# 2
# Life Course Data in Criminology

## 2.1 Criminology life course studies

### *2.1.1 Background*

An important question in criminology is the study of the way that people's level of criminal activity varies through their lives. Can it be said that there are "career criminals" of different kinds? Are there particular patterns of persistence in the levels of crimes committed by individuals? These issues have been studied by criminologists for many years. Of continuing importance is the question of whether there are distinct subgroups or clusters within the population, or whether observed criminal behaviors are part of a continuum. Naturally, one pattern of particular interest is "desistance', the discontinuation of regular offending.

The classic study Glueck and Glueck (1950) considered the criminal histories of 500 delinquent boys. The Gluecks and subsequent researchers (especially Sampson and Laub, 1993) carried out a prospective longitudinal study of the formation and development of criminal "careers" of the individuals in their sample. The subjects were initially interviewed at age around 14, and were followed up subsequently, both by personal interview and through FBI and police records. The main part of the data was collected by the Gluecks themselves over the period 1940 to 1965, but there are subsequent data right up to the present day, giving individual life course information up to age 70. These data are very unusual in providing long-term longitudinal information; most criminological data are cross-sectional or at best longitudinal only over restricted age ranges.

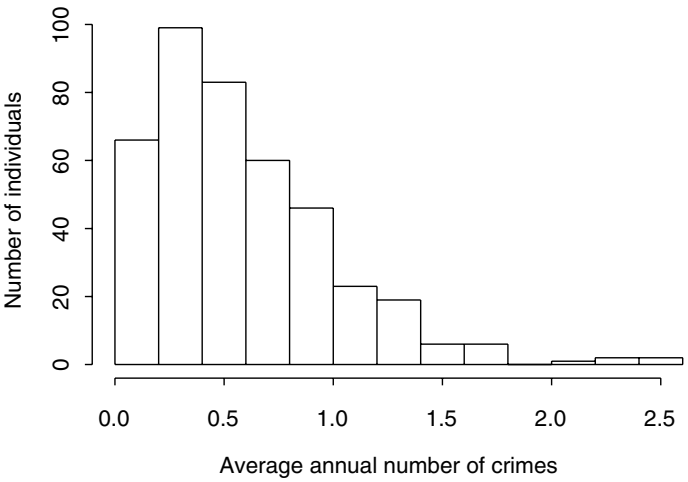


Figure 2.1. Histogram of the average annual number of arrests for each of 413 men over a 25-year time period.

The objective is to understand the pattern or trajectory through life of offending for the members of the sample. For each individual, the number of official arrests in each year of their life is recorded, starting in some cases as early as age 7. Obviously these are only a surrogate for the number of crimes committed, but they give a good indication of the general level of criminal activity. There is information on the type of crime and also on various concomitant information, but we do not consider this in detail.

### *2.1.2 The life course data*

We concentrate on a single set of data giving the numbers of arrests of 413 men over a 25-year period in each of their lives, from age 11 to age 35. These are the individuals for whom we have full information over this period. An immediate indication of the diversity within the group is given by considering the overall annual average number of arrests for each individual. Figure 2.1 shows that some of the men had only a low overall arrest rate, while others were clearly habitual offenders with 50 or more arrests registered in total. It is also clear that the distribution is highly skewed.

Another aspect is the high variability for each individual over time. Figure 2.2 shows the raw data for a typical individual. It can be seen that this person was arrested in connection with three offenses at age 11, one at age 14, and so on. The small numbers of crimes each year mean that

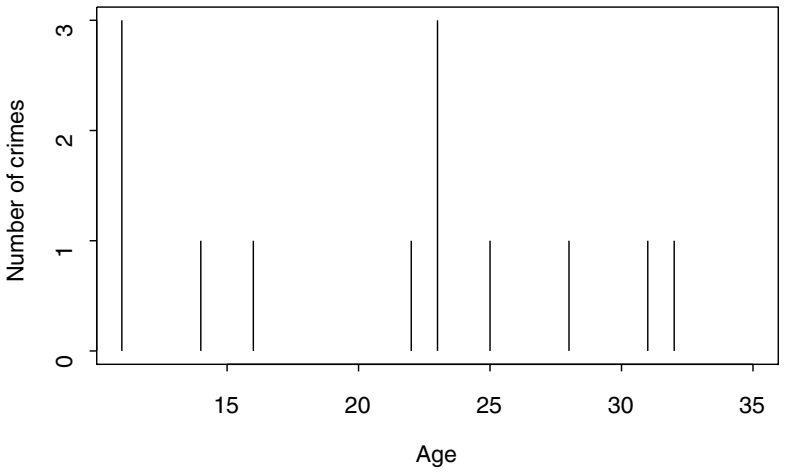


Figure 2.2. The record of a particular individual, showing the numbers of arrests at various ages. This individual was arrested for three offenses at age 11, one at age 14, and so on, but was not arrested at all in years 12, 13, 15, etc.

every individual is likely to show a sporadic pattern of some sort. Despite the very noisy nature of the data, one of our aims is to find ways of quantifying meaningful patterns in individuals that reflect variation in the wider population.

Our analysis raises a number of questions of broader importance in functional data analysis. The approach is to represent the criminal record of each subject by a single function of time, and then to use these functions for detailed analysis. But how should discrete observations be made into functional data in the first place? Does the functional nature of the data have any implications when producing smoothed estimates of quantities such as the overall mean curve? How can meaningful aspects of variation of the entire population be estimated and quantified in the presence of such large variability in individuals?

## 2.2 First steps in a functional approach

### *2.2.1 Turning discrete values into a functional datum*

We construct for each individual a function of time that represents his level of criminal activity. A simple approach would be to interpolate the raw

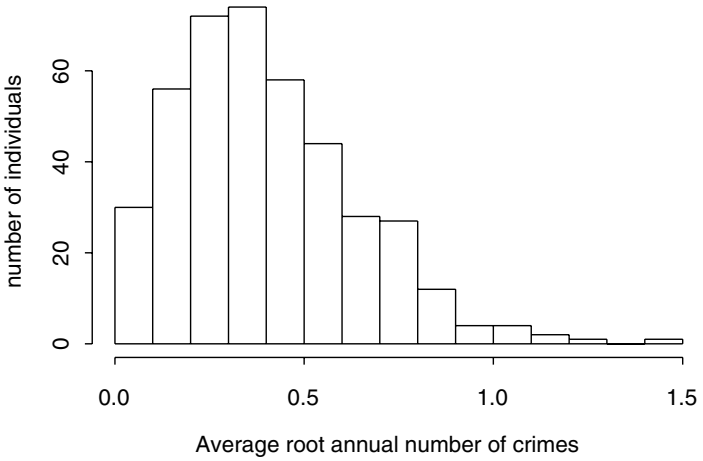


Figure 2.3. Histogram of the averages for each of 413 individuals of the square roots of annual tallies of arrests.

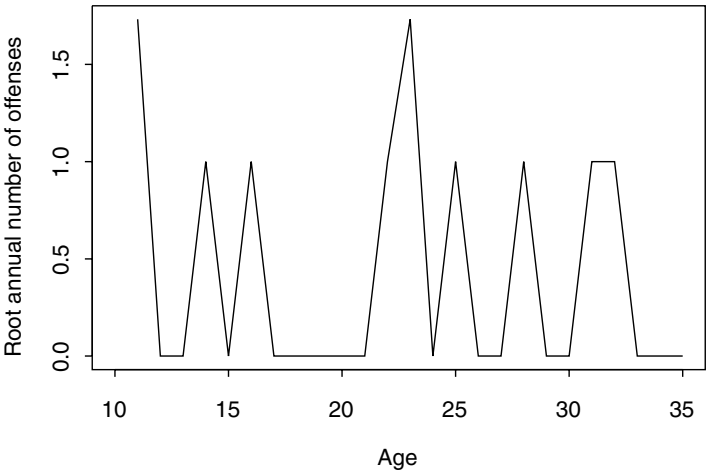


Figure 2.4. Linear interpolant of the square roots of the counts shown in Figure 2.2.

numbers of arrests in each year, but because of the skewness of the annual counts this would give inordinate weight to high values in the original data. In order to stabilize the variability somewhat, we start by taking the square root of the number of arrests each year. The rationale for this is partly pragmatic: if we plot a histogram of the averages across time of these square roots we see from Figure 2.3 that the skewness is somewhat reduced. In addition, if the numbers of arrests are Poisson counts, then the square root is the standard variance-stabilizing transformation.

One could conceivably smooth the square roots of annual counts to produce a functional observation for the individual considered in Figure 2.2. However, in order not to suppress any information at this stage, we interpolate linearly to produce the functional observation shown in Figure 2.4. We now throw away the original points and regard this function as a whole as being the datum for this individual. In the remainder of this chapter, we denote by $Y_1(t), Y_2(t), \ldots, Y_{413}(t)$ the 413 functional observations constructed from the square roots of the annual arrest count for the 413 individuals in the study.

### *2.2.2 Estimating the mean*

The next step in the analysis of the data is to estimate the mean function of the functional data. The natural estimator to begin with is simply the sample average defined in this case by

$$\bar{Y}(t) = \frac{1}{413} \sum_{i=1}^{413} Y_i(t).$$

The function $\bar{Y}(t)$ is plotted in Figure 2.5. It can be seen that, despite the large number of functions on which the mean is based, there is still some fluctuation in the result of a kind that is clearly not relevant to the problem at hand; there is no reason why 29-year olds commit fewer offenses than both 28- and 30-year olds for instance! Before embarking on a discussion of smoothing the mean function, it should be pointed out that this particular set of data has high local variability. In many other practical examples no smoothing will be necessary.

There are many possible approaches to the smoothing of the curve in Figure 2.5, and the one we use is a *roughness penalty* method. We measure the roughness, or variability, of a curve $g$ by the integrated squared second derivative of $g$. Our estimate of the overall mean is then the curve $m_\lambda(t)$ that minimizes the penalized squared error

$$S_\lambda(g) = \int \{g(t) - \bar{Y}(t)\}^2 dt + \lambda \int \{g''(t)\}^2 dt. \tag{2.1}$$

Here the *smoothing parameter* $\lambda \geq 0$ controls the trade-off between closeness of fit to the average of the data, as measured by the first integral in

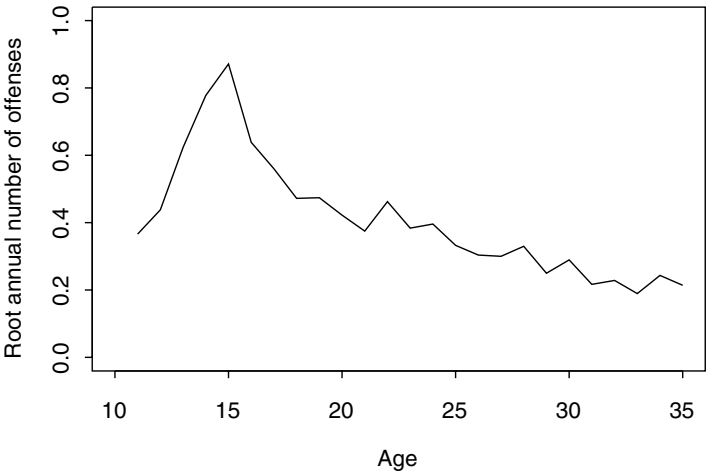


Figure 2.5. The sample mean function of the criminology functional data.

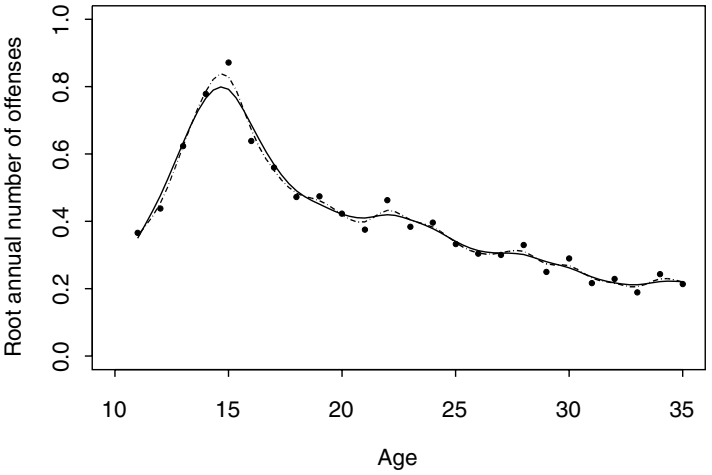


Figure 2.6. Estimate of the overall mean of the square root of the number of arrests per year. Points: raw means of the data. Dashed curve: roughness penalty smooth, $\lambda = 2 \times 10^{-7}$, cross-validation choice. Solid curve: roughness penalty smooth, $\lambda = 10^{-6}$, subjective adjustment.

(2.1) and the variability of the curve, as measured by the second integral. Both integrals are taken over the range of the parameter $t$, in this case from 11 to 35. If $\lambda = 0$ then the curve $m_\lambda(t)$ is equal to the sample mean curve $\bar{Y}(t)$. As $\lambda$ increases, the curve $m_\lambda(t)$ gets closer to the standard linear regression fit to the values of $\bar{Y}(t)$.

In practice, the smoothing parameter $\lambda$ has to be chosen to obtain a curve $m_\lambda(t)$ that is reasonably faithful to the original sample average but eliminates obviously extraneous variability. In practice, it is often easiest to choose the smoothing parameter subjectively, but in some circumstances an automatic choice of smoothing parameter may be useful, if only as a starting point for further subjective adjustment. An approach to this automatic choice using a method called *cross-validation* is discussed in Section 2.6. In Figure 2.6 we give the smoothed mean curve obtained by an automatic choice of smoothing, and also the effect of a subjective adjustment to this automatic choice. For the remainder of our analysis, this subjectively smoothed curve is used as an estimate of the overall mean function. We use the subjectively smoothed curve rather than the initial automatic choice because of the need to have a firm stable reference curve against which to judge individuals later in the analysis. In constructing this reference, we want to be sure that spurious variability is kept to a minimum.

## 2.3 Functional principal component analyses

### *2.3.1 The basic methodology*

What are the types of variability between the boys in the sample? There is controversy among criminologists as to whether there are distinct criminal groups or types. Some maintain that there are, for instance, specific groups of high offenders, or persistent offenders. Others reject this notion and consider that there is a continuum of levels and types of offending.

Principal components analysis (PCA) is a standard approach to the exploration of variability in multivariate data. PCA uses an eigenvalue decomposition of the variance matrix of the data to find directions in the observation space along which the data have the highest variability. For each principal component, the analysis yields a *loading vector* or *weight vector* which gives the direction of variability corresponding to that component. For details, see any standard multivariate analysis textbook, such as Johnson and Wichern (2002).

In the functional context, each principal component is specified by a *principal component weight function* $\xi(t)$ defined over the same range of $t$ as the functional data. The principal component scores of the individuals in the sample are the values $z_i$ given by

$$z_i = \int \xi(t) Y_i(t) dt. \tag{2.2}$$

The aim of simple PCA is to find the weight function $\xi_1(t)$ that maximizes the variance of the principal component scores $z_i$ subject to the constraint

$$\int \xi(t)^2 dt = 1. \tag{2.3}$$

Without a constraint of this kind, we could make the variance as large as we liked simply by multiplying $\xi$ by a large quantity.

The second-, third-, and higher-order principal components are defined in the same way, but with additional constraints. The second component function $\xi_2(t)$ is defined to maximize the variance of the principal component scores subject to the constraint (2.3) and the additional constraint

$$\int \xi_2(t)\xi_1(t)dt = 0. \tag{2.4}$$

In general, for the $j$th component we require the additional constraints

$$\int \xi_j(t)\xi_1(t)dt = \int \xi_j(t)\xi_2(t)dt = \ldots = \int \xi_j(t)\xi_{j-1}(t) = 0, \tag{2.5}$$

which will ensure that all the estimated principal components are mutually orthogonal.

In the case of the criminology data, the approach just described corresponds approximately to the following; the approximation is due to the approximation of the integrals by sums in (2.2) through (2.5).

1. Regard each of the functional data as a vector in 25-dimensional space, by reading off the values at each year of the individual's age.

2. Carry out a standard PCA on the resulting data set of 413 observations in 25-dimensional space.

3. Interpolate each principal component weight vector to give a weight function.

In Figure 2.7 the results of this approach are illustrated. For each of the first three principal components, three curves are plotted. The dashed curve is the overall smoothed mean, which is the same in all cases. The other two curves show the effect of adding and subtracting a suitable multiple of the principal component weight function.

It can be seen that the first principal component corresponds to the overall level of offending from about age 15 to age 35. All the components have a considerable amount of local variability, and in the case of the second component, particularly, this almost overwhelms any systematic effect. Clearly some smoothing is appropriate, not surprisingly given the high variability of the data.

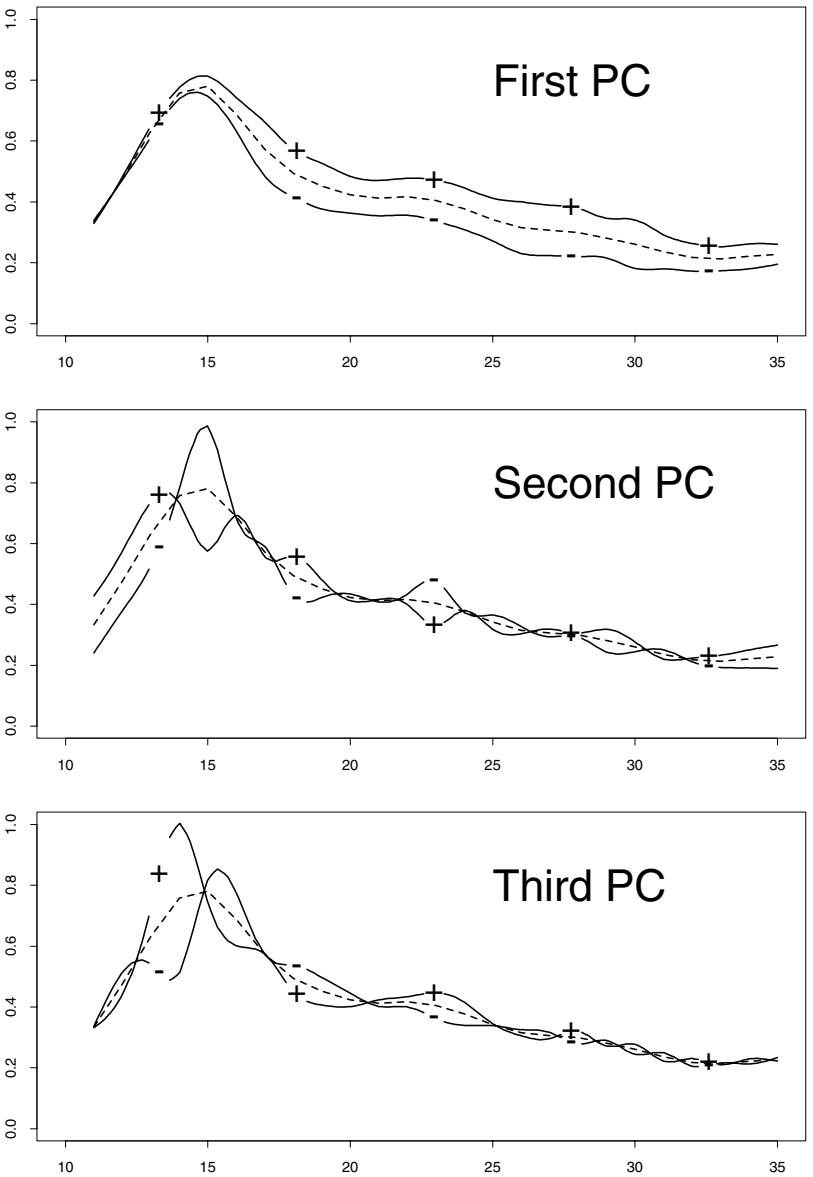


Figure 2.7. The effect of the first three unsmoothed principal components of the criminology data. In each graph, the dashed curve is the overall mean, and the solid curves are the mean ± a suitable multiple of the relevant principal component weight function. The + and − signs show which curve is which.

### 2.3.2 Smoothing the PCA

Smoothing a functional principal component analysis is not just a matter of smoothing the components produced by a standard PCA. Rather, we return to the original definition of principal components analysis and incorporate smoothing into that. Let us consider the leading principal component first of all.

To obtain a smoothed functional PCA, we take account of the need not only to control the size of $\xi$, but also to control its roughness. With this in mind, we replace the constraint (2.3) by a constraint that takes roughness into account as well. Thus, the first smoothed principal component weight function is the function $\xi_1(t)$ that maximizes the variance of the principal component scores subject to the constraint

$$\int \{\xi(t)\}^2 dt + \alpha \int \{\xi''(t)\}^2 dt = 1. \tag{2.6}$$

As usual, the parameter $\alpha \geq 0$ controls the amount of smoothing inherent in the procedure.

A roughness penalty is also incorporated into the additional constraints on the second-, third-, and higher-order smoothed principal components. The second component function $\xi_2(t)$ is now defined to maximize the variance of the principal component scores subject to (2.6) and the additional constraint

$$\int \xi_2(t)\xi_1(t)dt + \alpha \int \xi_2''(t)\xi_1''(t)dt = 0. \tag{2.7}$$

For the $j$th component we require constraints analogous to (2.5), but with corresponding extra terms taking the roughness penalty into account. This will ensure that the estimated components satisfy the condition

$$\int \xi_i(t)\xi_j(t)dt + \alpha \int \xi_i''(t)\xi_j''(t)dt = 0$$

for all $i$ and $j$ with $i \neq j$.

There are some attractive features to this approach to defining a smoothed principal components analysis. First, when $\alpha = 0$, we recover the standard unsmoothed PCA of the data. Second, despite the recursive nature of their definition, the principal components can be found in a single linear algebra calculation; details are given in Section 2.5.3.

### 2.3.3 Smoothed PCA of the criminology data

The first three principal component weight functions arising from a smoothed PCA are given in Figure 2.8. The smoothing parameter was chosen by subjective adjustment to the value $\alpha = 10^{-5}$. It can be seen that each of these components now has a clear interpretation.

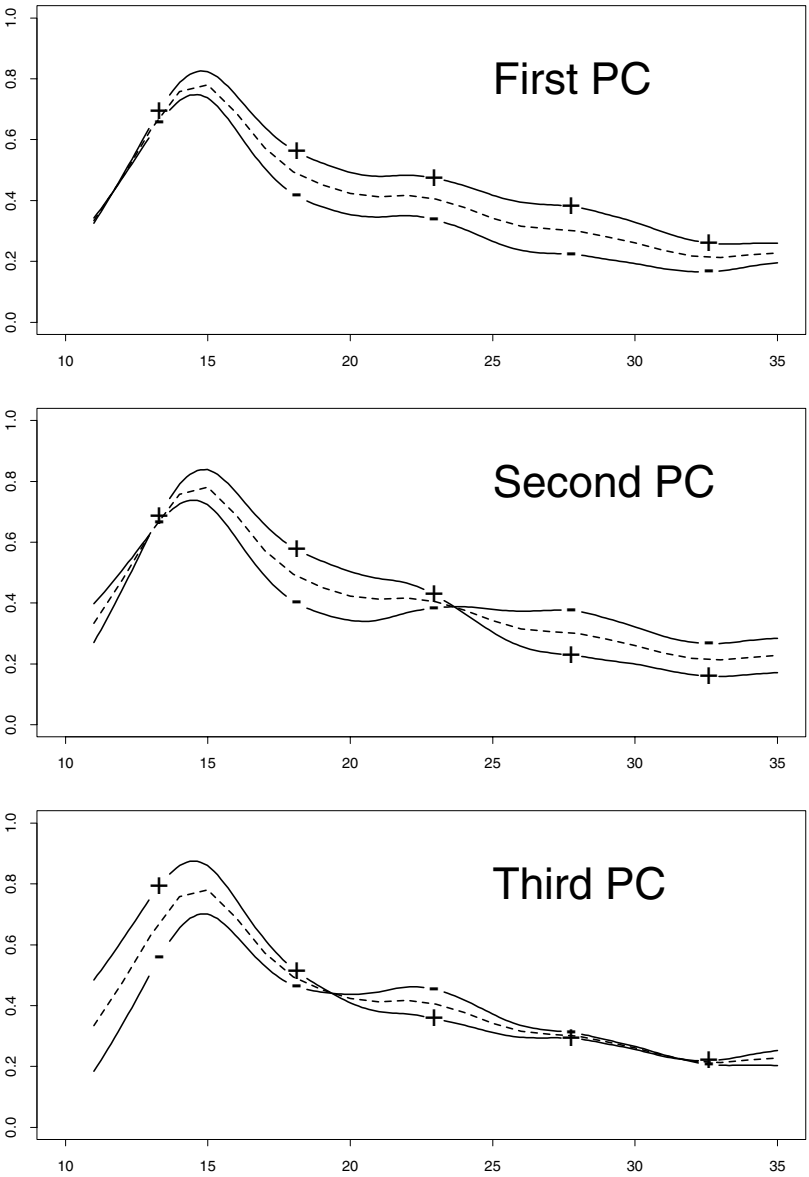


Figure 2.8. The effect on the mean curve of adding and subtracting a multiple of each of the first three smoothed functional principal components. The smoothing parameter was set to $\alpha = 10^{-5}$.

The first quantifies the general level of criminal activity throughout later adolescence and adulthood. A high scorer on this component would show especially above-average activity in the years from age 18 to age 30. It is interesting that this increased difference is not in the teenage years when the general level is very high anyway. High scorers on this component are above average during late adolescence but not markedly so; it is in their late teens and twenties that they depart most strongly from the mean. For this reason we call this component "Adult crime level."

The second component indicates a mode of variability corresponding to high activity up to the early twenties, then reforming to better than average in later years. High scorers are juvenile delinquents who then see the error of their ways and reform permanently. On the other hand those with large negative scores are well-behaved teenagers who then later take up a life of crime. We call this component "Long-term desistance."

The third component measures activity earlier in life. High scorers on this component are high offenders right from childhood through their teenage years. The component then shows a bounceback in the early twenties, later reverting to overall average behavior. This component is most affected by juvenile criminal activity and we call it "Juvenile crime level."

Sampson and Laub (1993, Chapter 1) place particular emphasis on early onset of delinquency and on adult desistance as important aspects of the life course often neglected by criminologists. Our analysis supports their claim, because the smoothed principal components analysis has picked out components corresponding to these features.

### *2.3.4 Detailed examination of the scores*

We now find the score of each of the 413 individuals in the sample on these three principal components, by integrating the weight function against the functional datum in each case. This gives each individual a score on each of the attributes "adult," "desistance," and "juvenile." These are plotted in pairs in Figure 2.9. There is essentially no correlation among these scores, so the three aspects can be considered as uncorrelated within the population.

However, the distribution of the first component, labeled "Adult" in the plots, is very skewed, with a long tail to the right; note that the mean of these scores is only 1.8. Even after taking the square root transformation, there are some individuals with very high overall rates of offending. If the overall score is low, then the values of "Desistance" are tightly clustered, but this is not the case for higher levels. This is for the simple reason that individuals with low overall crime rates have no real scope either to desist strongly, or to increase strongly. Because the overall rate cannot be negative, there are, essentially, constraints on the size of the second component in terms of that of the first, and these are visible in the plot. What the plot shows is that individuals with high overall rates can equally

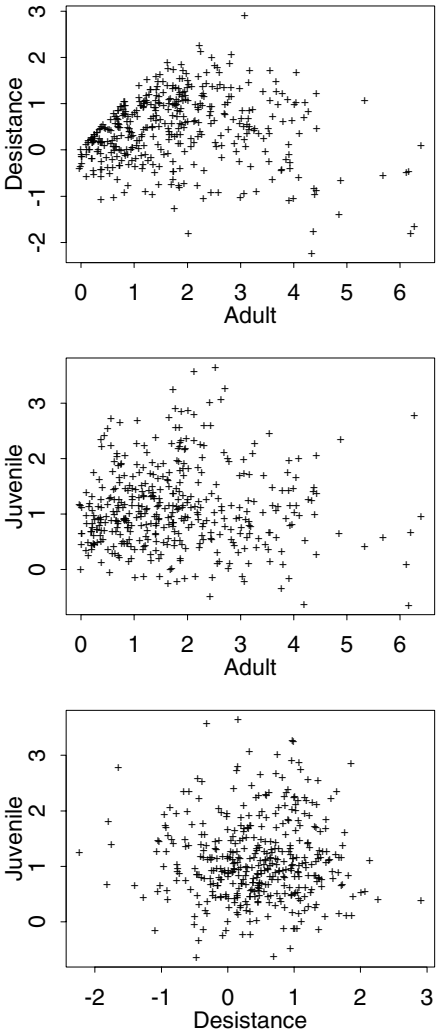


Figure 2.9. Plots of the first three principal components scores of the criminology life course data. The mean of the Adult scores is about 1.8.

well be strong desisters or strong "late developers," The same variability of behavior is not possible among low offenders.

The second and third components have symmetric unimodal distributions, and the third plot gives the kind of scatter one would expect from an uncorrelated bivariate normal distribution. The second plot of course shows the skewness of the "Adult" variable, but otherwise shows no very distinctive features.

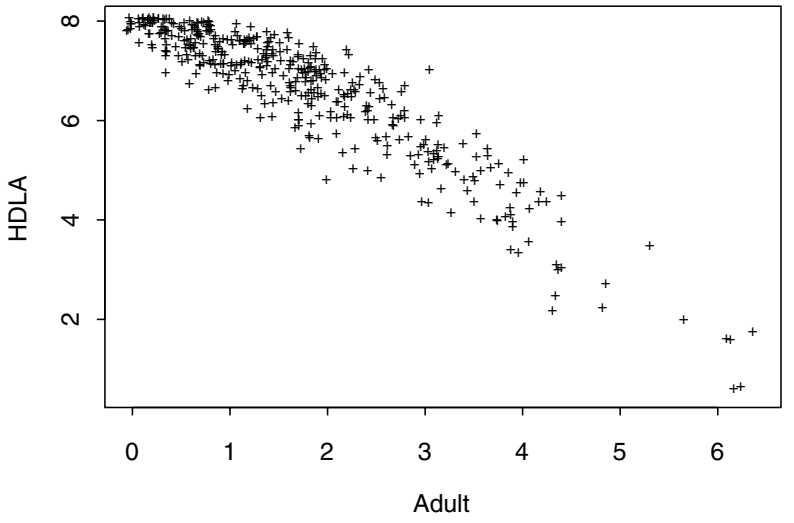


Figure 2.10. High desistance/low adult score plotted against Adult score.

Let us return to the plot of Adult against Desistance scores. An important issue in criminology is the existence of distinct groups of individuals in the population. There is no suggestion in this plot of a cluster of high-crime individuals even though there is a long tail in the distribution. However, there does appear to be a preponderance of cases near the upper boundary of the plotted points toward the left of the picture. These are all individuals with low adult crime rates and with nearly the maximum possible desistance for their adult crime scores. In order to identify these cases, we introduce a high desistance/low adult (HDLA) score, defined by

$$\text{HDLA} = 0.7 \times (\text{Desistance score}) - (\text{Adult score}) + 8.$$

A plot of the HDLA score against the Adult score is given in Figure 2.10. The multiple of 0.7 in the definition of HDLA was chosen to make the boundary at the top of this plot horizontal. The arbitrary constant 8 was added to make all the scores positive. We can see that there is a range of values of Adult scores for which HDLA is near its maximum value. A histogram of the HDLA values is given in Figure 2.11. Although the individuals with HDLA values near the maximum do not form a separate group, there is certainly a strong tendency for a cluster to form near this value. What do the trajectories of such individuals look like?

Ignoring all other variability, we examine the raw data of the 36 individuals with HDLA scores above 7.87. These are plotted in Figure 2.12. The individual trajectories cannot be easily distinguished, but the message

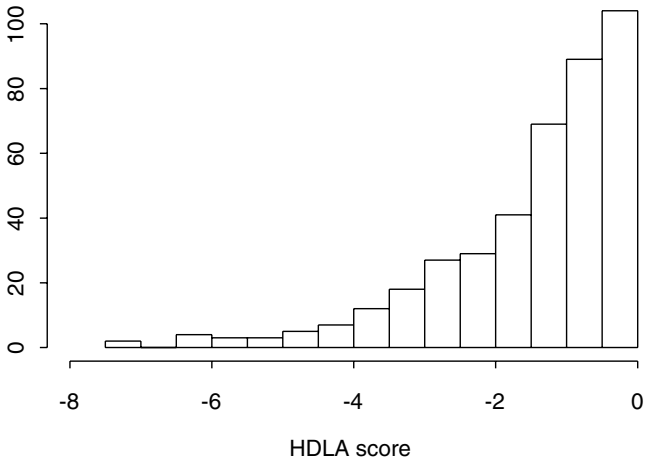


Figure 2.11. Histogram of the HDLA scores.

is clear: these are individuals who give up crime altogether by their late teens, even though earlier on they may have been quite high offenders. This is confirmed by Figure 2.13, which compares the HDLA score to the last age at which any offense is committed. A small number of individuals have very high HDLA scores but still offend very sporadically later in life. Thus the HDLA score is a more robust measure of almost total desistance than is the simple statistic of the last age at which any offense is committed.

## 2.4 What have we seen?

Constructing functional observations from discrete data is not always straightforward, and it is often preferable to transform the original data in some way. In the case of the criminology life course data, a square root of the original annual counts gave good results.

A key feature of the life course data is the high variability of the individual functional data. Even though there are over 400 curves, the sample mean curve still contains noticeable spurious fluctuation. A roughness penalty smoothing approach gives a natural way of incorporating smoothing into the estimation of the mean. In the functional context, some guidance as to the appropriate value of the smoothing parameter can be obtained by a cross-validation method discussed in more detail below.

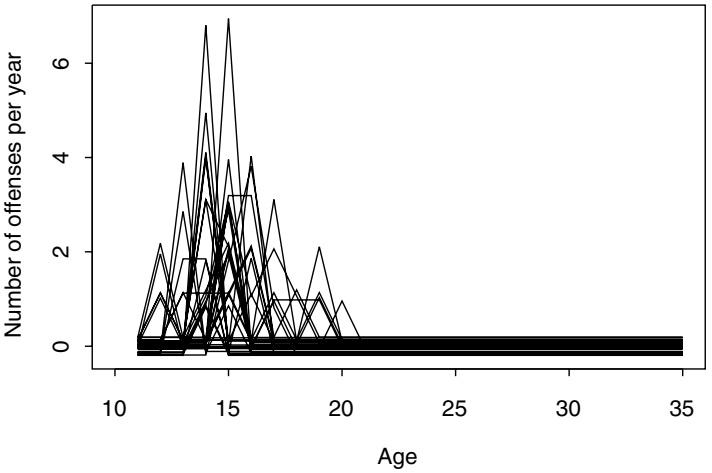


Figure 2.12. Raw data for the individuals with HDLA scores above 0.27. The data have been slightly jittered in order to separate the lines.

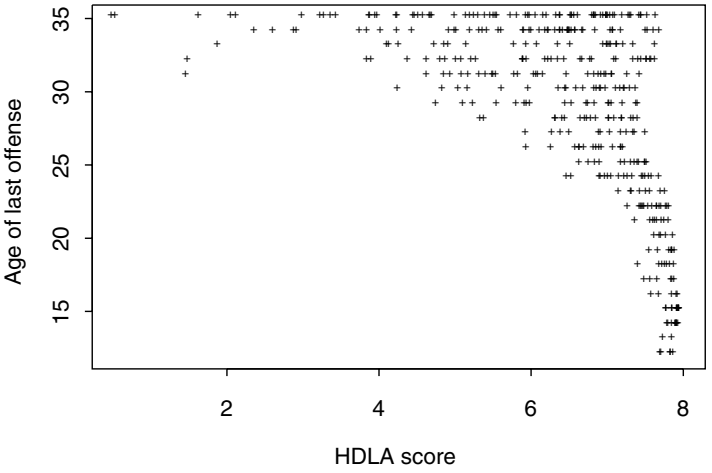


Figure 2.13. Age of last recorded offense plotted against HDLA scores. The individuals with highest HDLA scores correspond closely to those who give up crime altogether by age 20.

Without some smoothing, a functional principal components analysis of these data does not give very meaningful results. However, good results can be obtained by incorporating a roughness penalty into the size constraint of the principal component weight functions. The various principal components have immediate interpretations in terms of the original criminological issues, and can be used to build a composite score, the high desistance/low adult score, which brings out particular features of importance. There is no real evidence of strong grouping within the original data.

At this point, we have finished the specific task of analyzing the criminology data, but our discussion has raised two particular matters that are worth exploring in more detail. A general matter is the way that functional observations are stored and processed. A more specific issue is the cross-validation approach to the choice of smoothing parameter when estimating the mean. Some readers may wish to skip these sections, especially Section 2.5.2 onwards.

## 2.5 How are functions stored and processed?

### *2.5.1 Basis expansions*

In the example we have considered, we could simply store all the original values at the 25 evaluation points, since these points are the same for each individual in the sample. However, there are several reasons for considering other approaches. First, it is in the spirit of functional data analysis that we wish to specify the whole function, not just its value at a finite number of points. Second, it is important to have a method that can generalize to the case where the evaluation points are not the same for every individual in the sample. Third, we will often wish to be able to evaluate the derivatives of a functional datum or other function we are considering.

A good way of storing functional observations is in terms of a suitable *basis*. A basis is a standard set of functions, denoted $\beta_1(t), \beta_2(t), \ldots, \beta_m(t)$, for example, such that any function of interest can be expanded in terms of the functions $\beta_j(t)$. If a functional datum $x(t)$ is written

$$x(t) = \sum_{j=1}^{m} \xi_j \beta_j(t) \tag{2.8}$$

then the vector of $m$ coefficients $\xi = (\xi_1, \ldots, \xi_m)$ specifies the function.

Storing functional data in terms of an appropriate basis is a key step in most functional data analyses. Very often, the basis is defined implicitly within the procedure and there is no need for the user to be aware of it. For example, our treatment of the criminology life course data used a very simple basis, the *polygonal basis* made up of triangular functions like the ones shown in Figure 2.14. In mathematical terms, the basis functions $\delta_i(t)$

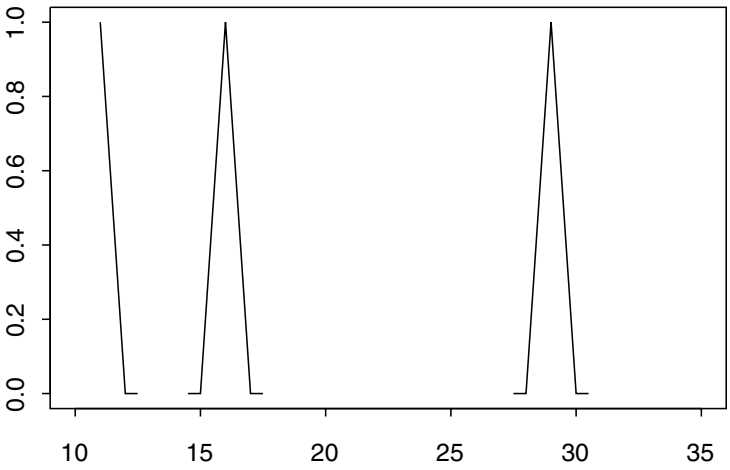


Figure 2.14. Three triangular basis functions. The functions are zero outside the range plotted.

are defined for $i = 1, 2, \ldots, 25$ and $11 \leq t \leq 35$ by setting $t_i = i + 10$ and

$$\delta_i(t) = \begin{cases} 1 - |t - t_i| & \text{if } |t - t_i| < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.9}$$

The coefficients $\xi_j$ of a particular function are, in this case, exactly the values $x(j + 10)$ of the function at the evaluation points. In between these points the function is interpolated linearly.

Because the basis functions $\delta_j(t)$ are not themselves everywhere smooth, they will not give rise to smooth basis expansions either. A good basis for the representation of smooth functions is a basis of B-splines, as plotted in Figure 2.15. B-splines are a flexible and numerically stable basis that is very commonly used. Except near the boundaries, the B-splines we use are all identical bell-shaped curves. The nonzero part of each B-spline is a piecewise cubic polynomial, with four cubic pieces fitting together smoothly to give a curve that has jumps only in its third derivative.

In the following sections, we give more details of the calculations involving basis expansions. These are intended for readers who are interested in the way that the basis expansions are used in practice and might wish to reconstruct the calculations for themselves. The algorithms are not explained in detail, but the more mathematically sophisticated reader not willing to take the results on trust should have no difficulty in reconstructing the arguments underlying them.

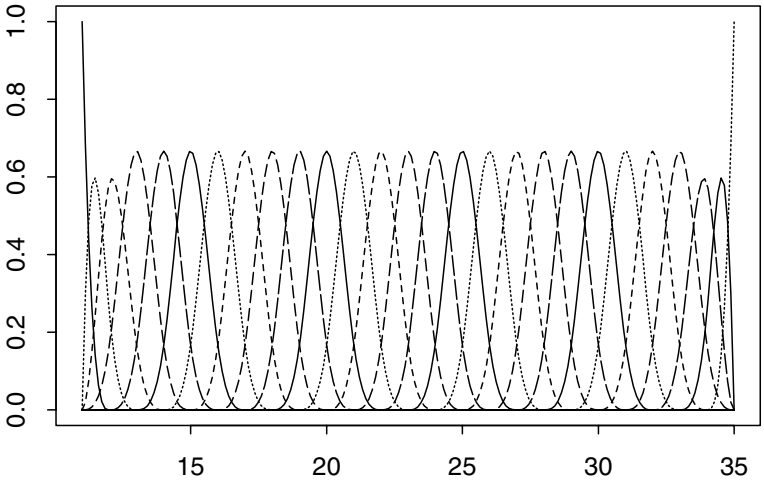


Figure 2.15. A B-spline basis that can be used to represent smooth functions

The first step is to use discrete observations of a function to obtain a basis representation. Then we move to the ways in which the smoothed mean estimation and the smoothed principal components analysis are carried out for a set of functional data held in basis representation form. The life course data are used as a concrete example, but the general principles can be extended widely. Some of this material is discussed in more detail in Ramsay and Silverman (1997) but it is convenient to draw it all together here. Some additional material, including S-PLUS software, is given in the Web page corresponding to this chapter.

### *2.5.2 Fitting basis coefficients to the observed data*

Consider the criminology data for a single individual in the sample. In our case the function corresponding to that individual is specified at the 25 points corresponding to ages from 11 to 35, and a triangular basis is used. More generally we will have values $x_1, x_2, \ldots, x_n$ at $n$ *evaluation points* $t_1, t_2, \ldots, t_n$, and we will have a more general set of basis functions $\beta_j(t)$. Define the $n \times m$ matrix $B$ to have elements

$$B_{ij} = \beta_j(t_i),$$

so that if the coefficient vector is $\xi$ then the vector of values at the evaluation points is $B\xi$.

There are now two cases to consider.[1] If there are no more basis functions than evaluation points, so that $m \leq n$, then we can fit the basis functions by least squares, to minimize the sum of squares of deviations between $x_k$ and $\sum_j \xi_j \beta_j(t_k)$. By standard statistical least squares theory, setting

$$\xi = (B'B)^{-1}B'x$$

will then specify the coefficients completely. If $m = n$ the resulting expansion $x(t) = \sum_j \xi_j \beta_j(t)$ will interpolate the values $x_i$ exactly, whereas if $m < n$ the expansion will be a smoothed version of the original data. In the criminology data example, the matrix $B$ is the identity matrix and so we simply set $\xi = x$.

On the other hand, if there are more basis functions than evaluation points, there will be many choices of $\xi$ that will interpolate the given values exactly, so that

$$x_k = \sum_{j=1}^{m} \xi_j \beta_j(t_k) \text{ for each } k = 1, 2, \ldots, n, \tag{2.10}$$

which can be written in vector form as $B\xi = x$. In order to choose between these, we choose the parameters that minimize the roughness of the curve, suitably quantified. For instance, if a B-spline basis is used, we can use the roughness penalty $\int \{x''(t)\}^2 dt$. Define the matrix $K$ by

$$K_{ij} = \int \beta_i''(t)\beta_j''(t)dt. \tag{2.11}$$

Then the roughness is equal to $\xi' K \xi$, so we choose the coefficient vector $\xi$ to minimize $\xi' K \xi$ subject to the constraint $B\xi = x$. If a triangular basis is used, we could use a roughness penalty based on first derivatives, but the principle is the same.

One specific feature of the general approach we have described is that it does not matter if the various functional data in the sample are not observed at the same evaluation points—the procedure will refer all the different functional data to the same basis, regardless of the evaluation points at which each has been observed.

### *2.5.3 Smoothing the sample mean function*

Now we move on to the calculation of the smoothed overall mean and to smoothed principal components analysis. In all cases, it is assumed that we have a set of functional data $Y_1(t), Y_2(t), \ldots, Y_n(t)$ expanded in terms

[1] This discussion is subject to the technical condition that $B$ is of full rank. If, exceptionally, this is not so, then a roughness penalty approach can still be used to distinguish between different basis representations that fit the data equally well.

of a basis $\delta_1(t), \ldots, \delta_m(t)$. Thus there is an $n \times m$ matrix $A = (a_{ij})$ of coefficients such that

$$Y_i(t) = \sum_{j=1}^{m} a_{ij}\delta_j(t).$$

If we let $\bar{a}_j = n^{-1}\sum_i a_{ij}$, then we have

$$\bar{Y}(t) = \sum_{j=1}^{m} \bar{a}_j\delta_j(t).$$

Because the basis functions $\delta_j(t)$ may not be sufficiently smooth to allow the appropriate roughness penalty to be defined, we may wish to use a different basis $\beta_k(t)$ of size $M$ when expanding the estimated mean curve. Given an $M$-vector $\gamma$ of coefficients, consider the function $g$ with these basis function coefficients in the new basis:

$$g(t) = \sum_{j=1}^{m} \gamma_j\beta_j(t).$$

Define the matrices $J$ and $L$ by

$$J_{ij} = \int \beta_i(t)\beta_j(t)dt \quad \text{and} \quad L_{ij} = \int \beta_i(t)\delta_j(t)dt$$

and the matrix $K$ by (2.11) above.

From these definitions it follows that

$$\int \{g(t) - \bar{Y}(t)\}^2 dt + \lambda \int g''(t)^2 dt = \int \bar{Y}(t)^2 dt + \gamma' J\gamma + \lambda\gamma' K\gamma - 2\gamma' L\bar{a}.$$

By standard linear algebra, this expression is minimized when $\gamma$ is the vector of coefficients $\gamma^{(\lambda)}$ given by

$$(J + \lambda K)\gamma^{(\lambda)} = L\bar{a}. \tag{2.12}$$

Solving equation (2.12) to find $\gamma^{(\lambda)}$, we can conclude that

$$m_\lambda(t) = \sum_{j=1}^{m} \gamma_j^{(\lambda)}\beta_j(t).$$

### *2.5.4 Calculations for smoothed functional PCA*

Now consider the smoothed functional principal components analysis as discussed in Section 2.3. Suppose that $\xi(t)$ is a possible principal component weight function, and that the vector $f$ gives the coefficients of the basis expansion of $\xi(t)$ in terms of the $\beta_j(t)$, so that

$$\xi(t) = \sum_{j=1}^{m} f_j\beta_j(t).$$

The vector of principal component scores of the data is then

$$\left(\int \xi(t)Y_i(t)dt\right) = AL'f. \tag{2.13}$$

Let $V$ be the sample variance matrix of the basis coefficients of the functional data, so that

$$V_{jk} = (n-1)^{-1}\sum_{i=1}^{n}(a_{ij} - \bar{a}_j)(a_{ik} - \bar{a}_k).$$

The variance of the principal component scores is then $f'LVL'f$. On the other hand, the constraint (2.6) on the size and roughness of $\xi(t)$ is given by

$$\int \xi(t)^2 dt + \alpha \int \xi''(t)^2 dt = f'(J + \alpha K)f = 1. \tag{2.14}$$

To find the leading smoothed principal component, we need to maximize the quadratic form $f'LVL'f$ subject to the constraint (2.14). There are several ways of doing this, but the following approach works well.

**Step 1** Use the Choleski decomposition to find a matrix $U$ such that $J + \alpha K = U'U$.

**Step 2** Write $g = Uf$ so that $f'(J + \alpha K)f = g'g$. Define $g^{(1)}$ to be the leading eigenvector of the matrix $(U^{-1})'LVL'U^{-1}$. Normalize $g^{(1)}$ to have length 1, so that $g^{(1)}$ maximizes $(U^{-1}g)'LVL'U^{-1}g$ subject to the constraint $g'g = 1$. Set $f^{(1)} = U^{-1}g^{(1)}$. Then $f^{(1)}$ is the basis coefficient vector of the leading smoothed principal component weight function.

**Step 3** More generally, let $g^{(j)}$ be the $j$th normalized eigenvector of $(U^{-1})'LVL'U^{-1}$. Then $U^{-1}g^{(j)}$ is the basis coefficient vector of the $j$th smoothed principal component weight function.

## 2.6 Cross-validation for estimating the mean

In classical univariate statistics, the mean of a distribution is the least squares predictor of observations from the distribution, in the sense that if $\mu$ is the population mean, and $X$ is a random observation from the distribution, then $E\{(X-\mu)^2\} < E\{(X-a)^2\}$ for any other number $a$. So one way of evaluating an estimate of $\mu$ is to take a number of new observations from the distribution, and see how well they are predicted by the value yielded by our estimate. In the one-dimensional case this may not be a very important issue, but in the functional case, we can use this insight to guide our choice of smoothing parameter.

In an ideal world, we would measure the efficacy of prediction by comparing the estimated mean curve to new functional observations. However, it would take 25 years or more to collect new data! (And, even if we were prepared to wait, the social context would have changed in such a way as to make it impossible to assume the new data came from the same distribution as the original data.) Therefore we have to manufacture the "new observation" situation from our existing data.

The way we do this is to leave each function out in turn from the estimation of the mean. The function left out plays the role of "new data." To be precise, let $m_\lambda^{-i}(t)$ be the smoothed sample mean calculated with smoothing parameter $\lambda$ from all the data except $Y_i(t)$. To see how well $m_\lambda^{-i}$ predicts $Y_i$, we calculate

$$\int \{m_\lambda^{-i}(t) - Y_i(t)\}^2 dt.$$

To avoid edge effects, the integral is taken over a slightly smaller range than that of the data; we integrate over $12 \leq t \leq 34$, but in this case the results are not much affected by this restriction. We now cycle through the whole functional data set and add these integrals together to produce a single measure of the efficacy of the smoothing parameter $\lambda$. This quantity is called the *cross-validation score* $\mathrm{CV}(\lambda)$; in our case

$$\mathrm{CV}(\lambda) = \sum_{i=1}^{413} \int_{12}^{34} \{m_\lambda^{-i}(t) - Y_i(t)\}^2 dt.$$

The smaller the value of $\mathrm{CV}(\lambda)$, the better the performance of $\lambda$ as measured by the cross-validation method.

A plot of the cross-validation score for the criminology data is shown in Figure 2.16. The smoothing parameter value selected by minimizing this score is $\lambda = 2 \times 10^{-7}$. As noted in Figure 2.6, the use of this smoothing parameter yields an estimated mean with some remaining fluctuations that are presumably spurious, and in our context it is appropriate to adjust the smoothing parameter upward a little. In general, it is advisable to use automatic methods such as cross-validation as a guide rather than as a rigid rule.

Before leaving the subject of cross-validation, it is worth pointing out the relation between the cross-validation method we have described here and the standard cross-validation method used in nonparametric regression. In nonparametric regression, we are interested in estimating a curve from a sample $(t_i, X_i)$ of numerical observations $X_i$ taken at time points $t_i$, and a cross-validation score for a particular smoothing procedure can be found by omitting the $X_i$ one at a time. In the functional case, however, we omit the functional data one at a time, and so the various terms in the cross-validation score relate to the way that a whole function $Y_i(t)$ is predicted from the other functions in the data set.

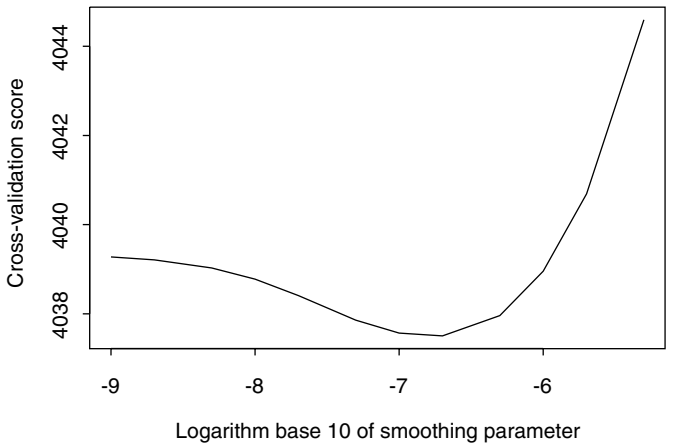


Figure 2.16. Cross-validation score for the estimation of the mean of the criminology data. The smoothing parameter is plotted on a logarithmic scale, and the minimum value is attained at $\lambda = 2 \times 10^{-7}$.

## 2.7 Notes and bibliography

Glueck and Glueck (1950) describe in detail the way in which the original sample of 500 delinquent boys was constructed and the initial part of the data collection, a process which they continued throughout their careers. A fascinating account of the original collection and processing of the life course data, and the way they were rediscovered, reconstructed, and reinforced is given by Sampson and Laub (1993). Sampson and Laub also describe the methodological controversies within the criminological research community which underlie the interest in the longitudinal analysis of these data.

A general discussion of roughness penalty methods is given in Ramsay and Silverman (1997, Chapter 4), and for a fuller treatment including bibliography the reader is referred to Green and Silverman (1994). The idea of smoothing using roughness penalties has a very long history, going back in some form to the nineteenth century, and certainly to Whittaker (1923). An important early reference to the use of cross-validation to guide the choice of smoothing parameter is Craven and Wahba (1979). In the functional context, the idea of leaving out whole data curves is discussed by Rice and Silverman (1991). The smoothing method for functional principal components analysis described in Section 2.3 is due to Silverman (1996). See also Ramsay and Silverman (1997, Chapter 7).