Nonparametric time dependent principal components analysis.

T. Prvan* A. W. Bowman†

(Received 15 August 2001; revised August 2002)

Abstract

Principal Component Analysis (PCA) is a popular data reduction technique widely used in data mining. It is common to ignore any existing time component of the data when performing PCA. One way of incorporating this dimension is to perform PCA for the data at each such point. The disadvantage of this approach is that there may not be enough data at each time point. We overcome this by using a smoothed covariance or correlation matrix and by the choice of bandwidth we control the amount of neighbouring data contributing to the calculation. Permutations are used to construct reference bands to test whether there is a time effect. If there is a time effect then performing PCA as a data reduction technique is inappropriate. Nonetheless the

*Department of Statistics, DEFS, Macquarie University, Sydney, NSW 2109, Australia. mailto:tprvan@efs.mq.edu.au

†Department of Statistics, University of Glasgow, Glasgow, G12 8QW, UK.

[0]See http://anziamj.austms.org.au/V44/CTAC2001/Prva for this article,

 Published 1 April 2003. ISSN 1446-8735

smoothed loadings of the principal components deemed to account for most of the variation in the data may give one insight into the structure of the data. The techniques are illustrated using aircraft development data.

Contents

1 Introduction

Principal Component Analysis (PCA) is one of the best known approaches to data reduction. The original set of variables is processed to produce a new and smaller set of variables designed to contain as much information as possible. PCA is a linear transformation which locates directions of maximum variance in the original data and rotates the data along these axes. The eigenvectors associated with the ordered eigenvalues (largest to smallest) of the covariance or correlation matrix give the loadings (coefficients) for each principal component (PC). A correlation matrix is always used if the variables are not all on the same scale otherwise a covariance matrix can be used. Usually the first PC contains most of the information.

PCA is used widely in data mining as a data reduction technique. Most data mining observations have a time component. If there is a time component, it is reasonable to suspect that these sources of variation vary over time. Our proposed nonparametric time dependent PCA should be able to detect whether this is the case. If there is a time effect, then performing PCA without taking this into account may not be ideal. PCA would then be an inappropriate data reduction technique.

Principal component analysis may also be used to detect structure in the relationships among variables. Again, if there is a time component it is reasonable to suspect that the sources of variation will vary over time. Changes in the loadings of the variables over time reveal another dimension about the data.

The aim of this article is to introduce a method of nonparametric smoothing for time dependent PCA, and to explore its role in identifying sources of variation as well as their variation over time. Smoothed covariance and correlation matrices are developed by generalising the parametric formulae to incorporate weights which introduce the smoothing. The form of the estimators is discussed in Section 2. Reference bands for testing whether there is a time effect are proposed in Section 3. This is followed by further discussion in Section 4.

2 Nonparametric time dependent PCA

Principal component analysis can be performed on either the covariance or correlation matrix. In biology there is a preference for the covariance matrix because the measurements being considered are usually on the same scale. Besides identifying the sources of variation, interest lies in interpreting the principal components. When

measurements are not in the same units, the correlation matrix must be used. Now that the principal components are for standardized variables, the principal components may be less easy to interpret directly (Jolliffe [5]). We now want to evaluate the covariance or correlation matrix at time t. Suppose our data is of the form $(t_1, \boldsymbol{x}_1), \ldots, (t_n, \boldsymbol{x}_n)$ where $\boldsymbol{x}_i \in \mathbb{R}^p$. If the time t corresponds to a data time point we could just calculate the sample covariance or correlation matrix at this data point if we have more than one observed vector of values. Our estimator for the covariance matrix would be

$$\begin{aligned} S(t) &= \frac{1}{(\text{no. of } t_i = t)} \sum_{\text{all } t_i = t} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}(t))(\boldsymbol{x}_i - \overline{\boldsymbol{x}}(t))^T \\ &= \frac{1}{(\text{no. of } t_i = t)} A(t)\,, \end{aligned}$$

where

$$A(t) = \sum_{\text{all } t_i = t} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}(t))(\boldsymbol{x}_i - \overline{\boldsymbol{x}}(t))^T\,.$$

The correlation matrix would be

$$\begin{aligned} R(t) &= D(1/s_i(t))S(t)D(1/s_i(t)) \\ &= D(1/\sqrt{a_{ii}(t)})A(t)D(1/\sqrt{a_{ii}(t)})\,, \end{aligned}$$

where $D(.)$ denotes the diagonal matrix containing the reciprocals of the standard deviations or the square roots of the diagonal elements of $A(t)$.

What happens if t does not coincide with a data point? According to the above approach, we could not calculate the covariance or correlation matrix. A way around this is to consider a weighted average of observations in the neighbourhood of t, with those closer to t contributing more to the calculation of the quantity of interest.

That is,

$$\begin{aligned} S_w(t) &= \frac{1}{\sum_{i=1}^n w_i(t)} \sum_{i=1}^n w_i(t)(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_w(t))(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_w(t))^T \\ &= \frac{1}{\sum_{i=1}^n w_i(t)} A_w(t)\,, \end{aligned}$$

where

$$A_w(t) = \sum_{i=1}^n w_i(t)(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_w(t))(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_w(t))^T\,,$$

$$\overline{\boldsymbol{x}}_w(t) = \frac{1}{\sum_{i=1}^n w_i(t)} \sum_{i=1}^n w_i(t)\boldsymbol{x}_i\,,$$

and

$$R_w(t) = D(1/\sqrt{[A_w(t)]_{ii}})A_w(t)D(1/\sqrt{[A_w(t)]_{ii}})\,.$$

We have used $df = \sum_{i=1}^n w_i(t)$ instead of $df = \sum_{i=1}^n w_i(t) - 1$ in the calculation of the weighted covariance matrix, because if the data is sparse around the time point being considered, it is conceivable that the sum of the weights could be less than 1. This is not an issue when working with smoothed correlation matrices since the df cancel out. Normal kernels will be used as the weights with

$$w_i(t) = w\left(\frac{t_i - t}{h}\right)\,.$$

This centres a normal distribution with standard deviation h around the point t_i. The further t is from t_i the less weight the point is given. The bandwidth h, also known as the smoothing parameter, controls the degree of smoothing applied to the data.

The nonparametric time dependent principal component analysis technique then consists of performing PCA at each time t considered for the smoothed covariance or correlation matrix. Typically

TABLE 1: recorded aircraft information.

t	-	year of first manufacture
$x1$	-	total engine power (kw)
$x2$	-	wing span (metres)
$x3$	-	length (metres)
$x4$	-	maximum take-off weight (kg)
$x5$	-	maximum speed (km/hr)
$x6$	-	range (km)

this will be a uniformly dense grid in $[t_1, t_n]$. As in any smoothing procedure, the choice of bandwidth h will have important effects on the resultant estimator. For this reason, in the data set considered, different values of the bandwidth were used and a value chosen subjectively from inspection of principal component loadings versus time plots. Too big a bandwidth masks curvature in the data while too small a bandwidth displays more sampling variation.

Example: Aircraft data

An aircraft data set will be used to illustrate nonparametric time dependent PCA. The progress of modern technology is very rapid and it is an issue of current concern to develop ways of monitoring trends in technology so that informed decisions can be taken by government, commercial organisations or other bodies. To develop such techniques it is sensible to begin with a simple example where data can be obtained easily and where the technology itself is well understood. A fuller description of the data on aircraft design can be found in Bowman and Azzalini [1].

Seven pieces of information have been recorded on a wide variety

of aircraft, and are listed in Table 1. We have $n = 709$ complete sets of measurements for years 1914 through to 1984.

Since the scales are different for the six variables of interest, logarithms of the data have been taken. Figure 1 displays weighted means for logarithms of the data, with logarithms of the data superimposed. The bandwidth $h = 7$ was chosen after experimenting with many different choices.

Figure 2 displays the results obtained from nonparametric time dependent PCA applied to the aircraft data using correlation matrices since the six measurements are not all on the same scale. Recall that principal component analysis is used to find the linear combinations of variables with large variance. The Proportions versus Year of Manufacture plot displays the proportion of the total variability explained by each of the six PCs over time. From this plot see that the first two PCs account for 70% to 75% of the total variability while the first three PCs account for 83% to 90% of the total variability. Principal components loadings versus Year of Manufacture (time) plots are used to detect structure in the relationships between variables over time. The first PC has all its loadings positive even though they vary over time so PC1 is a weighted average which varies over time. The loadings for the logarithm of maximum speed ($\log x5$) are much smaller than the other loadings most of the time so we conclude that maximum speed has a negligible contribution to variability. PC2 always has the logarithm of total engine power ($\log x1$) and logarithm of maximum speed ($\log x5$) with positive loadings and logarithm of of wingspan ($\log x2$) with negative loadings so we conclude that PC2 always has total engine power and maximum speed contrasting with wingspan. The other variables' loadings change sign over time and their absolute values are smaller so they contribute much less to variability. PC3 has the logarithm of range ($\log x6$) loading over time always being much

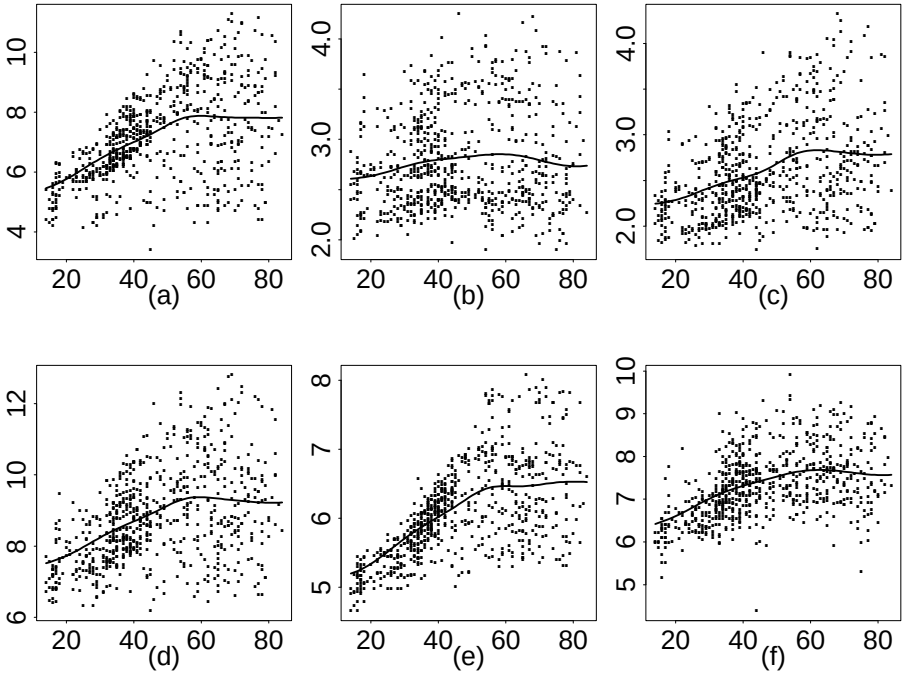

FIGURE 1: Smoothed means of logarithms of Aircraft data. The bandwidth $h = 7$ was used. (a) Weighted mean of logarithm of total engine power (kw) versus Year of Manufacture. (b) Weighted mean of logarithm of wing span (m) versus Year of Manufacture. (c) Weighted mean of logarithmm of length (m) versus Year of Manufacture. (d) Weighted mean of logarithm of take off weight (kg) versus Year of Manufacturer. (e) Weighted mean of logarithm of maximum speed (km/hr) versus Year of Manufacture. (f) Weighted mean of logarithm of range (km) versus Year of Manufacture.

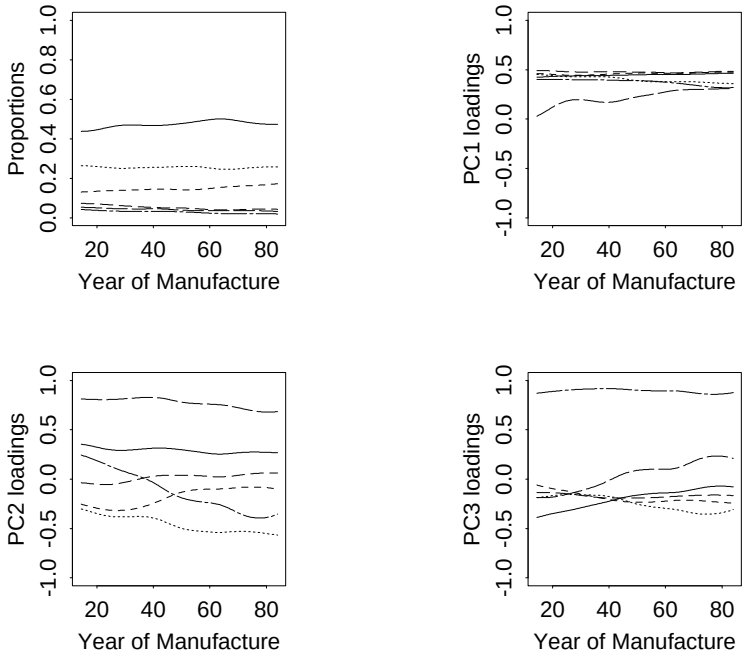

FIGURE 2: Graphical summaries for a time dependent PCA for the transformed Aircraft data. The bandwidth $h = 7$ was used. (Legend for Proportions plot: PC1 - solid line, PC2 - dotted line, PC3 - short dashed line, PC4 -medium dashed line, PC5 - long dashed line, and PC6 - remaining line. Legend for loadings plots: $\log x1$ - solid line, $\log x2$ - dotted line, $\log x3$ - short dashed line, $\log x4$ - medium dashed line, $\log x5$ - long dashed line, and $\log x6$ - remaining line.)

greater in size to the loadings of the other variables so we conclude that range dominates the contribution to variability.

In the next section we see that variability is not constant over time.

3 Reference bands assess whether there is a time effect

The reference model of interest is $H_0 : \Sigma(t) = \Sigma$ or $H_0 : R(t) = R$; that is, the null hypothesis is that the covariance matrix or correlation matrix remains constant over time. Hence the loadings (coefficients of the PCs) would remain constant over time. There is the added complication that loadings which are eigenvector entries, scaled so that the eigenvector has a particular length (in S-Plus length of the eigenvector is 1), are unique only up to a sign change. Care must be taken to ensure that when considering a particular PC we standardize our loadings output so that all of the loadings of one of the variables are always positive.

Assessing whether the covariance matrix or correlation matrix has changed over time is difficult to do graphically, so we concentrate on assessing whether the coefficients (loadings) of the PCs that account for most of the variation vary over time. If these do not vary over time, then the covariance or correlation matrix from which they are derived should also not vary over time.

It is not always immediately clear from the PC loadings plots versus time, whether we have a time effect or if the relationships exhibited can be attributed to sampling variation. As noted by Bowman and Wright [2] confidence bands are difficult to construct

because of the bias inherent in all forms of smoothing. Reference bands are an alternative way of investigating whether there is a genuine time relationship. Reference bands indicate where a nonparametric curve should be expected to lie if a particular hypothesis of interest holds (for more details see Bowman and Young [3] and Bowman and Azzalini [1]). We implement reference bands through suitable resampling techniques.

The idea behind testing for no time effect using reference bands is simple. If there was no time effect then permuting the values of $\boldsymbol{x}_i$ that are associated with t_i should lead to similar PC loadings after we have taken the sign into account as discussed above, as well as differences due to location and scale. We first standardize the data and then obtain the nonparametric time dependent PCA for this data. The PC loadings are plotted against time for the PCs we consider to account for most of the variability in the data. As discussed earlier, the sign of the loadings needs to be taken into account so that we always have the loading of one particular variable of the PC under consideration always positive. To obtain the reference bands we resample from this "new" data and perform nonparametric time dependent PCA each time we resample. We retain the same relationship regarding sign of loading of the particular variable of the PC being considered as for the unpermuted standardized data. These reference bands for the loadings of each PC considered should indicate where the nonparametric curve for the loadings is likely to be when the hypothesis of no time effect is correct. When obtaining the standardised data, a much smaller smoothing parameter is used in the calculation of the smoothed mean vector and smoothed covariance matrix than in the nonparametric time dependent principal component analysis. The kth entry of the standardized data

vector at time t_i is

$$\frac{\boldsymbol{e}_k^T(\boldsymbol{x}(t_i) - \boldsymbol{x}_w(t_i))}{\sqrt{\boldsymbol{e}_k^T S_w(t_i)\boldsymbol{e}_k}\sqrt{(\sum_{j=1}^n w_j(t_i) + 1)/\sum_{j=1}^n w_j(t_i)}},$$

where $\boldsymbol{e}_k \in \mathbb{R}^p$ has a 1 in the kth position and zeros elsewhere. If the reference bands, obtained from the resamples from the standardized data, for each PC loading envelop the PC loading for the unpermuted standardized data, then we conclude there is no time effect. Typically 100 resamples or more are used in constructing the reference bands.

For the aircraft data, logarithms of the data except for time were taken, and we then used $h = 1$ to obtain the standardized data. Nonparametric time dependent PCA was performed on the standardized data. Since the first three PCs of the unstandardized data accounted for 83% to 90% of the variation, we plotted the loadings for the first three PCs for the standardized data. The reference bands were obtained by permuting the standardized data and performing nonparametric time dependent PCA using $h = 7$. This was done 100 times to give 100 reference bands for the loadings of the first three first PCs. In all there were 18 plots. The reference bands for the absence of time effect for the PC loadings for the first three PCs (Figure 3, Figure 4, Figure 5) do not always envelop the PC loadings for the standardized logarithm of the original aircraft data so time has an effect on variability. This is not unexpected.

4 Discussion

With the advent of modern computing, it has become possible to implement straightforward ideas which require substantial computing. The method presented here is one such example. A literature

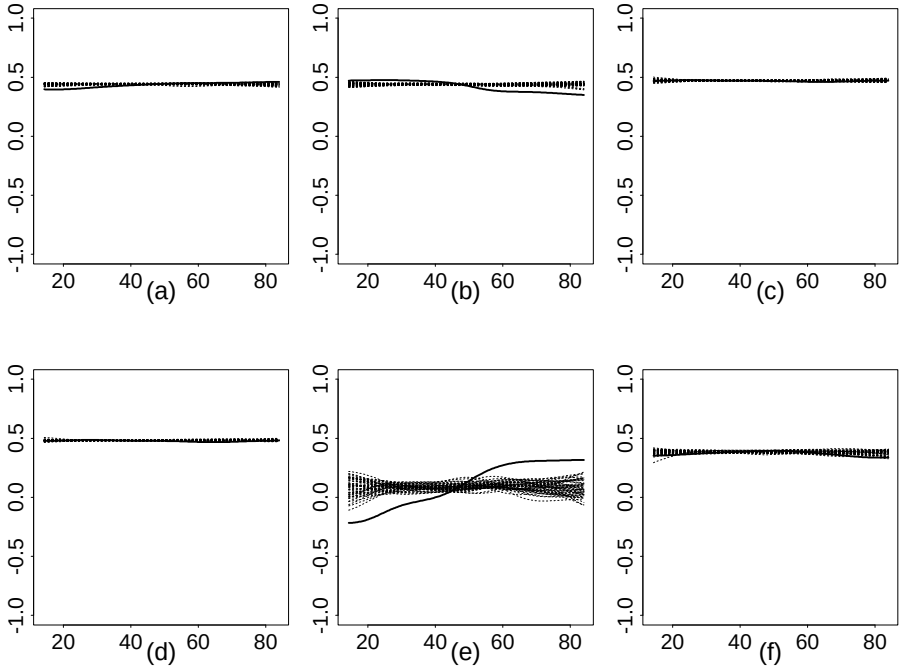

FIGURE 3: Reference bands for PC1 loadings. The bandwidth $h = 7$ was used. (a) $\log x1$ loading versus time, (b) $\log x2$ loading versus time, (c) $\log x3$ loading versus time, (d) $\log x4$ loading versus time, (e) $\log x5$ loading versus time, (f) $\log x6$ loading versus time.

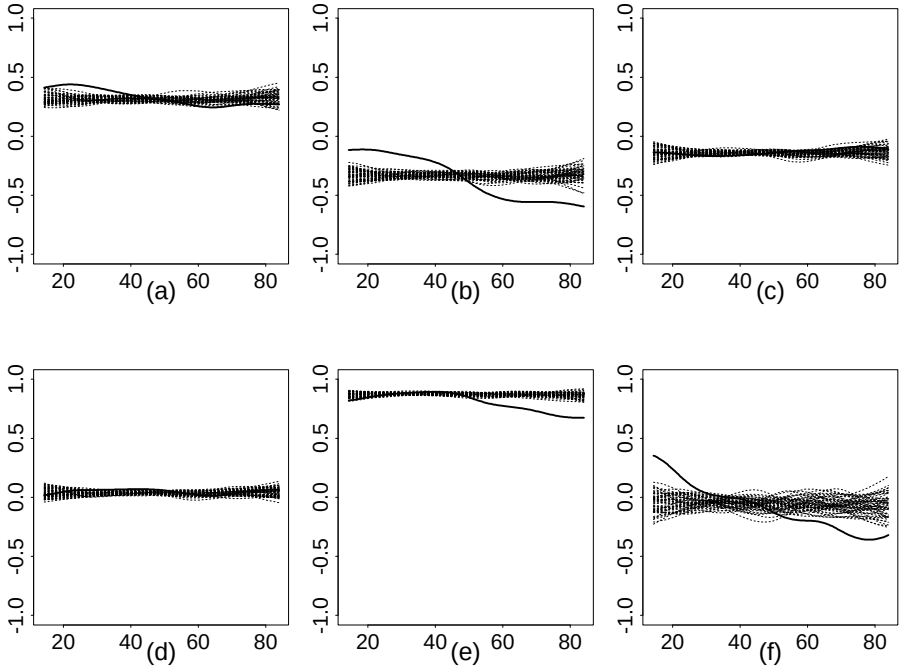

FIGURE 4: Reference bands for PC2 loadings. The bandwidth $h = 7$ was used.(a) $\log x1$ loading versus time, (b) $\log x2$ loading versus time, (c) $\log x3$ loading versus time, (d) $\log x4$ loading versus time, (e) $\log x5$ loading versus time, (f) $\log x6$ loading versus time.

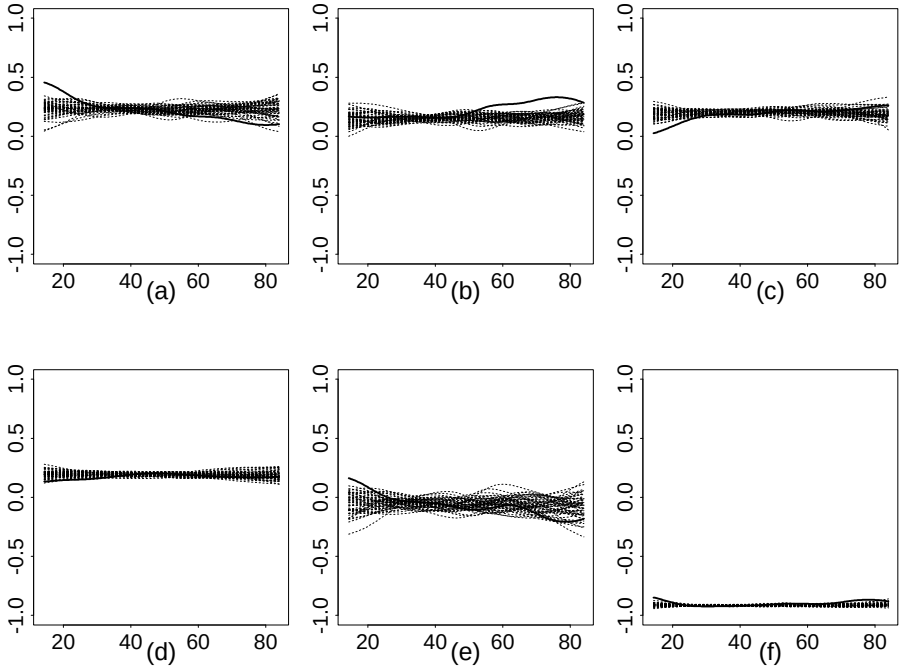

FIGURE 5: Reference bands for PC3 loadings. The bandwidth $h = 7$ was used.(a) $\log x1$ loading versus time, (b) $\log x2$ loading versus time, (c) $\log x3$ loading versus time, (d) $\log x4$ loading versus time, (e) $\log x5$ loading versus time, (f) $\log x6$ loading versus time.

search revealed that Rice and Silverman [6] developed a smoothed principal component analysis for functional data. Our approach differs from Rice and Silverman [6] and related approaches [7, e.g.]) in that instead of their assuming that the data $X_1(t), \ldots, X_n(t)$ are drawn from a stochastic process X on some bounded interval, we consider the vectors of standard multivariate analysis have an extra dimension such as time. We have measurements on several variables at many time points, whereas Silverman [7] is measuring the same variable for each of the n realisations of the stochastic process X.

One popular use of principal component analysis is as a data reduction tool. In data mining, PCA is explicitly used as a method of data compression and is also known as the Karhunen-Loeve, or K-L method (Han and Kamber [4, p.123]). If the data is collected over time, it is important to assess whether it is appropriate to ignore time and use the leading principal components to collapse the dimensionality of the data. Nonparametric time dependent principal component analysis provides a method for assessing whether it is sensible to reduce the data on the basis of PCA, ignoring time.

Another type of application where PCA has been found useful is in identifying the most important sources of variation in anatomical measurements in various species (Jolliffe [5]). If there is a factor such as time, PCA is usually performed ignoring it (e.g. aged animals) or at best performing separate PCAs on subgroups of the data (e.g. pups, yearlings, subadults, adults). If there is a time component, it is reasonable to suspect that these sources of variation vary over time. In the period of little growth before a growth spurt one would expect the variability to be less. Nonparametric time dependent PCA should be able to detect this. The loadings of the variables change over time would reveals another dimension about growth.

References

[1] Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations.* Oxford: Oxford University Press. C632, C637

[2] Bowman, A. W. and Wright, E. M. (2000). Graphical Exploration of Covariate Effects on Survival Data Through Nonparametric Quantile Curves. *Biometrics*, **56**, 563–570. C636

[3] Bowman, A. W. and Young, S. G. (1996). Graphical comparison of nonparametric curves. *Applied Statistics* **45**, 83–98. C637

[4] Han, J. and Kamber, M. (2001). *Data Mining: Concepts and Techniques.* San Diego, CA: Academic Press. C642

[5] Jolliffe, I. T. (1986). Principal Component Analysis. Virginia: Springer-Verlag New York Inc. C630, C642

[6] Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B.* **53** 233–243. C642

[7] Silverman, B. W. (1996). Smoothed functional principal components by norm. *Ann. Statist.* **24** no. 1, 1–24. C642