12

Principal Component Analysis for Time Series and Other Non-Independent Data

12.1 Introduction

In much of statistics it is assumed that the n observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are independent. This chapter discusses the implications for PCA of non-independence among $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Much of the chapter is concerned with PCA for time series data, the most common type of non-independent data, although data where $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are measured at n points in space are also discussed. Such data often have dependence which is more complicated than for time series. Time series data are sufficiently different from ordinary independent data for there to be aspects of PCA that arise only for such data, for example, PCs in the frequency domain.

The results of Section 3.7, which allow formal inference procedures to be performed for PCs, rely on independence of $\mathbf{x}_1$, $\mathbf{x}_2, \ldots, \mathbf{x}_n$, as well as (usually) on multivariate normality. They cannot therefore be used if more than very weak dependence is present between $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. However, when the main objective of PCA is descriptive, not inferential, complications such as non-independence do not seriously affect this objective. The effective sample size is reduced below n, but this reduction need not be too important. In fact, in some circumstances we are actively looking for dependence among $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. For example, grouping of observations in a few small areas of the two-dimensional space defined by the first two PCs implies dependence between those observations that are grouped together. Such behaviour is actively sought in cluster analysis (see Section 9.2) and is often welcomed as a useful insight into the structure of the data, rather than decried as an undesirable feature.

We have already seen a number of examples where the data are time series, but where no special account is taken of the dependence between observations. Section 4.3 gave an example of a type that is common in atmospheric science, where the variables are measurements of the same meteorological variable made at p different geographical locations, and the n observations on each variable correspond to different times. Section 12.2 largely deals with techniques that have been developed for data of this type. The examples given in Section 4.5 and Section 6.4.2 are also illustrations of PCA applied to data for which the variables (stock prices and crime rates, respectively) are measured at various points of time. Furthermore, one of the earliest published applications of PCA (Stone, 1947) was on (economic) time series data.

In time series data, dependence between the $\mathbf{x}$ vectors is induced by their relative closeness in time, so that $\mathbf{x}_h$ and $\mathbf{x}_i$ will often be highly dependent if $|h-i|$ is small, with decreasing dependence as $|h-i|$ increases. This basic pattern may in addition be perturbed by, for example, seasonal dependence in monthly data, where decreasing dependence for increasing $|h-i|$ is interrupted by a higher degree of association for observations separated by exactly one year, two years, and so on.

Because of the emphasis on time series in this chapter, we need to introduce some of its basic ideas and definitions, although limited space permits only a rudimentary introduction to this vast subject (for more information see, for example, Brillinger (1981); Brockwell and Davis (1996); or Hamilton (1994)). Suppose, for the moment, that only a single variable is measured at equally spaced points in time. Our time series is then $\ldots x_{-1}, x_0, x_1, x_2, \ldots$. Much of time series analysis is concerned with series that are stationary, and which can be described entirely by their first- and second-order moments; these moments are

$$\begin{aligned} \mu &= E(x_i), && i = \ldots, -1, 0, 1, 2, \ldots \\ \gamma_k &= E[(x_i - \mu)(x_{i+k} - \mu)], && i = \ldots, -1, 0, 1, 2, \ldots \\ & && k = \ldots, -1, 0, 1, 2, \ldots, \end{aligned} \tag{12.1.1}$$

where μ is the mean of the series and is the same for all x_i in stationary series, and γ_k, the kth autocovariance, is the covariance between x_i and x_{i+k}, which depends on k but not i for stationary series. The information contained in the autocovariances can be expressed equivalently in terms of the power spectrum of the series

$$f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma_k e^{-ik\lambda}, \tag{12.1.2}$$

where $i = \sqrt{-1}$ and λ denotes angular frequency. Roughly speaking, the function $f(\lambda)$ decomposes the series into oscillatory portions with different frequencies of oscillation, and $f(\lambda)$ measures the relative importance of these portions as a function of their angular frequency λ. For example, if a

series is almost a pure oscillation with angular frequency λ_0, then $f(\lambda)$ is large for λ close to λ_0 and near zero elsewhere. This behaviour is signalled in the autocovariances by a large value of γ_k at $k = k_0$, where k_0 is the period of oscillation corresponding to angular frequency λ_0 (that is $k_0 = 2\pi/\lambda_0$), and small values elsewhere.

Because there are two different but equivalent functions (12.1.1) and (12.1.2) expressing the second-order behaviour of a time series, there are two different types of analysis of time series, namely in the time domain using (12.1.1) and in the frequency domain using (12.1.2).

Consider now a time series that consists not of a single variable, but p variables. The definitions (12.1.1), (12.1.2) generalize readily to

$$\mathbf{\Gamma}_k = E[(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_{i+k} - \boldsymbol{\mu})'], \tag{12.1.3}$$

where

$$\boldsymbol{\mu} = E[\mathbf{x}_i]$$

and

$$\mathbf{F}(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \mathbf{\Gamma}_k e^{-ik\lambda} \tag{12.1.4}$$

The mean $\boldsymbol{\mu}$ is now a p-element vector, and $\mathbf{\Gamma}_k$, $\mathbf{F}(\lambda)$ are $(p \times p)$ matrices.

Principal component analysis operates on a covariance or correlation matrix, but in time series we can calculate not only covariances between variables measured at the same time (the usual definition of covariance, which is given by the matrix $\mathbf{\Gamma}_0$ defined in (12.1.3)), but also covariances between variables at different times, as measured by $\mathbf{\Gamma}_k, k \neq 0$. This is in contrast to the more usual situation where our observations $\mathbf{x}_1, \mathbf{x}_2, \ldots$ are independent, so that any covariances between elements of $\mathbf{x}_i, \mathbf{x}_j$ are zero when $i \neq j$. In addition to the choice of which $\mathbf{\Gamma}_k$ to examine, the fact that the covariances have an alternative representation in the frequency domain means that there are several different ways in which PCA can be applied to time series data.

Before looking at specific techniques, we define the terms 'white noise' and 'red noise.' A white noise series is one whose terms are all identically distributed and independent of each other. Its spectrum is flat, like that of white light; hence its name. Red noise is equivalent to a series that follows a positively autocorrelated first-order autoregressive model

$$x_t = \phi x_{t-1} + \epsilon_t, \quad t = \ldots 0, 1, 2 \ldots,$$

where ϕ is a constant such that $0 < \phi < 1$ and $\{\epsilon_t\}$ is a white noise series. The spectrum of a red noise series decreases as frequency increases, like that of red light in the range of visual radiation.

The next section of this chapter describes a range of approaches based on PCA that have been used on time series data in atmospheric science. Although many are inspired by the special nature of the data commonly

encountered in this area, with observations corresponding to times and variables to spatial position, they are not necessarily restricted to such data.

Time series are usually measured at discrete points in time, but sometimes the series are curves. The analysis of such data is known as *functional data analysis* (functional PCA is the subject of Section 12.3). The final section of the chapter collects together a number of largely unconnected ideas and references concerning PCA in the context of time series and other non-independent data.

12.2 PCA-Related Techniques for (Spatio-) Temporal Atmospheric Science Data

It was noted in Section 4.3 that, for a common type of data in atmospheric science, the use of PCA, more often referred to as *empirical orthogonal function* (EOF) analysis, is widespread. The data concerned consist of measurements of some variable, for example, sea level pressure, temperature, ..., at p spatial locations (usually points on a grid) at n different times. The measurements at different spatial locations are treated as variables and the time points play the rôle of observations. An example of this type was given in Section 4.3. It is clear that, unless the observations are well-separated in time, there is likely to be correlation between measurements at adjacent time points, so that we have non-independence between observations. Several techniques have been developed for use in atmospheric science that take account of correlation in both time and space, and these will be described in this section. First, however, we start with the simpler situation where there is a single time series. Here we can use a principal component-like technique, called *singular spectrum analysis* (SSA), to analyse the autocorrelation in the series. SSA is described in Section 12.2.1, as is its extension to several time series, multichannel singular spectrum analysis (MSSA).

Suppose that a set of p series follows a multivariate first-order autoregressive model in which the values of the series at time t are linearly related to the values at time $(t-1)$, except for a multivariate white noise term. An estimate of the matrix defining the linear relationship can be subjected to an eigenanalysis, giving insight into the structure of the series. Such an analysis is known as *principal oscillation pattern* (POP) analysis, and is discussed in Section 12.2.2.

One idea underlying POP analysis is that there may be patterns in the maps comprising our data set, which travel in space as time progresses, and that POP analysis can help to find such patterns. Complex (Hilbert) empirical orthogonal functions (EOFs), which are described in Section 12.2.3, are designed to achieve the same objective. Detection of detailed oscillatory behaviour is also the aim of multitaper frequency-domain singular value decomposition, which is the subject of Section 12.2.4.

Some time series have cyclic behaviour with fixed periods, such as an annual or diurnal cycle. Modifications of POP analysis and PCA that take such cycles into account are discussed in Section 12.2.5. A brief discussion of examples and comparative studies is given in Section 12.2.6.

12.2.1 Singular Spectrum Analysis (SSA)

Like a number of other statistical techniques, SSA appears to have been 'invented' independently in a number of different fields. Elsner and Tsonis (1996) give references from the 1970s and 1980s from chaos theory, biological oceanography and signal processing, and the same idea is described in the statistical literature by Basilevsky and Hum (1979), where it is referred to as the 'Karhunen-Loève method,' a term more often reserved for continuous time series (see Section 12.3). Other names for the technique are 'Pisarenko's method' (Smyth, 2000) and 'singular systems analysis' (von Storch and Zwiers, 1999, Section 13.6); fortunately the latter has the same acronym as the most popular name. A comprehensive coverage of SSA is given by Golyandina et al. (2001).

The basic idea in SSA is simple: a principal component analysis is done with the variables analysed being lagged versions of a single time series variable. More specifically, our p variables are $x_t, x_{(t+1)}, \ldots, x_{(t+p-1)}$ and, assuming the time series is stationary, their covariance matrix is such that the (i, j)th element depends only on $|i - j|$. Such matrices are known as Töplitz matrices. In the present case the (i, j)th element is the autocovariance $\gamma_{|i-j|}$. Because of the simple structure of Töplitz matrices, the behaviour of the first few PCs, and their corresponding eigenvalues and eigenvectors (the EOFs) which are trigonometric functions (Brillinger, 1981, Section 3.7), can be deduced for various types of time series structure. The PCs are moving averages of the time series, with the EOFs providing the weights in the moving averages.

Töplitz matrices also occur when the p-element random vector $\mathbf{x}$ consists of *non-overlapping* blocks of p consecutive values of a single time series. If the time series is stationary, the covariance matrix $\mathbf{\Sigma}$ for $\mathbf{x}$ has Töplitz structure, with the well-known pattern of trigonometric functions for its eigenvalues and eigenvectors. Craddock (1965) performed an analysis of this type on monthly mean temperatures for central England for the period November 1680 to October 1963. The p (= 12) elements of $\mathbf{x}$ are mean temperatures for the 12 months of a particular year, where a 'year' starts in November. There is some dependence between different values of $\mathbf{x}$, but it is weaker than that between elements within a particular $\mathbf{x}$; between-year correlation was minimized by starting each year at November, when there was apparently evidence of very little continuity in atmospheric behaviour for these data. The sample covariance matrix does not, of course, have exact Töplitz structure, but several of the eigenvectors have approximately the form expected for such matrices.

Durbin (1984) uses the structure of the eigenvectors of Töplitz matrices in a different context. In regression analysis (see Chapter 8), if the dependent variable and the predictor variables are all time series, then the Töplitz structure for the covariance matrix of error terms in the regression model can be used to deduce properties of the least squares estimators of regression coefficients.

Returning to SSA, for a time series with an oscillatory component SSA has an associated pair of EOFs with identical eigenvalues. Coefficients of both EOFs have the same oscillatory pattern but are $\frac{\pi}{2}$ radians out of phase with each other. Although Elsner and Tsonis (1996) describe a number of other uses of SSA, the major application in atmospheric science has been to 'discover' dominant periodicities in a series (Allen and Smith, 1996). One advantage that SSA has in this respect over traditional spectral analysis is that the frequencies of oscillations detected by SSA can take any value in a given range rather than be restricted to a fixed set of frequencies. A disadvantage of SSA is, however, a tendency to find apparent periodicities when none exists. Allen and Smith (1996) address this problem with a carefully constructed hypothesis testing procedure for deciding whether or not apparent periodicities are statistically significant in the presence of red noise, a more realistic assumption than white noise for background variation in many climatological series. Allen and Smith (1997) discuss what they call a generalization of SSA for detecting 'signal' in the presence of red noise.

The way that SSA has been introduced above is in a 'population' context, although when talking about statistical significance in the last paragraph it is clear that we are discussing samples. A sample consists of a series $x_1, x_2, \ldots, x_n$, which is rearranged to give an $(n' \times p)$ matrix whose ith row is

$$\mathbf{x}_i' = (x_i, x_{(i+1)}, \ldots, x_{(i+p-1)}) \quad i = 1, 2, \ldots, n',$$

where $n' = n - p + 1$.

A practical consideration is the choice of p - large values of p allow longer-period oscillations to be resolved, but choosing p too large leaves too few observations, n', from which to estimate the covariance matrix of the p variables. Elsner and Tsonis (1996, Section 5.2) give some discussion of, and references to, the choice of p, and remark that choosing $p = \frac{n}{4}$ is a common practice.

Estimation of the covariance matrix raises further questions. If the n' observations on p variables are treated as an 'ordinary' data matrix, the corresponding covariance matrix will generally have all its elements unequal. If the series is stationary, the covariance between the ith and jth variables should only depend on $|i-j|$. Estimates can be constructed to be constrained to give this Töplitz structure for the covariance matrix. There is discussion of which estimates to use in Elsner and Tsonis (1996, Section 5.3). We now offer examples of SSA.

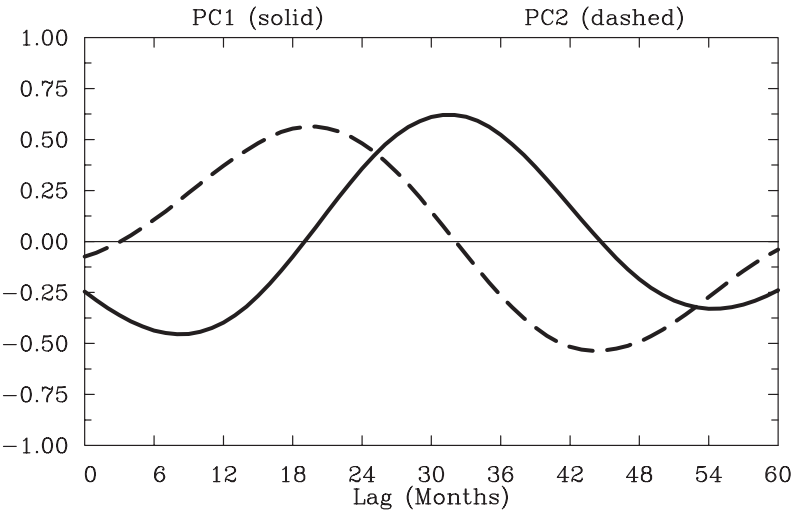

Figure 12.1. Plots of loadings for the first two components in an SSA with $p = 61$ of the Southern Oscillation Index data.

Southern Oscillation Index

The data considered here are monthly values of the Southern Oscillation Index (SOI) for the years 1876–2000, produced by the Australian Bureau of Meteorology's National Climate Centre. The number of observations in the series is therefore $n = 12 \times 125 = 1500$. The index is a measure of the East-West pressure gradient between Tahiti in the mid-Pacific Ocean and Darwin, Australia. It is a major source of climate variation. SSA was carried on the data with $p = 61$, and Figure 12.1 gives a plot of the loadings for the first two EOFs. Their eigenvalues correspond to 13.7% and 13.4% of the total variation. The closeness of the eigenvalues suggests a quasi-oscillatory pattern, and this is clearly present in the loadings of Figure 12.1. Note, however, that the relationship between the two EOFs is not a simple displacement by $\frac{\pi}{2}$. Figure 12.2 shows time series plots of the scores for the first two components (PCs) in the SSA. These again reflect the oscillatory nature of the components. A reconstruction of the series using only the first two PCs shown in Figure 12.3 captures some of the major features of the original series, but a large amount of other variability remains, reflecting the fact that the two components only account for 27.1% of the total variation.

Multichannel SSA

In multichannel SSA (MSSA) we have the more usual atmospheric science set-up of p spatial locations and n time points, but rather than finding a covariance matrix directly from the $(n \times p)$ data matrix, the data are rearranged into a larger $(n' \times p')$ matrix, where $n' = n - m + 1$, $p' = mp$

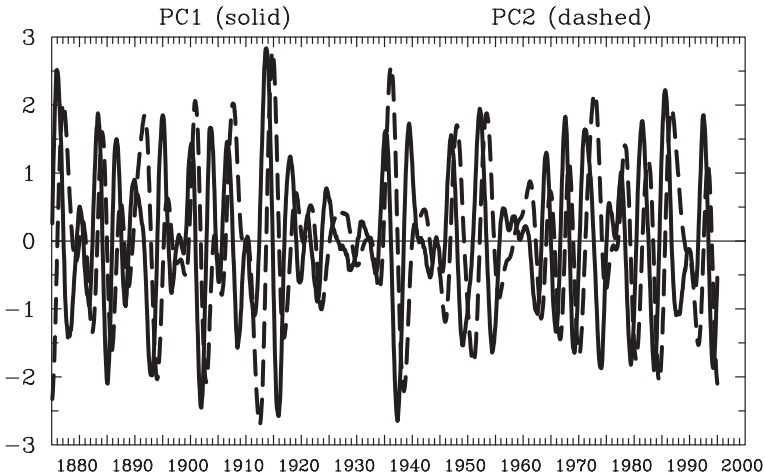

Figure 12.2. Plots of scores for the first two components in an SSA with $p = 61$ for Southern Oscillation Index data.

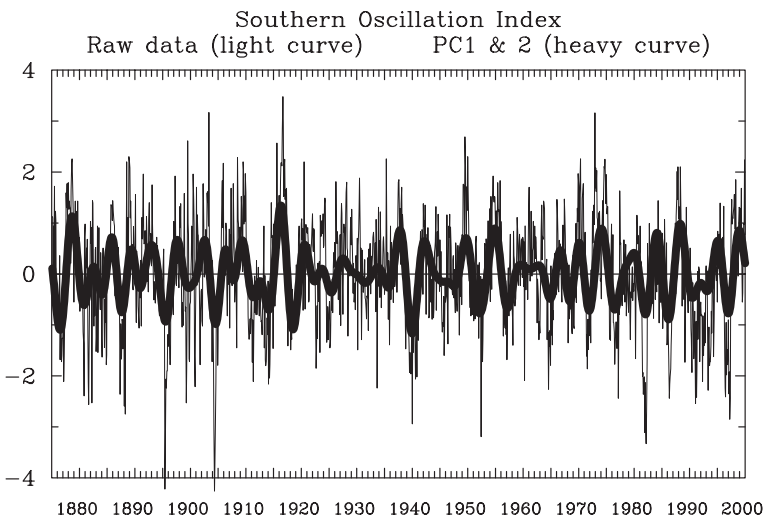

Figure 12.3. Southern Oscillation Index data together with a reconstruction using the first two components from an SSA with $p = 61$.

and a typical row of the matrix is

$$\mathbf{x}_i' = (x_{i1}, x_{(i+1)1}, \ldots, x_{(i+m-1)1}, x_{i2}, \ldots, x_{(i+m-1)2}, \ldots, x_{(i+m-1)p}),$$

$i = 1, 2, \ldots, n'$, where x_{ij} is the value of the measured variable at the ith time point and the jth spatial location, and m plays the same rôle in MSSA as p does in SSA. The covariance matrix for this data matrix has the form

$$\begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} & \cdots & \mathbf{S}_{1p} \\ \mathbf{S}_{21} & \mathbf{S}_{22} & \cdots & \mathbf{S}_{2p} \\ \vdots & \vdots & & \\ \mathbf{S}_{p1} & \mathbf{S}_{p2} & \cdots & \mathbf{S}_{pp} \end{bmatrix},$$

where $\mathbf{S}_{kk}$ is an $(m \times m)$ covariance matrix at various lags for the kth variable (location), with the same structure as the covariance matrix in an SSA of that variable. The off-diagonal matrices $\mathbf{S}_{kl}$, $k \neq l$, have (i, j)th element equal to the covariance between locations k and l at time lag $|i-j|$. Plaut and Vautard (1994) claim that the 'fundamental property' of MSSA is its ability to detect oscillatory behaviour in the same manner as SSA, but rather than an oscillation of a single series the technique finds oscillatory spatial patterns. Furthermore, it is capable of finding oscillations with the same period but different spatially orthogonal patterns, and oscillations with the same spatial pattern but different periods.

The same problem of ascertaining 'significance' arises for MSSA as in SSA. Allen and Robertson (1996) tackle this problem in a similar manner to that adopted by Allen and Smith (1996) for SSA. The null hypothesis here extends one-dimensional 'red noise' to a set of p independent AR(1) processes. A general multivariate AR(1) process is not appropriate as it can itself exhibit oscillatory behaviour, as exemplified in POP analysis (Section 12.2.2).

MSSA extends SSA from one time series to several, but if the number of time series p is large, it can become unmanageable. A solution, which is used by Benzi et al. (1997), is to carry out PCA on the $(n \times p)$ data matrix, and then implement SSA separately on the first few PCs. Alternatively for large p, MSSA is often performed on the first few PCs instead of the variables themselves, as in Plaut and Vautard (1994).

Although MSSA is a natural extension of SSA, it is also equivalent to extended empirical orthogonal function (EEOF) analysis which was introduced independently of SSA by Weare and Nasstrom (1982). Barnett and Hasselmann (1979) give an even more general analysis, in which different meteorological variables, as well as or instead of different time lags, may be included at the various locations. When different variables *replace* different time lags, the temporal correlation in the data is no longer taken into account, so further discussion is deferred to Section 14.5.

The general technique, including both time lags and several variables, is referred to as multivariate EEOF (MEEOF) analysis by Mote et al.

(2000), who give an example of the technique for five variables, and compare the results to those of separate EEOFs (MSSAs) for each variable. Mote and coworkers note that it is possible that some of the dominant MEEOF patterns may not be dominant in any of the individual EEOF analyses, and this may viewed as a disadvantage of the method. On the other hand, MEEOF analysis has the advantage of showing directly the connections between patterns for the different variables. Discussion of the properties of MSSA and MEEOF analysis is ongoing (in addition to Mote et al. (2000), see Monahan et al. (1999), for example). Compagnucci et al. (2001) propose yet another variation on the same theme. In their analysis, the PCA is done on the transpose of the matrix used in MSSA, a so-called T-mode instead of S-mode analysis (see Section 14.5). Compagnucci et al. call their technique *principal sequence pattern analysis*.

12.2.2 Principal Oscillation Pattern (POP) Analysis

SSA, MSSA, and other techniques described in this chapter can be viewed as special cases of PCA, once the variables have been defined in a suitable way. With the chosen definition of the variables, the procedures perform an eigenanalysis of a covariance matrix. POP analysis is different, but it is described briefly here because its results are used for similar purposes to those of some of the PCA-based techniques for time series included elsewhere in the chapter. Furthermore its core is an eigenanalysis, albeit not on a covariance matrix.

POP analysis was introduced by Hasselman (1988). Suppose that we have the usual $(n \times p)$ matrix of measurements on a meteorological variable, taken at n time points and p spatial locations. POP analysis has an underlying assumption that the p time series can be modelled as a multivariate first-order autoregressive process. If $\mathbf{x}_t'$ is the tth row of the data matrix, we have

$$(\mathbf{x}_{(t+1)} - \boldsymbol{\mu}) = \boldsymbol{\Upsilon}(\mathbf{x}_t - \boldsymbol{\mu}) + \boldsymbol{\epsilon}_t, \quad t = 1, 2, \ldots, (n-1), \tag{12.2.1}$$

where $\boldsymbol{\Upsilon}$ is a $(p \times p)$ matrix of constants, $\boldsymbol{\mu}$ is a vector of means for the p variables and $\boldsymbol{\epsilon}_t$ is a multivariate white noise term. Standard results from multivariate regression analysis (Mardia et al., 1979, Chapter 6) lead to estimation of $\boldsymbol{\Upsilon}$ by $\hat{\boldsymbol{\Upsilon}} = \mathbf{S}_1\mathbf{S}_0^{-1}$, where $\mathbf{S}_0$ is the usual sample covariance matrix for the p variables, and $\mathbf{S}_1$ has (i, j)th element equal to the sample covariance between the ith and jth variables at lag 1. POP analysis then finds the eigenvalues and eigenvectors of $\hat{\boldsymbol{\Upsilon}}$. The eigenvectors are known as principal oscillation patterns (POPs) and denoted $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_p$. The quantities $z_{t1}, z_{t2}, \ldots, z_{tp}$ which can be used to reconstitute $\mathbf{x}_t$ as $\sum_{k=1}^{p} z_{tk}\mathbf{p}_k$ are called the POP coefficients. They play a similar rôle in POP analysis to that of PC scores in PCA.

One obvious question is why this technique is called principal *oscillation* pattern analysis. Because $\hat{\boldsymbol{\Upsilon}}$ is not symmetric it typically has a

mixture of real and complex eigenvectors. The latter occur in pairs, with each pair sharing the same eigenvalue and having eigenvectors that are complex conjugate pairs. The real eigenvectors describe non-oscillatory, non-propagating damped patterns, but the complex eigenvectors represent damped oscillations and can include standing waves and/or spatially propagating waves, depending on the relative magnitudes of the real and imaginary parts of each complex POP (von Storch et al., 1988).

As with many other techniques, the data may be pre-processed using PCA, with $\mathbf{x}$ in equation (12.2.1) replaced by its PCs. The description of POP analysis in Wu et al. (1994) includes this initial step, which provides additional insights.

Kooperberg and O'Sullivan (1996) introduce and illustrate a technique which they describe as a hybrid of PCA and POP analysis. The analogous quantities to POPs resulting from the technique are called Predictive Oscillation Patterns (PROPs). In their model, $\mathbf{x}_t$ is written as a linear transformation of a set of underlying 'forcing functions,' which in turn are linear functions of $\mathbf{x}_t$. Kooperberg and O'Sullivan (1996) find an expression for an upper bound for forecast errors in their model, and PROP analysis minimizes this quantity. The criterion is such that it simultaneously attempts to account for as much as possible of $\mathbf{x}_t$, as with PCA, and to reproduce as well as possible the temporal dependence in $\mathbf{x}_t$, as in POP analysis.

In an earlier technical report, Kooperberg and O'Sullivan (1994) mention the possible use of canonical correlation analysis (CCA; see Section 9.3) in a time series context. Their suggestion is that a second group of variables is created by shifting the usual measurements at p locations by one time period. CCA is then used to find relationships between the original and time-lagged sets of variables.

12.2.3 Hilbert (Complex) EOFs

There is some confusion in the literature over the terminology 'complex EOFs,' or 'complex PCA.' It is perfectly possible to perform PCA on complex numbers, as well as real numbers, whether or not the measurements are made over time. We return to this general version of complex PCA in Section 13.8. Within the time series context, and especially for meteorological time series, the term 'complex EOFs' has come to refer to a special type of complex series. To reduce confusion, von Storch and Zwiers (1999) suggest (for reasons that will soon become apparent) referring to this procedure as Hilbert EOF analysis. We will follow this recommendation. Baines has suggested removing ambiguity entirely by denoting the analysis as 'complex Hilbert EOF analysis.' The technique seems to have originated with Rasmusson et al. (1981), by whom it was referred to as *Hilbert singular decomposition*.

Suppose that $\mathbf{x}_t$, $t = 1, 2, \ldots, n$ is a p-variate time series, and let

$$\mathbf{y}_t = \mathbf{x}_t + i\mathbf{x}_t^H, \tag{12.2.2}$$

where $i = \sqrt{-1}$ and $\mathbf{x}_t^H$ is the Hilbert transform of $\mathbf{x}_t$, defined as

$$\mathbf{x}_t^H = \sum_{s=0}^{\infty} \frac{2}{(2s+1)\pi}(\mathbf{x}_{(t+2s+1)} - \mathbf{x}_{(t-2s-1)}).$$

The definition assumes that $\mathbf{x}_t$ is observed at an infinite number of times $t = \ldots, -1, 0, 1, 2, \ldots$. Estimation of $\mathbf{x}_t^H$ for finite samples, to use in equation (12.2.2), is discussed by von Storch and Zwiers (1999, Section 16.2.4) and Bloomfield and Davis (1994).

If a series is made up of oscillatory terms, its Hilbert transform advances each oscillatory term by $\frac{\pi}{2}$ radians. When $\mathbf{x}_t$ is comprised of a single periodic oscillation, $\mathbf{x}_t^H$ is identical to $\mathbf{x}_t$, except that it is shifted by $\frac{\pi}{2}$ radians. In the more usual case, where $\mathbf{x}_t$ consists of a mixture of two or more oscillations or pseudo-oscillations at different frequencies, the effect of transforming to $\mathbf{x}_t^H$ is more complex because the phase shift of $\frac{\pi}{2}$ is implemented separately for each frequency.

A Hilbert EOF (HEOF) analysis is simply a PCA based on the covariance matrix of $\mathbf{y}_t$ defined in (12.2.2). As with (M)SSA and POP analysis, HEOF analysis will find dominant oscillatory patterns, which may or may not be propagating in space, that are present in a standard meteorological data set of p spatial locations and n time points.

Similarly to POP analysis, the eigenvalues and eigenvectors (HEOFs) are complex, but for a different reason. Here a covariance matrix is analysed, unlike POP analysis, but the variables from which the covariance matrix is formed are complex-valued. Other differences exist between POP analysis and HEOF analysis, despite the similarities in the oscillatory structures they can detect, and these are noted by von Storch and Zwiers (1999, Section 15.1.7). HEOF analysis maximizes variances, has orthogonal component scores and is empirically based, all attributes shared with PCA. In direct contrast, POP analysis does not maximize variance, has non-orthogonal POP coefficients (scores) and is model-based. As in ordinary PCA, HEOFs may be simplified by rotation (Section 11.1), and Bloomfield and Davis (1994) discuss how this can be done.

There is a connection between HEOF analysis and PCA in the frequency domain which is discussed in Section 12.4.1. An example of HEOF analysis is now given.

Southern Hemisphere Sea Surface Temperature

This example and its figures are taken, with permission, from Cai and Baines (2001). The data are sea surface temperatures (SSTs) in the Southern Hemisphere. Figure 12.4 gives a shaded contour map of the coefficients in the first four (ordinary) PCs for these data (the first four EOFs), to-

gether with the variance accounted for by each PC. Figure 12.5 displays similar plots for the real and imaginary parts of the first Hilbert EOF (labelled CEOFs on the plots). It can be seen that the real part of the first Hilbert EOF in Figure 12.5 looks similar to EOF1 in Figure 12.4. There are also similarities between EOF3 and the imaginary part of the first Hilbert EOF.

Figures 12.6, 12.7 show plots of time series (scores), labelled temporal coefficients in 12.7, for the first and third ordinary PCs, and the real and imaginary parts of the first HEOF, respectively. The similarity between the first PC and the real part of the first HEOF is obvious. Both represent the same oscillatory behaviour in the series. The imaginary part of the first HEOF is also very similar, but lagged by $\frac{\pi}{2}$, whereas the scores on the third EOF show a somewhat smoother and more regular oscillation. Cai and Baines (2001) note that the main oscillations visible in Figures 12.6, 12.7 can be identified with well-known El Niño-Southern Oscillation (ENSO) events. They also discuss other physical interpretations of the results of the HEOF analysis relating them to other meteorological variables, and they provide tests for statistical significance of HEOFs.

The main advantage of HEOF analysis over ordinary PCA is its ability to identify and reconstruct propagating waves, whereas PCA only finds standing oscillations. For the first and second Hilbert EOFs, and for their sum, this propagating behaviour is illustrated in Figure 12.8 by the movement of similar-valued coefficients from west to east as time progresses in the vertical direction.

12.2.4 Multitaper Frequency Domain-Singular Value Decomposition (MTM SVD)

In a lengthy paper, Mann and Park (1999) describe MTM-SVD, developed earlier by the same authors. It combines multitaper spectrum estimation methods (MTM) with PCA using the singular value decomposition (SVD). The paper also gives a critique of several of the other techniques discussed in the present section, together with frequency domain PCA which is covered in Section 12.4.1. Mann and Park (1999) provide detailed examples, both real and artificial, in which MTM-SVD is implemented.

Like MSSA, POP analysis, HEOF analysis and frequency domain PCA, MTM-SVD looks for oscillatory behaviour in space and time. It is closest in form to PCA of the spectral matrix $\mathbf{F}(\lambda)$ (Section 12.4.1), as it operates in the frequency domain. However, in transferring from the time domain to the frequency domain MTM-SVD constructs a set of different tapered Fourier transforms (hence the 'multitaper' in its name). The frequency domain matrix is then subjected to a singular value decomposition, giving 'spatial' or 'spectral' EOFs, and 'principal modulations' which are analogous to principal component scores. Mann and Park (1999) state that the

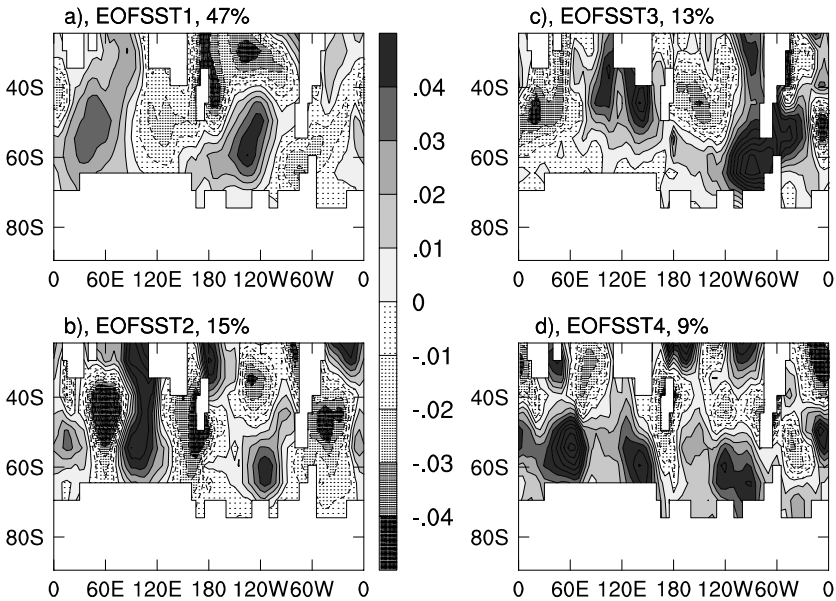

Figure 12.4. The first four EOFs for Southern Hemisphere SST.

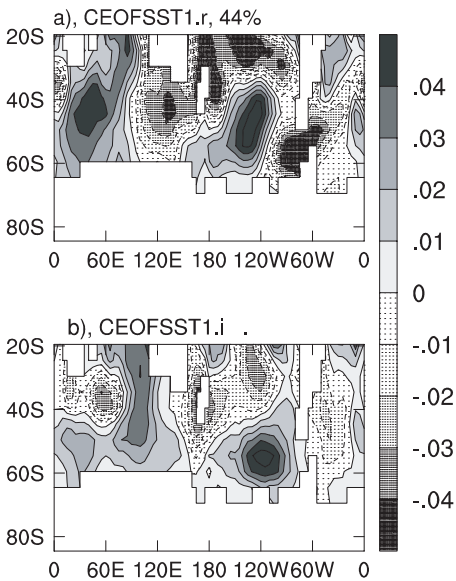

Figure 12.5. Real and imaginary parts of the first Hilbert EOF for Southern Hemisphere SST.

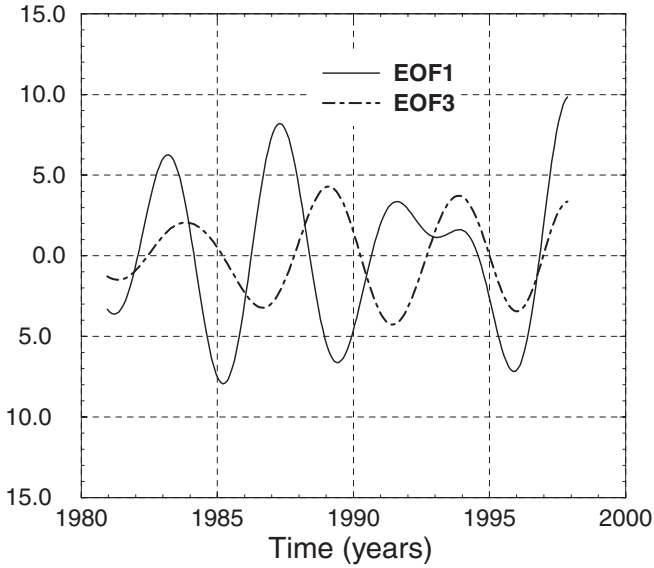

Figure 12.6. Plots of temporal scores for EOF1 and EOF3 for Southern Hemisphere SST.

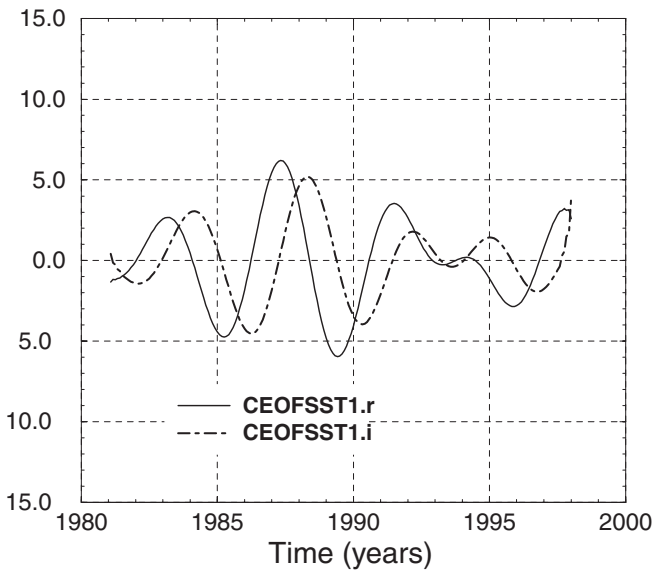

Figure 12.7. Plots of temporal scores for real and imaginary parts of the first Hilbert EOF for Southern Hemisphere SST.

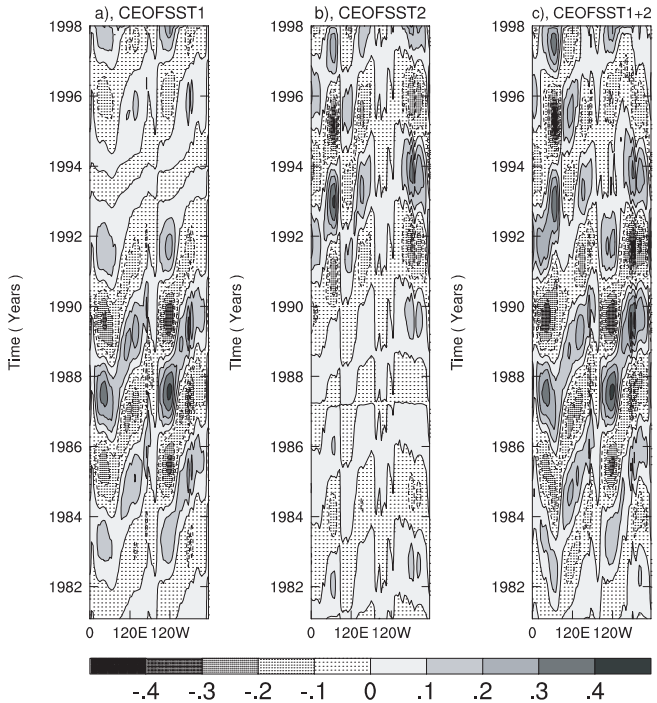

Figure 12.8. Propagation of waves in space and time in Hilbert EOF1, Hilbert EOF2, and the sum of these two Hilbert EOFs.

distinction between MTM-SVD and standard frequency domain PCA is that the former provides a *local* frequency domain decomposition of the different spectral estimates given by the multitapers, whereas the latter produces a *global* frequency domain decomposition over the spectral estimates. Mann and Park (1999) describe tests for the statistical significance of the oscillations found by MTM-SVD using a bootstrap approach, which, it is claimed, is effective for a general, smoothly varying, coloured noise background, and is not restricted to red noise as in Allen and Smith (1996) and Allen and Robertson (1996).

12.2.5 Cyclo-Stationary and Periodically Extended EOFs (and POPs)

The assumption of temporal stationarity is implicit for most of the methods described in this chapter. In meteorological data there is often a cycle with a fixed period, most commonly the annual cycle but sometimes a diurnal cycle. There may then be stationarity for times at the same point in the cycle but not across different points in the cycle. For example, for monthly data the probability distribution may be same every April but different in

April than in June. Similarly, the joint distribution for April and August, which have a four-month separation, may be the same in different years, but different from the joint distribution for June and October, even though the latter are also separated by four months. Such behaviour is known as *cyclo-stationarity*, and both PCA and POP analysis have been modified for such data.

The modification is easier to explain for POP analysis than for PCA. Suppose that τ is the length of the cycle; for example $\tau = 12$ for monthly data and an annual cycle. Then for a time series of length $n = n'\tau$ equation (12.2.1) is replaced by

$$(\mathbf{x}_{(s\tau+t+1)} - \boldsymbol{\mu}_{(t+1)}) = \boldsymbol{\Upsilon}_t(\mathbf{x}_{(s\tau+t)} - \boldsymbol{\mu}_t) + \boldsymbol{\epsilon}_{(s\tau+t)}, \qquad (12.2.3)$$
$$t = 0, 1, \ldots, (\tau - 2);\ s = 0, 1, 2, \ldots, (n' - 1),$$

with $(t + 1)$ replaced by 0, and s replaced by $(s + 1)$ on the left-hand side of (12.2.3) when $t = (\tau - 1)$ and $s = 0, 1, 2, \ldots, (n' - 2)$. Here the mean $\boldsymbol{\mu}_t$ and the matrix $\boldsymbol{\Upsilon}_t$ can vary within cycles, but not between cycles. Cyclo-stationary POP analysis estimates $\boldsymbol{\Upsilon}_0, \boldsymbol{\Upsilon}_1, \ldots, \boldsymbol{\Upsilon}_{(\tau-1)}$ and is based on an eigenanalysis of the product of those estimates $\hat{\boldsymbol{\Upsilon}}_0\hat{\boldsymbol{\Upsilon}}_1 \ldots \hat{\boldsymbol{\Upsilon}}_{(\tau-1)}$.

The cyclo-stationary variety of PCA is summarized by Kim and Wu (1999) but is less transparent in its justification than cyclo-stationary POP analysis. First, vectors $\mathbf{a}_{0t}, \mathbf{a}_{1t}, \ldots, \mathbf{a}_{(\tau-1)t}$ are found such that $\mathbf{x}_t = \sum_{j=0}^{\tau-1} \mathbf{a}_{jt} e^{\frac{2\pi i j t}{\tau}}$, and a new vector of variables is then constructed by concatenating $\mathbf{a}_{0t}, \mathbf{a}_{1t}, \ldots, \mathbf{a}_{(\tau-1)t}$. Cyclo-stationary EOFs (CSEOFs) are obtained as eigenvectors of the covariance matrix formed from this vector of variables. Kim and Wu (1999) give examples of the technique, and also give references explaining how to calculate CSEOFs.

Kim and Wu (1999) describe an additional modification of PCA that deals with periodicity in a time series, and which they call a *periodically extended* EOF technique. It works by dividing a series of length $n = n'\tau$ into n' blocks of length τ. A covariance matrix, $\mathbf{S}^{EX}$, is then computed as

$$\begin{bmatrix} \mathbf{S}_{11}^{EX} & \mathbf{S}_{12}^{EX} & \cdots & \mathbf{S}_{1n'}^{EX} \\ \mathbf{S}_{21}^{EX} & \mathbf{S}_{22}^{EX} & \cdots & \mathbf{S}_{2n'}^{EX} \\ \vdots & \vdots & & \\ \mathbf{S}_{n'1}^{EX} & \mathbf{S}_{n'2}^{EX} & \cdots & \mathbf{S}_{n'n'}^{EX} \end{bmatrix},$$

in which the (i, j)th element of $\mathbf{S}_{kl}^{EX}$ is the sample covariance between the measurements at the ith location at time k in each block and at the jth location at time l in the same block. The description of the technique in Kim and Wu (1999) is sketchy, but it appears that whereas in ordinary PCA, covariances are calculated by averaging over *all* time points, here the averaging is done over times at the same point within each block *across* blocks. A similar averaging is implicit in cyclo-stationary POP analysis. Examples of periodically extended EOFs are given by Kim and Wu (1999).

12.2.6 Examples and Comparisons

Relatively few examples have been given in this section, due in part to their complexity and their rather specialized nature. Several of the techniques described claim to be able to detect stationary and propagating waves in the standard type of spatio-temporal meteorological data. Each of the methods has proponents who give nice examples, both real and artificial, in which their techniques appear to work well. However, as with many other aspects of PCA and related procedures, caution is needed, lest enthusiasm leads to 'over-interpretation.' With this caveat firmly in mind, examples of SSA can be found in Elsner and Tsonis (1996), Vautard (1995); MSSA in Plaut and Vautard (1994), Mote et al. (2000); POPs in Kooperberg and O'Sullivan (1996) with cyclo-stationary POPs in addition in von Storch and Zwiers (1999, Chapter 15); Hilbert EOFs in Horel (1984), Cai and Baines (2001); MTM-SVD in Mann and Park (1999); cyclo-stationary and periodically extended EOFs in Kim and Wu (1999). This last paper compares eight techniques (PCA, PCA plus rotation, extended EOFs (MSSA), Hilbert EOFs, cyclo-stationary EOFs, periodically extended EOFs, POPs and cyclo-stationary POPs) on artificial data sets with stationary patterns, with patterns that are stationary in space but which have amplitudes changing in time, and with patterns that have periodic oscillations in space. The results largely confirm what might be expected. The procedures designed to find oscillatory patterns do not perform well for stationary data, the converse holds when oscillations are present, and those techniques devised for cyclo-stationary data do best on such data.

12.3 Functional PCA

There are many circumstances when the data are curves. A field in which such data are common is that of chemical spectroscopy (see, for example, Krzanowski et al. (1995), Mertens (1998)). Other examples include the trace on a continuously recording meteorological instrument such as a barograph, or the trajectory of a diving seal. Other data, although measured at discrete intervals, have an underlying continuous functional form. Examples include the height of a child at various ages, the annual cycle of temperature recorded as monthly means, or the speed of an athlete during a race.

The basic ideas of PCA carry over to this continuous (functional) case, but the details are different. In Section 12.3.1 we describe the general set-up in functional PCA (FPCA), and discuss methods for estimating functional PCs in Section 12.3.2. Section 12.3.3 presents an example. Finally, Section 12.3.4 briefly covers some additional topics including curve registration, bivariate FPCA, smoothing, principal differential analysis, prediction, discrimination, rotation, density estimation and robust FPCA.

A key reference for functional PCA is the book by Ramsay and Silverman (1997), written by two researchers in the field who, together with Besse (see, for example, Besse and Ramsay, 1986), have been largely responsible for bringing these ideas to the attention of statisticians. We draw heavily on this book in the present section. However, the ideas of PCA in a continuous domain have also been developed in other fields, such as signal processing, and the topic is an active research area. The terminology 'Karhunen-Loève expansion' is in common use in some disciplines to denote PCA in a continuous domain. Diamantaras and Kung (1996, Section 3.2) extend the terminology to cover the case where the data are discrete time series with a theoretically infinite number of time points.

In atmospheric science the Karhunen-Loève expansion has been used in the case where the continuum is spatial, and the different observations correspond to different discrete times. Preisendorfer and Mobley (1988, Section 2d) give a thorough discussion of this case and cite a number of earlier references dating back to Obukhov (1947). Bouhaddou et al. (1987) independently consider a spatial context for what they refer to as 'principal component analysis of a stochastic process,' but which is PCA for a (two-dimensional spatial) continuum of variables. They use their approach to approximate both a spatial covariance function and the underlying spatial stochastic process, and compare it with what they regard as the less flexible alternative of kriging. Guttorp and Sampson (1994) discuss similar ideas in a wider review of methods for estimating spatial covariance matrices. Durbin and Knott (1972) derive a special case of functional PCA in the context of goodness-of-fit testing (see Section 14.6.2).

12.3.1 The Basics of Functional PCA (FPCA)

When data are functions, our usual data structure $x_{ij},\ i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, p$ is replaced by $x_i(t),\ i = 1, 2, \ldots, n$ where t is continuous in some interval. We assume, as elsewhere in the book, that the data are centred, so that a mean curve $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i(t)$ has been subtracted from each of the original curves $\tilde{x}_i(t)$. Linear functions of the curves are now integrals instead of sums, that is $z_i = \int a(t)x_i(t)dt$ rather than $z_i = \sum_{j=1}^{p} a_j x_{ij}$. In 'ordinary' PCA the first PC has weights $a_{11}, a_{21}, \ldots, a_{p1}$, which maximize the sample variance $\text{var}(z_i)$, subject to $\sum_{j=1}^{p} a_{j1}^2 = 1$. Because the data are centred,

$$\text{var}(z_{i1}) = \frac{1}{n-1}\sum_{i=1}^{n} z_{i1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\Big[\sum_{j=1}^{p} a_{j1}x_{ij}\Big]^2.$$

Analogously for curves, we find $a_1(t)$ which maximizes

$$\frac{1}{n-1}\sum_{i=1}^{n} z_{i1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[\int a_1(t)x_i(t)\,dt\right]^2,$$

subject to $\int a_1^2(t)dt = 1$. Here, and elsewhere, the integral is over the range of values of t for which the data are observed. Subsequent FPCs are defined successively, as with ordinary PCA, to maximise

$$\text{var}(z_{ik}) = \frac{1}{n-1}\sum_{i=1}^{n}\left[\int a_k(t)x_i(t)\,dt\right]^2,$$

subject to $\int a_k^2(t)dt = 1$; $\int a_k(t)a_h(t)dt = 0$, $k = 2, 3, \ldots$; $h = 1, 2, \ldots$; $h < k$.

The sample covariance between $x(s)$ and $x(t)$ can be defined as $S(s,t) = \frac{1}{(n-1)}\sum_{i=1}^{n} x_i(s)x_i(t)$, with a corresponding definition for correlation, and to find the functional PCs an eigenequation is solved involving this covariance function. Specifically, we solve

$$\int S(s,t)a(t)dt = la(s). \tag{12.3.1}$$

Comparing (12.3.1) to the eigenequation $\mathbf{Sa} = l\mathbf{a}$ for ordinary PCA, the pre-multiplication of a vector of weights by a matrix of covariances on the left-hand side is replaced by an integral operator in which the covariance function is multiplied by a weight *function* $a(t)$ and then integrated. In ordinary PCA the number of solutions to the eigenequation is usually p, the number of variables. Here the p variables are replaced by a infinite number of values for t but the number of solutions is still finite, because the number of curves n is finite. The number of non-zero eigenvalues $l_1, l_2, \ldots$, and corresponding functions $a_1(t), a_2(t), \ldots$ cannot exceed $(n-1)$.

12.3.2 Calculating Functional PCs (FPCs)

Unless the curves are all fairly simple functional forms, it is not possible to solve (12.3.1) exactly. Ramsay and Silverman (1997, Section 6.4) give three computational methods for FPCA, but in most circumstances they represent *approximate* solutions to (12.3.1). In the first of the three, the data are discretized. The values $x_i(t_1), x_i(t_2), \ldots, x_i(t_p)$ form the ith row of an $(n \times p)$ data matrix which is then analysed using standard PCA. The times $t_1, t_2, \ldots, t_p$ are usually chosen to be equally spaced in the range of continuous values for t. To convert the eigenvectors found from this PCA into functional form, it is necessary to renormalize the eigenvectors and then interpolate them with a suitable smoother (Ramsay and Silverman, 1997, Section 6.4.1).

A second approach assumes that the curves $x_i(t)$ can be expressed linearly in terms of a set of G basis functions, where G is typically less than n. If $\phi_1(t), \phi_2(t), \ldots, \phi_G(t)$ are the basis functions, then $x_i(t) = \sum_{g=1}^{G} c_{ig}\phi_g(t)$, $i = 1, 2, \ldots, n$ or, in matrix form, $\mathbf{x}(t) = \mathbf{C}\boldsymbol{\phi}(t)$, where $\mathbf{x}'(t) = (x_1(t), x_2(t), \ldots, x_n(t))$, $\boldsymbol{\phi}'(t) = (\phi_1(t), \phi_2(t), \ldots, \phi_G(t))$ and $\mathbf{C}$ is an $(n \times G)$ matrix with (i,g)th element c_{ig}.

The sample covariance between $x(s)$ and $x(t)$ can be written

$$\frac{1}{n-1}\mathbf{x}'(s)\mathbf{x}(t) = \frac{1}{n-1}\boldsymbol{\phi}'(s)\mathbf{C}'\mathbf{C}\boldsymbol{\phi}(t).$$

Any eigenfunction $a(t)$ can be expressed in terms of the basis functions as $a(t) = \sum_{g=1}^{G} b_g\phi_g(t) = \boldsymbol{\phi}'(t)\mathbf{b}$ for some vector of coefficients $\mathbf{b}' = (b_1, b_2, \ldots, b_G)$. The left-hand-side of equation (12.3.1) is then

$$\begin{aligned}\int S(s,t)a(t)\,dt &= \int \frac{1}{n-1}\boldsymbol{\phi}'(s)\mathbf{C}'\mathbf{C}\boldsymbol{\phi}(t)\boldsymbol{\phi}'(t)\mathbf{b}\,dt \\ &= \frac{1}{n-1}\boldsymbol{\phi}'(s)\mathbf{C}'\mathbf{C}\left[\int \boldsymbol{\phi}(t)\boldsymbol{\phi}'(t)\,dt\right]\mathbf{b}.\end{aligned}$$

The integral is a $(G\times G)$ matrix $\mathbf{W}$ whose (g,h)th element is $\int \phi_g(t)\phi_h(t)dt$. If the basis is orthogonal, $\mathbf{W}$ is simply the identity matrix $\mathbf{I}_G$. Hence choosing an orthogonal basis in circumstances where such a choice makes sense, as with a Fourier basis for periodic data, gives simplified calculations. In general (12.3.1) becomes

$$\frac{1}{n-1}\boldsymbol{\phi}'(s)\mathbf{C}'\mathbf{C}\mathbf{W}\mathbf{b} = l\boldsymbol{\phi}'(s)\mathbf{b}$$

but, because this equation must hold for all available values of s, it reduces to

$$\frac{1}{n-1}\mathbf{C}'\mathbf{C}\mathbf{W}\mathbf{b} = l\mathbf{b}. \tag{12.3.2}$$

When $\int a^2(t)dt = 1$ it follows that

$$1 = \int a^2(t)\,dt = \int \mathbf{b}'\boldsymbol{\phi}(t)\boldsymbol{\phi}'(t)\mathbf{b}\,dt = \mathbf{b}'\mathbf{W}\mathbf{b}.$$

If $a_k(t)$ is written in terms of the basis functions as $a_k(t) = \sum_{g=1}^{G} b_{kg}\phi_g(t)$, with a similar expression for $a_l(t)$, then $a_k(t)$ is orthogonal to $a_l(t)$ if $\mathbf{b}_k'\mathbf{W}\mathbf{b}_l = 0$, where $\mathbf{b}_k' = (b_{k1}, b_{k2}, \ldots, b_{kG})$, and $\mathbf{b}_l'$ is defined similarly.

In an eigenequation, the eigenvector is usually normalized to have unit length (norm). To convert (12.3.2) into this form, define $\mathbf{u} = \mathbf{W}^{\frac{1}{2}}\mathbf{b}$. Then $\mathbf{u}'\mathbf{u} = 1$ and (12.3.2) can be written

$$\frac{1}{n-1}\mathbf{W}^{\frac{1}{2}}\mathbf{C}'\mathbf{C}\mathbf{W}^{\frac{1}{2}}\mathbf{u} = l\mathbf{u}. \tag{12.3.3}$$

Equation (12.3.3) is solved for l and $\mathbf{u}$, $\mathbf{b}$ is obtained as $\mathbf{W}^{-\frac{1}{2}}\mathbf{u}$, and finally $a(t) = \boldsymbol{\phi}'(t)\mathbf{b} = \boldsymbol{\phi}'(t)\mathbf{W}^{-\frac{1}{2}}\mathbf{u}$.

The special case where the basis is orthogonal has already been mentioned. Here $\mathbf{W} = \mathbf{I}_G$, so $\mathbf{b} = \mathbf{u}$ is an eigenvector of $\frac{1}{n-1}\mathbf{C}'\mathbf{C}$. Another special case, noted by Ramsay and Silverman (1997), occurs when the data curves themselves are taken as the basis. Then $\mathbf{C} = \mathbf{I}_n$ and $\mathbf{u}$ is an eigenvector of $\frac{1}{n-1}\mathbf{W}$.

The third computational method described by Ramsay and Silverman (1997) involves applying numerical quadrature schemes to the integral on the left-hand side of (12.3.1). Castro et al. (1986) used this approach in an early example of functional PCA. Quadrature methods can be adapted to cope with irregularly spaced data (Ratcliffe and Solo, 1998). Aguilera et al. (1995) compare a method using a trigonometric basis with one based on a trapezoidal scheme, and find that the behaviour of the two algorithms is similar except at the extremes of the time interval studied, where the trapezoidal method is superior.

Preisendorfer and Mobley (1988, Section 2d) have two interesting approaches to finding functional PCs in the case where t represents spatial position and different observations correspond to different discrete times. In the first the eigenequation (12.3.1) is replaced by a dual eigenequation, obtained by using a relationship similar to that between $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}\mathbf{X}'$, which was noted in the proof of Property G4 in Section 3.2. This dual problem is discrete, rather than continuous, and so is easier to solve. Its eigenvectors are the PC scores for the continuous problem and an equation exists for calculating the eigenvectors of the original continuous eigenequation from these scores.

The second approach is similar to Ramsay and Silverman's (1997) use of basis functions, but Preisendorfer and Mobley (1988) also compare the basis functions and the derived eigenvectors (EOFs) in order to explore the physical meaning of the latter. Bouhaddou et al. (1987) independently proposed the use of interpolating basis functions in the implementation of a continuous version of PCA, given what is necessarily a discrete-valued data set.

12.3.3 Example - 100 km Running Data

Here we revisit the data that were first introduced in Section 5.3, but in a slightly different format. Recall that they consist of times taken for ten 10 km sections by 80 competitors in a 100 km race. Here we convert the data into speeds over each section for each runner. Ignoring 'pitstops,' it seems reasonable to model the speed of each competitor through the race as a continuous curve. In this example the horizontal axis represents position in (one-dimensional) space rather than time. Figure 12.9 shows the speed for each competitor, with the ten discrete points joined by straight lines. Despite the congestion of lines on the figure, the general pattern of slowing down is apparent. Figure 12.10 shows the coefficients for the first three ordinary PCs of these speed data. In Figure 12.11 the piecewise linear plots of Figure 12.10 are smoothed using a spline basis. Finally, Figure 12.12 gives the eigenfunctions from a FPCA of the data, using spline basis functions, implemented in S-Plus.

There are strong similarities between Figures 12.10–12.12, though some differences exist in the details. The first PC, as with the 'time taken' version

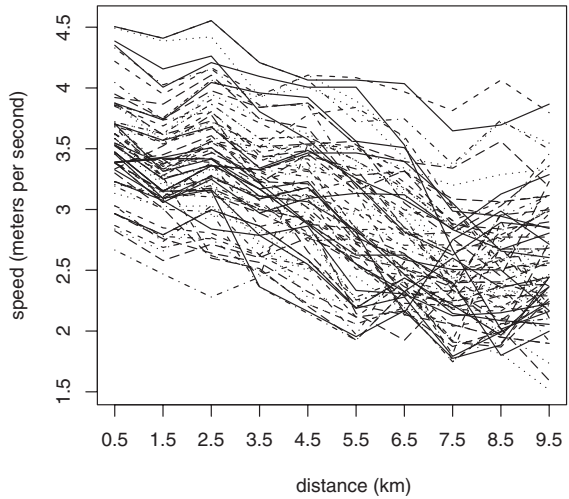

Figure 12.9. Plots of speed for 80 competitors in a 100 km race.

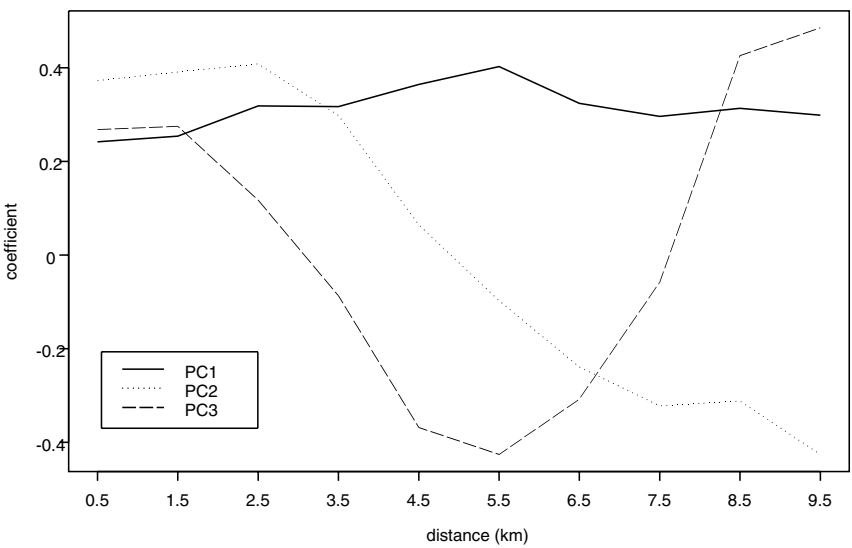

Figure 12.10. Coefficients for first three PCs from the 100 km speed data.

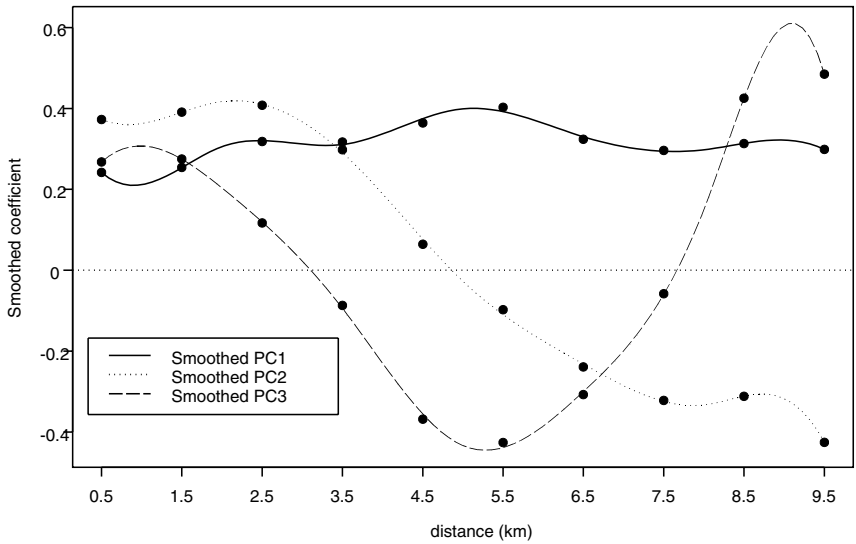

Figure 12.11. Smoothed version of Figure 12.10 using a spline basis; dots are coefficients from Figure 12.10.

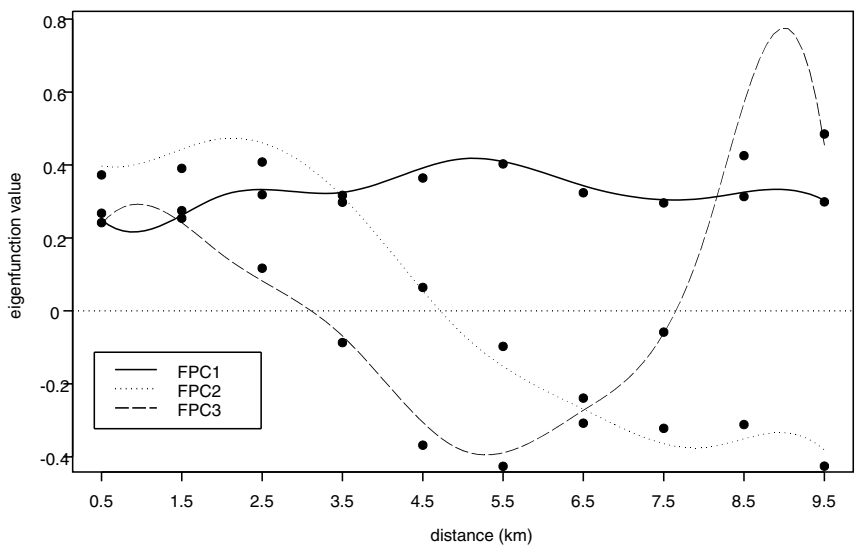

Figure 12.12. Coefficients (eigenfunctions) for the first three components in a functional PCA of the 100 km speed data using a spline basis; dots are coefficients from Figure 12.10.

of the data in Section 5.3, is a measure of overall speed throughout the race. The second component in all cases represents the degree to which runners slow down during the race. The pattern in Figure 12.10 is very much like that in Table 5.2. Figures 12.11 and 12.12 have a similar shape for this component, although the absolute values of the eigenfunction in Figure 12.12 are slightly larger than the coefficients in Figure 12.11. The reason for this is that the speed for the first 10 km section is plotted at distance equal to 0.5 km, the speed for the second section at 1.5 km and so on. The eigenfunctions for the FPCs are then calculated over the interval 0.5 to 9.5, a range of 9, whereas there are 10 elements in the ordinary PC eigenvectors, leading to different normalizations. Calculating the FPCs over the range 0 to 10 leads to erratic behaviour at the ends of the interval. The pattern for the third PC is again similar in all three figures, except at the extreme right-hand end.

12.3.4 Further Topics in FPCA

This subsection collects together briefly a number of topics that are relevant to FPCA. As in the rest of this section, the main reference is Ramsay and Silverman (1997).

Curve Registration

In some examples where the data are functions, the individual curves $x_i(t)$ may be observed over different ranges for t. Most procedures for analysing functional data assume the same range of t for all curves. Furthermore, it is often the case that the horizontal extent of the curve is of less importance than its shape. For example, suppose that the trajectory of underwater dives made by seals is of interest (see Schreer et al. (1998)). Although the lengths of the dives may be informative, so are their shapes, and to compare these the curves should be aligned to start and finish at the same respective times. This process is known as *curve registration*. Another example arises when outlines of sliced carrots are examined in order to determine to which variety of carrot an individual specimen belongs (Horgan et al., 2001). More complicated situations are discussed by Ramsay and Silverman (1997, Chapter 5), in which several landmarks along the curves must be aligned, leading to differential stretching and expansion of 'time' for different parts of the curves. Capra and Müller (1997) use an accelerated time model to align mortality curves for medflies. They refer to their accelerated time as 'eigenzeit.'

Sometimes alignment is desirable in the vertical, as well the horizontal, direction. Consider again the dive trajectories of seals. If the shapes of the dives are more important than their depths, then the dives can be aligned vertically at their deepest points.

To analyse only the registered curves, without taking into account of what was done in the registration process, is to throw away information.

Keeping one or more parameters for each curve, defining how registration was done, leads to what Ramsay and Silverman (1997) called 'mixed data.' Each observation for such data consists of a curve, together with p other 'ordinary' variables. Ramsay and Silverman (1997, Chapter 8) discuss the analysis of such data.

Bivariate FPCA

In other cases, the data are not 'mixed' but there is more than one curve associated with each individual. An example involving changes of angles in both hip and knee during a gait cycle is described by Ramsay and Silverman (1997, Section 6.5). They discuss the analysis of bivariate curves from this example using bivariate FPCA. Suppose that the two sets of curves are $x_1(t), x_2(t), \ldots, x_n(t)$; $y_1(t), y_2(t), \ldots, y_n(t)$. Define a bivariate covariance function $S(s,t)$ as

$$\begin{bmatrix} S_{XX}(s,t) & S_{XY}(s,t) \\ S_{YX}(s,t) & S_{YY}(s,t) \end{bmatrix},$$

where $\mathbf{S}_{XX}(s,t)$, $\mathbf{S}_{YY}(s,t)$ are covariance functions defined, as earlier, for $(x(s), x(t))$ and $(y(s), y(t))$, respectively, and $\mathbf{S}_{XY}(s,t)$ has elements that are covariances between $x(s)$ and $y(t)$. Suppose that

$$z_{Xi} = \int a_X(t) x_i(t)\, dt, \quad z_{Yi} = \int a_Y(t) y_i(t) dt.$$

Finding $a_X(t)$, $a_Y(t)$ to maximize $\frac{1}{n-1}\sum_{i=1}^{n}(z_{Xi}^2 + z_{Yi}^2)$ leads to the eigenequations

$$\int S_{XX}(s,t) a_X(t)\, dt + \int S_{XY}(s,t) a_Y(t)\, dt = l a_X(t)$$
$$\int S_{YX}(s,t) a_X(t)\, dt + \int S_{YY}(s,t) a_Y(t)\, dt = l a_Y(t).$$

This analysis can be extended to the case of more than two curves per individual.

Smoothing

If the data are not smooth, the weighting functions $a(t)$ in FPCA may not be smooth either. With most curves, an underlying smoothness is expected, with the superimposed roughness being due to noise that should ideally be removed. Ramsay and Silverman (1997, Chapter 7) tackle this problem. Their main approach incorporates a roughness penalty into FPCA's variance-maximizing problem. The second derivative of a curve is often taken as a measure of its roughness and, if $D^2 a(t)$ represents the second derivative of $a(t)$, a smooth curve requires a small value of $D^2 a(t)$. Ramsay and Silverman's approach is to maximize

$$\frac{\frac{1}{n-1}\sum_{i=1}^{n}[\int a(t) x_i(t) dt]^2}{\int a^2(t) dt + \lambda \int (D^2 a(t))^2 dt} \tag{12.3.4}$$

subject to $\int a^2(t)dt = 1$, where λ is a tuning parameter. Taking $\lambda = 0$ simply gives unsmoothed FPCA; as λ increases, so does the smoothness of the optimal $a(t)$. Ramsay and Silverman (1997, Section 7.3) show that solving the optimization problem reduces to solving the eigenequation

$$\int S(s,t)a(t)dt = l(1 + \lambda D^4)a(s), \tag{12.3.5}$$

where $D^4a(s)$ is the fourth derivative of $a(s)$. In their Section 7.4 they give some computational methods for (approximately) solving equation (12.3.5), with the choice of λ discussed in Section 7.5. Minimizing (12.3.4) by solving (12.3.5) implies a form of orthogonality between successive $a_k(t)$ that is different from the usual one. Ocaña et al. (1999) interpret this as a different choice of norm and inner product, and discuss the choice of norm in FPCA from a theoretical point of view. If (12.3.5) is solved by using basis functions for the data, the number of basis functions and the smoothing parameter λ both need to be chosen. Ratcliffe and Solo (1998) address the problem of choosing both simultaneously and propose an improved cross-validation procedure for doing so.

An alternative to incorporating smoothing directly into FPCA is to smooth the data first and then conduct unsmoothed FPCA on the smoothed curves. The smoothing step uses the well-known idea of minimizing the sum of squares between fitted and observed curves, supplemented by a penalty function based on lack of smoothness (a roughness penalty). The quantity we wish to minimize can be written succicntly as

$$\|x(t) - \hat{x}(t)\|^2 + \lambda\|\hat{x}(t)\|^2, \tag{12.3.6}$$

where $x(t)$, $\hat{x}(t)$ denote observed and fitted curves, respectively, $\|\,.\,\|$ denotes an appropriately chosen norm, not necessarily the same in the two parts of the expression, and λ is a tuning parameter. The second norm is often related to the second derivative, $D^2[\hat{x}(t)]$, which is a commonly used measure of roughness.

Besse et al. (1997) use a similar approach in which they optimize a criterion which, like (12.3.6), is a sum of two terms, one a measure of fit and the other a roughness penalty with a tuning parameter λ. Besse et al.'s model assumes that, apart from noise, the curves lie in a q-dimensional space. A version $\hat{R}_q$ of the criterion R_q, suggested by Besse and de Falguerolles (1993) and discussed in Section 6.1.5, is minimized with respect to q and λ simultaneously. If smoothing of the data is done using splines, then the PCA on the resulting interpolated data is equivalent to a PCA on the original data, but with a different metric (Besse, 1988). We return to this topic in Section 14.2.2.

A more complicated approach to smoothing followed by PCA, involving different smoothing parameters for different FPCs, is discussed by Ramsay and Silverman (1997, Section 7.8). Grambsch et al. (1995) also use a smoothing process—though a different one—for the data, followed by PCA

on the smoothed data. Kneip (1994) independently suggested smoothing the data, followed by PCA on the smoothed data, in the context of fitting a model in which a small number of functions is assumed to underlie a set of regression curves. Theoretical properties of Kneip's (1994) procedure are examined in detail in his paper. Champely and Doledec (1997) use lo(w)ess (locally weighted scatterplot smoothing—Cleveland, 1979, 1981) to fit smooth trend and periodic curves to water quality data, and then apply FPCA separately to the trend and periodic curves.

Principal Differential Analysis

Principal differential analysis (PDA), a term coined by Ramsay (1996) and discussed in Ramsay and Silverman (1997, Chapter 14) is another method of approximating a set of curves by a smaller number of functions. Although PDA has some similarities to FPCA, which we note below, it concentrates on finding functions that have a certain type of smoothness, rather than maximizing variance. Define a linear differential operator

$$L = w_0 I + w_1 D + \ldots + w_{m-1} D^{m-1} + D^m,$$

where D^i, as before, denotes the ith derivative operator and I is the identity operator. PDA finds weights $w_0, w_1, \ldots, w_{m-1}$ for which $[Lx_i(t)]^2$ is small for each observed curve $x_i(t)$. Formally, we minimize $\sum_{i=1}^{n} \int [Lx_i(t)]^2 \, dt$ with respect to $w_0, w_1, \ldots, w_{m-1}$.

Once $w_0, w_1, \ldots, w_{m-1}$ and hence L are found, any curve satisfying $Lx(t) = 0$ can be expressed as a linear combination of m linearly independent functions spanning the null space of the operator L. Any observed curve $x_i(t)$ can be approximated by expanding it in terms of these m functions. This is similar to PCA, where the original data can be approximated by expanding them in terms of the first few (say m) PCs. The difference is that PCA finds an m-dimensional space with a least squares fit to the original data, whereas PDA finds a space which penalizes roughness. This last interpretation follows because $Lx_i(t)$ is typically rougher than $x_i(t)$, and PDA aims to make $Lx_i(t)$, or rather $[Lx_i(t)]^2$, as small as possible when averaged over i and t. An application of PDA to the study of variations in handwriting is presented by Ramsay (2000).

Prediction and Discrimination

Aguilera et al. (1997, 1999a) discuss the idea of predicting a continuous time series by regressing functional PCs for the future on functional PCs from the past. To implement this methodology, Aguilera et al. (1999b) propose cutting the series into a number of segments n of equal length, which are then treated as n realizations of the same underlying process. Each segment is in turn divided into two parts, with the second, shorter, part to be predicted from the first. In calculating means and covariance functions, less weight is given to the segments from the early part of the

series than to those closer to the present time from where the future is to be predicted. Besse et al. (2000) also use functional PCA to predict time series, but in the context of a smoothed first order autoregessive model. The different replications of the function correspond to different years, while the function ranges over the months within years. A 'local' version of the technique is developed in which the assumption of stationarity is relaxed.

Hall et al. (2001) advocate the use of functional PCA as a dimension-reducing step in the context of discriminating between different types of radar signal. Although this differs from the usual set-up for PCA in discriminant analysis (see Section 9.1) because it notionally has an infinite number of variables in a continuum, there is still the possibility that some of the later discarded components may contain non-trivial discriminatory power.

Rotation

As with ordinary PCA, the interpretation of FPCA may be improved by rotation. In addition to the conventional rotation of coefficients in a subset of PCs (see Section 11.1), Ramsay and Silverman (1997, Section 6.3.3) suggest that the coefficients $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m$ of the first m eigenfunctions with respect to a chosen set of basis functions, as defined in Section 12.3.2, could also be rotated to help interpretation. Arbuckle and Friendly (1977) propose a variation on the usual rotation criteria of Section 11.1 for variables that are measurements at discrete points on a continuous curve. They suggest rotating the results of an ordinary PCA towards smoothness rather than towards simplicity as usually defined (see Section 11.1).

Density Estimation

Kneip and Utikal (2001) discuss functional PCA as a means of examining common structure and differences in a set of probability density functions. The densities are first estimated from a number of data sets using kernel density estimators, and these estimates are then subjected to functional PCA. As well as a specific application, which examines how densities evolve in data sets collected over a number of years, Kneip and Utikal (2001) introduce some new methodology for estimating functional PCs and for deciding how many components should be retained to represent the density estimates adequately. Their paper is followed by published discussion from four sets of discussants.

Robustness

Locantore et al. (1999) consider robust estimation of PCs for functional data (see Section 10.4).

12.4 PCA and Non-Independent Data—Some Additional Topics

In this section we collect together a number of topics from time series, and other contexts in which non-independent data arise, where PCA or related techniques are used. Section 12.4.1 describes PCA in the frequency domain, and Section 12.4.2 covers growth curves and longitudinal data. In Section 12.4.3 a summary is given of another idea (optimal fingerprints) from climatology, though one that is rather different in character from those presented in Section 12.2. Section 12.4.4 discusses spatial data, and the final subsection provides brief coverage of a few other topics, including non-independence induced by survey designs.

12.4.1 PCA in the Frequency Domain

The idea of PCA in the frequency domain clearly has no counterpart for data sets consisting of independent observations. Brillinger (1981, Chapter 9) devotes a whole chapter to the subject (see also Priestley et al. (1974)). To see how frequency domain PCs are derived, note that PCs for a p-variate random vector $\mathbf{x}$, with zero mean, can be obtained by finding $(p \times q)$ matrices $\mathbf{B}$, $\mathbf{C}$ such that

$$E[(\mathbf{x} - \mathbf{Cz})'(\mathbf{x} - \mathbf{Cz})]$$

is minimized, where $\mathbf{z} = \mathbf{B}'\mathbf{x}$. This is equivalent to the criterion that defines Property A5 in Section 2.1 It turns out that $\mathbf{B} = \mathbf{C}$ and that the columns of $\mathbf{B}$ are the first q eigenvectors of $\mathbf{\Sigma}$, the covariance matrix of $\mathbf{x}$, so that the elements of $\mathbf{z}$ are the first q PCs for $\mathbf{x}$. This argument can be extended to a time series of p variables as follows. Suppose that our series is $\ldots \mathbf{x}_{-1}$, $\mathbf{x}_0$, $\mathbf{x}_1$, $\mathbf{x}_2, \ldots$ and that $E[\mathbf{x}_t] = \mathbf{0}$ for all t. Define

$$\mathbf{z}_t = \sum_{u=-\infty}^{\infty} \mathbf{B}'_{t-u}\mathbf{x}_u,$$

and estimate $\mathbf{x}_t$ by $\sum_{u=-\infty}^{\infty} \mathbf{C}_{t-u}\mathbf{z}_u$, where

$$\ldots \mathbf{B}_{t-1}, \mathbf{B}_t, \mathbf{B}_{t+1}, \mathbf{B}_{t+2}, \ldots, \mathbf{C}_{t-1}, \mathbf{C}_t, \mathbf{C}_{t+1}, \mathbf{C}_{t+2}, \ldots$$

are $(p \times q)$ matrices that minimize

$$E\Big[\Big(\mathbf{x}_t - \sum_{u=-\infty}^{\infty} \mathbf{C}_{t-u}\mathbf{z}_u\Big)^* \Big(\mathbf{x}_t - \sum_{u=-\infty}^{\infty} \mathbf{C}_{t-u}\mathbf{z}_u\Big)\Big],$$

where * denotes conjugate transpose. The difference between this formulation and that for ordinary PCs above is that the relationships between $\mathbf{z}$ and $\mathbf{x}$ are in terms of all values of $\mathbf{x}_t$ and $\mathbf{z}_t$, at different times, rather than between a single $\mathbf{x}$ and $\mathbf{z}$. Also, the derivation is in terms of general complex

series, rather than being restricted to real series. It turns out (Brillinger, 1981, p. 344) that

$$\mathbf{B}'_u = \frac{1}{2\pi}\int_0^{2\pi} \tilde{\mathbf{B}}(\lambda)e^{iu\lambda}d\lambda$$
$$\mathbf{C}_u = \frac{1}{2\pi}\int_0^{2\pi} \tilde{\mathbf{C}}(\lambda)e^{iu\lambda}d\lambda,$$

where $\tilde{\mathbf{C}}(\lambda)$ is a $(p \times q)$ matrix whose columns are the first q eigenvectors of the matrix $\mathbf{F}(\lambda)$ given in (12.1.4), and $\tilde{\mathbf{B}}(\lambda)$ is the conjugate transpose of $\tilde{\mathbf{C}}(\lambda)$.

The q series that form the elements of $\mathbf{z}_t$ are called the first q PC series of $\mathbf{x}_t$. Brillinger (1981, Sections 9.3, 9.4) discusses various properties and estimates of these PC series, and gives an example in Section 9.6 on monthly temperature measurements at 14 meteorological stations. Principal component analysis in the frequency domain has also been used on economic time series, for example on Dutch provincial unemployment data (Bartels, 1977, Section 7.7).

There is a connection between frequency domain PCs and PCs defined in the time domain (Brillinger, 1981, Section 9.5). The connection involves Hilbert transforms and hence, as noted in Section 12.2.3, frequency domain PCA has links to HEOF analysis. Define the vector of variables $\mathbf{y}_t^H(\lambda) = (\mathbf{x}'_t(\lambda), \mathbf{x}_t'^H(\lambda))'$, where $\mathbf{x}_t(\lambda)$ is the contribution to $\mathbf{x}_t$ at frequency λ (Brillinger, 1981, Section 4.6), and $\mathbf{x}_t^H(\lambda)$ is its Hilbert transform. Then the covariance matrix of $\mathbf{y}_t^H(\lambda)$ is proportional to

$$\begin{bmatrix} \mathrm{Re}(\mathbf{F}(\lambda)) & \mathrm{Im}(\mathbf{F}(\lambda)) \\ -\mathrm{Im}(\mathbf{F}(\lambda)) & \mathrm{Re}(\mathbf{F}(\lambda)) \end{bmatrix},$$

where the functions Re(.), Im(.) denote the real and imaginary parts, respectively, of their argument. A PCA of $\mathbf{y}_t^H$ gives eigenvalues that are the eigenvalues of $\mathbf{F}(\lambda)$ with a corresponding pair of eigenvectors

$$\begin{bmatrix} \mathrm{Re}(\tilde{C}_j(\lambda)) \\ \mathrm{Im}(\tilde{C}_j(\lambda)) \end{bmatrix}, \begin{bmatrix} -\mathrm{Im}(\tilde{C}_j(\lambda)) \\ \mathrm{Re}(\tilde{C}_j(\lambda)) \end{bmatrix},$$

where $\tilde{C}_j(\lambda)$ is the jth column of $\tilde{\mathbf{C}}(\lambda)$.

Horel (1984) interprets HEOF analysis as frequency domain PCA averaged over all frequency bands. When a single frequency of oscillation dominates the variation in a time series, the two techniques become the same. The averaging over frequencies of HEOF analysis is presumably the reason behind Plaut and Vautard's (1994) claim that it is less good than MSSA at distinguishing propagating patterns with different frequencies.

Preisendorfer and Mobley (1988) describe a number of ways in which PCA is combined with a frequency domain approach. Their Section 4e discusses the use of PCA after a vector field has been transformed into the frequency domain using Fourier analysis, and for scalar-valued fields

their Chapter 12 examines various combinations of real and complex-valued harmonic analysis with PCA.

Stoffer (1999) describes a different type of frequency domain PCA, which he calls the *spectral envelope*. Here a PCA is done on the spectral matrix $\mathbf{F}(\lambda)$ relative to the time domain covariance matrix $\mathbf{\Gamma}_0$. This is a form of generalized PCA for $\mathbf{F}(\lambda)$ with $\mathbf{\Gamma}_0$ as a metric (see Section 14.2.2), and leads to solving the eigenequation $[\mathbf{F}(\lambda) - l(\lambda)\mathbf{\Gamma}_0]\mathbf{a}(\lambda) = 0$ for varying angular frequency λ. Stoffer (1999) advocates the method as a way of discovering whether the p series $x_1(t), x_2(t), \ldots x_p(t)$ share common signals and illustrates its use on two data sets involving pain perception and blood pressure.

The idea of cointegration is important in econometrics. It has a technical definition, but can essentially be described as follows. Suppose that the elements of the p-variate time series $\mathbf{x}_t$ are stationary after, but not before, differencing. If there are one or more vectors $\boldsymbol{\alpha}$ such that $\boldsymbol{\alpha}'\mathbf{x}_t$ is stationary without differencing, the p series are cointegrated. Tests for cointegration based on the variances of frequency domain PCs have been put forward by a number of authors. For example, Cubadda (1995) points out problems with previously defined tests and suggests a new one.

12.4.2 Growth Curves and Longitudinal Data

A common type of data that take the form of curves, even if they are not necessarily recorded as such, consists of measurements of growth for animals or children. Some curves such as heights are monotonically increasing, but others such as weights need not be. The idea of using principal components to summarize the major sources of variation in a set of growth curves dates back to Rao (1958), and several of the examples in Ramsay and Silverman (1997) are of this type. Analyses of growth curves are often concerned with predicting future growth, and one way of doing this is to use principal components as predictors. A form of generalized PC regression developed for this purpose is described by Rao (1987).

Caussinus and Ferré (1992) use PCA in a different type of analysis of growth curves. They consider a 7-parameter model for a set of curves, and estimate the parameters of the model separately for each curve. These 7-parameter estimates are then taken as values of 7 variables to be analyzed by PCA. A two-dimensional plot in the space of the first two PCs gives a representation of the relative similarities between members of the set of curves. Because the parameters are not estimated with equal precision, a weighted version of PCA is used, based on the fixed effects model described in Section 3.9.

Growth curves constitute a special case of longitudinal data, also known as 'repeated measures,' where measurements are taken on a number of individuals at several different points of time. Berkey et al. (1991) use PCA to model such data, calling their model a 'longitudinal principal compo-

nent model.' As they are in Rao's (1987) generalized principal component regression, the PCs are used for prediction.

Growth curves are also the subject of James et al. (2000). Their objective is to model growth curves when the points at which the curves are measured are possibly irregular, different for different individuals (observations), and sparse. Various models are proposed for such data and the differences and connections between models are discussed. One of their models is the reduced rank model

$$x_i(t) = \mu(t) + \sum_{k=1}^{q} a_k(t) z_{ik} + \epsilon_i(t), \; i = 1, 2, \ldots, n, \tag{12.4.1}$$

where $x_i(t)$ represents the growth curve for the ith individual, $\mu(t)$ is a mean curve, $\epsilon_i(t)$ is an error term for the ith individual and $a_k(t), z_{ik}$ are curves defining the principal components and PC scores, respectively, as in Section 12.3.1. James et al. (2000) consider a restricted form of this model in which $\mu(t)$ and $a_k(t)$ are expressed in terms of a spline basis, leading to a model

$$\mathbf{x}_i = \mathbf{\Phi}_i \mathbf{b}_0 + \mathbf{\Phi}_i \mathbf{B} \mathbf{z} + \boldsymbol{\epsilon}_i, \; i = 1, 2, \ldots, n. \tag{12.4.2}$$

Here $\mathbf{x}_i, \boldsymbol{\epsilon}_i$ are vectors of values $x_i(t), \epsilon_i(t)$ at the times for which measurements are made on the ith individual; $\mathbf{b}_0, \mathbf{B}$ contain coefficients in the expansions of $\mu(t), a_k(t)$, respectively in terms of the spline basis; and $\mathbf{\Phi}_i$ consists of values of that spline basis at the times measured for the ith individual. When all individuals are measured at the same time, the subscript i disappears from $\mathbf{\Phi}_i$ in (12.4.2) and the error term has covariance matrix $\sigma^2 \mathbf{I}_p$, where p is the (common) number of times that measurements are made. James et al. (2000) note that the approach is then equivalent to a PCA of the spline coefficients in $\mathbf{B}$. More generally, when the times of measurement are different for different individuals, the analysis is equivalent to a PCA with respect to the metric $\mathbf{\Phi}_i' \mathbf{\Phi}_i$. This extends the idea of metric-based PCA described in Section 14.2.2 in allowing different (non-diagonal) metrics for different observations. James et al. (2000) discuss how to choose the number of knots in the spline basis, the choice of q in equation (12.4.1), and how to construct bootstrap-based confidence intervals for the mean function, the curves defining the principal components, and the individual curves.

As noted in Section 9.3.4, redundancy analysis can be formulated as PCA on the predicted responses in a multivariate regression. Van den Brink and ter Braak (1999) extend this idea so that some of the predictor variables in the regression are not included among the predictors on which the PCA is done. The context in which they implement their technique is where species abundances depend on time and on which 'treatment' is applied. The results of this analysis are called *principal response curves*.

12.4.3 Climate Change—Fingerprint Techniques

In climate change detection, the objective is not only to discover whether change is taking place, but also to explain any changes found. If certain causes are suspected, for example increases in greenhouse gases, variations in solar output, or volcanic activity, then changes in these quantities can be built into complex atmospheric models, and the models are run to discover what happens to parameters of interest such as the global pattern of temperature. Usually changes in one (or more) of the potential causal variables will manifest themselves, according to the model, in different ways in different geographical regions and for different climatic variables. The predicted patterns of change are sometimes known as 'fingerprints' associated with changes in the causal variable(s). In the detection of climate change it is usually more productive to attempt to detect changes that resemble such fingerprints than to search broadly over a wide range of possible changes. The paper by Hasselmann (1979) is usually cited as the start of interest in this type of climate change detection, and much research has been done subsequently. Very similar techniques have been derived via a number of different approaches. Zwiers (1999) gives a good, though not exhaustive, summary of several of these techniques, together with a number of applications. North and Wu (2001) describe a number of recent developments, again with applications.

The basic idea is that the observed data, which are often values of some climatic variable x_{tj}, where t indexes time and j indexes spatial position, can be written

$$x_{tj} = s_{tj} + e_{tj}.$$

Here s_{tj} is the deterministic response to changes in the potential causal variables (the signal), and e_{tj} represents the stochastic noise associated with 'normal' climate variation.

Suppose that an optimal detection variable A_t at time t is constructed as a linear combination of the observed data x_{sj} for $s = t, (t-1), \ldots, (t-l+1)$ and $j = 1, 2, \ldots, m$, where m is the number of spatial locations for which data are available. The variable A_t can be written as $A_t = \mathbf{w}'\mathbf{x}$, where $\mathbf{x}$ is an ml-vector of observed measurements at times $t, (t-1), \ldots, (t-l+1)$, and all m spatial locations, and $\mathbf{w}$ is a vector of weights. Then $\mathbf{w}$ is chosen to maximize the signal to noise ratio

$$\frac{[E(A_t)]^2}{\text{var}(A_t)} = \frac{[\mathbf{w}'\mathbf{s}_t]^2}{\mathbf{w}'\boldsymbol{\Sigma}_e\mathbf{w}},$$

where $\mathbf{s}_t$ is an ml-vector of known signal at time t and $\boldsymbol{\Sigma}_e$ is the spatio-temporal covariance matrix of the noise term. It is straightforward to show that the optimal detector, sometimes known as the *optimal fingerprint*, has the form $\hat{\mathbf{w}} = \boldsymbol{\Sigma}_e^{-1}\mathbf{s}_t$. The question then arises: Where does PCA fit into this methodology?

Underlying the answer to this question is the fact that $\mathbf{\Sigma}_e$ needs to be estimated, usually from a 'control run' of the atmospheric model in which no changes are made to the potential causal variables. Because the dimension ml of $\mathbf{\Sigma}_e$ is large, it may be nearly singular, and estimation of $\mathbf{\Sigma}_e^{-1}$, which is required in calculating $\hat{\mathbf{w}}$, will be unstable. To avoid this instability, the estimate of $\mathbf{\Sigma}_e$ is replaced by the first few terms in its spectral decomposition (Property A3 of Section 2.1), that is by using its first few PCs. Allen and Tett (1999) discuss the choice of how many PCs to retain in this context. 'First few' is perhaps not quite the right phrase in some of these very large problems; North and Wu (2001) suggest keeping up to 500 out of 3600 in one of their studies.

Expressing the optimal fingerprint in terms of the PCs of the estimate of $\mathbf{\Sigma}_e$ also has advantages in interpretation, as in principal component regression (Chapter 8), because of the uncorrelatedness of the PCs (Zwiers, 1999, equations (13), (14)).

12.4.4 Spatial Data

Many of the techniques discussed earlier in the chapter have a spatial dimension. In Section 12.2, different points in space mostly correspond to different 'variables.' For MSSA, variables correspond to combinations of time lag and spatial position, and Plaut and Vautard (1994) refer to the output of MSSA as 'space-time EOFs.' North and Wu (2001) use the same term in a different situation where variables also correspond to space-time combinations, but with time in separated rather than overlapping blocks.

The possibility was also raised in Section 12.3 that the continuum of variables underlying the curves could be in space rather than time. The example of Section 12.3.3 is a special case of this in which space is one-dimensional, while Bouhaddou et al. (1987) and Guttorp and Sampson (1994) discuss estimation of continuous (two-dimensional) spatial covariances. Extended EOF analysis (see Sections 12.2.1, 14.5) has several variables measured at each spatial location, but each *space* × *variable* combination is treated as a separate 'variable' in this type of analysis. Another example in which variables correspond to location is described by Boyles (1996). Here a quality control regime includes measurements taken on a sample of n parts from a production line. The measurements are made at p locations forming a lattice on each manufactured part. We return to this example in Section 13.7.

In some situations where several variables are measured at each of n spatial locations, the different spatial locations correspond to different *observations* rather than variables, but the observations are typically not independent. Suppose that a vector $\mathbf{x}$ of p variables is measured at each of n spatial locations. Let the covariance between the elements x_j and x_k of $\mathbf{x}$ for two observations which are separated in space by a vector $\mathbf{h}$ be the (j,k)th element of a matrix $\mathbf{\Sigma}_{\mathbf{h}}$. This expression assumes second order sta-

tionarity in the sense that the covariances do not depend on the location of the two observations, only on the vector joining them. On the other hand, it does not require isotropy—the covariance may depend on the direction of $\mathbf{h}$ as well as its length. The *intrinsic correlation model* (Wackernagel, 1995, Chapter 22) assumes that $\mathbf{\Sigma_h} = \rho_{\mathbf{h}}\mathbf{\Sigma}$. Because all terms in $\mathbf{\Sigma}$ are multiplied by the same spatial factor, this factor cancels when correlations are calculated from the covariances and the correlation between x_j and x_k does not depend on $\mathbf{h}$. Wackernagel (1995, Chapter 22) suggests testing whether or not this model holds by finding principal components based on sample covariance matrices for the p variables. Cross-covariances between the resulting PCs are then found at different separations $\mathbf{h}$. For $k \neq l$, the kth and lth PCs should be uncorrelated for different values of $\mathbf{h}$ under the intrinsic correlation model, because $\mathbf{\Sigma_h}$ has the same eigenvectors for all $\mathbf{h}$.

An extension of the intrinsic correlation model to a 'linear model of coregionalization' is noted by Wackernagel (1995, Chapter 24). In this model the variables are expressed as a sum of $(S + 1)$ spatially uncorrelated components, and the covariance matrix now takes the form

$$\mathbf{\Sigma_h} = \sum_{u=0}^{S} \rho_{u\mathbf{h}} \mathbf{\Sigma}_u.$$

Wackernagel (1995, Chapter 25) suggests that separate PCAs of the estimates of the matrices $\mathbf{\Sigma}_0, \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_S$ may be informative, but it is not clear how these matrices are estimated, except that they each represent different spatial scales.

As well as dependence on $\mathbf{h}$, the covariance or correlation between x_j and x_k may depend on the nature of the measurements made (point measurements, averages over an area where the area might be varied) and on the size of the domain. Vargas-Guzmán et al. (1999) discuss each of these aspects, but concentrate on the last. They describe a procedure they name *growing scale PCA*, in which the nature of the measurements is fixed and averaging (integration) takes place with respect to $\mathbf{h}$, but the size of the domain is allowed to vary continuously. As it does, the PCs and their variances also evolve continuously. Vargas-Guzmán et al. illustrate this technique and also the linear model of coregionalization with a three-variable example. The basic form of the PCs is similar at all domain sizes and stabilizes as the size increases, but there are visible changes for small domain sizes. The example also shows the changes that occur in the PCs for four different areal extents of the individual measurements. Buell (1975) also discussed the dependence of PCA on size and shape of spatial domains (see Section 11.4), but his emphasis was on the shape of the domain and, unlike Vargas-Guzmán et al. (1999), he made no attempt to produce a continuum of PCs dependent on size.

Kaplan et al. (2001) consider optimal interpolation and smoothing of spatial field data that evolve in time. There is a basic first-order autore-

gressive structure relating the vector of true values of the field at time $(t+1)$ to that at time t, as in equation (12.2.1), but in addition the values are measured with error, so that there is a second equation, relating the observed field $\mathbf{y}_t$ to the true field $\mathbf{x}_t$, which is

$$\mathbf{y}_t = \boldsymbol{\Xi}\mathbf{x}_t + \boldsymbol{\xi}_t, \tag{12.4.3}$$

where $\boldsymbol{\Xi}$ is a $(p \times p)$ matrix and $\boldsymbol{\xi}_t$ is a vector of observational errors. In the most general case the matrices $\boldsymbol{\Upsilon}$ and $\boldsymbol{\Xi}$ in equations (12.2.1), (12.4.1) may depend on t, and the covariance matrices of the error terms $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\xi}_t$ need not be proportional to the identity matrix, or even diagonal. There are standard procedures for optimal interpolation and smoothing, but these become computationally prohibitive for the large data sets considered by Kaplan and researchers. They therefore suggest projecting the data onto the first few principal components of the covariance matrix of the anomaly field, that is, the field constructed by subtracting the long-term climatic mean from the measurements at each spatial location in the field. Kaplan et al. (2001) describe a number of subtleties associated with the approach. The main problem is the need to estimate the covariance matrices associated with $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\xi}_t$. Difficulties arise because of possible non-stationarity in both means and covariance, and because of changing spatial coverage over time, among other things, but Kaplan and his coworkers propose solutions to overcome the difficulties.

Among the procedures considered computationally 'extremely expensive' by Kaplan et al. (2001) is the Kalman filter. Wikle and Cressie (1999) propose a form of the Kalman filter in which there is an additional non-dynamic term that captures small-scale spatial variability. Apart from this term, their model is similar to that of Kaplan et al. and they, too, suggest dimension reduction using principal components. Wikle and Cressie's model has space defined continuously but the principal components that give their dimension reduction are derived from predicted values of the spatial process on a regular grid.

In a spatial discriminant analysis example, Storvik (1993) suggests linearly transforming the data, but not to PCs. Instead, he finds linear functions $\mathbf{a}_k'\mathbf{x}$ of $\mathbf{x}$ that successively minimize autocorrelation between $\mathbf{a}_k'\mathbf{x}(s)$ and $\mathbf{a}_k'\mathbf{x}(s+\Delta)$, where the argument for $\mathbf{x}$ denotes spatial position and Δ is a fixed distance apart in space. It turns out that the functions are derived via an eigenanalysis of $\mathbf{S}^{-1}\mathbf{S}_\Delta$, where $\mathbf{S}$ is the usual sample covariance matrix for $\mathbf{x}$ and $\mathbf{S}_\Delta$ is the sample covariance matrix for $\mathbf{x}(s)-\mathbf{x}(s+\Delta)$. This analysis has some resemblance to the procedure for POP analysis (Section 12.2.2), but in a spatial rather than time series context.

12.4.5 Other Aspects of Non-Independent Data and PCA

A different type of non-independence between observations is induced in sample surveys for which the survey design is more complex than simple

random sampling. Skinner et al. (1986) consider what happens to PCA when the selection of a sample of observations on p variables $\mathbf{x}$ depends on a vector of covariates $\mathbf{z}$. They use Taylor expansions to approximate the changes (compared to simple random sampling) in the eigenvalues and eigenvectors of the sample covariance matrix when samples are chosen in this way. Skinner et al. present a simulation study in which stratified samples are taken, based on the value of a single covariate z, whose correlations with 6 measured variables range from 0.48 to 0.60. Substantial biases arise in the estimates of the eigenvalues and eigenvectors for some of the stratification schemes examined, and these biases are well-approximated using the Taylor expansions. By assuming multivariate normality, Skinner and coworkers show that a maximum likelihood estimate of the covariance matrix can be obtained, given knowledge of the survey design, and they show that using the eigenvalues and eigenvectors of this estimate corrects for the biases found for the usual covariance matrix. Tortora (1980) presents an example illustrating the effect of a disproportionate stratified survey design on some of the rules for variable selection using PCA that are described in Section 6.3.

Similar structure to that found in sample surveys can arise in observational as well as design-based data sets. For example, Konishi and Rao (1992) consider multivariate data, which could be genetic or epidemiological, for samples of families. There are correlations between members of the same family, and families may be of unequal sizes. Konishi and Rao (1992) propose a model for the correlation structure in such familial data, which they use to suggest estimates of the principal components underlying the data.

Solow (1994) notes that when a trend is present in a set of time series, it is often a major source of variation and will be found by an early PC. However, he argues that to identify trends it is better to look for a linear combination of the p series that has maximum autocorrelation, rather than maximum variance. Solow (1994) solves this maximization problem by finding the smallest eigenvalues and corresponding eigenvectors of $\mathbf{S}^{-1}\mathbf{S}_D$, where $\mathbf{S}_D$ is the covariance matrix of first differences of the series. This is the same eigenproblem noted in the Section 12.4.4, which Storvik (1993) solved in a spatial setting, but its objectives are different here so that autocorrelation is maximized rather than minimized. Solow (1994) also modifies his procedure to deal with compositional data, based on the approach suggested by Aitchison (1983) for PCA (see Section 13.3).

Peña and Box (1987) consider a factor analysis model for time series in which p time series are assumed to be derived linearly from a smaller number m of underlying factors. Each factor follows an autoregressive moving-average (ARMA) model. Peña and Box (1987) use the eigenvalues and eigenvectors of covariance matrices between the series measured simultaneously (essentially a PCA) and at various lags to deduce the structure of their factor model. They claim that considerable simplification is possible

compared to a multivariate ARMA model involving all p of the original series.

Wold (1994) suggests exponentially weighted moving principal components in the context of process control (see Section 13.7), and Diamantaras and Kung (1996, Section 3.5) advocate PCs based on weighted covariance matrices for multivariate time series, with weights decreasing exponentially for less recent observations.

Yet another rôle for PCs in the analysis of time series data is presented by Doran (1976). In his paper, PCs are used to estimate the coefficients in a regression analysis of one time series variable on several others. The idea is similar to that of PC regression (see Section 8.1), but is more complicated as it involves the frequency domain. Consider the distributed lag model, which is a time series version of the standard regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ of equation (8.1.1), with the time series structure leading to correlation between elements of $\boldsymbol{\epsilon}$. There is a decomposition of the least squares estimator of the regression coefficients $\boldsymbol{\beta}$ rather like equation (8.1.8) from PC regression, except that the eigenvalues are replaced by ratios of spectral density estimates for the predictors (signal) and error (noise). Doran (1976) suggests using an estimate in which the terms corresponding to the smallest values of this signal to noise ratio are omitted from the least squares decomposition.