9

Regularized principal components analysis

9.1 Introduction

In this chapter, we discuss the application of smoothing to functional principal components analysis. In Chapter 5 we have already seen that smoothing methods are useful in functional data analysis in preprocessing the data to obtain functional observations. The emphasis in this chapter is somewhat different, in that we incorporate the smoothing into the principal components analysis itself.

Our discussion provides a further insight into the way the method of regularization, discussed in Chapter 5, can be used rather generally in functional data analysis. The basic idea is to put into practice, in any particular context, the philosophy of combining a measure of goodness-of-fit with a roughness penalty.

Consideration of the third component in Figure 8.1 indicates that some smoothing may be appropriate when estimating functional principal components. A more striking example is provided by the pinch force data discussed in Section 1.5.2. Rather than smoothing the data initially, consider the data in Figure 9.1, which consists of the original records of the force exerted by the thumb and forefinger during each of 20 brief squeezes or pinches. The observed records are not very smooth, and consequently the principal component curves in Figure 9.2 show substantial variability. There is a clear need for smoothing or regularizing of the estimated principal component curves. In this chapter, we develop a method for smoothed principal component analysis, but first of all the application of the method

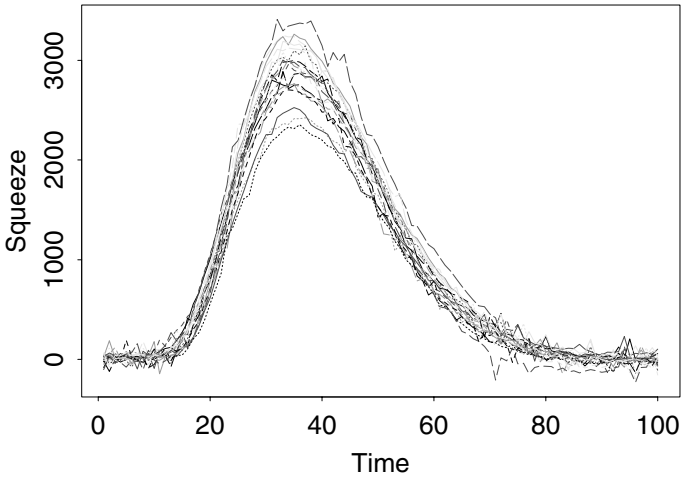

Figure 9.1. The aligned original recordings of the force relative to a baseline value exerted during each of 20 brief pinches.

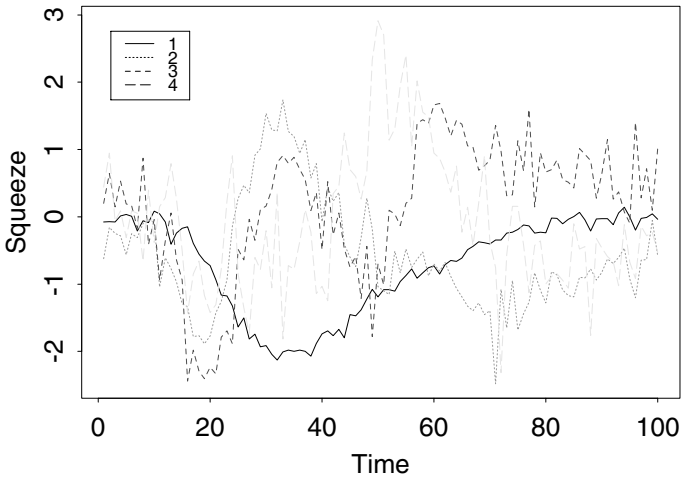

Figure 9.2. The first four principal component curves for the pinch force data without regularization.

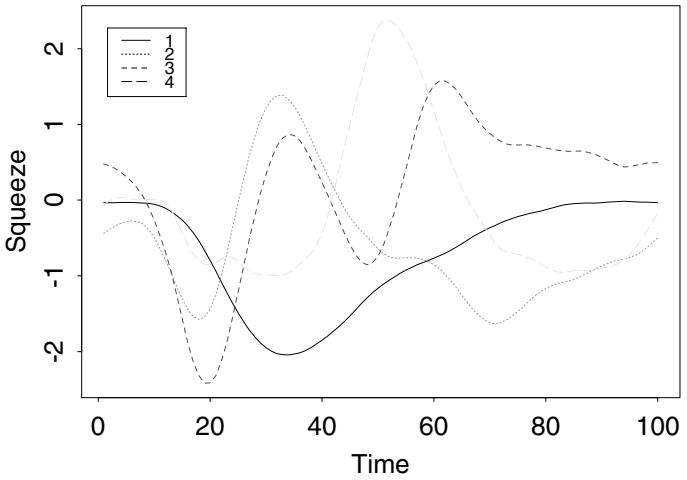

Figure 9.3. The first four smoothed principal components for the pinch force data, smoothed by the method of Section 9.3. The smoothing parameter is chosen by cross-validation.

to the pinch force data is demonstrated. In subsequent sections, the method is defined in detail and various aspects of its implementation are discussed.

9.2 The results of smoothing the PCA

Figure 9.3 shows the effect of applying principal components analysis using the method for smoothed PCA set out subsequently in this chapter. The method incorporates a smoothing parameter λ to control the amount of smoothing applied, and this has been chosen by a cross-validation method set out in Section 9.3.3. The smoothing method achieves the aim of removing the considerable roughness in the raw principal component curves in Figure 9.2.

Figure 9.4 displays the effects on the mean curve of adding and subtracting a multiple of each of the first four smoothed principal components. The first component corresponds to an effect whereby the shape of the impulse is not substantially changed, but its overall scale is increased. The second component (with appropriate sign) corresponds roughly to a compression in the overall time scale during which the squeeze takes place. Both of these effects were removed in the analysis of Ramsay, Wang and Flanagan (1995) before any detailed analysis was carried out. It is, however, interesting to

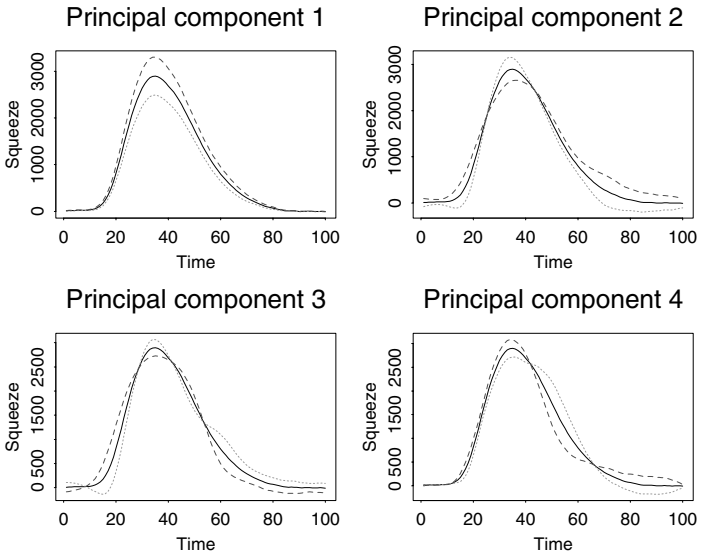

Figure 9.4. The effect on the overall mean curve of adding and subtracting a suitable multiple of each of the first four smoothed principal component curves provided in Figure 9.3.

note that they occur as separate components and therefore are essentially uncorrelated with one another, and with the effects found subsequently. The third component corresponds to an effect whereby the main part takes place more quickly but the tail after the main part is extended to the right. The fourth component corresponds to a higher peak correlated with a tail-off that is faster initially but subsequently slower than the mean. The first and second effects are transparent in their interest, and the third and fourth are of biomechanical interest in indicating how the system compensates for departures from the (remarkably reproducible) overall mean. The smoothing we have described makes the effects very much clearer than they are in the raw principal component plot.

The estimated variances σ^2 indicate that the four components displayed respectively explain 86.2%, 6.7%, 3.5% and 1.7% of the variability in the original data, with 1.9% accounted for by the remaining components. The individual principal component scores indicate that there is one curve with a fairly extreme value of principal component 2 (corresponding to moving more quickly than average through the cycle) but this curve is not unusual in other respects.

9.3 The smoothing approach

9.3.1 Estimating the leading principal component

Our smoothed PCA approach is based on a roughness penalty idea, as discussed in Chapter 7. Suppose ξ is a possible principal component curve. As in standard spline smoothing, we usually penalize the roughness of ξ by its integrated squared second derivative over the interval of interest, $\texttt{PEN}_2(\xi) = \|D^2\xi\|^2$.

Consider, first, the estimation of the leading principal component. In an unsmoothed functional PCA as described in Chapter 8, we work with the sample variance $\texttt{var} \int \xi x_i$ of the principal component scores $\int \xi x_i$ over the observations x_i. The first principal component weight function is chosen to maximize $\texttt{var} \int \xi x_i$ subject to the constraint $\|\xi\|^2 = 1$. As explained in Section 8.2.4, this maximization problem is solved by finding the leading solution of the eigenfunction equation $V\xi = \rho\xi$.

However, maximizing this sample variance is not our only aim. We also want to prevent the roughness $\texttt{PEN}_2(\xi) = \int \xi''(t)^2 dt$ of the estimated principal component ξ from being too large. The key to the roughness penalty approach is to make explicit this possible conflict. As usual in the roughness penalty method, the trade-off is controlled by a smoothing parameter $\lambda \geq 0$ which regulates the importance of the roughness penalty term.

Given any possible principal component function ξ with $\|\xi\|^2 = 1$, one way of penalizing the sample variance $\texttt{var} \int \xi x_i$ is to divide it by $\{1 + \lambda \times \texttt{PEN}_2(\xi)\}$. This gives the *penalized sample variance*

$$\texttt{PCAPSV}(\xi) = \frac{\texttt{var} \int \xi x_i}{\|\xi\|^2 + \lambda \times \texttt{PEN}_2(\xi)}. \tag{9.1}$$

Increasing the roughness of ξ while maintaining λ fixed decreases $\texttt{PCAPSV}(\xi)$, as defined in (9.1), since $\texttt{PEN}_2(\xi)$ increases. Moreover, PCAPSV reverts to the raw sample variance as $\lambda \to 0$. On the other hand, the larger the value of λ, the more the penalized sample variance is affected by the roughness of ξ. In the limit $\lambda \to \infty$, the component ξ is forced to be of the form $\xi = a$ in the periodic case and $\xi = a + bt$ in the nonperiodic case, for some constants a and b.

9.3.2 Estimating subsequent principal components

Of course, it is usually of interest not merely to estimate the leading principal component, but also to estimate the other components. The way our procedure works is to estimate each ξ_j to maximize the penalized variance $\texttt{PCAPSV}(\xi)$ as defined in (9.1), subject to two constraints. The first constraint is the usual requirement that $\|\xi_j\|^2 = 1$. Secondly, we impose a

modified form of orthogonality to the previously estimated components

$$\int \xi_j(s)\xi_k(s)ds + \int D^2\xi_j(s)D^2\xi_k(s)ds = 0 \text{ for } k = 1, \ldots, j-1. \quad (9.2)$$

The use of the modified orthogonality condition (9.2) means that we can find the estimates of all the required principal components by solving a single eigenvalue problem, and this will be explained in Section 9.4, where practical algorithms are discussed. Silverman (1996) provides a detailed investigation of the theoretical advantages of this approach.

9.3.3 Choosing the smoothing parameter by cross-validation

How should the smoothing parameter λ be chosen? It is perfectly adequate for many purposes to choose the smoothing parameter subjectively, but we can also use a cross-validation approach to choose the amount of smoothing automatically. Some general remarks about the use of automatic methods for choosing smoothing parameters are found in Section 3.1 of Green and Silverman (1994).

To consider how a cross-validation score could be calculated, suppose that x is an observation from the population. Then, by the optimal basis property discussed in Section 8.2.3, the principal components have the property that, for each m, an expansion in terms of the functions $\xi_1, \ldots, \xi_m$ can explain more of the variation in x than any other collection of m functions. To quantify the amount of variation in x accounted for by these functions, we define x^* to be the projection of x onto the subspace spanned by $\xi_1, \ldots, \xi_m$ and let ζ_m be the residual component $x - x^*$. Thus, ζ_m is the component of x orthogonal to the functions $\xi_1, \ldots, \xi_m$.

If we wish to consider the efficacy of the first m components, then a measure to consider is $E\|\zeta_m\|^2$; in order not to be tied to a particular m, we can, for example, minimize $\sum_m E\|\zeta_m\|^2$. In both cases, we do not have new observations x to work with, and the usual cross-validation paradigm has to be used, as follows:

1. Subtract the overall mean from the observed data x_i.
2. For a given smoothing parameter λ, let $\xi_j^{[i]}(\lambda)$ be the estimate of ξ_j obtained from all the data except x_i.
3. Define $\zeta_m^{[i]}(\lambda)$ to be the component of x_i orthogonal to the subspace spanned by $\{\xi_j^{[i]}(\lambda) : j = 1, \ldots, m\}$.
4. Combine the $\zeta_m^{[i]}(\lambda)$ to obtain the cross-validation scores

$$\mathtt{CV}_m(\lambda) = \sum_{i=1}^{n} \|\zeta_m^{[i]}(\lambda)\|^2 \quad (9.3)$$

and hence

$$\mathtt{CV}(\lambda) = \sum_{m=1}^{\infty} \mathtt{CV}_m(\lambda). \tag{9.4}$$

In practice, we would of course truncate the sum in (9.4) at some convenient point. Indeed, given n data curves, we can estimate at most $n-1$ principal components, and so the sum must be truncated at $m = n-1$ if not at a smaller value.

5. Minimize $\mathtt{CV}(\lambda)$ to provide the choice of smoothing parameter.

Clearly there are other possible ways of combining the $\mathtt{CV}_m(\lambda)$ to produce a cross-validation score to account for more than one value of m, but we restrict attention to $\mathtt{CV}(\lambda)$ as defined in (9.4).

In the pinch force data example considered Section 9.2, it was found satisfactory to calculate the cross-validation score on a grid (on a logarithmic scale) of values of the smoothing parameter λ and pick out the minimum. The grid can be quite coarse, since small changes in the numerical value of λ do not make very much difference to the smoothed principal components. For this example, we calculated the cross-validation scores for $\lambda = 0$ and $\lambda = 1.5^{i-1}$ for $i = 1, \ldots, 30$, and we attained the minimum of $\mathtt{CV}(\lambda)$ by setting $\lambda = 37$.

9.4 Finding the regularized PCA in practice

In practice, the smoothed principal components are most easily found by working in terms of a suitable basis. First of all, consider the periodic case, for which it is easy to set out an algorithm based on Fourier series.

9.4.1 The periodic case

Suppose, for simplicity, that $\mathcal{T}$ is the interval $[0,1]$ and that periodic boundary conditions are valid for all the functions we are considering. In particular, this means that the data $x_i(s)$ themselves are regarded as being periodic. Let $\{\phi_\nu\}$ be the series of Fourier functions defined in (3.7). For each j, define $\omega_{2j-1} = \omega_{2j} = 2\pi j$. Given any periodic function x, we can expand x as a Fourier series with coefficients $c_\nu = \int x\phi_\nu$, so that

$$x(s) = \sum_\nu c_\nu \phi_\nu(s) = \mathbf{c}'\boldsymbol{\phi}(s).$$

The operator D^2 has the useful property that, for each ν,

$$D^2\phi_\nu = -\omega_\nu^2 \phi_\nu,$$

meaning that we can also expand D^2x as

$$D^2x(s) = -\sum_\nu \omega_\nu^2 c_\nu \phi_\nu(s).$$

By standard orthogonality properties of trigonometric functions, the ϕ_ν are orthonormal, and it follows that the roughness penalty $\|D^2x\|^2$ can be written as a weighted sum of squares of the coefficients c_ν:

$$\|D^2x\|^2 = \int (-\sum_\nu \omega_\nu^2 c_\nu \phi_\nu)(-\sum_\nu \omega_\nu^2 c_\nu \phi_\nu) = \sum_\nu \omega_\nu^4 c_\nu^2.$$

Now proceed by expanding the data functions to sufficient terms in the basis to approximate them closely. We can use a fast Fourier transform on a finely discretized version of the observed data functions to do this efficiently. Denote by $\mathbf{c}_i$ the vector of Fourier coefficients of the observation $x_i(s)$, so that $x_i(s) = \mathbf{c}_i'\boldsymbol{\phi}(s)$ where $\boldsymbol{\phi}$ is the vector of basis functions. Let $\mathbf{V}$ be the covariance matrix of the vectors $\mathbf{c}_i$, and let $\mathbf{S}$ be the diagonal matrix with entries

$$S_{\nu\nu} = (1 + \lambda\omega_\nu^4)^{-1/2}.$$

The matrix $\mathbf{S}$ then corresponds to a smoothing operator S.

Let $\mathbf{y}$ be the vector of coefficients of any potential principal component curve ξ, so that

$$\xi(s) = \sum_\nu y_\nu \phi_\nu(s) = \mathbf{y}'\boldsymbol{\phi}(s). \tag{9.5}$$

In terms of Fourier coefficients, we have

$$\texttt{PCAPSV}(\xi) = \frac{\mathbf{y}'\mathbf{V}\mathbf{y}}{\mathbf{y}'\mathbf{S}^{-2}\mathbf{y}}. \tag{9.6}$$

Furthermore, if $\mathbf{y}_{(j)}$ denotes the vector of Fourier coefficients of the curve ξ_k, then the constraint (9.2) can be written as $\mathbf{y}_{(j)}'\mathbf{S}^{-2}\mathbf{y}_{(k)} = 0$ for $k = 1, \ldots, j-1$.

It follows from standard arguments in linear algebra that the estimates specified in Section 9.3 have Fourier coefficients that satisfy the eigenvector equation

$$\mathbf{V}\mathbf{y} = \rho\mathbf{S}^{-2}\mathbf{y}, \tag{9.7}$$

which can be rewritten

$$(\mathbf{S}\mathbf{V}\mathbf{S})(\mathbf{S}^{-1}\mathbf{y}) = \rho(\mathbf{S}^{-1}\mathbf{y}). \tag{9.8}$$

The matrix $\mathbf{SVS}$ is the covariance matrix of the vectors $\mathbf{S}\mathbf{c}_i$, the Fourier coefficient vectors of the original data smoothed by the application of the smoothing operator S.

To find the solutions of (9.8), suppose that $\mathbf{u}$ is an eigenvector of $\mathbf{SVS}$ with eigenvalue ρ. Finding the eigenvectors and eigenvalues of $\mathbf{SVS}$ corresponds precisely to carrying out an unsmoothed PCA of the *smoothed* data

$\mathbf{Sc}_i$. Then it is apparent that any multiple of $\mathbf{Su}$ is a solution of (9.8) for the same ρ. Because we require $\|\mathbf{y}\|^2 = 1$, renormalize and set $\mathbf{y} = \mathbf{Su}/\|\mathbf{Su}\|$. The functional principal component ξ corresponding to $\mathbf{y}$ is then computed from (9.5).

Putting these steps together gives the following procedure for carrying out the smoothed principal component analysis of the original data:

1. Compute the coefficients $\mathbf{c}_i$ for the expansion of each sample function x_i in terms of basis $\boldsymbol{\phi}$.
2. Operate on these coefficients by the smoothing operator S.
3. Carry out a standard PCA on the resulting smoothed coefficient vectors $\mathbf{Sc}_i$.
4. Apply the smoothing operator S to the resulting eigenvectors $\mathbf{u}$, and renormalize so that the resulting vectors $\mathbf{y}$ have unit norm.
5. Compute the principal component function ξ from (9.5).

9.4.2 The nonperiodic case

Now turn to the nonperiodic case, where Fourier expansions are no longer appropriate because of the boundary conditions. Suppose that $\{\phi_\nu\}$ is a suitable basis for the space of smooth functions $\mathcal{S}$ on $[0, 1]$. Possible bases include B-splines on a fine mesh, or possibly orthogonal polynomials up to some degree. In either case, we choose the dimensionality of the basis to represent the functions $x_i(s)$ well. As in the discussion of the periodic case, let $\mathbf{c}_i$ be the vector of coefficients of the data function $x_i(s)$ in the basis $\{\phi_\nu\}$. Let $\mathbf{V}$ be the covariance matrix of the vectors $\mathbf{c}_i$.

Define $\mathbf{J}$ to be the matrix $\int \boldsymbol{\phi}\boldsymbol{\phi}'$, whose elements are $\int \phi_j \phi_k$ and $\mathbf{K}$ the matrix whose elements are $\int D^2\phi_j D^2\phi_k$. The penalized sample variance can be written as

$$\texttt{PCAPSV} = \frac{\mathbf{y}'\mathbf{JVJy}}{\mathbf{y}'\mathbf{Jy} + \lambda \mathbf{y}'\mathbf{Ky}} \tag{9.9}$$

and the eigenequation corresponding to (9.7) is given by

$$\mathbf{JVJy} = \rho(\mathbf{J} + \lambda\mathbf{K})\mathbf{y}. \tag{9.10}$$

Now perform a factorization $\mathbf{LL}' = \mathbf{J} + \lambda\mathbf{K}$ and define $\mathbf{S} = \mathbf{L}^{-1}$. We can find a suitable matrix $\mathbf{L}$ by an SVD or by Choleski factorization, in which case $\mathbf{L}$ is a lower triangular matrix. The equation (9.10) can now be written as

$$(\mathbf{SJVJS}')(\mathbf{L}'\mathbf{y}) = \rho\mathbf{L}'\mathbf{y}.$$

We can now work through stages corresponding to those for the periodic case. The algorithm obtained is as follows:

1. Expand the observed data x_i with respect to the basis $\boldsymbol{\phi}$ to obtain coefficient vectors $\mathbf{c}_i$.
2. Solve $\mathbf{L}d_i = \mathbf{J}\mathbf{c}_i$ for each i to find the vectors $d_i = \mathbf{S}\mathbf{J}\mathbf{c}_i$.
3. Carry out a standard PCA on the coefficient vectors d_i.
4. Apply the smoothing operator $\mathbf{S}'$ to the resulting eigenvectors $\mathbf{u}$ by solving $\mathbf{L}'\mathbf{y} = u$ in each case, and renormalize so that the resulting vectors $\mathbf{y}$ have $\mathbf{y}'\mathbf{J}\mathbf{y} = 1$.
5. Transform back to find the principal component functions ξ using (9.5).

If we use a B-spline basis and define $\mathbf{L}$ by a Choleski factorization, then the matrices $\mathbf{J}$, $\mathbf{K}$ and $\mathbf{L}$ are all band matrices, and by using appropriate linear algebra routines, we can carry out all the calculations extremely economically. Even in the full matrix case, especially if not too many basis functions are used, the computations are reasonably fast because $\mathbf{S}$ never has to be found explicitly.

9.5 Alternative approaches

In this section, we discuss two alternative approaches to smoothed functional PCA.

9.5.1 Smoothing the data rather than the PCA

In this section, we compare the method of regularized principal components analysis with an approach akin to that discussed earlier in the book. Instead of carrying out our smoothing step within the PCA, we smooth the data first, and then carry out an unsmoothed PCA. This approach to functional PCA was taken by Besse and Ramsay (1986), Ramsay and Dalzell (1991) and Besse, Cardot and Ferraty (1997). Of course, conceivably any smoothing method can be used to smooth the data, but to make a reasonable comparison, we use a roughness penalty smoother based on integrated squared second derivative. For simplicity, let us restrict our attention to the case of periodic boundary conditions.

Suppose that x is a data curve, and that we regard x as the sum of a smooth curve and a noise process. We would obtain the roughness penalty estimate of the smooth curve by minimizing

$$\texttt{PENRSS} = \|x - g\|^2 + \lambda \|D^2 g\|^2$$

over g in $\mathcal{S}$. As usual, λ is a smoothing parameter that controls the trade-off between fidelity to the data and smoothing. This is a generalization of the

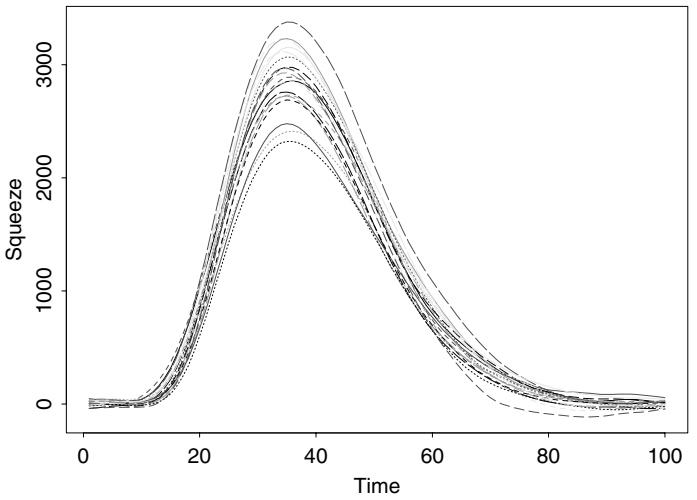

Figure 9.5. The pinch force data curves, smoothed by a roughness penalty method with the same smoothing parameter as used for the smoothed PCA, and with the baseline pressure subtracted.

spline smoothing method discussed in Chapter 5 to the case of functional data.

Consider an expansion of x and g in terms of Fourier series as in Section 9.4.1, and let $\mathbf{c}$ and $\mathbf{d}$ be the resulting vectors of coefficients. Then

$$\texttt{PENRSS} = \|\mathbf{c} - \mathbf{d}\|^2 + \lambda \sum_{\nu} \omega_\nu^4 d_\nu^2,$$

and hence the coefficients of the minimizing g satisfy

$$\mathbf{d} = \mathbf{S}^2\mathbf{c}, \tag{9.11}$$

where $\mathbf{S}$ is as defined in Section 9.4.1. Note that this demonstrates that the smoothing operator $\mathbf{S}$ used twice in the algorithm set out in Section 9.4.1 can be regarded as a half-spline-smooth, since $\mathbf{S}^2$ is the operator corresponding to classical spline smoothing.

Now let us consider the effect of smoothing the data by the operator $\mathbf{S}^2$ using the same smoothing parameter $\lambda = 37$ as in the construction of Figures 9.3 and 9.4. The effect of this smoothing on the data is illustrated in Figure 9.5. Figure 9.6 shows the first four principal component curves of the smoothed data. Although the two methods do not give identical results, the differences between them are too small to affect any interpretation.

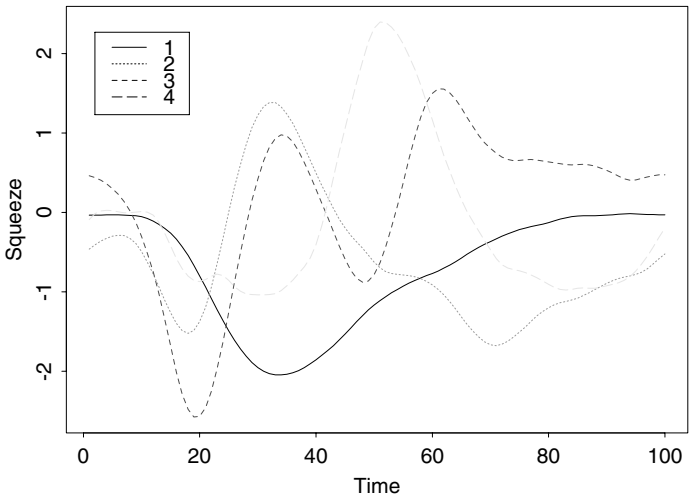

Figure 9.6. The first four principal component curves of the smoothed data as shown in Figure 9.5.

However, this favorable comparison depends rather crucially on the way in which the data curves are smoothed, and in particular on the match between the smoothing level implied in (9.11) and the smoothing level used for the PCA itself. For example, we tried smoothing the force functions curves individually, selecting the smoothing parameters by the generalized cross-validation approach used in the S-PLUS function `smooth.spline`. The result was much less successful, in the sense that the components were far less smooth. The reason appears to be that this smoothing technique tended to choose much smaller values of the smoothing parameter λ.

Kneip (1994) considers several aspects of an approach that first smooths the data and then extracts principal components. Under a model where the data are corrupted by a white noise error process, he investigates the dependence of the quality of estimation of the principal components on both sample size and sampling rate. In an application based on economics data, he shows that smoothing is clearly beneficial in a practical sense.

9.5.2 A stepwise roughness penalty procedure

Another approach to the smoothing of functional PCA was set out by Rice and Silverman (1991). They considered a stepwise procedure incorporating the roughness penalty in a different way. Their proposal requires a separate smoothing parameter λ_j for each principal component. The prin-

cipal components are estimated successively, the estimate $\xi_j^\dagger$ of ξ_j being found by maximizing $\mathtt{var} \int \xi x_i - \lambda_j \|D^2 \xi\|^2$ subject to the conventional orthonormality conditions $\|\xi\|^2 = 1$ and $\int \xi \xi_k^\dagger = 0$ for $k = 1, \ldots, j-1$.

This approach is computationally more complicated because a separate eigenproblem has to be posed and solved for each principal component; for more details, see the original paper. Theoretical results in Pezzulli and Silverman (1993) and Silverman (1996) also suggest that the procedure described in Section 9.3 is likely to be advantageous under conditions somewhat milder than those for the Rice-Silverman procedure.

9.5.3 A further approach

Yao, Müller, Clifford, Dueker, Follet, Lin, Bucholz and Vogel (2003) regularize the principal component scores f_{im} by shrinking them towards zero.