A Modified Principal Component Technique Based on the LASSO

Ian T. JOLLIFFE, Nickolay T. TRENDAFILOV, and Mudassir UDDIN

In many multivariate statistical techniques, a set of linear functions of the original p variables is produced. One of the more difficult aspects of these techniques is the interpretation of the linear functions, as these functions usually have nonzero coefficients on all p variables. A common approach is to effectively ignore (treat as zero) any coefficients less than some threshold value, so that the function becomes simple and the interpretation becomes easier for the users. Such a procedure can be misleading. There are alternatives to principal component analysis which restrict the coefficients to a smaller number of possible values in the derivation of the linear functions, or replace the principal components by "principal variables." This article introduces a new technique, borrowing an idea proposed by Tibshirani in the context of multiple regression where similar problems arise in interpreting regression equations. This approach is the so-called LASSO, the "least absolute shrinkage and selection operator," in which a bound is introduced on the sum of the absolute values of the coefficients, and in which some coefficients consequently become zero. We explore some of the properties of the new technique, both theoretically and using simulation studies, and apply it to an example.

Key Words: Interpretation; Principal component analysis; Simplification.

1. INTRODUCTION

Principal component analysis (PCA), like several other multivariate statistical techniques, replaces a set of p measured variables by a small set of derived variables. The derived variables, the principal components, are linear combinations of the p variables. The dimension reduction achieved by PCA is especially useful if the components can be readily interpreted, and this is sometimes the case; see, for example, Jolliffe (2002, chap 4). In other examples, particularly where a component has nontrivial loadings on a substantial

Ian T. Jolliffe is Professor, Department of Mathematical Sciences, University of Aberdeen, Meston Building, King's College, Aberdeen AB24 3UE, Scotland, UK (E-mail: itj@maths.abdn.ac.uk). Nickolay T. Trendafilov is Senior Lecturer, Faculty of Computing, Engineering and Mathematical Sciences, University of the West of England, Bristol, BS16 1QY, UK (E-mail: Nickolay.Trendafilov@uwe.ac.uk). Mudassir Uddin is Associate Professor, Department of Statistics, University of Karachi, Karachi-75270, Pakistan (E-mail: mudassir2000@hotmail.com).

Journal of Computational and Graphical Statistics, Volume 12, Number 3, Pages 531–547
DOI: 10.1198/1061860032148

proportion of the p variables, interpretation can be difficult, detracting from the value of the analysis.

A number of methods are available to aid interpretation. Rotation, which is commonplace in factor analysis, can be applied to PCA, but has its drawbacks (Jolliffe 1989, 1995). A frequently used informal approach is to ignore all loadings smaller than some threshold absolute value, effectively treating them as zero. This can be misleading (Cadima and Jolliffe 1995). A more formal way of making some of the loadings zero is to restrict the allowable loadings to a small set of values; for example, $-1, 0, 1$ (Hausman 1982). Vines (2000) described a variation on this theme. One further strategy is to select a subset of the variables themselves, which satisfy similar optimality criterion to the principal components, as in McCabe's (1984) "principal variables."

This article introduces a new technique which shares an idea central to both Hausman's (1982) and Vines's (2000) work. This idea is that we choose linear combinations of the measured variables which successively maximizes variance, as in PCA, but we impose extra constraints, which sacrifices some variance in order to improve interpretability. In our technique the extra constraint is in the form of a bound on the sum of the absolute values of the loadings in that component. This type of bound has been used in regression (Tibshirani 1996), where similar problems of interpretation occur, and is known there as the LASSO (least absolute shrinkage and selection operator). As with the methods of Hausman (1982) and Vines (2000), and unlike rotation, our technique usually produces some exactly zero loadings in the components. In contrast to Hausman (1982) and Vines (2000) it does not restrict the nonzero loadings to a discrete set of values. This article shows, through simulations and an example, that the new technique is a valuable additional tool for exploring the structure of multivariate data.

Section 2 establishes the notation and terminology of PCA, and introduces an example in which interpretation of principal components is not straightforward. The most usual approach to simplifying interpretation, the rotation of PCs, is shown to have drawbacks. Section 3 introduces the new technique and describes some of its properties. Section 4 revisits the example of Section 2, and demonstrates the practical usefulness of the technique. A simulation study, which investigates the ability of the technique to recover known underlying structures in a dataset, is summarized in Section 5. The article ends with further discussion in Section 6, including some modifications, complications, and open questions.

2. A MOTIVATING EXAMPLE

Consider the classic example, first introduced by Jeffers (1967), in which a PCA was done on the correlation matrix of 13 physical measurements, listed in Table 1, made on a sample of 180 pitprops cut from Corsican pine timber.

Let $\mathbf{x}_i$ be the vector of 13 variables for the ith pitprop, where each variable has been standardized to have unit variance. What PCA does, when based on the correlation matrix, is to find linear functions $\mathbf{a}_1'\mathbf{x}, \mathbf{a}_2'\mathbf{x}, \ldots, \mathbf{a}_p'\mathbf{x}$ which successively have maximum sample variance, subject to $\mathbf{a}_h'\mathbf{a}_k = 0$ for $k \geq 2$, and $h < k$. In addition, a normalization constraint

Table 1. Definitions of Variables in Jeffers' Pitprop Data

Variable	*Definition*
x_1	Top diameter in inches
x_2	Length in inches
x_3	Moisture content, % of dry weight
x_4	Specific gravity at time of test
x_5	Oven-dry specific gravity
x_6	Number of annual rings at top
x_7	Number of annual rings at bottom
x_8	Maximum bow in inches
x_9	Distance of point of maximum bow from top in inches
x_{10}	Number of knot whorls
x_{11}	Length of clear prop from top in inches
x_{12}	Average number of knots per whorl
x_{13}	Average diameter of the knots in inches

$\mathbf{a}_k'\mathbf{a}_k = 1$ is necessary to get a bounded solution. The derived variable $\mathbf{a}_k'\mathbf{x}$ is the kth principal component (PC). It turns out that $\mathbf{a}_k$, the vector of coefficients or loadings for the kth PC is the eigenvector of the sample correlation matrix $\mathbf{R}$ corresponding to the kth largest eigenvalue l_k. In addition the sample variance of $\mathbf{a}_k'\mathbf{x}$ is equal to l_k. Because of the successive maximization property, the first few PCs will often account for most of the sample variation in all the standardized measured variables. In the pitprop example, Jeffers (1967) was interested in the first six PCs, which together account for 87% of the total variance. The loadings in each of these six components are given in Table 2, together with the individual and cumulative percentage of variance in all 13 variables, accounted for by $1, 2, \ldots, 6$ PCs.

PCs are easiest to interpret if the pattern of loadings is clear-cut, with a few large

Table 2. Loadings for Correlation PCA for Jeffers' Pitprop Data

	Component					
Variable	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
x_1	0.404	0.212	−0.219	−0.027	−0.141	−0.086
x_2	0.406	0.180	−0.245	−0.025	−0.188	−0.111
x_3	0.125	0.546	0.114	0.015	0.433	0.120
x_4	0.173	0.468	0.328	0.010	0.361	−0.090
x_5	0.057	−0.138	0.493	0.254	−0.122	−0.560
x_6	0.284	−0.002	0.476	−0.153	−0.269	0.032
x_7	0.400	−0.185	0.261	−0.125	−0.176	0.030
x_8	0.294	−0.198	−0.222	0.294	0.203	0.103
x_9	0.357	0.010	−0.202	0.132	−0.117	0.103
x_{10}	0.379	−0.252	−0.120	−0.201	0.173	−0.019
x_{11}	−0.008	0.187	0.021	0.805	−0.302	0.178
x_{12}	−0.115	0.348	0.066	−0.303	−0.537	0.371
x_{13}	−0.112	0.304	−0.352	−0.098	−0.209	−0.671
Simplicity factor (varimax)	0.059	0.103	0.082	0.397	0.086	0.266
Variance (%)	32.4	18.2	14.4	8.9	7.0	6.3
Cumulative Variance (%)	32.4	50.7	65.0	74.0	80.9	87.2

Table 3. Loadings for Rotated Correlation PCA, Using the Varimax Criterion, for Jeffers' Pitprop Data.

	Component					
Variable	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
x_1	−0.019	0.074	0.043	−0.027	−0.519	−0.077
x_2	−0.018	0.015	0.048	−0.024	−0.540	−0.102
x_3	−0.024	0.705	−0.128	0.003	−0.059	0.107
x_4	0.029	0.689	0.112	0.001	0.014	−0.087
x_5	0.258	0.009	0.477	0.218	0.205	−0.524
x_6	−0.185	0.061	0.604	−0.005	−0.032	0.012
x_7	0.031	−0.069	0.512	−0.102	−0.151	0.092
x_8	0.440	−0.042	−0.072	0.083	−0.221	0.239
x_9	0.097	−0.058	0.045	0.094	−0.408	0.141
x_{10}	0.271	−0.054	0.129	−0.367	−0.216	0.135
x_{11}	0.057	−0.022	−0.029	0.882	−0.137	0.075
x_{12}	−0.776	−0.056	0.091	0.079	−0.123	0.145
x_{13}	−0.120	−0.049	−0.280	−0.077	−0.269	−0.748
Simplicity factor (varimax)	0.362	0.428	0.199	0.595	0.131	0.343
variance (%)	13.0	14.6	18.4	9.7	23.9	7.6
cumulative variance (%)	13.0	27.6	46.0	55.7	79.6	87.2

(absolute) values and many small loadings in each PC. Although Jeffers (1967) makes an attempt to interpret all six components, some are, to say the least, messy and he ignores some intermediate loadings. For example, PC2 has the largest loadings on x_3, x_4, with small loadings on x_6, x_9, but a whole range of intermediate values on other variables.

A traditional way to simplify loadings is by rotation. If $\mathbf{A}$ is the (13×6) matrix whose kth column is $\mathbf{a}_k$, then $\mathbf{A}$ is post-multiplied by a matrix $\mathbf{T}$ to give rotated loadings $\mathbf{B} = \mathbf{AT}$. If $\mathbf{b}_k$ is the kth column of $\mathbf{B}$ then $\mathbf{b}_k^{'}\mathbf{x}$ is the kth rotated component. The matrix $\mathbf{T}$ is chosen so as to optimize some simplicity criterion. Various criteria have been proposed, all of which attempt to create vectors of loadings whose elements are close to zero or far from zero, with few intermediate values. The idea is that each variable should be either clearly important or clearly unimportant in a rotated component, with as few cases as possible of borderline importance. Varimax is the most widely used rotation criterion and, like most other such criteria, it tends to drive at least some of the loadings in each component towards zero. This is not the only possible type of simplicity. A component whose loadings are all roughly equal is easy to interpret but will be avoided by most standard rotation criteria. It is difficult to envisage any criterion which could encompass all possible types of simplicity, and we concentrate here on simplicity as defined by varimax.

Table 3 gives the rotated loadings for six components in the correlation PCA of the pitprop data, together with the percentage of total variance accounted for by each rotated PC (RPC). The rotation criterion used in Table 3 is varimax (Krzanowski and Marriott 1995, p. 138), which is the most frequent choice (often the default in software), but other criteria give similar results. Varimax rotation aims to maximize the sum, over rotated components, of a criterion which takes values between zero and one. A value of zero occurs when all loadings in the component are equal, whereas a component with only one nonzero loading produces a value of unity. This criterion, or "simplicity factor," is given for each component in Tables 2 and 3, and it can be seen that its values are larger for most of the rotated components than

they are for the unrotated components.

There are, however, a number of disadvantages associated with rotation. In the context of interpreting the results in this example, we note that we have lost the "successive maximization of variance" property of the unrotated components, so what we are interpreting after rotation are not the "most important sources of variation" in the data. The RPC with the highest variance appears arbitrarily as the fifth, and this accounts for 24% of the total variation, compared to 32% in the first unrotated PC. In addition, a glance at the loadings and simplicity factors for the RPCs shows that, more generally, those components which are easiest to interpret among the six in Table 3 are those which have the smallest variance. RPC5 is still rather complicated. Other problems associated with rotation were discussed by Jolliffe (1989, 1995). A simplified component technique (SCoT), in which the two steps of RPCA (PCA, followed by rotation) are combined into one, was discussed by Jolliffe and Uddin (2000). The technique is based on a similar idea proposed by Morton (1989) in the context of projection pursuit. It maximizes variance but adds a penalty function which is a multiple of one of the simplicity criteria, such as varimax. SCoT has some advantages compared to standard rotation, but shares a number of its disadvantages.

The next section introduces an alternative to rotation which has some clear advantages over rotated PCA and SCoT. A detailed comparison of the new technique, SCoT, rotated PCA, and Vines' (2000) simple components is given for an example involving sea surface temperatures in Jolliffe, Uddin, and Vines (2002).

3. MODIFIED PCA BASED ON THE LASSO

Tibshirani (1996) studied the difficulties involved in the interpretation of multiple regression equations. These problems may occur due to the instability of the regression coefficients in the presence of collinearity, or simply because of the large number of variables included in the regression equation. Some current alternatives to least squares regression, such as shrinkage estimators, ridge regression, principal component regression, or partial least squares, handle the instability problem by keeping all variables in the equation, whereas variable selection procedures find a subset of variables and keep only the selected variables in the equation. Tibshirani (1996) proposed a new method, the "least absolute shrinkage and selection operator" LASSO, which is a compromise between variable selection and shrinkage estimators. The procedure shrinks the coefficients of some of the variables not simply *towards* zero, but *exactly* to zero, giving an implicit form of variable selection. LeBlanc and Tibshirani (1998) extended the idea to regression trees. Here we adapt the LASSO idea to PCA.

3.1 THE LASSO APPROACH IN REGRESSION

In standard multiple regression we have the equation

$$y_i = \alpha + \sum_{j=1}^{p} \beta_j x_{ij} + e_i, \quad i = 1, 2, \ldots, n$$

where $y_1, y_2, \ldots, y_n$ are measurements on a response variable y, x_{ij}, $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, p$, are corresponding values of p predictor variables, $e_1, e_2, \ldots, e_n$ are error terms and $\alpha, \beta_1, \beta_2, \ldots, \beta_p$ are parameters in the regression equation. In least squares regression, these parameters are estimated by minimizing the residual (or error) sum of squares,

$$\sum_{i=1}^{n} \left(y_i - \alpha - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

The LASSO imposes an additional restriction on the coefficients, namely

$$\sum_{j=1}^{p} |\beta_j| \leq t$$

for some "tuning parameter" t. For suitable choices of t this constraint has the interesting property that it forces some of the coefficients in the regression equation to zero. An equivalent way of deriving LASSO estimates is to minimize the residual sum of squares with the addition of a penalty function based on $\sum_{j=1}^{p} |\beta_j|$. Thus, we minimize

$$\sum_{i=1}^{n} \left(y_i - \alpha - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

for some multiplier λ. For any given value of t in the first LASSO formulation there is a value of λ in the second formulation that gives equivalent results.

3.2 THE LASSO APPROACH IN PCA (SCOTLASS)

PCA on a correlation matrix finds linear combinations $\mathbf{a}_k'\mathbf{x} (k = 1, 2, \ldots, p)$, of the p measured variables $\mathbf{x}$, each standardized to have unit variance, which successively have maximum variance

$$\mathbf{a}_k' \mathbf{R}\, \mathbf{a}_k, \tag{3.1}$$

subject to

$$\mathbf{a}_k' \mathbf{a}_k = 1 \quad \text{and (for } k \geq 2) \quad \mathbf{a}_h' \mathbf{a}_k = 0, \quad h < k. \tag{3.2}$$

The proposed method of LASSO-based PCA performs the maximization under the extra constraints

$$\sum_{j=1}^{p} |\mathrm{a_{kj}}| \leq t \tag{3.3}$$

for some tuning parameter t, where $\mathrm{a_{kj}}$ is the jth element of the kth vector $\mathbf{a}_k, (k = 1, 2, \ldots, p)$. We call the new technique SCoTLASS (Simplified Component Technique-LASSO).

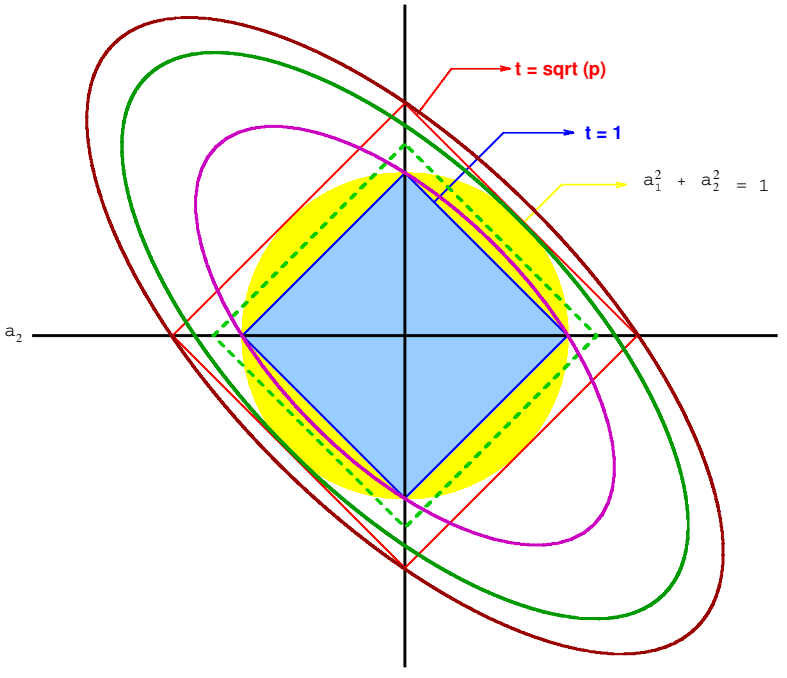

Figure 1. The Two-Dimensional SCoTLASS.

3.3 SOME PROPERTIES

SCoTLASS differs from PCA in the inclusion of the constraints defined in (3.3), so a decision must be made on the value of the tuning parameter, t. It is easy to see that

(a) for $t \geq \sqrt{p}$, we get PCA;
(b) for $t < 1$, there is no solution; and
(c) for $t = 1$, we must have exactly one nonzero a_{kj} for each k.

As t decreases from $\sqrt{p}$, we move progressively away from PCA and eventually to a solution where only one variable has a nonzero loading on each component. All other variables will shrink (not necessary monotonically) with t and ultimately reach zero. Examples of this behavior are given in the next section.

The geometry of SCoTLASS in the case when $p = 2$ is shown in Figure 1 where we plot the elements a_1, a_2 of the vector $\mathbf{a}$. For PCA, in Figure 1 the first component $\mathbf{a}_1'\mathbf{x}$, where $\mathbf{a}_1' = (a_{11}, a_{12})$, corresponds to the point on circumference of the shaded circle $(\mathbf{a}_1'\mathbf{a}_1 = 1)$ which touches the "*largest*" possible ellipse $\mathbf{a}_1'\mathbf{R}\mathbf{a}_1 = \text{constant}$.

For SCoTLASS with $1 < t < \sqrt{p}$ we are restricted to the part of the circle $\mathbf{a}_1'\mathbf{a}_1 = 1$ inside the dotted square $\sum_{j=1}^{2} |a_{1j}| \leq t$.

For $t = 1$, corresponding to the inner shaded square, the optimal (only) solutions are on the axes.

Figure 1 shows a special case, in which the axes of the ellipses are at 45° to the a_1, a_2 axes, corresponding to equal variances for x_1, x_2 (or a correlation matrix). This gives two

optimal solutions for SCoTLASS in the first quadrant, symmetric about the 45° line. If PCA or SCoTLASS is done on a covariance, rather than correlation matrix (see Remark 1 in Section 6), with unequal variances, there will be a unique solution in the first quadrant.

3.4 IMPLEMENTATION

PCA reduces to an easily implemented eigenvalue problem, but the extra constraint in SCoTLASS means that it needs numerical optimization to estimate parameters and suffers from the problem of many local optima. A number of algorithms have been tried to implement the technique including simulated annealing (Goffe, Ferrier, and Rogers 1994), but the results reported in the following example were derived using the projected gradient approach (Chu and Trendafilov 2001; Helmke and Moore 1994). The LASSO inequality constraint (3.3) in the SCoTLASS problem (3.1)–(3.3) is eliminated by making use of an exterior penalty function. Thus, the SCoTLASS problem (3.1)–(3.3) is transformed into a new maximization problem subject to the equality constraint (3.2). The solution of this modified maximization problem is then found as an ascent gradient vector flow onto the p-dimensional unit sphere following the standard projected gradient formalism (Chu and Trendafilov 2001; Helmke and Moore 1994). Detailed consideration of this solution will be reported separately.

We have implemented SCoTLASS using MATLAB. The code requires as input the correlation matrix $\mathbf{R}$, the value of the tuning parameter t and the number of components to be retained (m). The MATLAB code returns a loading matrix and calculates a number of relevant statistics. To achieve an appropriate solution, a number of parameters of the projected gradient method (e.g., starting points, absolute and relative tolerances) also need to be defined.

4. EXAMPLE: PITPROPS DATA

Here we revisit the pitprop data from section 2. We have studied many other examples, not discussed here, and in general the results are qualitatively similar. Table 4 gives loadings for SCoTLASS with $t = 2.25, 2.00, 1.75$ and 1.50. Table 5 gives variances, cumulative variances, "simplicity factors," and number of zero loadings, for the same values of t, as well as corresponding information for PCA ($t = \sqrt{13}$) and RPCA. The simplicity factors are values of the varimax criterion for each component.

It can be seen that as the value of t is decreased, the simplicity of the components increases as measured by the number of zero loadings and by the varimax simplicity factor, although the increase in the latter is not uniform. The increase in simplicity is paid for by a loss of variance retained. By $t = 2.25, 1.75$ the percentage of variance accounted for by the first component in SCoTLASS is reduced from 32.4 for PCA to 26.7, 19.6, respectively. At $t = 2.25$ this is still larger than the largest contribution (23.9) achieved by a single component in RPCA. The comparison with RPCA is less favorable when the variation accounted for by all six retained components is examined. RPCA necessarily retains the

Table 4. Loadings for SCoTLASS for Four Values of t Based on the Correlation Matrix for Jeffers' Pitprop Data

		Component					
Technique	*Variable*	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
SCoTLASS	x_1	0.558	0.085	−0.093	−0.107	0.056	0.017
($t = 2.25$)	x_2	0.580	0.031	−0.087	−0.147	0.073	0.047
	x_3	0.000	0.647	−0.129	0.215	−0.064	−0.101
	x_4	0.000	0.654	−0.000	0.211	−0.080	0.127
	x_5	−0.000	0.000	0.413	−0.000	0.236	0.747
	x_6	0.001	0.208	0.529	−0.022	−0.108	0.033
	x_7	0.266	−0.000	0.385	0.000	−0.121	0.020
	x_8	0.104	−0.098	0.000	0.584	0.127	−0.188
	x_9	0.372	−0.000	−0.000	0.019	0.142	−0.060
	x_{10}	0.364	−0.154	0.000	0.212	−0.296	0.000
	x_{11}	−0.000	0.099	−0.000	0.000	0.879	−0.156
	x_{12}	−0.000	0.241	−0.001	−0.699	−0.044	−0.186
	x_{13}	−0.000	0.026	−0.608	−0.026	−0.016	0.561
SCoTLASS	x_1	0.623	0.041	−0.049	0.040	0.051	−0.000
($t = 2.00$)	x_2	0.647	0.076	−0.001	0.072	0.059	−0.007
	x_3	0.000	0.000	−0.684	−0.128	−0.054	0.106
	x_4	0.000	−0.001	−0.670	−0.163	−0.063	−0.100
	x_5	−0.000	−0.267	0.000	0.000	0.228	−0.772
	x_6	0.000	−0.706	−0.044	0.011	−0.003	0.001
	x_7	0.137	−0.539	0.001	−0.000	−0.096	0.001
	x_8	0.001	−0.000	0.053	−0.767	0.095	0.098
	x_9	0.332	0.000	0.000	−0.001	0.065	0.014
	x_{10}	0.254	−0.000	0.124	−0.277	−0.309	−0.000
	x_{11}	0.000	0.000	−0.065	0.000	0.902	0.194
	x_{12}	−0.000	0.000	−0.224	0.533	−0.069	0.137
	x_{13}	0.000	0.364	−0.079	0.000	0.000	−0.562
SCoTLASS	x_1	0.664	−0.000	0.000	−0.025	0.002	−0.035
($t = 1.75$)	x_2	0.683	−0.001	0.000	−0.040	0.001	−0.018
	x_3	0.000	0.641	0.195	0.000	0.180	−0.030
	x_4	0.000	0.701	0.001	0.000	−0.000	−0.001
	x_5	−0.000	0.000	−0.000	0.000	−0.887	−0.056
	x_6	0.000	0.293	−0.186	0.000	−0.373	0.044
	x_7	0.001	0.107	−0.658	−0.000	−0.051	0.064
	x_8	0.001	−0.000	−0.000	0.735	0.021	−0.168
	x_9	0.283	−0.000	−0.000	0.000	−0.000	−0.001
	x_{10}	0.113	−0.000	−0.001	0.388	−0.017	0.320
	x_{11}	0.000	0.000	0.000	−0.000	−0.000	−0.923
	x_{12}	−0.000	0.001	0.000	−0.554	0.016	0.004
	x_{13}	0.000	−0.000	0.703	0.001	−0.197	0.080
SCoTLASS	x_1	0.701	−0.000	−0.001	−0.001	0.001	−0.000
($t = 1.50$)	x_2	0.709	−0.001	−0.001	−0.001	0.001	−0.000
	x_3	0.001	0.698	−0.068	0.002	−0.001	0.001
	x_4	0.001	0.712	−0.001	0.002	−0.001	−0.001
	x_5	−0.000	0.000	0.001	−0.001	0.093	−0.757
	x_6	0.000	0.081	0.586	−0.031	0.000	0.000
	x_7	0.001	0.000	0.807	−0.001	0.000	0.002
	x_8	0.001	−0.000	0.001	0.044	−0.513	−0.001
	x_9	0.079	0.000	0.001	0.001	−0.001	−0.000
	x_{10}	0.002	−0.000	0.027	0.660	−0.000	−0.002
	x_{11}	0.000	0.001	−0.000	−0.749	−0.032	−0.000
	x_{12}	−0.000	0.001	−0.000	−0.001	0.853	0.083
	x_{13}	0.000	0.000	−0.001	0.000	−0.001	0.648

Table 5. Simplicity Factor, Variance, Cumulative Variance and Number of Zero Loadings for Individual Components in PCA, RPCA, and SCoTLASS for Four Values of t, Based on the Correlation Matrix for Jeffers' Pitprop Data

		Component					
Technique	*Measure*	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
PCA	Simplicity factor (varimax)	0.059	0.103	0.082	0.397	0.086	0.266
(= SCoTLASS with	Variance (%)	32.4	18.2	14.4	8.9	7.0	6.3
$t = \sqrt{13}$)	Cumulative variance (%)	32.4	50.7	65.1	74.0	80.9	87.2
RPCA	Simplicity factor (varimax)	0.362	0.428	0.199	0.595	0.131	0.343
	Variance (%)	13.0	14.6	18.4	9.7	23.9	7.6
	Cumulative variance (%)	13.0	27.6	46.0	55.7	79.6	87.2
SCoTLASS	Simplicity factor (varimax)	0.190	0.312	0.205	0.308	0.577	0.364
($t = 2.25$)	Variance (%)	26.7	17.2	15.9	9.7	8.9	6.7
	Cumulative variance (%)	26.7	43.9	59.8	69.4	78.4	85.0
	Number of zero loadings	6	3	5	3	0	1
SCoTLASS	Simplicity factor (varimax)	0.288	0.301	0.375	0.387	0.646	0.412
($t = 2.00$)	Variance (%)	23.1	16.4	16.2	11.2	8.9	6.5
	Cumulative variance (%)	23.1	39.5	55.8	67.0	75.9	82.3
	Number of zero loadings	7	6	2	4	1	2
SCoTLASS	Simplicity factor (varimax)	0.370	0.370	0.388	0.360	0.610	0.714
($t = 1.75$)	Variance (%)	19.6	16.0	13.2	13.0	9.2	9.1
	Cumulative variance (%)	19.6	35.6	48.7	61.8	71.0	80.1
	Number of zero loadings	7	7	7	7	3	0
SCoTLASS	Simplicity factor (varimax)	0.452	0.452	0.504	0.464	0.565	0.464
($t = 1.50$)	Variance (%)	16.1	14.9	13.8	10.2	9.9	9.6
	Cumulative variance (%)	16.1	31.0	44.9	55.1	65.0	74.5
	Number of zero loadings	5	7	2	1	3	5

same total percentage variation (87.2) as PCA, but SCoTLASS drops to 85.0 and 80.1 for $t = 2.25, 1.75$, respectively. Against this loss, SCoTLASS has the considerable advantage of retaining the successive maximisation property. At $t = 2.25$, apart from switching of components 4 and 5, the SCoTLASS components are nicely simplified versions of the PCs, rather than being something different as in RPCA. A linked advantage is that if we decide to look only at five components, then the SCoTLASS components will simply be the first five in Table 4, whereas RPCs based on $(m - 1)$ retained components are not necessarily similar to those based on m.

A further "plus" for SCoTLASS is the presence of zero loadings, which aids interpretation, but even where there are few zeros the components are simplified compared to PCA. Consider specifically the interpretation of the second component, which we noted earlier was "messy" for PCA. For $t = 1.75$ this component is now interpreted as measuring mainly moisture content and specific gravity, with small contributions from numbers of annual rings, and all other variables negligible. This gain in interpretability is achieved by reducing the percentage of total variance accounted for from 18.2 to 16.0. For other components, too, interpretation is made easier because in the majority of cases the contribution of a variable is clearcut. Either it is important or it is not, with few equivocal contributions.

Table 6. Specified Eigenvectors of a Six-Dimensional Block Structure

	Eigenvectors					
Variable	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
x_1	0.096	−0.537	0.759	−0.120	0.335	−0.021
x_2	0.082	−0.565	−0.599	0.231	0.511	−0.013
x_3	0.080	−0.608	−0.119	−0.119	−0.771	0.016
x_4	0.594	0.085	−0.074	−0.308	0.069	0.731
x_5	0.584	0.096	−0.114	−0.418	0.052	−0.678
x_6	0.533	0.074	0.180	0.805	−0.157	−0.069
Variance	1.8367	1.640	0.751	0.659	0.607	0.506

For $t = 2.25, 2.00, 1.75, 1.50$ the number of the zeros is as follows: 18, 22, 31, and 23. It seems surprising that we obtain fewer zeros with $t = 1.50$ than with $t = 1.75$, that is, the solution with $t = 1.75$ appears to be simpler than the one with $t = 1.50$. In fact this impression is misleading (see also the next paragraph). The explanation of this anomaly is in the projected gradient method used for numerical solution of the problem, which approximates the LASSO constraint with a certain smooth function and thus the zero-loadings produced may be also approximate. One can see that the solution with $t = 1.50$ contains a total of 56 loadings with less than 0.005 magnitude, compared to 42 in the case $t = 1.75$.

Another interesting comparison is in terms of average varimax simplicity over the first six components. This is 0.343 for RPCA compared to 0.165 for PCA. For $t = 2.25, 2.00, 1.75, 1.50$ the average simplicity is 0.326, 0.402, 0.469, 0.487, respectively. This demonstrates, that although the varimax criterion is not an explicit part of SCoTLASS, by taking t small enough we can do better than RPCA with respect to its own criterion. This is achieved by moving outside the space spanned by the retained PCs, and hence settling for a smaller amount of overall variation retained.

5. SIMULATION STUDIES

One question of interest is whether SCoTLASS is better at detecting underlying simple structure in a data set than is PCA or RPCA. To investigate this question we simulated data from a variety of known structures. Because of space constraints, only a small part of the results is summarized here; further details can be found in Uddin (1999).

Given a vector $\mathbf{l}$ of positive real numbers and an orthogonal matrix $\mathbf{A}$, we can attempt to find a covariance matrix or correlation matrix whose eigenvalues are the elements of $\mathbf{l}$, and whose eigenvectors are the column of $\mathbf{A}$. Some restrictions need to be imposed on $\mathbf{l}$ and $\mathbf{A}$, especially in the case of correlation matrices, but it is possible to find such matrices for a wide range of eigenvector structures. Having obtained a covariance or correlation matrix it is straightforward to generate samples of data from multivariate normal distributions with the given covariance or correlation matrix. We have done this for a wide variety of eigenvector structures (principal component loadings), and computed the PCs, RPCs, and SCoTLASS components from the resulting sample correlation matrices. Various structures have been

Table 7. Specified Eigenvectors of a Six-Dimensional Intermediate Structure

	Eigenvectors					
Variable	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
x_1	0.224	−0.509	0.604	0.297	−0.327	0.361
x_2	0.253	−0.519	−0.361	−0.644	−0.341	−0.064
x_3	0.227	−0.553	−0.246	0.377	0.608	−0.267
x_4	0.553	0.249	−0.249	−0.052	0.262	0.706
x_5	0.521	0.254	−0.258	0.451	−0.509	−0.367
x_6	0.507	0.199	0.561	−0.384	0.281	−0.402
Variance	1.795	1.674	0.796	0.618	0.608	0.510

investigated, which we call block structure, intermediate structure and uniform structure. Tables 6–8 give one example of each type of structure. The structure in Table 6 has blocks of nontrivial loadings and blocks of near-zero loadings in each underlying component. Table 8 has a structure in which all loadings in the first two components have similar absolute values and the structure in Table 7 is intermediate to those of Tables 6 and 8. An alternative approach to the simulation study would be to replace the near-zero loadings by exact zeros and the nearly equal loadings by exact equalities. However, we feel that in reality underlying structures are never quite that simple so we perturbed them a little.

It might be expected that if the underlying structure is simple, then sampling variation is more likely to take sample PCs away from simplicity than to enhance this simplicity. It is of interest to investigate whether the techniques of RPCA and SCoTLASS which increase simplicity compared to the sample PCs will do so in the direction of the true underlying structure. The closeness of a vector of loadings from any of these techniques to the underlying true vector is measured by the angle between the two vectors of interest. These angles are given in Tables 9–11 for single simulated datasets from three different types of six-dimensional structure; they typify what we found in other simulations. Three values of t (apart from that for PCA) are shown in the tables. Their exact values are unimportant, and are slightly different in different tables. They are chosen to illustrate typical behavior in our simulations.

The results illustrate that, for each structure, RPCA is perhaps surprisingly, and certainly disappointingly, bad at recovering the underlying structure. SCoTLASS, on the other hand,

Table 8. Specified Eigenvectors of a Six-Dimensional Uniform Structure

	Eigenvectors					
Variable	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
x_1	−0.455	0.336	−0.087	0.741	−0.328	0.125
x_2	−0.439	0.370	−0.212	−0.630	−0.445	−0.175
x_3	−0.415	0.422	0.378	−0.110	0.697	0.099
x_4	0.434	0.458	0.040	−0.136	−0.167	0.744
x_5	0.301	0.435	−0.697	0.114	0.356	−0.306
x_6	0.385	0.416	0.563	0.104	−0.234	−0.545
Variance	1.841	1.709	0.801	0.649	0.520	0.480

Table 9. Angles Between the Underlying Vectors and the Sample Vectors of PCA, RPCA, and SCoTLASS With Various Values of t, for a Specified "Block" Structure of Correlation Eigenvectors

		Vectors			
Technique	*t*	*(1)*	*(2)*	*(3)*	*(4)*
PCA	$\sqrt{6}$	12.9	12.0	15.4	79.5
RPCA		37.7	45.2	45.3	83.1
SCoTLASS	$t = 2.00$	12.9	12.0	15.4	78.6
	$t = 1.82$	11.9	11.4	13.8	73.2
	$t = 1.75$	9.4	10.0	12.5	85.2

is capable of improvement over PCA. For example, for $t = 1.75$ it not only improves over PCA in terms of angles in Table 9, but it also has 3, 3, and 2 zero loadings in its first three components, thus giving a notably simpler structure. None of the methods manages to reproduce the underlying structure for component 4 in Table 9.

The results for intermediate structure in Table 10 are qualitatively similar to those in Table 9, except that SCoTLASS does best for higher values of t than in Table 9. For uniform structure (Table 11) SCoTLASS does badly compared to PCA for all values of t. This is not unexpected because, although uniform structure is simple in its own way, it is not the type of simplicity which SCoTLASS aims for. It is also the case that the varimax criterion is designed so that it stands little chance of finding uniform structure. Other rotation criteria, such as quartimax, can in theory find uniform vectors of loadings, but they were tried and also found to be unsuccessful in our simulations. It is probable that a uniform structure is more likely to be found by the techniques proposed by Hausman (1982) or Vines (2000). Although SCoTLASS will usually fail to find such structures, their existence may be indicated by a large drop in the variance explained by SCoTLASS as decreasing values of t move it away from PCA.

6. DISCUSSION

A new technique, SCoTLASS, has been introduced for discovering and interpreting the major sources of variability in a dataset. We have illustrated its usefulness in an example, and have also shown, through simulations, that it is capable of recovering certain types of

Table 10. Angles Between the Underlying Vectors and the Sample Vectors of PCA, RPCA, and SCoTLASS With Various Values of t, for a Specified "Intermediate" Structure of Correlation Eigenvectors

		Vectors			
Technique	*t*	*(1)*	*(2)*	*(3)*	*(4)*
PCA	$\sqrt{6}$	13.7	15.1	23.6	78.3
RPCA		53.3	42.9	66.4	77.7
SCoTLASS	$t = 2.28$	5.5	10.4	23.7	78.3
	$t = 2.12$	9.5	12.0	23.5	77.8
	$t = 2.01$	17.5	19.1	22.5	71.7

Table 11. Angles Between the Underlying Vectors and the Sample Vectors of PCA, RPCA, and SCoTLASS with Various Values of t, for a Specified "Uniform" Structure of Correlation Eigenvectors

		Vectors			
Technique	*t*	*(1)*	*(2)*	*(3)*	*(4)*
PCA	$\sqrt{6}$	6.0	6.4	23.2	31.8
RPCA		52.0	54.1	39.4	43.5
SCoTLASS	$t = 2.15$	24.3	24.6	23.4	30.7
	$t = 1.91$	34.5	34.6	21.8	23.5
	$t = 1.73$	41.0	41.3	22.5	19.0

underlying structure. It is preferred in many respects to rotated principal components, as a means of simplifying interpretation compared to principal component analysis. Although we are convinced of the value of SCoTLASS, there are a number of complications and open questions which are now listed as a set of remarks.

Remark 1. In this article we have carried out the techniques studied on correlation matrices. Although it is less common in practice, PCA and RPCA can also be implemented on covariance matrices. In this case PCA successively finds uncorrelated linear functions of the *original, unstandardized variables.* SCoTLASS can also be implemented in this case, the only difference being that the sample correlation matrix $\mathbf{R}$ is replaced by the sample covariance matrix $\mathbf{S}$ in equation (3.1). We have investigated covariance-based SCoTLASS, both for real examples, and using simulation studies. Some details of its performance are different from the correlation-based case, but qualitatively they are similar. In particular, there are a number of reasons to prefer SCoTLASS to RPCA.

Remark 2. In PCA, the constraint $\mathbf{a}_h'\mathbf{a}_k = 0$ (orthogonality of vectors of loadings) is equivalent to $\mathbf{a}_h'\mathbf{R}\mathbf{a}_k = 0$ (different components are uncorrelated). This equivalence is special to the PCs and is a consequence of the $\mathbf{a}_k$ being eigenvectors of $\mathbf{R}$. When we rotate the PC loadings we lose at least one of these two properties (Jolliffe 1995). Similarly, in SCoTLASS, if we impose $\mathbf{a}_h'\mathbf{a}_k = 0$ we no longer have uncorrelated components. For example, Table 12 gives the correlations between the six SCoTLASS components when $t = 1.75$, for the pitprop data. Although most of the correlations in Table 12 are small in absolute value, there are also nontrivial ones (r_{12}, r_{14}, r_{34}, and r_{35}).

Table 12. Correlation Matrix for the First Six SCoTLASS Components for $t = 1.75$ using Jeffers' Pitprop Data

	Correlation matrix					
Components	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
(1)	1.000	0.375	−0.234	0.443	−0.010	0.061
(2)		1.000	−0.114	−0.076	−0.145	−0.084
(3)			1.000	−0.438	0.445	−0.187
(4)				1.000	−0.105	0.141
(5)					1.000	−0.013
(6)						1.000

It is possible to replace the constraint $\mathbf{a}_h'\mathbf{a}_k = 0$ in SCoTLASS by $\mathbf{a}_h'\mathbf{R}\,\mathbf{a}_k = 0$, thus choosing to have uncorrelated components rather than orthogonal loadings, but this option is not explored in the present article.

Remark 3. The choice of t is clearly important in SCoTLASS. As t decreases, simplicity increases, but variation explained decreases, and we need to achieve a suitable tradeoff between these two properties. The correlation between components noted in the previous remark is another aspect of any tradeoff. Correlations are small for t close to $\sqrt{p}$, but have the potential to increase as t decreases. It might be possible to construct a criterion which defines the "best tradeoff," but there is no unique construction, because of the difficulty of deciding how to measure simplicity and how to combine variance, simplicity, and correlation. At present, it seems best to compute the SCoTLASS components for several values of t, and judge subjectively at what point a balance between these various aspects is achieved. In our example, we used the same value of t for all components in a data set, but varying t for different components is another possibility.

Remark 4. Our algorithms for SCoTLASS are slower than those for PCA. This is because SCoTLASS is implemented subject to an extra restriction on PCA and we lose the advantage of calculation via the singular value decomposition which makes the PCA algorithm fast. Sequential-based PCA with an extra constraint requires a good optimizer to produce a global optimum. In the implementation of SCoTLASS, a projected gradient method is used which is globally convergent and preserves accurately both the equality and inequality constraints. It should be noted that as t is reduced from $\sqrt{p}$ downwards towards unity the CPU time taken to optimize the objective function remains generally the same (11 sec on average for 1GHz PC), but as t decreases the algorithm becomes progressively prone to hit local minima and thus more (random) starts are required to find a global optimum. Osborne, Presnell, and Turlach (2000) gave an efficient procedure, based on convex programming and a dual problem, for implementing the LASSO in the regression context. Whether or not this approach can be usefully adapted to SCoTLASS will be investigated in further research. Although we are reasonably confident that our algorithm has found global optima in the example of Section 4 and in the simulations, there is no guarantee. The jumps that occur in some coefficients, such as the change in $a_{3,6}$ from -0.186 to 0.586 as t decreases from 1.75 to 1.50 in the pitprop data, could be due to one of the solutions being a local optimum. However, it seems more likely to us that it is caused by the change in the nature of the earlier components, which together with the orthogonality constraint imposed on the third component, opens up a different range of possibilities for the latter component. There is clearly much scope for further work on the implementation of SCoTLASS.

Remark 5. In a number of examples, not reported here, several of the nonzero loadings in the SCoTLASS components are exactly equal, especially for large values of p and small values of t. At present we have no explanation for this, but it deserves further investigation.

Remark 6. The way in which SCoTLASS sets some coefficients to zero is different in concept from simply truncating to zero the smallest coefficients in a PC. The latter attempts to approximate the PC by a simpler version and can have problems with that approximation, as shown by Cadima and Jolliffe (1995). SCoTLASS looks for simple sources of variation and, like PCA, aims for high variance, but because of simplicity considerations the simplified components can, in theory, be moderately different from the PCs. We seek to *replace* PCs, rather than *approximate* them, although because of the shared aim of high variance, the results will often not be too different.

Remark 7. There are a number of other recent developments which are relevant to interpretation problems in multivariate statistics. Jolliffe (2002), Chapter 11 reviews these in the context of PCA. Vines's (2000) use of only a discrete number of values for loadings has already been mentioned. It works well in some examples, but for the pitprops data the components 4–6 are rather complex. A number of aspects of the strategy for selecting a subset of variables were explored by Cadima and Jolliffe (1995, 2001), and by Tanaka and Mori (1997). The LASSO has been generalized in the regression context to so-called bridge estimation, in which the constraint $\sum_{j=1}^{p} |\beta_j| \leq t$ is replaced by $\sum_{j=1}^{p} |\beta_j|^{\gamma} \leq t$ where γ is not necessarily equal to unity—see, for example, Fu (1998). Tibshirani (1996) also mentioned the nonnegative garotte, due to Breiman (1995), as an alternative approach. Translation of the ideas of the bridge and nonnegative garotte to the context of PCA, and comparison with other techniques, would be of interest in future research.

ACKNOWLEDGMENTS

An earlier draft of this article was prepared while the first author was visiting the Bureau of Meteorology Research Centre (BMRC), Melbourne, Australia. He is grateful to BMRC for the support and facilities provided during his visit, and to the Leverhulme Trust for partially supporting the visit under their Study Abroad Fellowship scheme. Comments from two referees and the editor have helped to improve the clarity of the article.

[Received November 2000. Revised July 2002.]

REFERENCES

Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garotte," *Technometrics*, 37, 373–384.

Cadima, J., and Jolliffe, I. T. (1995), "Loadings and Correlations in the Interpretation of Principal Components," *Journal of Applied Statistics*, 22, 203–214.

——— (2001), "Variable Selection and the Interpretation of Principal Subspaces," *Journal of Agricultural, Biological, and Environmental Statistics*, 6, 62–79.

Chu, M. T., and Trendafilov, N. T. (2001), "The Orthogonally Constrained Regression Revisited," *Journal of Computational and Graphical Statistics*, 10, 1–26.

Fu, J. W. (1998), "Penalized Regression: The Bridge Versus the Lasso," *Journal of Computational and Graphical Statistics*, 7, 397–416.

Goffe, W. L., Ferrier, G. D., and Rogers, J. (1994), "Global Optimizations of Statistical Functions with Simulated Annealing," *Journal of Econometrics*, 60, 65–99.

Hausman, R. (1982), "Constrained Multivariate Analysis," in *Optimization in Statistics*, eds. S. H. Zanckis and J. S. Rustagi, Amsterdam: North Holland, pp. 137–151.

Helmke, U., and Moore, J. B. (1994), *Optimization and Dynamical Systems*, London: Springer.

Jeffers, J. N. R. (1967), "Two Case Studies in the Application of Principal Component Analysis," *Applied Statistics*, 16, 225–236.

Jolliffe, I. T. (1989), "Rotation of Ill-Defined Principal Components," *Applied Statistics*, 38, 139–147.

——— (1995), "Rotation of Principal Components: Choice of Normalization Constraints," *Journal of Applied Statistics*, 22, 29–35.

——— (2002), *Principal Component Analysis* (2nd ed.), New York: Springer-Verlag.

Jolliffe, I. T., and Uddin, M. (2000), "The Simplified Component Technique—An Alternative to Rotated Principal Components," *Journal of Computational and Graphical Statistics*, 9, 689–710.

Jolliffe I. T., Uddin, M., and Vines, S. K. (2002), "Simplified EOFs—Three Alternatives to Rotation," *Climate Research*, 20, 271–279.

Krzanowski, W. J., and Marriott, F. H. C. (1995), *Multivariate Analysis, Part II*, London: Arnold.

LeBlanc, M., and Tibshirani, R. (1998), "Monotone Shrinkage of Trees," *Journal of Computational and Graphical Statistics*, 7, 417–433.

Morton, S. C. (1989), "Interpretable Projection Pursuit," Technical Report 106, Department of Statistics, Stanford University.

McCabe, G. P. (1984), "Principal Variables," *Technometrics*, 26, 137–144.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000), "On the LASSO and its Dual," *Journal of Computational and Graphical Statistics*, 9, 319–337.

Tanaka, Y., and Mori, Y. (1997), "Principal Component Analysis Based on a Subset of Variables: Variable Selection and Sensitivity Analysis," *American Journal of Mathematical and Management Sciences*, 17, 61–89.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288.

Uddin, M. (1999), "Interpretation of Results from Simplified Principal Components," Ph.D. thesis, University of Aberdeen, Aberdeen, Scotland.

Vines, S. K. (2000), "Simple Principal Components," *Applied Statistics*, 49, 441–451.