Dimension reduction via principal variables

J A Cumming and D A Wooff

Department of Mathematical Sciences, Durham University, Science Laboratories, Stockton Road, Durham, DH1 3LE, United Kingdom

Received March 2006; revised October 2006

Abstract

For many large-scale data sets it is necessary to reduce dimensionality to the point where further exploration and analysis can take place. Principal variables are a subset of the original variables and preserve, to some extent, the structure and information carried by the original variables. Dimension reduction using principal variables is considered and a novel algorithm for determining such principal variables is proposed. This method is tested and compared with eleven other variable selection methods from the literature in a simulation study and is shown to be highly effective. Extensions to this procedure are also developed, including a method to determine longitudinal principal variables for repeated measures data, and a technique for incorporating utilities in order to modify the selection process. The method is further illustrated with real data sets, including some larger UK data relating to patient outcome after total knee replacement.

Key words: Variable selection, Principal components, Partial correlation, Partial covariance, Utility, Longitudinal data, Repeated measures

1 Introduction

The focus of this paper is on dimension reduction of a multivariate data set, constrained by the need to retain only a subset of the original variables for further investigation. Following [1], we shall term such a subset the *principal variables* (PVs) for the data set, under specified criteria. The particular advantage of employing PVs for dimension reduction is that any remaining variables can be discarded. Principal component analysis (PCA) is highly effective at

Email addresses: `j.a.cumming@durham.ac.uk` (J A Cumming), `d.a.wooff@durham.ac.uk` (D A Wooff).

providing a reduced-dimension representation, but typically still requires all the original variables. McCabe [1] discusses the deficiencies of PCA methods in this context, and lays the groundwork for generating PVs under a number of optimality criteria. The selection of subsets of original variables, for example for applications in regression modelling, has a long and ongoing history [2]. The approach described here is not tied to any such application, but was motivated by the need to reduce dimension before carrying out longitudinal chain graph modelling. Such a requirement is increasingly a feature of many application areas, particularly those involving very large data sets with thousands of variables.

We begin in Section 3 by describing existing techniques for selecting PVs. In Section 4 we show that the usual spectral decomposition leads to an alternative criterion for variable selection. In Section 5 we describe a stepwise algorithm for variable selection using this criterion. In Section 6 we show how the method can be extended to take into account utilities expressed over the variables, and we discuss constructing PVs in the situation where there are repeated measurements on the full set of variables. In Section 7, scree-type plots are introduced to help decide on the appropriate number of PVs to describe adequately the original data set. Simulation studies showing the methodology to be generally superior to extant methods are shown in Section 8. An example from the literature and a case study are provided in Section 9. Algorithms are presented in an appendix.

2 Preliminaries

In what follows, we suppose that we collect n observations on a $p-$dimensional measurement vector into the $n \times p$ data matrix $\boldsymbol{X}$. Suppose that the sample covariance and correlation matrices are respectively $\boldsymbol{\Sigma}$ and $\boldsymbol{R}$. Our aim is to select some subset of m PVs, $m < p$, which best (in some sense) represents the original variables. Suppose that we partition the set of variables $\boldsymbol{V}$ into subsets $\boldsymbol{V}_{(1)}, \boldsymbol{V}_{(2)}$. It will be helpful to consider partitioning $\boldsymbol{\Sigma}$ correspondingly as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Then, the partial covariance matrix for $\boldsymbol{V}_{(2)}$ given $\boldsymbol{V}_{(1)}$ is

$$\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}, \tag{1}$$

and the partial correlation matrix, $\boldsymbol{R}_{22.1}$, is obtained by scaling $\boldsymbol{\Sigma}_{22.1}$ so that diagonal elements are unity. If we begin with a correlation matrix $\boldsymbol{R}$, then we further define the *unscaled* partial correlation matrix:

$$\tilde{\boldsymbol{S}}_{22.1} = \boldsymbol{R}_{22} - \boldsymbol{R}_{21}\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}. \tag{2}$$

3 Existing techniques

Jolliffe [3,4] discusses various ad hoc techniques for PV selection based on PCA. His methods **B1**, **B2**, **B4** associate a single variable with each of the PCs of $\boldsymbol{\Sigma}$ (or, by choice, $\boldsymbol{R}$). Jolliffe's method **B3** was shown in [3] to dismiss the wrong variables from data of a very simple type; it is not considered here. The associated variable is that which has the largest absolute loading in the PC under consideration. Methods **B1** and **B2** correspond to a backward elimination procedure where the variables associated with the final components are excluded until only m remain: method **B1** iteratively repeats the PCA on each remaining subset of variables, whereas method **B2** depends only on an initial PCA. Method **B4** corresponds to forward selection by associating and retaining variables with high loadings in absolute value on the first m PCs. These methods were shown to be both fast and efficient via a simulation study. However, the mechanism of selection can be seriously unreliable because it depends only on loadings, while neglecting not only variances of variables and components but also the patterns of correlations between them [5].

McCabe [1] considered twelve apparently different criteria used to drive variable selection processes, and showed that all twelve correspond to one of the following four criteria:

$$\mathbf{M1}.\, \max|\boldsymbol{\Sigma}_{11}| \equiv \min|\boldsymbol{\Sigma}_{22\cdot1}| \equiv \min\prod_i \lambda_i, \tag{3}$$

$$\mathbf{M2}.\, \min \operatorname{tr}(\boldsymbol{\Sigma}_{22\cdot1}) \equiv \min\sum_i \lambda_i, \tag{4}$$

$$\mathbf{M3}.\, \min||\boldsymbol{\Sigma}_{22\cdot1}||^2 \equiv \min\sum_i \lambda_i^2, \tag{5}$$

$$\mathbf{M4}.\, \max\sum_{i=1}^{k} \rho_i^2, \text{with } k = \min(m, p-m). \tag{6}$$

Here, $\boldsymbol{\Sigma}_{11}$ is the covariance matrix for the selected subset of PVs; $\boldsymbol{\Sigma}_{22\cdot1}$ is the conditional covariance matrix of the variables not selected given those selected; $|\boldsymbol{A}|$ and $\operatorname{tr}(\boldsymbol{A})$ are the determinant and trace of the matrix $\boldsymbol{A}$, respectively; $||\boldsymbol{A}||^2$ is the squared norm $(\sum\sum a_{ij}^2)$; λ_i are the eigenvalues of $\boldsymbol{\Sigma}_{22\cdot1}$; and the ρ_i are the canonical correlations between the variables not selected and those selected. As McCabe points out, after the selection of the principal variables $\boldsymbol{\Sigma}_{22\cdot1}$ represents the information left in the remaining unselected variables and so it is quite plausible that three of the optimality criteria should be functions of this matrix. Only for **M2** can we find near-optimal subsets by stepwise selection using results from [6] involving variable variances and squared multiple correlation coefficients - we call this method **M2S**. Optimal subsets under the other criteria can be found only by exhaustive evaluation of all candidate subsets. This rapidly becomes computationally infeasible as the number of variables increases.

Recent work by Al-Kandari and Jolliffe [7,8] has investigated and compared the performance of some of the McCabe and Jolliffe methods discussed above on simulated data of varying degrees of dependence among the variables. They also introduced methods $\mathbf{M}_1^*, \mathbf{M}_3^*$, which were variants of McCabe's **M1**, **M3** whose criteria replaced the partial covariance matrix by a partial correlation matrix. They evaluated the selection methods on these different artificial data sets and assessed them using a variety of different performance measures. Their results showed that the efficiency of the various selection methods is dependent on the performance criterion and furthermore that it may not be wise to rely on a single method for variable selection.

Beale et al.[9] discuss a method for discarding variables based upon multiple correlation (method **A1**). They suggested retaining the subset of m variables which maximises the minimum multiple correlation between the m selected variables and any of the remaining variables. At the time this method was found to be too slow to be practically useful [4,9], as it requires exhaustive enumeration and evaluation of all subsets of size m. Jolliffe instead proposed an alternative stepwise version (**A2**) whereby at each stage the variable with the highest multiple correlation with the remaining variables was excluded until only m variables remain [4].

Krzanowski [10] proposed an approach (**KP**) based on Procrustes Analysis: this method intends explicitly to preserve the multivariate *structure* of the original data in the final variable subset, rather than selecting a set which seeks to maximise some variance measure over the variables. To assess variable subsets, the data points were first transformed to principal component space using a PCA of all the variables. They were then transformed to a PCA-space based on a reduced subset of the variables, and the sum of squared differences between data points in these two configurations was used to evaluate the variable subset.

De Falguerolles et al.[11] proposed a method based on graphical Gaussian models (**DF**). This seeks to choose as a subset those variables which appear to form a highly connected hub within a graphical model: the selected variables then should have many connections to other variables and should leave the unselected variables conditionally independent given those selected. To do this they seek the variable subset that minimises the deviance of this hypothesised model from the saturated model via the expression:

$$D^2 = -n \log \left(\left| \tilde{\boldsymbol{S}}_{22\cdot1} \right| / \left| \operatorname{diag}(\tilde{\boldsymbol{S}}_{22\cdot1}) \right| \right),$$

where n is the sample size and $\tilde{\boldsymbol{S}}_{22.1}$ is defined in (2). Such an approach is likely to miss important structure: for example, independent variables are unconnected with others, and will be dismissed from selection.

4 An alternative criterion

Consider the $(p \times p)$ correlation matrix $\boldsymbol{R}$, which we shall assume full rank for convenience. It can be expressed in terms of its spectral decomposition as:

$$\boldsymbol{R} = \sum_{i=1}^{p} \lambda_i \boldsymbol{a}_i \boldsymbol{a}_i^T = \boldsymbol{A} \boldsymbol{\Lambda} \boldsymbol{A}^T,$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ are the ordered eigenvalues of $\boldsymbol{R}$ and $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p$ the associated eigenvectors. $\boldsymbol{A}$ is then the $(p \times p)$ orthonormal matrix whose columns are the $\boldsymbol{a}_i$ and $\boldsymbol{\Lambda}$ is the $(p \times p)$ diagonal matrix with entries λ_i. Suppose that we focus on criterion **M3**, which utilises $||\boldsymbol{R}||^2$. We may write this in a number of different but equivalent forms, two being:

$$\begin{aligned} ||\boldsymbol{R}||^2 &= \sum_{i=1}^{p} \lambda_i^2 && (7) \\ &= \sum_{j=1}^{p} \sum_{i=1}^{p} (\lambda_i a_{ji})^2 = \sum_{j=1}^{p} h_j = \sum_{j=1}^{p} (\sum_{i=1}^{p} r_{ij}^2), && (8) \end{aligned}$$

where

$$h_j = \sum_{i=1}^{p} r_{ij}^2 = \sum_{i=1}^{p} (\lambda_i a_{ji})^2. \qquad (9)$$

Decomposition (7) makes plain that the first PC provides the linear combination of original variables with maximum contribution to $||\boldsymbol{R}||^2$, and with the remaining PCs giving progressively less. This is the basis for Jolliffe's variable selection methods **B1**, **B2**, **B4**: method **B4** selects variables which have high loadings on the most important PCs, and methods **B1** and **B2** reject variables with high loadings in the least important PCs. Decomposition (8) suggests that we may instead examine the values $h_1, \ldots, h_p$, which are the sum of the squared correlations between variable v_j and other variables. Large values of h_j are obtained when variable v_j has, on average, high loadings on important PCs. As the values $\{h_j\}_{j=1}^{p}$ combine information from both the eigenvalues and the loadings, we would expect choices based on them to be more robust to the sensitivity issues raised by selection via single loadings on important PCs [5].

5 Stepwise selection using h and partial covariance

A simple reduced-dimension representation is obtained by choosing the m variables with highest values of h_j, such that $\sum_{j=1}^{m} h_j$ reaches some predetermined proportionate threshold. A more sophisticated approach is to select the best variable $v_{(1)}$, compute the partial covariance matrix for the unselected variables, given $v_{(1)}$, and apply the same strategy iteratively until the specified threshold is reached. By using the partial covariances of the variables under consideration given those already selected, we eliminate the effects of those chosen variables from subsequent analysis and so compensate for them when determining the next variable for selection. This allows us to account for and exploit the multivariate correlation structure that may exist among the variables. By iteratively taking the partial covariance given the subset of selected variables, we ensure that at each stage of selection process we choose a variable which captures aspects of the variation that are not represented by those variables already selected.

The stepwise selection scheme is as follows. We typically begin with the correlation matrix, $\boldsymbol{R}$, rather than the raw covariance matrix, $\boldsymbol{\Sigma}$, in order to remove initial scale effects. We determine the h_j values for each variable v_j for $j = 1, \ldots, p$. We identify that variable with the largest h_j value and select it. We then form the unscaled partial correlation matrix $\tilde{\boldsymbol{S}}_{22.1}$ using (2) for the remaining variables given the variable(s) we have already selected. The process then repeats: we calculate h values, identify candidate variables, select the best variable and compute a partial correlation matrix for the remaining candidates. The full algorithm is presented in Figure A.1, and labelled **H** henceforth.

Once the first variable is selected and we have transformed to partial form, it is not sensible to re-scale the resulting partial covariance matrix to correlation form. The reason for this is that the rescaling artificially inflates the h scores for variables highly correlated with those already selected. The consequence is that correlated blocks of variables would then be selected in preference to independent variables, whereas our desire is to reduce dimension and preserve structure - further details may be found in [12]. Consequently, the **H** method does not dismiss uncorrelated variables. In general, it selects a single variable from each block of correlated variables before proceeding to select the uncorrelated variables. Continuing selection past this point would introduce redundancies into the selected subset. Rescaling at each stage to partial correlation form (as in methods $\mathbf{M}_1^*, \mathbf{M}_3^*$) rather than using $\tilde{\boldsymbol{S}}$ tends to dismiss the uncorrelated variables in favour of selecting many variables from a correlated block.

6 Extensions

6.1 Repeated measures principal variables

For some applications we obtain repeated measurements on the variables of interest. For example, patients are routinely monitored for the same set of variables before, during, and after treatment. We can apply a dimension reduction technique at each time point, but it is often preferable to determine a single subset of *longitudinal* PVs to represent the full time spectrum.

One of the main difficulties in generating temporal PCs is to account for both the temporal and multivariate nature of the data, namely correlation structure among variables at one time point and their associations between time points. Methods based on stationary time series [4] are typically inappropriate: stationarity is often an unreasonable assumption, and there are typically too few time points. Berkey et al. [13] discuss a *longitudinal principal components* regression model which performs PCA on the various observations of a single variable over time and then uses the resulting PCs as predictors in a linear model. However, this method ignores the multivariate nature of the data. *Functional data analysis* [14] can be used to determine functional, and thus temporal, forms for PCs, but only where there are many observations.

Here we adopt a method proposed by Prvan and Bowman [15], based on the nonparametric time-dependent PCA. Suppose we have a data matrix $\boldsymbol{X}$ containing n cases $\boldsymbol{x}_i$, $i = 1, \ldots, n$. Each x_i is observed at a time point $t_i \in T$, where T is the set of all time points at which data is observed. In the case of repeated measures data, the same case will be repeatedly observed at different time points. First, we choose a focal time θ. We associate with each $\boldsymbol{x}_i$ a weight ω_i defined as:

$$\omega_i = \omega(t_i, \theta, \sigma) = \phi\left(\frac{t_i - \theta}{\sigma}\right), \tag{10}$$

where σ is a bandwidth parameter and $\phi(\cdot)$ is the standard Normal density function. We then construct a weighted mean $\bar{\boldsymbol{x}}_\omega(\theta)$ and a weighted covariance matrix $\boldsymbol{S}_\omega(\theta)$ as:

$$\bar{\boldsymbol{x}}_\omega(\theta) = \frac{1}{\sum_{i=1}^n \omega_i} \sum_{i=1}^n \omega_i \boldsymbol{x}_i, \tag{11}$$

$$\boldsymbol{S}_\omega(\theta) = \frac{1}{\sum_{i=1}^n \omega_i} \sum_{i=1}^n \omega_i (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_\omega(\theta))(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_\omega(\theta))^T. \tag{12}$$

The weighted correlation matrix $\boldsymbol{R}_\omega(\theta)$ is found by re-scaling $\boldsymbol{S}_\omega(\theta)$ in the usual way. This yields a matrix of correlations over the variables in the data relative to a particular time point θ. This method allows all the data to influence the value of the mean and variance at time θ. Data observed at a time close to θ have a strong influence on $\bar{\boldsymbol{x}}_\omega(\theta)$ and $\boldsymbol{S}_\omega(\theta)$, whereas data observed at more distant times exert a weaker influence. The magnitude of the effect of temporally adjacent data and the distance in time over which it applies is governed by the bandwidth parameter σ. The choice of σ is typically subjective and is based on the plots discussed in [15].

To determine an overall subset of longitudinal PVs, we construct $\boldsymbol{R}_\omega(t_i)$ for each time point t_i for which we make a measurement. We calculate h_{j,t_i} (9) for each variable v_j at each time point t_i. A simple guide to the selection value of each variable across all time points is then the total $h_j^T = \sum_i h_{j,t_i}$. All of the correlation matrices are then updated to partial covariance form given the variable we have selected. The full temporal algorithm is detailed in Figure A.2, and labelled henceforth as **HT**.

This procedure is reasonable for determining longitudinal PVs if we believe that the multivariate structure is preserved across time points, excepting random fluctuation. This could be checked formally by sphericity-type tests [2], as long as we were prepared to make further distributional assumptions. Informally, we can examine the plots used for choosing σ [15] to assess this. When the multivariate structure changes over time, more PVs will need to be extracted to adequately represent the data at all time points. Furthermore, if the sample size is not constant at all time points, it may then also be advisable to weight the h statistics to reflect this. The precise impact of variation in structure and sample sizes across time points is an area for future research.

6.2 Utility information

For some applications, it may be useful for external information to influence the selection process. Such information typically attaches to specific variables. For example, in a medical context a clinician will have an opinion on the relative merit of particular measurements for diagnosis or monitoring of a patient: some measurements may be deemed clinically more useful than others; some may be easier to measure than others. It is straightforward to incorporate such information into the selection algorithms presented here.

We will suppose that information concerning the desirability of retaining a variable in the final subset is expressed on a utility scale. Suppose that utility u_j is the utility for retaining variable v_j. A simple way of modifying the selection process is then to replace h_j by $h_j^U = u_j h_j$ and to select at each stage the variable with maximum h_j^U. When there are multiple utility scales, we employ *utilitarianism* [16], whereby individual utilities are simply summed. Thus from several utility vectors $\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(r)}$, we construct $\boldsymbol{u}^* = \sum_k \boldsymbol{u}^{(k)}$ to aggregate the individual components and then consider the product $u_j^* h_j$ for selection. This extends easily to constructing longitudinal PVs.

7 Diagnostics: scree plots

The scree plot [17] is a popular method for assessing the number of PCs which represent non-random structure in the data. The variances for the PCs are plotted in descending order of magnitude and one looks for the PC beyond which the variances decrease in a linear fashion. The components beyond this point are taken to be consonant with random noise. An equivalent plot can be drawn for the selection procedure outlined above. We have a choice of diagnostics to plot. First, if we select variable $v_{(\ell)}$ at stage ℓ with an associated value of $h_{(\ell)}$, then we may plot $h_{(\ell)}$ vs. ℓ. Note that if we start with a correlation matrix $\boldsymbol{R}$, then variables with $h_{(\ell)} < 1$ convey less information than a single independent quantity. Secondly, we may plot $\mathrm{tr}\left(\tilde{\boldsymbol{S}}_{22\cdot 1}^{(\ell)}\right)$ or $||\tilde{\boldsymbol{S}}_{22\cdot 1}^{(\ell)}||^2$ as the selection procedure continues. Thirdly, we may plot the cumulative proportion of the total variability explained. This is expressible in terms of $\mathrm{tr}\left(\tilde{\boldsymbol{S}}_{22\cdot 1}\right)$. We thus plot at stage ℓ:

$$\pi_\ell = \left(1 - \mathrm{tr}\left(\tilde{\boldsymbol{S}}_{22\cdot 1}^{(\ell)}\right) / \mathrm{tr}\left(\boldsymbol{R}\right)\right), \qquad (13)$$

where $\boldsymbol{R}$ is the initial correlation matrix, and $\tilde{\boldsymbol{S}}_{22\cdot 1}^{(\ell)}$ is the matrix (2) given selection of $v_{(1)}, \ldots, v_{(\ell)}$. Equivalently, we could produce a similar plot using $|| \cdot ||^2$ instead of $\mathrm{tr}(\cdot)$ since the squared norm is intimately related to the selection process.

8 Simulation results

8.1 Simple models

A Monte Carlo simulation study was performed to compare some of the selection methods discussed in Section 3. This study was the same as that proposed and used by Jolliffe [3], and later Krzanowski [10]. The study was performed in four parts, each part testing performance on a simulated data set conforming to a different pre-determined model. Each part was then repeated 500 times, each time re-simulating the data to allow sample variation. Models were constructed so that some variables were linear combinations of the others plus random noise: such variables are redundant. Given this knowledge of the data structure, we may determine which variables should be selected. Each returned subset may thus be classified as "Best", "Good", "Moderate" or "Bad": see [3] for details of the model specification and subset classification. All calculations were run in R for Windows version 2.1.1 [18] on a Pentium IV 2.4GHz PC with 1.5Gb RAM.

For each iteration of the simulation on a particular model, a data set of size $n = 100$ was generated containing between six and ten variables. These data conformed to the models *I–IV* defined by Jolliffe [3]. The structure of these models is such that the first contains a set of three pairs of variables; the second is a pair, a triple and a single independent variable; the third is effectively three pairs though with stronger correlations elsewhere; and the fourth is composed of a single variable, a pair, a triple and a quadruple. The first three models have three substantive variables, whereas the fourth has four.

For each generated data set, we applied twelve of the different selection methods: 11 taken from those in Section 3, plus the method **H** proposed here. The results are presented in Table 1 in the form of the percentage of different classes of subsets returned by the various selection methods over 500 simulations under the four models. Some of the methods (**M1**, **M2**, **M3**, **A1**, **DF**) are exhaustive in nature and hence require significant, normally infeasible, computation to enumerate and evaluate all subsets. Method $\mathbf{M}_3^*$ proposed in [8] was also tested on these data but returned a "Bad" subset for every run on each of the models; those results are therefore omitted from the table. This failure is likely attributable to the problems of re-scaling to correlation form discussed in Section 5.

Method **M1** appears inferior to **M2**, **M3**. With regard to methods based on PCA, method **B1** shows a poorer overall performance than **B2** due to its failures under Model *IV* and relatively poor performance on Model *III*. This is unexpected as **B1** repeated the PCA at each stage, which is assumed preferable to performing it only once as in **B2**. However, it is likely the case that re-performing the PCA on the reduced subset of variables is not sufficient to accommodate for the removal of each variable. Performing the PCA on the partial covariance/correlation matrix may serve to boost the performance here. However, both methods retain subsets that are "Good" or better for more than 90% of the simulations. Method **B4**, a forward selection method, is slightly more inconsistent.

Method **A1** consistently ignores single independent variables due to their low predictive power and negligible impact on the multiple correlation and so displays poor performance on models *II* and *IV*. Methods **DF** and **M2S** display

Table 1. Percentage performance of selection methods for the simulated data of Models *I–IV*

Model	*Subset*	*Selection method*											
	type	**M1**	**M2**	**M3**	**M2S**	**B1**	**B2**	**B4**	**A1**	**A2**	**KP**	**DF**	**H**
I	Best	100	100	100	57.8	100	99.4	74.2	100	100	100	100	100
	Bad	0	0	0	42.2	0	0.6	25.8	0	0	0	0	0
II	Best	29.8	99.4	99.4	4.4	0.2	0.6	97.0	0	0.2	4.6	22.4	99.2
	Good	70.2	0.6	0.6	12.0	99.8	99.4	3.0	0	99.8	95.4	0	0.8
	Bad	0	0	0	83.6	0	0	0	100	0	0.0	77.6	0
III	Best	47.6	100	99.8	8.4	38.6	76.6	35.2	35.6	38.4	70.8	51.6	32.8
	Good	52.4	0	0.2	17.8	61.4	0.2	4.8	64.4	61.6	29.2	48.4	67.2
	Mod.	0	0	0	72.4	0	20.6	55.6	0	0	0	0	0
	Bad	0	0	0	1.4	0	2.6	4.4	0	0	0	0	0
IV	Best	100	100	100	54.2	86	99.8	100	0	100	100	0	100
	Bad	0	0	0	45.8	14	0.2	0	100	0	0	100	0
Overall	Best	69.4	99.9	99.8	31.2	56.2	69.1	76.6	33.9	59.7	68.9	43.5	83.0
	Good	30.6	0.1	7.5	0.2	40.3	24.9	2.0	16.1	40.3	31.1	12.1	17.0
	Mod.	0.0	0.0	0.0	18.1	0.0	5.2	13.9	0.0	0.0	0.0	0	0.0
	Bad	0.0	0.0	0.0	43.2	3.5	0.8	7.5	50.0	0.0	0.0	44.4	0.0

similar behaviour, though less consistently rejecting the independent variable. **A2** performs quite well and with consistency. Krzanowski's Procrustes method **KP** is generally very good, rather better indeed than originally reported [10], and seems the best stepwise method for Model III.

The proposed method **H** performs generally very well: in 83% of cases producing the "Best" subsets, and never returning a subset worse than "Good", and with performance better than that of McCabe's optimal solution **M1**. Its performance on Model III is not quite as good as for some other methods. However, **H** is overall the best stepwise variable selection method of those studied.

8.2 Structured models

Krzanowski's **KP** method aims at determining a subset which preserves the original structural features of the data. To test this method he performed a modified version of the simulation performed above by building additional structure into the data and examining whether the selected subsets contained variables which conveyed this structure. Following the configuration of the models from the previous study, additional structure (single outlier, weak groups, strong groups) was added to a different number of variables (1,2,3) in the data (see [10] for details). Variable selection was then carried out on these modified data, selecting subsets of the appropriate dimension. Each subset was then examined to determine how many of the structure-bearing variables were present. Running this simulation 100 times for each combination of the type of structure, amount of structure and model type using method **H** generated the results presented in Table 2. Krzanowski's results are presented alongside for comparison.

The reasoning behind performing this simulation with structured data is to ascertain whether the 'structure' inserted into several variables is preserved in the reduced subset through the selection of these structure-laden variables. As the strength of the structure in the data increases, it becomes increasingly likely for method **KP** to select structure-bearing variables, especially for weak and strong groups. However, for the **H** method we find that for both weak and strong groups on two or three variables typically only one structure-bearing variable is selected. The effect of adding this structure to the data is equivalent to there being an underlying latent variable representing the structure with the variables themselves being noisy realisations of the latent variable. Procedure **H** detects the introduced structure as a correlated block, extracts a single representative variable and leaves the other structure variables effectively redundant and undesirable for selection. The results also show that of the methods examined, **KP** is quite sensitive to contamination by outliers.

Table 2
The number of times a structure-bearing variable is selected for each model with additional structure.

Type	*Num. str.*	*Num. of*	*Model*								*Overall*	
	vars. in	*str. vars.*	*I*		*II*		*III*		*IV*			
	data	*selected*	**H**	**KP**	**H**	**KP**	**H**	**KP**	**H**	**KP**	**H**	**KP**
Single Outlier	1	0	47	0	48	0	2	9	0	0	24.25	2.25
		1	53	100	52	100	98	91	100	100	75.75	97.75
	2	0	0	0	0	0	3	2	0	0	0.75	0.5
		1	99	6	98	15	83	75	55	97	83.75	48.25
		2	1	94	2	85	14	23	45	3	15.5	51.25
	3	0	0	0	0	0	0	0	0	0	0	0
		1	98	6	44	20	78	6	0	0	55.0	8
		2	2	84	49	72	21	84	100	100	43.0	85
		3	0	10	7	8	1	10	0	0	2.0	7
Weak Groups	1	0	46	0	48	0	0	0	0	0	23.5	0
		1	54	100	52	100	100	100	100	100	76.5	100
	2	0	0	0	0	0	0	0	0	0	0	0
		1	100	0	100	0	100	0	100	0	100	0
		2	0	100	0	100	0	100	0	100	0	100
	3	0	0	0	0	0	0	0	0	0	0	0
		1	100	0	100	0	100	0	93	0	98.25	0
		2	0	1	0	0	0	11	7	100	1.75	28
		3	0	99	0	100	0	89	0	0	0	72
Strong Groups	1	0	42	0	32	0	0	0	0	0	18.5	0
		1	58	100	68	100	100	100	100	100	81.5	100
	2	0	0	0	0	0	0	0	0	0	0	0
		1	100	0	100	0	100	0	100	0	100	0
		2	0	100	0	100	0	100	0	100	0	100
	3	0	0	0	0	0	0	0	0	0	0	0
		1	100	0	100	0	100	0	100	0	100	0
		2	0	0	0	0	0	0	0	0	0	0
		3	0	100	0	100	0	100	0	100	0	100

9 Variable selection applied to real data

9.1 Aphids data

This data set consists of 19 variables measured on 40 winged aphids: these data have often been examined in a variable reduction setting [19,10,4,20]. A correlation plot [21] is shown in Figure 1. The minimum number of variables required to describe adequately the data was suggest by Krzanowski [10] to be four using a cross-validatory approach. We shall thus extract four PVs: the

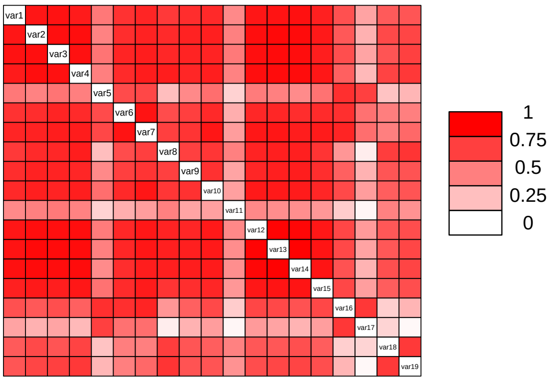


Fig. 1. Correlation plot for the correlation matrix of the aphids data set.

Table 3
Selected four-variable subsets for Jeffer's aphid data using various selection methods

	Variable														% tr $(\boldsymbol{R})$	% $\|\|\boldsymbol{R}\|\|^2$	
	2	3	5	6	7	8	9	11	12	13	14	16	17	18	19		
M1							×	×		×			×		×	82.2	98.3
M2			×					×		×				×		89.8	99.8
M3			×					×		×				×		89.8	99.8
M2S		×	×					×					×			86.0	99.3
B1			×				×	×							×	81.7	98.1
B2	×		×			×		×								85.5	99.3
B4			×					×		×			×			89.1	99.7
A1					×				×	×	×					79.6	98.6
A2			×	×				×					×			78.2	96.6
KP			×						×		×			×		86.3	99.5
DF					×					×		×		×		86.4	99.5
H			×					×		×			×			89.1	99.7

results are summarised in Table 3, with corresponding value of the percentage variance in terms of tr $(\boldsymbol{R})$. Overall, there is little practical difference in the subsets returned by the different methods, with each returning a subset that represents 80–90% of the variation. The exhaustive procedures **M2**, **M3** do best, closely followed by stepwise procedures **H**, and **B4**.

9.2 Knee replacement data

This data set consists of data on 599 patients who underwent a total knee replacement procedure between 1987 and 1997 [22]. There are 19 repeated measurement variables representing patient status (see Table 4). These measurements were collected through a series of up to four consultations between the patient and clinician and occurred pre-operatively and then at one, five and ten years post-operatively. Not all patients remained in the study up to the 10-year point. The number of patients under study were (599, 599, 239, 86) at (the pre-operative assessment, one year, five years, ten years). Many of the variables are ordinal, for example pain is scored on a five-point scale. We have assumed in such cases that it is reasonable to treat these as interval-scale

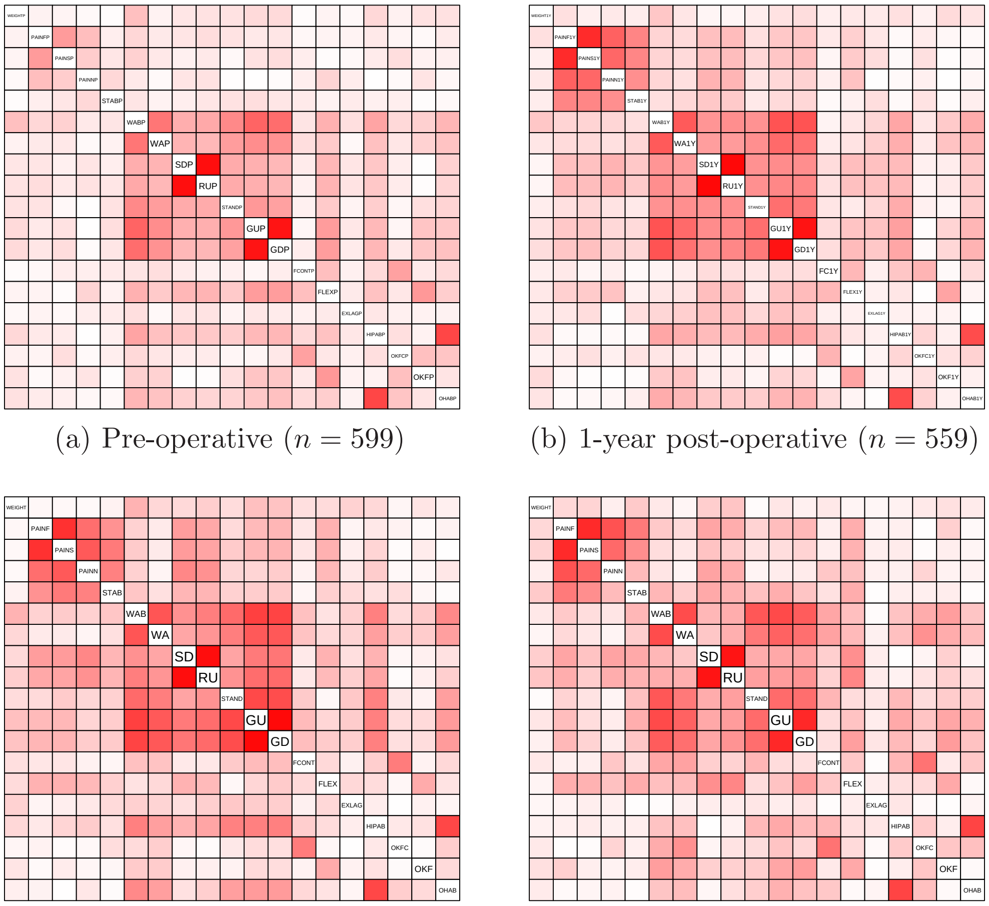
Table 4
Variables for the knees data, with two utility scores.

Name	Util Ease	Util Use	Description
Weight	9	7	Patient weight
PainF	7	8	Pain frequency
PainS	6	9	Pain severity
PainN	7	7	Night pain
Stab	5	6	Stability
WAb	8	5	Walking ability
WA	8	3	Walking aids
SD	6	7	Sitting down
RU	6	8	Rising up
Stand	5	4	Standing
GU	4	7	Going up stairs
GD	4	5	Going down stairs
FCont	4	5	Fixed contracture
Flex	4	7	Flexion
ExLag	4	5	Extension Lag
HipAb	4	4	Hip Abduction
OKF	4	4	Flexion of other knee
OKFC	4	6	Fixed contracture of other knee
OHAB	4	4	Abduction of other hip

data.

Correlation plots for the data at each time point are presented in Figure 2. Visually, the correlation structure displays some similarity over the four time points, perhaps with more noise at later time points, possibly due to the reduction in the sample size. There are two or three blocks of higher correlation. This is unsurprising: the clinical measurements contain several near surrogates, for example there are some different walking-ability measurements, several measurements of different aspects of pain, in addition to technical measurements such as angles of knee flexion. Thus, we would expect to see blocks of variables with high internal correlation.

Assessments of the intrinsic dimensionality (for a review, see [23]) of the data were varied. Kaiser's rule [24] using eigenvalue thresholds of 1 and 0.7 gave dimensions of 7 and 16 respectively. At least 75% of the variation is explained by the first 4 PCs at all time points. A scree plot [17] shows that the first PC is very important, and that there is a linear trend in the plot starting at the 7th component, suggestion dimension 6 or 7. The 'broken stick' method [25], suggests an intrinsic dimensionality of 7. Velicer's method [26] suggests



(a) Pre-operative ($n = 599$)

(b) 1-year post-operative ($n = 559$)

(c) 5-year post-operative ($n = 239$)

(d) 10-years post-operative ($n = 86$)

Fig. 2. Absolute value correlation plots of the repeated measurements in the knees data observed at each of the four time points

a dimension of 13. For illustration, we will extract variable subsets of size 7.

9.2.1 Reducing the pre-operative data

We first illustrate basic variable selection. Subsets derived from the alternative methods are shown in Table 5, together with the corresponding percentages of tr $(\boldsymbol{R})$ and $||\boldsymbol{R}||^2$ that are explained, and the computer time involved. Method **A1** was excluded from this analysis due to its prohibitively slow execution time and its previously poor performance. For the stepwise selection methods **M2S**, **B4**, **H**, and **A2** the variables are listed in the order of their selection, the results for the other variable selection or reduction methods are unordered. The exhaustive methods **M1**, **M2**, **M3**, and **DF** required the enumeration and evaluation of all 50388 possible seven-variable subsets. The **KP** method requires computation for many $(n \times n)$ matrices, which in this example has over 350000 elements.

Most of the extracted subsets convey more than 55% of $\mathrm{tr}\,(\boldsymbol{R})$and 75% of $||\boldsymbol{R}||^2$. The exceptions are methods **M1**, **B1** and **A2**: these have previously been judged rather inferior to other methods. Otherwise, there appears no great difference between the quality of solutions returned by the exhaustive optimal methods and the stepwise non-optimal methods. There is much overlap in the composition of the subsets returned. In terms of the variable groupings that were seen in the correlation plots in Figure 2, we observe that all subsets contain one variable from the tightly correlated pair (*Sitting Down*, *Rising Up*) and all methods but the poorly performing **M1**, **M2S**, **B1** and **A2** return a subset containing one variable from the pair (*Going Up Stairs*, *Going Down Stairs*). Similarly, all methods return at least one of the three pain score measurements with methods **M2S**, **A2** and **KP** both returning two. Method **H** has marginally the best performance of the non-exhaustive methods.

9.2.2 Longitudinal extraction

As a precursor to finding longitudinal PVs, we first apply stepwise variable selection procedure **H** separately to the data set for each time point. The aim is to provide baselines to which we may compare the final longitudinal PVs. The results of these variable selections are presented in the top four rows of Table 6. There is substantial agreement between the variables selected at each time point, suggesting that the underlying multivariate structure is preserved over time. The merit of these subsets, as defined by the percentage trace or squared norm of the original correlation matrix, is listed in the final two columns of Table 6. We can see that the extracted subsets for later time points explain a greater percentage of variation at that time point. A clinical interpretation is that the better performance for the post-operative data is likely due to the fact that the pain scores are somewhat more correlated post-operatively.

Next, the temporal selection method **HT** was applied to all four time points simultaneously to obtain a single subset for all the data. Visual examination of smoothing plots for the temporal PCA (see Section 6.1) suggested using a smoothing bandwidth of $\sigma = 1$ year. The longitudinal PVs extracted are listed in the bottom row of Table 6. The performance of this subset is listed in terms of the percentage variation of the *original* (unsmoothed) correlation matrices of the data at each time point. The performance on the data from each time point is encouraging, and slightly exceeds that of the individual subsets for the 1- and 5-year data. This is likely due to certain combinations of variables yielding better results than would be expected when running a simple stepwise procedure. Performing an exhaustive search would, of course, identify these best combinations, but would suffer from the problems involved with exhaustive methods. The matrix of partial variances of the remaining variables of the knees data given the selected 7-variable subset is illustrated

Table 5. The selected 7-variable subsets of the pre-operative knees data using various selection methods

	Variable									
Method	*1*	*2*	*3*	*4*	*5*	*6*	*7*	% tr $(\boldsymbol{R})$	% $\|\|\boldsymbol{R}\|\|^2$	*Time (s)*
M1	Weight	PainF	Stab	SD	FCont	ExLag	OKF	50.8	65.0	120.45
M2	PainS	PainN	RU	GU	FCont	OKF	OHAb	58.5	82.0	108.97
M3	PainS	PainN	RU	GU	FCont	OKF	OHAb	58.5	82.0	88.15
M2S	WAb	SD	OKF	PainS	OHAb	ExLag	PainN	55.7	78.6	17.52
B1	Weight	PainF	Stab	SD	FCont	ExLag	OKF	50.8	65.0	0.03
B2	Weight	PainN	SD	GD	FCont	OKF	OHAb	57.5	80.3	0.02
B4	GU	OKFC	PainS	OHAB	SD	ExLag	Weight	57.7	80.4	0.00
A2	Weight	PainS	Stab	SD	FCont	ExLag	OKF	50.9	65.5	9.14
DF	PainS	WAb	SD	GU	Flex	HipAb	OKFC	56.4	80.7	122.29
KP	Weight	PainS	PainN	SD	GU	FCont	OHAb	55.6	79.2	17.55
H	GU	RU	OHAb	PainS	FCont	OKF	ExLag	58.4	81.7	0.15

Table 6. The PVs selected for the different time points of the knees data using method **H**, longitudinal PVs selected using method **HT**, and longitudinal utility-weighted PVs.

		Selected Variables									
Time	*Utility*	*1*	*2*	*3*	*4*	*5*	*6*	*7*		%tr ($\boldsymbol{R}$)	$\%\|\|\boldsymbol{R}\|\|^2$
Pre-op	×	GU	RU	OHAb	PainS	FCont	OKF	ExLag		58.4	81.7
1-year	×	GU	PainF	SD	OHAb	OKF	FCont	Weight		62.8	87.9
5-years	×	GD	PainS	OKFC	OKF	ExLag	OHAb	Weight		67.5	91.0
10-years	×	GU	PainF	FCont	RU	OHAb	ExLag	OKF		70.0	92.0
All	×	GU	PainS	OKFC	SD	OHAb	ExLag	OKF	Pre-op:	57.7	80.8
									1-year:	65.2	89.2
									5-years:	69.4	92.8
									10-years:	67.6	90.7
All	Ease+Use	PainF	WAb	Weight	RU	Flex	ExLag	OHAb	Pre-op:	55.2	78.2
									1-year:	62.6	87.0
									5-years:	64.3	88.7
									10-years:	66.4	89.2

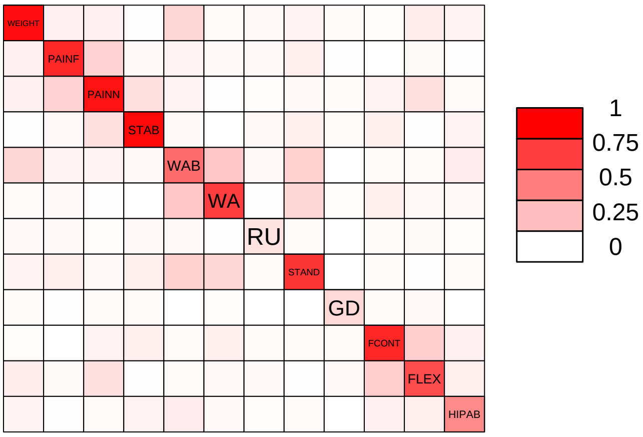


Fig. 3. Correlation plot of the partial covariance matrix of the remaining variables of the knees data given the seven chosen variables.

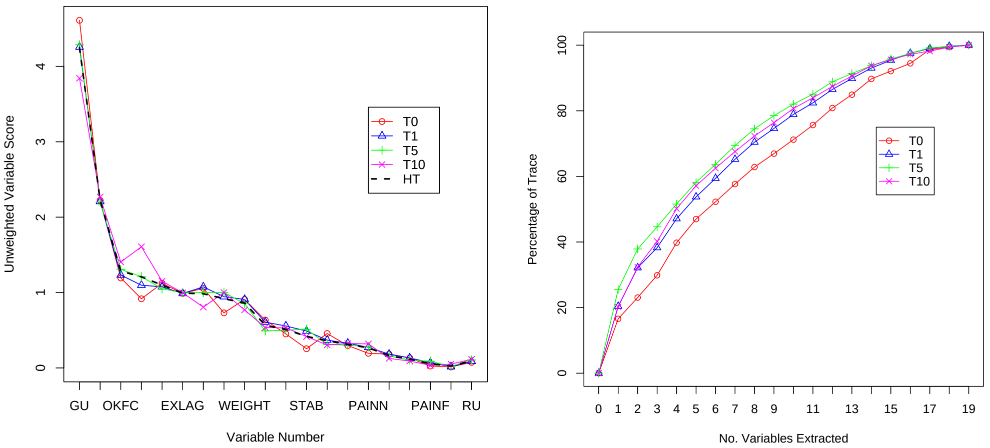
in Figure 3. It is approximately diagonal, reflecting the fact that most of the covariances have been captured, and suggesting that these variables are approximately conditionally uncorrelated given the variables we have selected.

Plots illustrating the temporal selection process are given in Figure 4. The first graph plots the score of each selected variable, with the different time points represented by different lines. The progress of the overall score (the solid black dashed line) is of a sharp initial decrease followed by a straightening out. The progress for the individual time points is more noisy. The interpretation is, for example, that the fourth variable is clearly a good choice for the 10-year data (indicated by $\times$) and a poorer choice for the pre-operative data ($\circ$). The percentage trace plot is constructed using the original correlation matrices for the data rather than the temporally smoothed matrices in order to more adequately assess the performance at the different time points. The information gain is rapid for the first few variables extracted, and then tails off. We commented above on the higher percentages of variation explained for the 10-year measurements.

9.2.3 Utility-based longitudinal PVs

A clinician provided two sets of utilities regarding these data. In each case, the utility scale is 0–10, with 0 representing undesirability: such a utility would prevent a variable from selection. A score of 10 would force a variable into the selection. The utilities are given in Table 4. The first utility (Util Ease) is a subjective measure of the ease with which the measurement is collected. The second (Util Use) is the perceived clinical usefulness of each measurement for patient diagnosis and monitoring.

Application of the temporal variable selection procedure to the knees data using these utilities yield the variable sets given in the final row of Table



(a) h scores

(b) Percentage of $\mathrm{tr}\,(\mathbf{\Sigma})$ explained

Fig. 4. Two scree plots produced from application of temporal variable selection procedure **HT** to the knees data

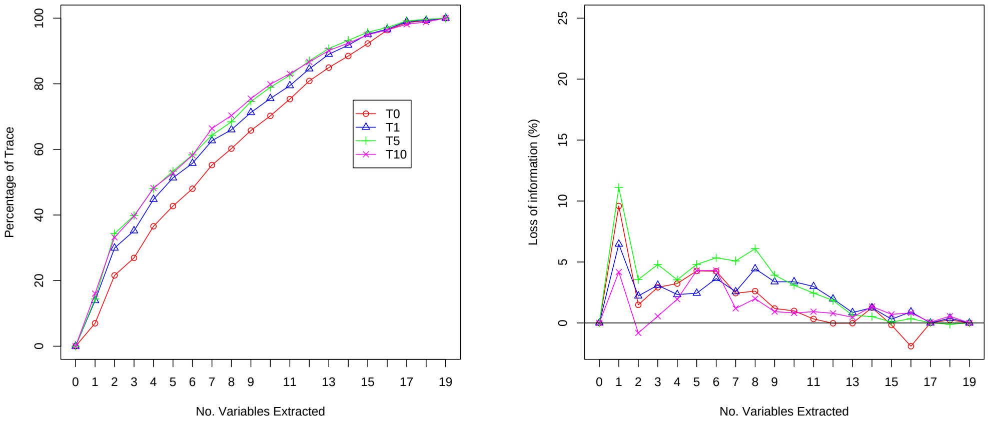
6. Comparison with the other temporal results demonstrates that there are differences between these variable subsets and those obtained without utility information. This is expected, as the selection procedure now balances the raw gain in information with a variable's desirability.

In terms of performance, we typically find that there is a lower percentage trace and squared norm explained by these variables when compared to the original longitudinal subset. The utility information is, to an extent, overriding the information from the data to allow the procedure to select a variable with lower information content, but higher utility. Typically the percentage trace is reduced by between 1% and 8%. We can, of course, extract further variables if we need to reach a pre-specified variance reduction threshold.

The percentage trace plot in Figure 5(a) shows a pattern similar to that of Figure 4(b), although with a shallower and less smooth ascent attributable to the use of utilities to adjust our selections. The plot in Figure 5(b) displays the difference between the percentage of the squared norm we explain when using the standard temporal method and the utility-based method - essentially the "loss" of information due to the use of utilities to adjust the selection process. In this example, this loss is negligible.

10 Discussion and Concluding Remarks

We have introduced a novel method for the extraction of PVs which largely outperforms extant methods. Further, the proposed method is easily extended to the determination of PVs for repeated measures data, and to the incorpora-



(a) Percentage Trace Explained

(b) Loss of Variability

Fig. 5. Plots for the application of utilities in variable selection from the knees data.

tion of utilities concerning the desirability of retaining certain variables. The methodology is straightforward to apply and interpret. Whilst the new basic method is an improvement on others, its advantage is in its extensibility and in its ethos, which is to focus attention on actual variables. Our results imply that there will be only minor differences between the proportions of variation explained by the first m PCs and the best m PVs, selected by a reasonable method such as $\mathbf{H}$, and that such differences diminish as m increases.

Whether or not the determination of PVs is the best approach to dimension reduction for large data sets depends on purpose. For our purpose, which is to take the reduced-dimension representation as input to graphical modelling of the relationships between pre-operative and post-operative measurements (and for which the partial covariance structure is directly relevant [27]), such dimension reduction appears the only possibility. Elsewhere, huge data sets are being constructed in areas such as credit scoring, complex manufacturing, and image analysis for astronomical data. For which the recent developments concerning CUR decompositions may form a possible means to dimension reduction [28]. These are likely to be cruder approximations, but may have potential for enormous data sets.

Acknowledgements

We are grateful to Dr John Egan of E-Tech Ltd. and Prof. Paul Gregg of James Cook University Hospital Middlesbrough for providing the knee-replacement data. We are also grateful to the constructive comments of the two anonymous referees, which have lead to notable improvement of this paper. This research was funded under EPSRC Industrial CASE studentship quota 02303277, via

the Smith Institute.

References

[1] G. P. McCabe, Principal variables, Technometrics 26 (2) (1984) 137–144.

[2] W. J. Krzanowski, F. H. C. Marriott, Multivariate Analysis I: Distributions, ordination and inference, Vol. I of Kendall's Library of Statistics, Arnold Publishers, 1994.

[3] I. T. Jolliffe, Discarding variables in principal component analysis. I: Artificial data, Applied Statistics 21 (2) (1972) 160–173.

[4] I. T. Jolliffe, Principal Component Analysis, 2nd Edition, Springer-Verlag, New York, 2002.

[5] J. Cadima, I. T. Jolliffe, Loadings and correlations in the interpretation of principal components, Journal of Applied Statistics 22 (2) (1995) 203–214.

[6] M. Okamoto, Optimality of principal components, in: P. R. Krishnaiah (Ed.), Multivariate Analysis II, Academic Press, 1969, pp. 673–685.

[7] N. M. Al-Kandari, I. T. Jolliffe, Variable selection and interpretation of covariance principal components, Communications in Statistics – Simulation and Computation 30 (2) (2001) 339–354.

[8] N. M. Al-Kandari, I. T. Jolliffe, Variable selection and interpretation in correlation principal components, Environmetrics 16 (2005) 659–672.

[9] E. M. Beale, M. G. Kendall, D. W. Mann, The discarding of variables in multivariate analysis, Biometrika 54 (1967) 357–366.

[10] W. J. Krzanowski, Selection of variables to preserve multivariate data structure, using principal components, Applied Statistics 36 (1) (1987) 22–33.

[11] A. de Falguerolles, S. Jmel, Un critère de choix de variables en analyse en composantes principales fondé sur des modèles graphiques gaussiens particuliers, Canadian Journal of Statistics 21 (3) (1993) 239–256.

[12] J. A. Cumming, Clinical decision support, Ph.D. thesis, Durham University (2006).

[13] C. S. Berkey, N. M. Laird, I. Valadian, J. Gardner, Modelling adolescent blood pressure patterns and their prediction of adult pressures, Biometrics 47 (3) (1991) 1005–1018.

[14] J. O. Ramsay, B. W. Silverman, Applied Functional Data Analysis: Methods and Case Studies, Springer-Verlag, New York, 2002.

[15] T. Prvan, A. W. Bowman, Nonparametric time dependent principal components analysis, The Australian & New Zealand Industrial and Applied Mathematics Journal 44 (2003) C627–C643.

[16] W. Bossert, J. A. Weymark, Utility in social choice, in: S. Barberà, P. J. Hammond, C. Seidl (Eds.), Handbook of Utility Theory, Vol. 2: Extensions, Kluwer Academic Publishers, 2004, Ch. 20, pp. 1099–1177.

[17] R. B. Cattell, The scree test for the number of factors, Multivariate Behavioral Research 1 (1966) 245–276.

[18] R Development Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria (2005).

[19] J. N. R. Jeffers, Two case studies in the application of principal component analysis, Applied Statistics 16 (1967) 225–236.

[20] I. T. Jolliffe, Discarding variables in principal component analysis. II: Real data, Applied Statistics 22 (1) (1973) 21–31.

[21] M. Friendly, Corrgrams: Exploratory displays for correlation matrices, The American Statistician 56 (4) (2002) 316–324.

[22] A. W. McCaskie, D. J. Deehan, T. P. Green, K. R. Lock, J. R. Thompson, W. M. Harper, P. J. Gregg, Randomised, prospective study comparing cemented and cementless total knee replacement: Results of press-fit condylar total knee replacement at five years, Journal of Bone & Joint Surgery - British Volume 80 (6) (1998) 971–975.

[23] P. R. Peres-Neto, D. A. Jackson, K. M. Somers, How many principal components? Stopping rules for determining the number of non-trivial axes revisited, Computational Statistics & Data Analsysis 49 (4) (2005) 974–997.

[24] H. F. Kaiser, The application of electronic computers to factor analysis, Educational and Psychological Measurement 20 (1960) 141–151.

[25] S. Frontier, Étude de la decroissance des valeurs propres dans une analyze en composantes principales: comparison avec le modèle de baton brisé, Journal of Experimental Marine Biology and Ecology 25 (1976) 341–347.

[26] W. F. Velicer, Determining the number of components from a matrix of partial correlations, Psychometrika 41 (1976) 321–327.

[27] J. Whittaker, Graphical Models In Applied Mathematical Multivariate Statistics, Wiley, Chichester, 1990.

[28] P. Drineas, R. Kannan, M. W. Mahoney, Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition, Tech. Rep. YALEU/DCS/TR-1271, Yale University (2004).

A Algorithms

A.1 ***H****: The iterative algorithm using h values and partial covariance*

(1) Let $V_1^{(\ell)}$ be the set of selected variables at stage ℓ. Similarly, let $V_2^{(\ell)}$ be the set of unselected variables at stage ℓ. Furthermore, denote the variable selected at stage ℓ by $v^{(\ell)}$.

(2) Set $V_1^{(1)} = \emptyset$, $V_2^{(1)} = V$, and $\tilde{\boldsymbol{S}}_{22\cdot 1}^{(1)} = \boldsymbol{R}$.

(3) For $\ell = 1, \ldots, m$

  (a) Calculate h_j $(j = 1, \ldots, p - \ell)$ from $\tilde{\boldsymbol{S}}_{22\cdot 1}^{(\ell)}$. Select variable $v^{(\ell)}$ with the largest h_j.

  (b) Set $V_2^{(\ell+1)} = V_2^{(\ell)} \setminus \{v^{(\ell)}\}$ and $V_1^{(\ell+1)} = V_1^{(\ell)} \cup \{v^{(\ell)}\}$.

  (c) Update $\tilde{\boldsymbol{S}}_{22\cdot 1}^{(\ell)}$ to $\tilde{\boldsymbol{S}}_{22\cdot 1}^{(\ell+1)}$ using:

$$\tilde{\boldsymbol{S}}_{22\cdot 1}^{(\ell+1)} = \tilde{\boldsymbol{S}}_{22}^{(\ell+1)} - \frac{\tilde{\boldsymbol{S}}_{21}^{(\ell+1)} \left(\tilde{\boldsymbol{S}}_{21}^{(\ell+1)}\right)^T}{\tilde{s}^{(\ell)}},$$

  where

$$\begin{aligned} \tilde{\boldsymbol{S}}_{22}^{(\ell+1)} &= \mathrm{Cov}[V_2^{(\ell+1)}], \\ \tilde{\boldsymbol{S}}_{21}^{(\ell+1)} &= \mathrm{Cov}[V_2^{(\ell+1)}, v^{(\ell)}], \end{aligned}$$

  and further $\tilde{\boldsymbol{S}}_{22}^{(\ell+1)}$, $\tilde{\boldsymbol{S}}_{21}^{(\ell+1)}$ are simply submatrices of $\tilde{\boldsymbol{S}}_{22\cdot 1}^{(\ell)}$, and $\tilde{s}_{(\ell)}$ is the partial variance of $v^{(\ell)}$ on the diagonal of $\tilde{\boldsymbol{S}}_{22\cdot 1}^{(\ell)}$.

A.2 ***HT****: The modified algorithm which incorporates a temporal aspect*

(1) Set $V_1^{(1)} = \emptyset$, $V_2^{(1)} = V$. For each time point t $(t = 1, \ldots, T)$: set $\boldsymbol{S}_{\omega,22\cdot 1}^{(1)}(t) = \boldsymbol{R}_\omega(t)$, which is the smoothed correlation matrix for time point t.

(2) For $\ell = 1, \ldots, m$

  (a) For each time point t $(t = 1, \ldots, T)$: using $\boldsymbol{S}_{\omega,22\cdot 1}^{(\ell)}(t)$ calculate the $h_{j,t}$ for each variable v_j $(j = 1, \ldots, p - \ell)$.

  (b) Calculate $h_j^T = \sum_t (h_{j,t})$. Select variable $v^{(\ell)}$ which maximises h_j^T.

  (c) Set $V_2^{(\ell+1)} = V_2^{(\ell)} \setminus \{v^{(\ell)}\}$ and $V_1^{(\ell+1)} = V_1^{(\ell)} \cup \{v^{(\ell)}\}$.

  (d) For each time point t: update $\boldsymbol{S}_{\omega,22\cdot 1}^{(\ell)}(t)$ to $\boldsymbol{S}_{\omega,22\cdot 1}^{(\ell+1)}(t)$ as above.