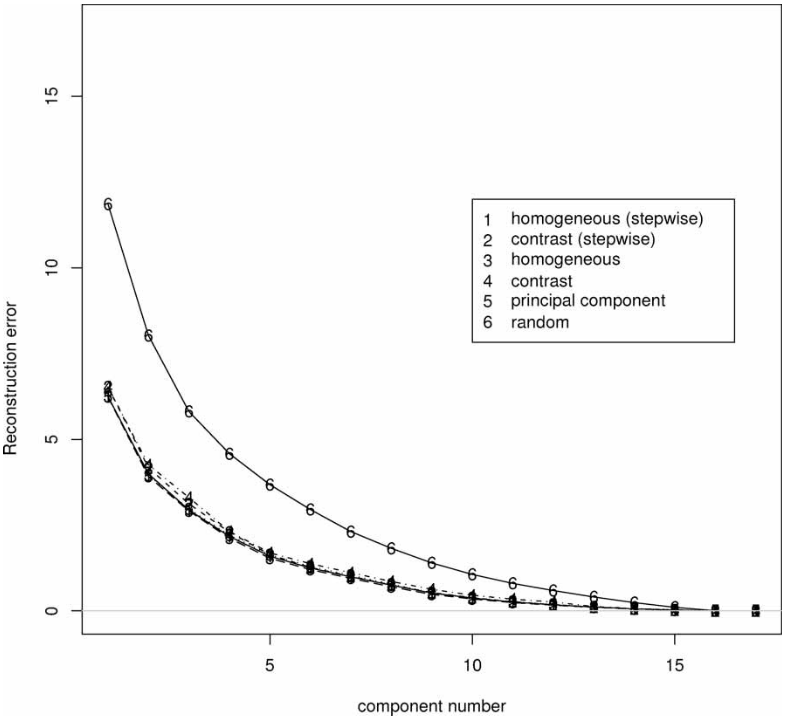

**Figure 3.** Reconstruction errors for homogeneous and contrast components in the cars dataset, compared to original principal components and reconstruction using random directions

cause the second term to shrink, leading to less sparse directions that are closer to the principal component direction. The $C1$ criterion could select the original principal component in some cases. Such a selection is impossible for $C2$. Otherwise these two criteria behave similarly.

The main goal of C1 and C2 is to provide a convenient way to index the possible $\alpha$. In practice, several values of the tuning constant $\eta$ could be tried.

The search for the sparse directions is similar to the homogeneous case. For fixed $k$, optimizing $C1$ or $C2$ is equivalent to minimizing $\theta$. The optimum will occur when the $k$ non-zero elements of $\alpha_i$ correspond to the $k$ coefficients of $\gamma_i$ with largest magnitude. Setting the $(p - k)$ elements of $\gamma_i$ with smallest absolute values to zero and renormalizing the resultant vector yields the sparse component $\alpha_i$ for fixed $k$. We loop over $k$ to find the best sparse component $\alpha_i$. Since the sparse components are real-valued, the identification of a reasonable scale of the $X$ columns will be less critical than with homogeneity constraints. Interpretation will still be aided by a reasonable scaling of $X$.

This approach amounts to setting the coefficients of the linear combination with smallest absolute values equal to zero. In some contexts, coefficient magnitude may not capture all available information about variable importance. Cadima & Jolliffe (1995) discuss this issue and present an alternative approach.

### *Sparsity Constraints Applied to the Cars Data*

We illustrate sparsity criterion $C1$ using the cars data. The first five sparse components when $\eta = 0.8$ and $\eta = 0.81$ are given in Table 3. With $\eta \leq 0.8$, the first sparse component is the principal component direction. The interpretation is simplified with $\eta = 0.81$, giving

**Table 3.** Cars data: sparse directions, and angles with corresponding principal component directions

| | $\eta = 0.8$ | | | | | $\eta = 0.81$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Min price | 0.23 | 0.40 | 0 | 0 | 0 | 0 | 0.40 | 0 | 0 | 0 |
| Price | 0.22 | 0.45 | 0 | 0 | 0 | 0 | 0.45 | 0 | 0 | 0 |
| Max price | 0.20 | 0.47 | 0 | 0 | −0.30 | 0 | 0.47 | 0 | 0 | 0 |
| City MPG | −0.27 | 0 | 0 | 0.55 | 0 | 0 | 0 | 0 | 0.55 | 0 |
| Highway MPG | −0.25 | 0 | 0 | 0.75 | 0 | 0 | 0 | 0 | 0.75 | 0 |
| Engine size | 0.28 | 0 | 0 | 0 | 0 | 0.45 | 0 | 0 | 0 | 0 |
| Horsepower | 0.24 | 0.31 | 0 | 0 | 0.39 | 0 | 0.31 | 0 | 0 | 0.41 |
| RPM | −0.14 | 0.44 | 0 | 0 | 0.87 | 0 | 0.44 | 0 | 0 | 0.91 |
| Rev./mile | −0.24 | 0 | 0.40 | 0 | 0 | 0 | 0 | 0.40 | 0 | 0 |
| Fuel tank capacity | 0.27 | 0 | 0 | 0 | 0 | 0.44 | 0 | 0 | 0 | 0 |
| Passengers | 0.19 | −0.34 | 0.54 | 0 | 0 | 0 | −0.35 | 0.54 | 0 | 0 |
| Length | 0.26 | 0 | 0 | 0.36 | 0 | 0 | 0 | 0 | 0.36 | 0 |
| Wheelbase | 0.27 | 0 | 0 | 0 | 0 | 0.44 | 0 | 0 | 0 | 0 |
| Width | 0.27 | 0 | 0 | 0 | 0 | 0.43 | 0 | 0 | 0 | 0 |
| Turn circle | 0.25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Rear seat room | 0.18 | 0 | 0.74 | 0 | 0 | 0 | 0 | 0.74 | 0 | 0 |
| Weight | 0.29 | 0 | 0 | 0 | 0 | 0.47 | 0 | 0 | 0 | 0 |
| Angle (deg.) | 0 | 21 | 31 | 35 | 30 | 51 | 21 | 31 | 35 | 34 |

non-zero coefficients in the first component corresponding directly to the car's size. Since car size is linearly related to all the variables, the first principal component explains much more variance than the sparse component.

We now examine criteria (a)–(d) for the sparse components. For criterion (a), Figure 1 plots the distribution of the angles between the sparse components $\alpha_i$ when $\eta = 0.8$ and $\eta = 0.81$. For $\eta = 0.8$, most of the angles are between 80 and 100 degrees. For $\eta = 0.81$, most of the angles are between 84 and 95 degrees. The median angle in both cases is 90 degrees.

For criterion (b), the correlations of the projected data are displayed in Figure 2. The projected data are much less correlated than the original data.

For criterion (c), the variances of the projected data for the first eight sparse components are given in Table 4, for $\eta = 0.5, 0.7, 0.8, 0.81, 0.91$. Eight components were chosen since they explain over 95% of the variation in the data. The variance in principal component direction $\gamma_i$ for sparse component $\alpha_i$ (i.e., $a_{ii}^2\lambda_i$) is given on the left of the table, for $i = 1,2,\ldots,8$. The variances of $X\gamma_1$ explained by the first sparse components obtained with $\eta = 0.81$ and $\eta = 0.9$ are much lower than $\lambda_1$. Other variances are reasonably close to variance of the corresponding principal component. As mentioned earlier, the sparse component captures variance in all the directions of $\Gamma$. For this reason, we give the total variance of $\alpha_1, \ldots, \alpha_8$ in the direction of $\gamma_i$ on the right part of the table, for $i = 1, \ldots, 10$. The ninth and tenth principal components are considered because of the non-orthogonality of the sparse components. This non-orthogonality means that variance in the direction of all principal components is captured. The sparse components corresponding to $\eta = 0.50, 0.70, 0.80$ over-explain the variance in the direction of $\gamma_1$, and the sparse components with $\eta = 0.81, 0.90$ under-explain the variance. In rows 2–8, the variance in the direction of the component is closer to the variance of the principal

**Table 4.** The variance of the cars data, projected onto sparse components, for $\eta = (0.5, 0.7, 0.8, 0.81, 0.9)$. Sparse component directions are denoted by $\alpha_i$, principal component directions by $\gamma_i$

| | | Variance of $X\alpha_i$ in $\gamma_i$ direction, for $i = 1, \ldots, 8$ (i.e., $a_{ii}^2 \mathrm{Var}(X\gamma_i)$) | | | | | Total variance of $X\alpha_1, \ldots, X\alpha_8$ in $\gamma_i$ direction for $i = 1, \ldots, 10$. (i.e. $\mathrm{Var}(X\gamma_i) \sum_{l=1}^{8} a_{li}^2$) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $i$ | Var(X$\gamma_i$) | 0.5 | 0.7 | 0.8 | 0.81 | 0.9 | 0.5 | 0.7 | 0.8 | 0.81 | 0.9 |
| 1 | 10.76 | 10.76 | 10.76 | 10.76 | 4.21 | 3.42 | 11.65 | 12.40 | 12.73 | 5.67 | 6.38 |
| 2 | 2.32 | 2.30 | 2.01 | 2.01 | 2.01 | 2.01 | 2.48 | 2.21 | 2.28 | 2.42 | 2.27 |
| 3 | 1.00 | 0.91 | 0.74 | 0.74 | 0.74 | 0.74 | 0.93 | 0.80 | 0.76 | 1.11 | 0.91 |
| 4 | 0.79 | 0.69 | 0.58 | 0.53 | 0.53 | 0.53 | 0.73 | 0.76 | 0.72 | 0.76 | 0.63 |
| 5 | 0.58 | 0.52 | 0.47 | 0.44 | 0.40 | 0.40 | 0.52 | 0.57 | 0.55 | 0.37 | 0.24 |
| 6 | 0.33 | 0.32 | 0.28 | 0.26 | 0.26 | 0.26 | 0.32 | 0.31 | 0.28 | 0.32 | 0.28 |
| 7 | 0.26 | 0.25 | 0.22 | 0.22 | 0.19 | 0.19 | 0.25 | 0.19 | 0.17 | 0.24 | 0.17 |
| 8 | 0.25 | 0.23 | 0.23 | 0.20 | 0.20 | 0.20 | 0.24 | 0.22 | 0.23 | 0.23 | 0.27 |
| 9 | 0.22 | | | | | | 0.00 | 0.04 | 0.07 | 0.03 | 0.05 |
| 10 | 0.13 | | | | | | 0.00 | 0.02 | 0.01 | 0.02 | 0.07 |

component, indicating that all but the first principal component are well approximated by the sparse components. The variances in the 9th and 10th rows are less than half the variances of the corresponding principal component, indicating that the first eight sparse components are not capturing much variance in the directions of $\gamma_9$ or $\gamma_{10}$. This small variance is reassuring, since it indicates that the sparse components are mostly capturing variances in the intended directions.

As $\eta$ increases, the reconstruction error (criterion (d)) will also increase. The reconstruction errors of sparse components with various $\eta$ are plotted in Figure 4. Values of $\eta =$

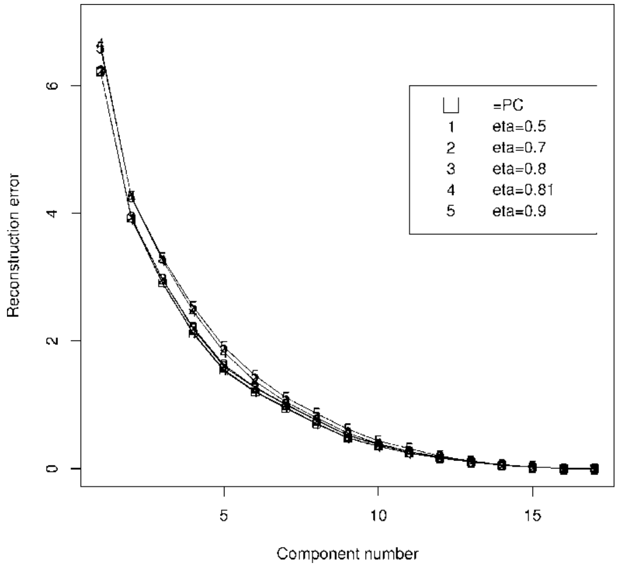

**Figure 4.** The reconstruction errors of the sparse and principal components

0.5, 0.7 and 0.8 give almost the same reconstruction error as the principal components, while $\eta = 0.81$ and 0.9 are just slightly worse. For $\eta = 0.5, 0.7, 0.8, 0.81, 0.9$, the corresponding percentages of non-zero coefficients are 51%, 38%, 32%, 27% and 26%, respectively. Even with sparse components, the data can be reconstructed with little error. This is an interesting contrast with the criteria for variance explained, which seem much more affected by the sparsity of the first principal component. This provides an illustration of our preference in the first section for reconstruction error over variance explained.

We now examine the sensitivity of the first component to $\eta$. Figure 5 shows the $C1$ criterion as a function of $k$, and its constituent parts, for five different values of the penalty parameter $\eta$. The $C1$ value plotted for each $k$ is the maximum $C1$ value over all $\alpha_1$ with $k$ non-zero coefficients. As the number of non-zero loadings ($k$) increases, the angle to the first principal component decreases by an almost constant amount. When the linear penalty $\eta k/p$ is added to $\theta/(2/\pi)$, small nonlinearities in $C1$ become apparent. When $\eta$ is near 0.80, these nonlinearities are responsible for instability of the optimal value of $k$. Does this instability mean that the $C1$ criterion should be discarded? Not necessarily. This example is one in which it would be difficult for an analyst to choose the number of non-zero components. The sensitivity of $C1$ draws attention to this difficulty, which we view as a positive characteristic.

## Stepwise Calculation of Interpretable Components

In the previous section, the interpretable directions $\alpha_i$ were chosen to be 'close' to fixed principal component directions. If some $\alpha_i$ are not very close to the corresponding direction, departures from properties (a)–(d) may be severe. This section offers one possible solution: a way to construct $\alpha_i$'s one at a time.

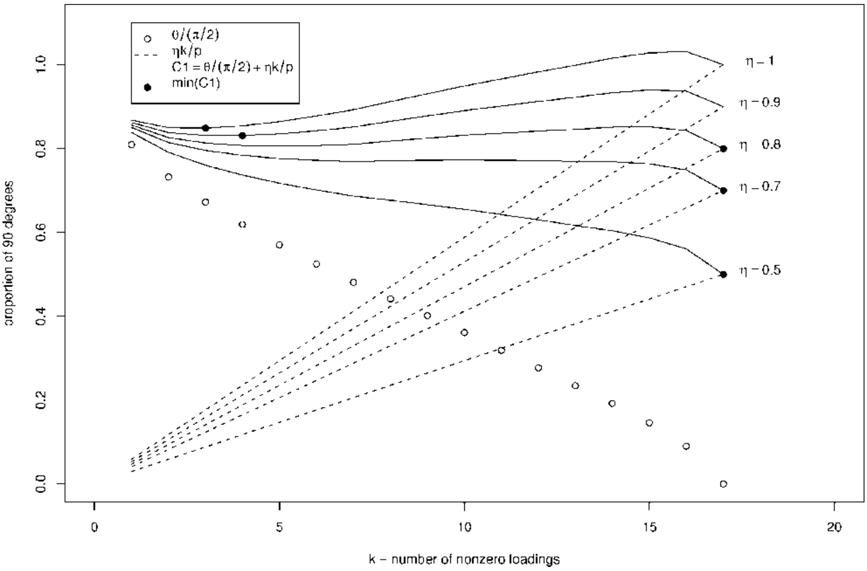

**Figure 5.** The $C1$ criterion for sparse components and its constituent parts. As the number of non-zero loadings ($k$) increases, the penalty $\eta k/p$ (- - -) increases linearly, and the angle $\theta$ to the principal component decreases. Penalty terms (- - -) and $C1$ criteria (–) are plotted for five different values of the tuning constant $\eta$. As the penalty constant $\eta$ increases, the minimum of $C1$ (denoted by •) shifts from $k = 17$ to a smaller value

The algorithm below capitalizes on a stepwise interpretation of principal components. If the first principal component direction is used to reconstruct the data, and this reconstruction is removed from the data, then the first principal component of the new data will simply be the second principal component of the original data. The algorithm is described below, followed by some comments.

```
1. Initialize:
   $F \leftarrow \tilde{X}$
   $A \leftarrow$ NULL
Repeat for $i = 1, \ldots, p$:
2. $\gamma \leftarrow$ first principal component direction of $\hat{\Sigma}_F = F'F/n$.
3. $A \leftarrow [A \,\vdots\, \alpha]$ where $\alpha$ is the interpretable direction corresponding to $\gamma$ (see Comment IV
   for details).
4. Update $F$ as the difference between the original data $\tilde{X}$ and its reconstruction using the
   directions in $A$:
   $F \leftarrow \tilde{X} - \tilde{X}A(A'\tilde{X}'\tilde{X}A)^{-1}A'\tilde{X}'\tilde{X}$
End repeat.
```

*Comments*

I In Step 4, if we set $T = \tilde{X}A$, the second term becomes $T(T'T)^{-1}T'\tilde{X}$. This term can be viewed as the prediction of $\tilde{X}$ using $T$ as predictor variables in a regression model.

II In Step 2, using the value of $F$ from Step 4, the $F'F/n$ term can be expressed as

$$F'F/n = \tilde{X}'\tilde{X}/n - \tilde{X}'\tilde{X}A(A'\tilde{X}'\tilde{X}A)^{-1}A'\tilde{X}'\tilde{X}/n = \hat{\Sigma} - \hat{\Sigma}A(A'\hat{\Sigma}A)^{-1}A'\hat{\Sigma}, \qquad (5)$$

where $\hat{\Sigma} = \tilde{X}'\tilde{X}/n$ is the sample covariance of $\tilde{X}$.

III Suppose $A$ was constructed from principal component directions instead of interpretable directions. That is, in Step 3, $\alpha = \gamma$. Then the algorithm would just identify the principal component directions $\gamma_1, \ldots, \gamma_p$.

This can be shown using the spectral decomposition of $\hat{\Sigma}$:

$$\hat{\Sigma} = \sum_{i=1}^{p} \gamma_i \lambda_i \gamma_i'$$

where $\lambda_i$ is the $i$th eigenvalue of $\hat{\Sigma}$. If $A = [\gamma_1, \ldots, \gamma_k]$, then the second term of equation (5) becomes

$$\hat{\Sigma}A(A'\hat{\Sigma}A)^{-1}A'\hat{\Sigma} = \sum_{i=1}^{k} \gamma_i \lambda_i \gamma_i'$$

That is, at the $k$th step, the spectral decomposition of $\hat{\Sigma}_F$ will be composed of exactly the last $p - k$ elements of the spectral decomposition of $\hat{\Sigma}$. Thus, the first eigenvector of $\hat{\Sigma}_F$ will be the $(k + 1)$th principal component direction.

IV In Step 3, the interpretable direction $\alpha$ corresponding to $\gamma$ is not calculated by directly applying constraints to $\gamma$. From Step 2, $\gamma$ is the first principal component direction of $\hat{\Sigma}_F$, but the linear combination $F\gamma$ is not equal to $\tilde{X}\gamma$ if $A$ is not constructed from principal component directions. Using the relation $F\gamma = \tilde{X}(\gamma - A(A'\tilde{X}'\tilde{X}A)^{-1}A'\tilde{X}'\tilde{X}\gamma)$, we transfer the linear combination back to $\tilde{X}$. Thus, $\alpha$ is calculated by applying constraints to the direction $\gamma - A(A'\tilde{X}'\tilde{X}A)^{-1}A'\tilde{X}'\tilde{X}\gamma$.

Since in each step, the linear effect of previously identified projections of $X$ has been partialed out, the resultant components tend to improve the uncorrelatedness of the projected data $T$. In the next section, we illustrate stepwise sparse components. An alternative to using just one of homogeneity, contrast or sparsity constraints (as described above) is explored later.

### *Stepwise Sparse Components for Cars Data*

The stepwise sparse components when $\eta = 0.8$ and $\eta = 0.81$ are given in Table 5. When $\eta = 0.8$, there is very little difference from the sparse components in Table 3. There are more differences when $\eta = 0.81$. These differences are probably due to the discrepancy between the first sparse direction and the first principal component direction. Subsequent directions adjust for this difference and find slightly different directions, rather than still attempting to come close to the principal component directions.

Performance criteria were examined for this example. Similar orthogonality of the $\alpha_i$'s and the uncorrelatedness of $T$ are found here, plotted in Figures 1 and 2. The variance of the projected data and reconstruction errors are very similar to results in the previous section.

### *Stepwise Best Components for Cars Data*

The stepwise procedure could be carried out separately for homogeneous, contrast, and sparse components. Instead, we calculate all three components in Step 3, and select the interpretable component with smallest angle to the direction $\gamma$. A single run of the

**Table 5.** Stepwise sparse components, cars dataset

| | $\eta = 0.8$ | | | | | $\eta = 0.81$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Min price | 0.23 | 0.40 | 0 | 0 | −0.27 | 0 | 0.43 | 0 | 0 | 0 |
| Price | 0.22 | 0.45 | 0 | 0 | −0.29 | 0 | 0.47 | 0 | 0 | 0 |
| Max price | 0.20 | 0.47 | 0 | 0 | −0.29 | 0 | 0.48 | 0 | 0 | 0 |
| City MPG | −0.27 | 0 | 0 | 0.54 | 0 | 0 | 0 | 0 | 0.51 | 0 |
| Highway MPG | −0.25 | 0 | 0 | 0.75 | 0 | 0 | 0 | 0 | 0.65 | 0 |
| Engine size | 0.28 | 0 | 0 | 0 | 0 | 0.45 | 0 | −0.25 | 0.25 | 0 |
| Horsepower | 0.24 | 0.31 | 0 | 0 | 0.37 | 0 | 0.33 | 0 | 0 | 0 |
| RPM | −0.14 | 0.44 | 0 | 0 | 0.78 | 0 | 0.35 | 0 | 0 | 0.67 |
| Rev./mile | −0.24 | 0 | 0.39 | 0 | 0 | 0 | 0 | 0.26 | 0 | 0.33 |
| Fuel tank capacity | 0.27 | 0 | 0 | 0 | 0 | 0.44 | 0 | 0 | 0 | 0.47 |
| Passengers | 0.19 | −0.35 | 0.53 | 0 | 0 | 0 | −0.25 | 0.53 | −0.22 | 0 |
| Length | 0.26 | 0 | 0 | 0.37 | 0 | 0 | 0 | 0 | 0.22 | 0 |
| Wheelbase | 0.27 | 0 | 0 | 0 | 0 | 0.44 | 0 | 0 | 0.24 | 0 |
| Width | 0.27 | 0 | 0 | 0 | 0 | 0.43 | −0.24 | −0.27 | 0.19 | 0.31 |
| Turn circle | 0.25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Rear seat room | 0.18 | 0 | 0.75 | 0 | 0 | 0 | 0 | 0.71 | 0.24 | 0 |
| Weight | 0.29 | 0 | 0 | 0 | 0 | 0.47 | 0 | 0 | 0 | 0.33 |

stepwise algorithm will then produce a combination of three types of interpretable components.

Here we use $\eta = 0.81$ for the sparse components computation. The 1st, 9th and 11th components were homogeneous, the 3rd, 7th and 12th were contrasts, and other components were sparse. The reconstruction errors are improved over sparse components with $\eta = 0.81$, and are almost exactly the same as for the original principal components.

## Data Mining Example

In this section we illustrate a more substantial application, with 200 variables and over 2000 observations. The proposed methodology is illustrated both as a dimension reduction technique, and in the context of logistic regression.

The dataset concerns a direct marketing application, in which a person responds or does not respond to a direct mailing. The data were provided by Gary Saarenvirta, and were used in the Statistical Society of Canada 2000 Case Study. A total of 2158 observations are available, broken into 1079 responders and 1079 non-responders. These data come from a larger set, with a response rate of about 1%. All responders in the larger set were used, and an equal number of non-responders were sampled at random, yielding the 2158 observations. The 200 explanatory variables consist of personal information such as gender, purchasing habits, etc, and demographic information for the census enumeration area in which the individuals live. The objective is to construct a model to predict response/non-response using the 200 variables. Most of the variables have been normalized.

Many teams participating in the case study found that linear methods such as logistic regression were among the more effective classification models for this data. Stepwise variable selection on the original predictors was attempted, but performance was generally better with low-dimensional projections as predictors. Successful projections included the first 10–40 principal components and partial least squares. Both a strength and weakness of projection methods is that all 200 predictors are combined with non-zero weights in the linear combinations. This gives predictions that may be more stable than variable selection, but it renders the model uninterpretable. Interpretable directions may offer a compromise between principal components and variable selection. If the interpretable directions allow good reconstruction of the original predictors, they may be more likely to perform well in a logistic regression.

We calculated the original homogeneous, contrast, and sparse directions, their corresponding stepwise version, and the stepwise best components. The homogeneity and contrast constraints have as many as 170 non-zero coefficients. Although the coefficients take restricted values, interpretation of so many non-zero coefficients will be difficult. We prefer to use the sparsity constraints only in this example.

We obtained the sparse components via the stepwise method described in the fourth section. Setting $\eta = 1.5$ gives less than 30 non-zero coefficients per direction in almost all sparse components. The reconstruction errors for 100 components (interpretable or principal) were less than 5% of the total variance. A similar comparison is given in Figure 6.

The first sparse component has 27 non-zero coefficients, all but one of which are positive. Absolute values of these coefficients range from 0.17 to 0.22 with most of them are 0.18 or 0.19, corresponding roughly to an average. The variable names are listed in Table 6. Only one variable has a negative coefficient – the number of taxfilers with income from \$1–\$14,999. All the variables are enumeration area variables, rather than personal variables. It seems all these variables measure wealth in the enumeration area.

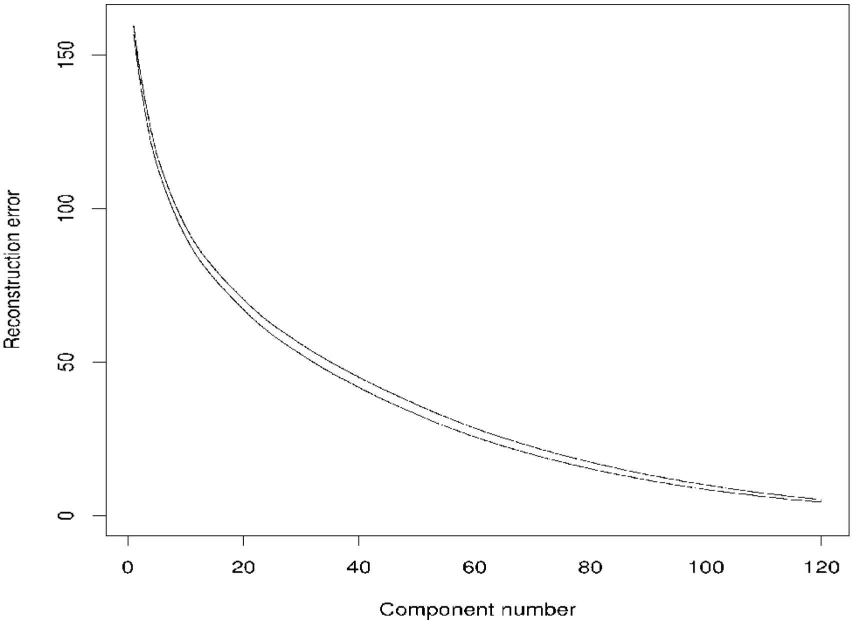

**Figure 6.** The reconstruction errors of the first 120 sparse components and the principal components, direct marketing example. The lower line corresponds to the principal components

The one variable with a negative coefficient measures the number of low income people, which also seems consistent with the idea that this component measures wealth. The methodology has identified a component with a clear interpretation, removing other less important variables and simplifying interpretation. Other components may be interpreted, but this will not be described here.

**Table 6.** Variables with nonzero coefficients in the 1st sparse component, direct marketing example

| | |
|---|---|
| Percentages | |
| % Family income >$50 K | % Household income >$50 K |
| % Males 15 + with income ≥$30 K | % Managerial & admin occupation |
| Dollar amounts | |
| Average family income | Median family income |
| Average household income | Median household income |
| Average income males 15 + | Median income males 15 + |
| Average total income | Average male total income |
| Average female total income | |
| Total counts of taxfilers | |
| With RRSP $4K-8K | With RRSP ≥$8K |
| Claiming donations | With income $1–$15K |
| With income $55K–$74K | With income $75K–$99K |
| With T4 earnings >$45K | With dividend income |
| Male with dividend income | With dividend income $1–$500 |
| With RRSP contributions | With RRSP contributions $1K–$4K |
| Male with RRSP contributions | Female with RRSP contributions |

For logistic regression, we use either the first 10 principal components or the first 10 sparse components as predictors. Ten instead of 100 components are used because interpretation of more than ten components is daunting.

The data were randomly divided into train and test sets containing 2/3 and 1/3 of the data respectively. Logistic regression models were fit to the training data using the two distinct groups of predictors.

Comparisons of fit are made via the gains chart in Figure 7. In this plot, each model generates a predicted probability of responding for every test case. The test cases are then ordered by decreasing probability of response. In the gains chart, this ordering corresponds to the horizontal axis. Suppose we choose to mail to the top proportion $p$ of test cases, or correspondingly select the leftmost proportion $p$ of the horizontal axis. At the horizontal value $p$, the height of the curve is the number of selected cases who actually responded divided by the total number of responders. A 45 degree line corresponds to random sampling of cases. The higher the curve above a 45 degree line, the more responders are selected early. Note that the best possible curve would be initially linear and then flat across the top of the chart, corresponding to selecting all the good cases first. In this problem, about 43% of the test cases are responders, so the best line would be flat after about $p = 0.43$.

In this case we see negligible difference in the gains between models based on either the first 10 principal or interpretable components. The interpretable components appear just as accurate for predicting response, while offering the advantage of interpretability.

## Discussion

In this section we make comparisons with the approaches for simplifying principal components mentioned in the first section, and discuss future work.

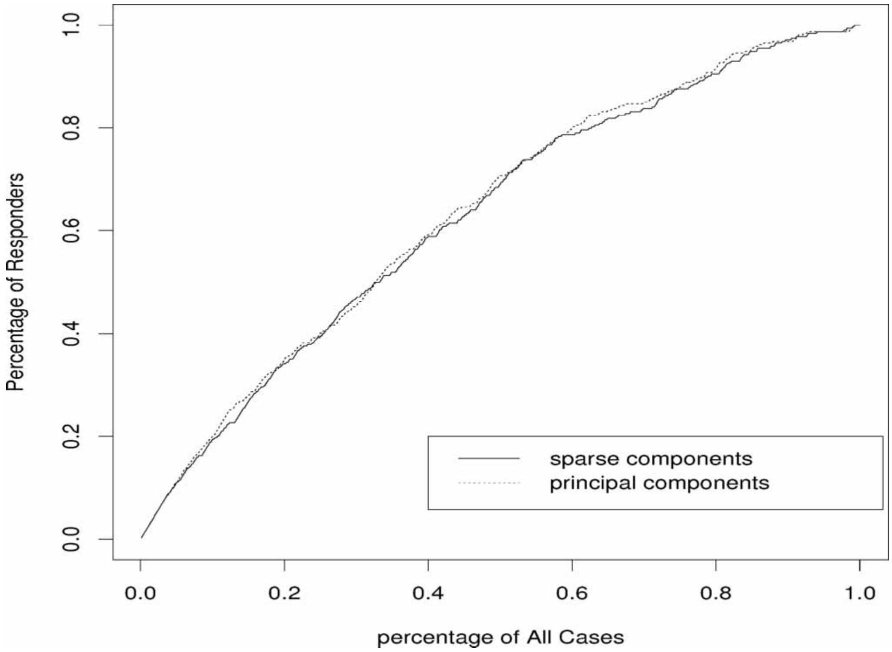

**Figure 7.** Gains chart for logistic regression applied to the direct marketing example

Previous approaches to simplifying principal components can be divided into two groups: those seeking loadings proportional to integers, and those with real loadings, many of which are zero or close to zero. Comparisons with homogeneous/contrast and sparse components are then natural.

Optimization is a challenge when loadings are constrained to be proportional to $\{0, \pm 1\}$ or more generally, integers, since for $q$ distinct possible loadings and $p$ dimensions, there are $q^p$ possible loadings. Hausman (1982) proposes a branch-and-bound search algorithm which identifies the constrained component maximizing variance, while being much faster than an exhaustive search. The solution involves a partial covariance matrix, like the stepwise search suggested in the fourth section. Although faster than an exhaustive search, such a branch-and-bound algorithm will not scale well to large problems like the data mining example in the fifth section. Performance limits may be similar to branch-and-bound algorithms for subset selection in regression (Furnival & Wilson, 1974) where entertaining more than 50 or 60 dimensions is infeasible.

To identify loadings proportional to integers, Vines (2000) uses "simplicity performing transformations", which are constrained rotations and rescalings applied to pairs of variables. A sequence of rotations are executed to maximize variance of the resultant variables, with the constraint that the loadings defining the new variables be proportional to integers. Since many searches over all pairs of variables are required, there will be as many as $\binom{p}{2} \approx p^2/2$ pairs of variables to be considered. For large dimension $p$, this may be challenging.

In contrast, our algorithm scales linearly in $p$, since the optimal loadings with $k$ non-zero elements are available in closed form for all three types of component. Thus, only a loop over $k$ from 1 to $p$ is necessary. While we cannot guarantee minimal variance, as Hausman does, the solutions are fast and, based on the examples, effective.

We compared our homogeneous components to Vines' simple principal components, and found very similar results, except in examples where the simple principal components were proportional to larger integers.

Among methods which produce real-valued but sparse loadings, rotation of principal components is the oldest. Once a group of important components are identified, they are rotated to increase interpretability, while keeping the orthogonality of the individual components. Jolliffe (1989) discusses this technique and argues that it is most effective when applied to groups of components whose variances are roughly equal. Jolliffe & Uddin (2000) introduce a one-step alternative to the two-step procedure of first finding the principal components and then rotation. In their procedure, they maximize the variance of each projection plus a penalty term, yielding coefficients close to 0 or 1. While this approach does not typically set any loadings to zero, it pushes them towards these values. Like our sparse components, Jolliffe & Uddin have a parameter that controls the influence of the penalty and the sparsity of the resultant components. A similar approach is proposed by Jolliffe *et al.* (2003), but with the constraint that the sum of absolute loadings in a component be less than some constant. This is similar to LASSO shrinkage (Tibshirani, 1996), and yields some loadings exactly equal to 0. Both methods proposed by Jolliffe & Uddin involve optimization in a space with many local optima, and could be more time consuming than our procedure.

All the approaches to simplified principal components mentioned above attempt to maximize the variance of the projected data. In the third section we saw that the reconstruction error and the variance of the projected data were quite different criteria. Although the interpretable components may not have as large a variance as the principal components, they were quite competitive in reconstructing the data. Thus, the fact that our interpretable components do not attempt to directly maximize variance is an important distinction, and not necessarily a disadvantage.

The current approach is also unified: for all three meanings of "interpretable", the same method of minimizing angle to the corresponding principal component is used to identify the interpretable components. This method can seek to be close to the original components, or proceed in a stepwise manner, partialling out the effect of each interpretable component as it is identified. We think that our approach should be complementary to previous work, in that we attempt to find interpretable directions by different methods.

In our approach, there remain some interesting issues for future research. For example, we have not developed a formal method for selecting the tuning parameter $\eta$ in sparsity criteria $C1$ and $C2$. We have chosen the $\eta$ value in examples by trying different values. By thinking of the problem in terms of reconstruction error, an analogy with regression can be made. This suggests that the choice of $\eta$ could be made via cross-validation, attempting to minimize (or find a sufficiently low value of) the reconstruction error on a test set.

Another interesting direction is the application of these techniques to methods other than principal components. Many adaptive modelling procedures (e.g. neural networks, projection pursuit regression, discriminant analysis, canonical correlation analysis, etc) are built upon linear combinations of variables, but are not interpretable. By constraining the linear combinations, models that are interpretable and still effective might be found.

## Acknowledgments

The authors are grateful to the helpful comments of Ian Jolliffe, Karen Vines, and an anonymous referee, which led to an improved version of the paper. We are also grateful to the organizers and participants involved in the 2000 SSC case study, which inspired the example in the fifth section. This research was supported by funding from MITACS and NSERC, and was conducted while the second author was visiting the University of Waterloo as a postdoctoral fellow.

## References

Cadima, J. & Jolliffe, I. T. (1995) Loadings and correlations in the interpretation of principal components, *Journal of Applied Statistics*, 22, pp. 203–214.

Furnival, G. M. & Wilson, R. W. Jr. (1974) Regression by Leaps and Bounds, *Technometrics*, 16, pp. 499–511.

Hausman, R. E. (1982) Constrained multivariate analysis, in: S. H. Zanckis & J. S. Rustagi (Eds) *Optimisation in Statistics*, pp. 137–151 (Amsterdam: North Holland).

Jolliffe, I. T. (1989) Rotation of ill-defined principal components, *Applied Statistics*, 38, pp. 139–147.

Jolliffe, I. T. & Uddin, M. (2000) The simplified component technique: an alternative to rotated principal components, *Journal of Computation and Graphical Statistics*, 9, pp. 689–710.

Jolliffe, I. T., Trendafilov, N. T. & Uddin, M. (2003) A modified principal component component technique based on the LASSO, *Journal of Computational and Graphical Statistics*, 12, pp. 531–547.

Lock, R. H. (1993) 1993 new car data, *Journal of Statistics Education*, 1. Available online at http://www.amstat.org/publications/jse/

Rao, C. R. (1965) *Linear Statistical Inference and Its Applications* (New York: Wiley).

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B*, 58(1), pp. 267–288.

Vines, S. K. (2000) Simple principal components, *Applied Statistics*, 49, pp. 441–451.