11

Canonical correlation and discriminant analysis

11.1 Introduction

11.1.1 The basic problem

In this chapter, we continue our consideration of exploratory approaches to functional data, specifically the case where we have observed *pairs* of functions $(X_i, Y_i), \quad i = 1, \ldots, N$, such as the hip and knee angles for the gait cycles of a number of children as discussed in Chapters 1 and 8. Suppose we wanted to know how variability in the knee angle cycle is related to that in the hip angle. In Section 8.5 we saw how principal components analysis can examine the variability in the two sets of curves taken together, but we did not explicitly address the issue of interaction between the two curves. In this chapter, we pursue a somewhat different emphasis by considering *canonical correlation analysis* (CCA), which seeks to investigate which modes of variability in the two sets of curves are most associated with one another.

In the functional context, canonical correlation analysis provides a pair of functions $(\xi(s), \eta(s))$ such that $\int \xi X_i$ and $\int \eta Y_i$ are well correlated with one another. We can think of $\xi(s)$ and $\eta(s)$ as the components of variation in the two curves that most account for the interaction between the hip and knee angles. Our method gives the curves shown in Figure 11.1. The values $\int \xi X_i$ and $\int \eta Y_i$ are called *canonical variates*, and the sample correlation between these variates is about 0.81 in this case.

In the figure, the curves ξ and η are rather similar, and the broad interpretation is that there is correlation between the two measurements $X_i(s)$

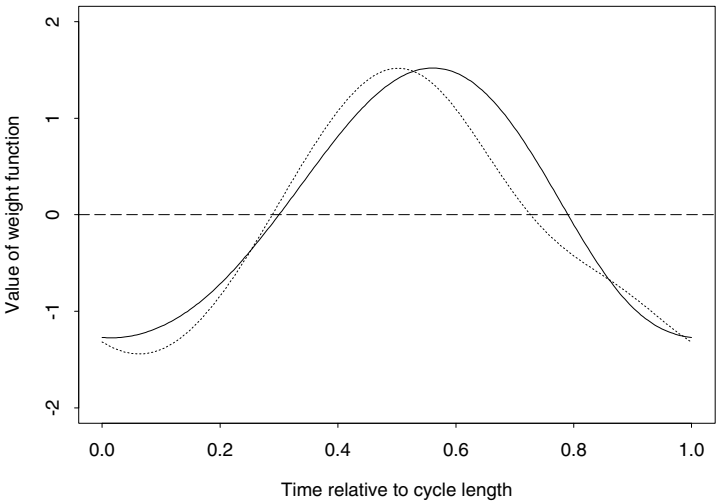

Figure 11.1. Estimated canonical variate weight functions for the gait data. Solid curve: weight function for hip observations; dotted curve: weight function for knee observations.

and $Y_i(s)$ at any particular time. But it is interesting that the extreme in the hip curve in the middle of the cycle occurs a little later than that in the knee curve, whereas the order of the extremes near the beginning of the cycle is reversed. This suggests that, in the middle of the cycle, high variability from the norm in the hip follows that in the knee; near the ends of the cycle, the effects occur in the opposite order. This may indicate a physical propagation of errors caused by the relevant strike of the heel at the beginning and in the middle of the cycle.

Having found these components of variability, we can go on to find further components of variation. Call the (ξ, η) we have already found (ξ_1, η_1). We can now look for another pair of functions (ξ_2, η_2) such that

- There is a high correlation between the variation in the hip angles described by a multiple of ξ_2 and that in the knee angles accounted for by η_2, but ...
- these effects are uncorrelated with the previously found contributions to variability corresponding to ξ_1 and η_1.

The functions ξ_2 and η_2 are shown in Figure 11.2. In this case the correlation between $\int \xi_2 X_i$ and $\int \eta_2 Y_i$ is about 0.72, only slightly lower than that for the first pair of canonical variates. The points at which the functions ξ_2 and η_2 cross the axis indicate conclusions similar to those outlined with respect to the leading variates. In the middle of the cycle the hip curve

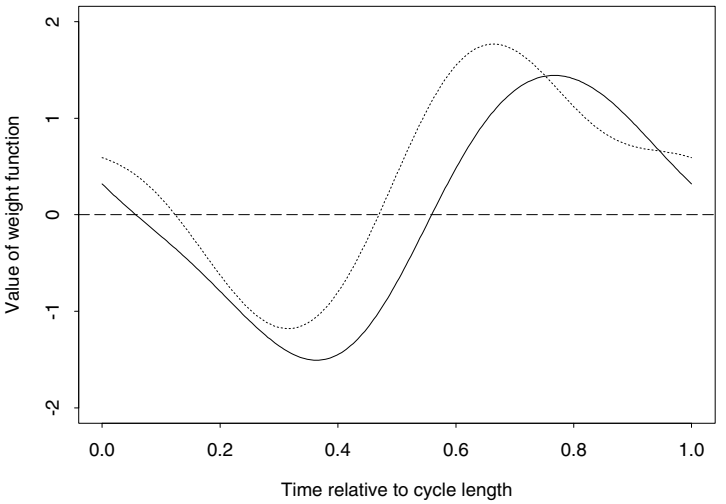

Figure 11.2. Second pair of smoothed canonical variate weight functions for the gait data. Solid curve: weight function for hip observations; dashed curve: weight function for knee observations.

crosses zero considerably later than the knee curve, whereas near the beginning of the cycle the hip curve crosses first. Put another way, we could roughly transform both the first and the second canonical variates to be identical for the hip and the knee by speeding up the hip cycle relative to the knee cycle in the first half of the cycle, and slowing it down in the second.

We shall see that the estimation of the weight functions as shown in Figures 11.1 and 11.2 is not quite straightforward and that an appropriate form of smoothing is essential. But first we review classical multivariate CCA; a fuller discussion can be found in most multivariate analysis textbooks, such as Anderson (1984). We then go on to develop our approach to functional CCA, largely based on the paper of Leurgans, Moyeed and Silverman (1993), and using the gait data as a running example. Another application is considered in Section 11.4. We shall see that some regularization is essential to obtain meaningful results, for reasons discussed briefly in Section 11.5. In Section 11.6, various algorithmic approaches and connections with other FDA topics are explored.

Finally, in Section 11.7, we present some extensions of the ideas of functional CCA to deal with problems of optimal scoring and discriminant analysis. This is based on work of Hastie, Buja and Tibshirani (1995).

11.2 Principles of classical canonical correlation analysis

Suppose we have n pairs of observed vectors (x_i, y_i), each x_i being a p-vector and each y_i being a q-vector. The object of canonical correlation analysis is to reduce the dimensionality of the data by finding the vectors $\mathbf{a}_1$ and $\mathbf{b}_1$ (p- and q-vectors respectively) for which the linear combinations $\mathbf{a}_1' x_i$ and $\mathbf{b}_1' y_i$ are as highly correlated as possible. The *canonical variates* $\mathbf{a}_1' x_i$ and $\mathbf{b}_1' y_i$ are the linear compounds of the original observations whose variability is most closely related in terms of correlation. The vectors $\mathbf{a}_1$ and $\mathbf{b}_1$ are called the *leading canonical variate weight vectors*.

Note that multiplying $\mathbf{a}_1$ and/or $\mathbf{b}_1$ by nonzero constants of the same sign does not alter the correlation. If the constants are opposite in sign, the correlation itself is reversed in sign but has the same magnitude. By convention, we choose $\mathbf{a}_1$ and $\mathbf{b}_1$ so that $\{\mathbf{a}_1' x_i\}$ and $\{\mathbf{b}_1' y_i\}$ both have sample variance equal to 1, and the correlation ρ_1 between the $\mathbf{a}_1' x_i$ and $\mathbf{b}_1' y_i$ is positive.

We can now go on to find subsidiary canonical variates. The jth pair of canonical variates is defined by a p-vector a_j and a q-vector b_j, chosen to maximize the sample correlation $\rho_j = \mathtt{corr}(\mathbf{a}_j' x_i, \mathbf{b}_j' y_i)$ subject to the constraints that

(a) $\mathtt{corr}(\mathbf{a}_j' x_i, \mathbf{a}_k' x_i) = 0$

(b) $\mathtt{corr}(\mathbf{b}_j' y_i, \mathbf{b}_k' y_i) = 0$

(c) $\mathtt{corr}(\mathbf{a}_j' x_i, \mathbf{b}_k' y_i) = 0,$

where in each case the correlations are the sample correlations as i takes the values $1, \ldots, n$.

11.3 Functional canonical correlation analysis

11.3.1 Notation and assumptions

We now return to the functional case, which is our main concern. As usual, assume that the N observed pairs of data curves (X_i, Y_i) are available for argument t in some finite interval $\mathcal{T}$, and that all integrals are taken over $\mathcal{T}$. Given functions $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, we define $\mathtt{ccorsq}(\boldsymbol{\xi}, \boldsymbol{\eta})$ to be the sample squared correlation of $\int \boldsymbol{\xi} X_i$ and $\int \boldsymbol{\eta} Y_i$, and therefore

$$\mathtt{ccorsq}(\boldsymbol{\xi}, \boldsymbol{\eta}) = \frac{\{\mathtt{cov}(\int \xi X_i, \int \eta Y_i)\}^2}{(\mathtt{var} \int \xi X_i)(\mathtt{var} \int \eta Y_i)}.$$

The use of a roughness penalty is central to our methodology. As usual we quantify the roughness of a function f by its integrated squared curvature $\|D^2 f\|^2 = \int (D^2 f)^2$.

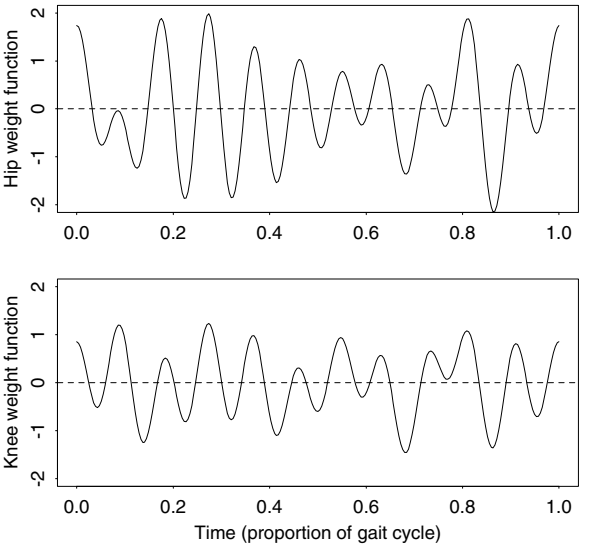

Figure 11.3. Unsmoothed canonical variate weight functions for the gait data that attain perfect correlation. Top panel: weight function for hip observations; bottom panel: weight function for knee observations.

11.3.2 The naive approach does not give meaningful results

For the moment concentrate on the leading canonical variates. We might imagine that the obvious way to proceed is simply to find functions $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ that maximize $\texttt{ccorsq}(\boldsymbol{\xi}, \boldsymbol{\eta})$. This would be equivalent to maximizing $\texttt{cov}(\int \xi X_i, \int \eta Y_i)$ subject to the constraints

$$\texttt{var}(\textstyle\int \xi X_i) = \texttt{var}(\int \eta Y_i) = 1. \tag{11.1}$$

However, simply carrying out this maximization does not produce a meaningful result. Figure 11.3 shows functions $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ that maximize the sample correlation `ccorsq` for the gait data example. The sample correlation achieved by these functions is 1. The functions displayed in Figure 11.3 do not give any meaningful information about the data and clearly demonstrate the need for a technique involving smoothing. In Section 11.5, we explain why this behavior is not specific to this particular data set but is an intrinsic property of CCA applied in the functional context.

A straightforward way of introducing smoothing is to modify the constraints (11.1) by adding roughness penalty terms to give

$$\texttt{var}(\textstyle\int \xi X_i) + \lambda \|D^2 \boldsymbol{\xi}\|^2 = \texttt{var}(\int \eta Y_i) + \lambda \|D^2 \boldsymbol{\eta}\|^2 = 1, \tag{11.2}$$

where λ is a positive smoothing parameter.

The effect of introducing the roughness penalty terms into the constraints is that, in evaluating particular candidates to be canonical variates, we

consider not only their variances, but also their roughness, and compare a weighted sum of these two quantities with the covariance term. The problem of maximizing the covariance $\mathtt{cov}(\int \xi X_i, \int \eta Y_i)$ subject to the constraints (11.2) is equivalent to maximizing the penalized squared sample correlation defined by

$$\mathtt{ccorsq}_\lambda(\boldsymbol{\xi}, \boldsymbol{\eta}) = \frac{\{\mathtt{cov}(\int \xi X_i, \int \eta Y_i)\}^2}{\{\mathtt{var}(\int \xi X_i) + \lambda \|D^2 \boldsymbol{\xi}\|^2\}\{\mathtt{var}(\int \eta Y_i) + \lambda \|D^2 \boldsymbol{\eta}\|^2\}}. \tag{11.3}$$

We refer to this procedure as *smoothed canonical correlation analysis*.

Our method of introducing smoothing or regularization is similar to the technique of ridge regression, which is often used in image processing and ill-posed problems to improve the conditioning of the variance matrices considered. The technique of ridge regression was applied to CCA by Vinod (1976). Multiplying the curves ξ and η by constants does not affect the value of the criterion $\mathtt{ccorsq}_\lambda(\boldsymbol{\xi}, \boldsymbol{\eta})$, and in the figures they are normalized to set $\int \xi^2 = \int \eta^2 = 1$.

11.3.3 Choice of the smoothing parameter

The larger the value of λ, the more emphasis is placed on the roughness penalty and the smaller will be the true correlation of the variates found by smoothed CCA. A good choice of the smoothing parameter is essential to give a pair of canonical variates with fairly smooth weight functions and a correlation that is not unreasonably low. The smoothing parameter can be chosen subjectively, but if we require an automatic procedure, a reasonable form of cross-validation is as follows:

Let $\mathtt{ccorsq}_\lambda^{-i}(\boldsymbol{\xi}, \boldsymbol{\eta})$ be the sample penalized squared correlation calculated as in (11.3) but with the observation (X_i, Y_i) omitted. Let $(\boldsymbol{\xi}_\lambda^{(-i)}, \boldsymbol{\eta}_\lambda^{(-i)})$ be the functions that maximize $\mathtt{ccorsq}_\lambda^{-i}(\boldsymbol{\xi}, \boldsymbol{\eta})$. The cross-validation score for λ is defined to be the squared correlation of the N pairs of numbers

$$(\int \boldsymbol{\xi}_\lambda^{(-i)} X_i, \int \boldsymbol{\eta}_\lambda^{(-i)} Y_i)$$

for $i = 1, \ldots, n$. We then choose λ to maximize this correlation. It is this choice of λ that was used for the gait data in Figures 11.1 and 11.2. The degree of smoothing chosen by cross-validation appears to be quite heavy, and to test the sensitivity of these conclusions, Leurgans, Moyeed and Silverman (1993) examined the first two pairs of canonical variates estimated with a value of λ reduced by a factor of 10. Though there was a little more variability in the canonical variate curves, the broad features remained the same.

Throughout this section, we have concentrated on the choice of smoothing parameter for the leading canonical variates. If we were particularly interested in the ideal smoothing parameter for a subsidiary canonical cor-

Table 11.1. Smoothed and unsmoothed sample correlations for the first three pairs of smoothed canonical variates for the gait data.

Canonical variates	Sample squared correlations	
	$\texttt{ccorsq}_\lambda(\boldsymbol{\xi}_\lambda, \boldsymbol{\eta}_\lambda)$	$\texttt{ccorsq}(\boldsymbol{\xi}_\lambda, \boldsymbol{\eta}_\lambda)$
First	0.755	0.810
Second	0.618	0.717
Third	0.141	0.198

relation, we could formulate a relevant cross-validation score. However, our practical experience has shown us that, although cross-validation works well for the leading canonical variate, its behavior is much more disappointing for subsequent canonical variates. We have found it to be more satisfactory simply to use the same value of λ for any subsidiary canonical variates considered.

We have used a single smoothing parameter λ for both ξ and η. It is possible to use separate smoothing parameters λ_1 and λ_2; the conceptual and algorithmic extensions are straightforward, but we have found a single smoothing parameter to be adequate in the examples we have considered.

11.3.4 The values of the correlations

Once the canonical variates have been found, we can consider the values of the correlations themselves. We can consider either the smoothed squared correlation $\texttt{ccorsq}_\lambda$ or the unsmoothed value $\texttt{ccorsq}$; there is no firm theoretical footing for the choice between them and in any case it would be a matter of some concern if the effect of smoothing was to make the values dramatically different.

For the gait data, Table 11.1 shows the values of the smoothed and unsmoothed squared correlations, and also includes corresponding values for the second and third pairs of smoothed canonical variates, estimated with the same λ. Table 11.1 shows that the second pair of canonical variates is almost as important as the first. On the other hand, the third pair of canonical variates have low estimated correlation, and we do not consider them further.

Before we leave the gait example, we note that scatterplots of the canonical variate scores $(\int \boldsymbol{\xi} X_i, \int \boldsymbol{\eta} Y_i)$ show that no particular curves have outlying scores for either of the first two canonical variates. In Section 8.5, we saw that the first principal component of variation in the hip curves alone corresponded to an overall vertical shift in the curves. If this shift were in any way correlated with a variation in the knee curves, the hip canonical variate curves would be more like constants than sine waves. Since this is not the case, we can see that this vertical shift is a property of the hip curves alone, independent of any variation in the knee angles.

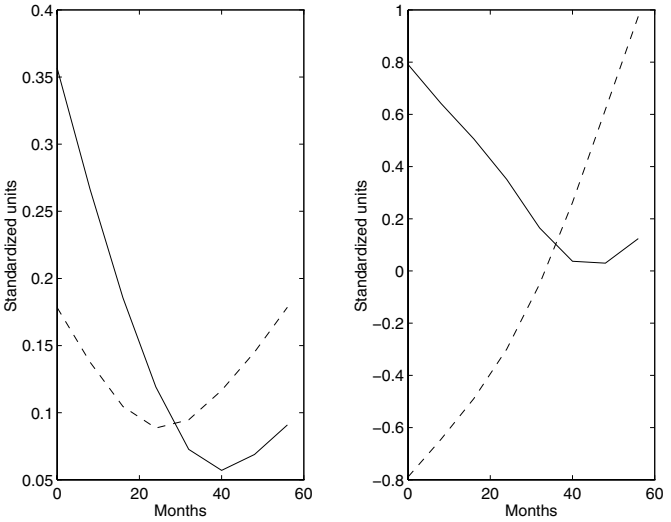

Figure 11.4. Smoothed canonical variate weight functions for the lupus data, from Buckheit et al. (1997). Left panel: results of CCA applied to GFR and KUC with solid curve corresponding to GFR and dashed curve to KUC. Right panel: results of CCA applied to GFR and GOP, with solid curve corresponding to GFR and dashed curve to GOP.

11.4 Application to the study of lupus nephritis

Buckheit, Olshen, Blouch and Myers (1997) applied functional CCA to renal physiology, in the study of diffuse proliferative lupus nephritis, and we present their results here as an illustration. The original paper should be consulted for further details; we are extremely grateful to Richard Olshen for his generosity in sharing and discussing this work with us prior to its publication.

They had available various measurements on a number of patients over a 60-month period. These include the glomerular filtration rate (GFR), the glomerular oncotic pressure (GOP) and the two-kidney ultrafiltration coefficient (KUC). They focused on nine patients labelled *progressors*, those whose kidney function, as measured by GFR was clearly declining over the period of study. The GFR measure is currently favored by clinicians as an overall indicator of progressive glomerular disease, a particular form of kidney degeneration, and therefore the progressors are the group suffering long-term kidney damage, likely to require eventual dialysis or transplantation. It is important to understand the kidney filtration dynamics in this disease, and this is facilitated by investigating the covariation between measured variables.

Within the progressor group, GFR and KUC tend to decrease considerably over the 60 month period, whereas the GOP measure increases somewhat. This contrasts with well-functioning kidneys, where an increase in GOP would be counteracted by an increase in KUC, resulting in steady GFR. Functional smoothed CCA was applied to explore variability and interaction effects in the progressor group. The correlations between GFR and each of KUC and GOP were investigated. Figure 11.4 shows the leading pairs of canonical variate weight functions. It is interesting that the linear functional of GFR most highly correlated with the other two variables is virtually the same in both cases.

To interpret the figure, remember that all patients concerned show an overall declining value of GFR. The U-shaped solid curves in the figure therefore correspond to a canonical variate where a positive value indicates a GFR record that starts at a value higher than average, but then declines more rapidly than average in the first 40 months, finally switching to a relatively less rapid decline in the last 20 months.

The left-hand panel shows that this variate is correlated with a similar effect for KUC, but the switch in rate happens earlier. This indicates not only that strong decline of GFR is associated with strong decline of KUC but also suggests that the pattern of GFR in some sense follows that of KUC, raising the hope that KUC could be used to predict future GFR behavior. On the other hand, the right-hand panel shows that this aspect of GFR behavior is correlated with an increase of GOP stronger than average over the entire time period. Thus, patients with rapidly increasing GOP are likely to be those whose GFR declines rapidly at first, though there may be some reduction in the rate of decline after about 36 months.

In broad terms, the CCA gives insights broadly consistent with those for the average behavior of the sample as a whole. It is interesting that the relationships between the variables are borne out on an individual level, not merely on an average level. Furthermore the detailed conclusions yielded by the CCA give important avenues for future thought and investigation concerning the way in which the variables interrelate. Of course, given the small sample size, any conclusions must be relatively tentative unless supported by other evidence.

11.5 Why is regularization necessary?

Apart from its importance as a practical method, canonical correlation analysis of functional data has an interesting philosophical aspect. In the principal components analysis context we have already seen that appropriately applied smoothing may improve the estimation accuracy. However, in most circumstances, we obtain reasonable estimates of the population principal components even if no smoothing is applied. By contrast, as we saw

in the gait example, in the context of functional CCA some regularization is absolutely essential to obtain meaningful results. This is the same conclusion that we will draw for the functional regression context discussed in Chapter 16. But in the canonical correlation case, the impact of smoothing is even more dramatic.

To understand the need for regularization, compare functional CCA with standard multivariate CCA. A standard condition of classical CCA is that $n > p + q + 1$ which ensures (with probability 1, under reasonable conditions) that the sample covariance matrix $\mathbf{V}_{12}$ of the n vectors (x_i, y_i) is nonsingular (see Eaton and Perlman,1973). In the functional case, p and q are essentially infinite, and so this condition cannot be fulfilled.

Furthermore, consider a sample $X_1, \ldots, X_N$ of functional data, and assume for the moment that the N curves are linearly independent. Now suppose that $z_1, \ldots, z_N$ is any real vector. By results that will be discussed in Chapter 16, it is possible to find a curve $\boldsymbol{\xi}$ such that, for some constant α_X, $z_i = \alpha_X + \int \boldsymbol{\xi} X_i$ for all i. Essentially, the reason for this is that we only have N constraints on ξ, but infinitely many degrees of freedom in the choice of ξ, because ξ is a function. Now suppose we have a second sample of curves Y_i, which may be correlated with the X_i in some way, and again are linearly independent. We can find a function $\boldsymbol{\eta}$ such that, for some constant α_Y, $z_i = \alpha_Y + \int \boldsymbol{\eta} Y_i$ for all i. This means that the given values z_i can be predicted perfectly either from the X_i or from the Y_i.

It follows that not only have we found functions $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ such that $\texttt{ccorsq}(\boldsymbol{\xi}, \boldsymbol{\eta}) = 1$, because the variates $\int \boldsymbol{\xi} X_i$ and $\int \boldsymbol{\eta} Y_i$ are perfectly correlated, but that we can prescribe the values z_i taken by the canonical variates to be whatever we please, up to a constant. In particular, we could start with *any* function $\boldsymbol{\xi}$, construct $z_i = \int \boldsymbol{\xi} X_i$, and then find a function $\boldsymbol{\eta}$ such that $\texttt{ccorsq}(\boldsymbol{\xi}, \boldsymbol{\eta}) = 1$. In this sense, every possible function can arise as a canonical variate weight function with perfect correlation!

Leurgans, Moyeed and Silverman (1993) discuss this result in greater detail. They demonstrate that the assumption of linear independence among the curves is a very mild one, and, by proving an appropriate consistency result, they show that regularization indeed makes meaningful estimates possible.

11.6 Algorithmic considerations

11.6.1 Discretization and basis approaches

There are several ways of carrying out our method of smoothed functional CCA numerically. For completeness, we present the methodology for the general case of different parameters λ_1 and λ_2. A direct approach is to set up a discrete version of the covariance `ccorsq` and of the constraints (11.2). Discretize the functions $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ and the covariance operators $v_{jk}(s, t)$ using

a fine grid, and replace the operator D^2 by a finite difference approximation. The problem then becomes one of maximizing a quadratic form subject to quadratic constraints, and it can be solved by standard numerical methods.

We can also use a basis for the functions X_i and Y_i, and for the weight functions $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$. Suppose that $\phi_1, \phi_2, \ldots, \phi_M$ is a suitable basis, which for simplicity we will assume is used for all of these four functions. As usual, define $\mathbf{K}$ to be the matrix with entries $\int (D^2\phi_j)(D^2\phi_k)$ and $\mathbf{J}$ the matrix with entries $\int \phi_j \phi_k$. If we use a Fourier or other orthonormal basis, then $\mathbf{J}$ is the identity matrix.

Define $\mathbf{C}$ and $\mathbf{D}$ to be the matrices of coefficients of the basis expansions of the X_i and Y_i respectively, meaning that

$$X_i = \sum_{\nu=1}^{M} c_{i\nu}\phi_\nu$$

and

$$Y_i = \sum_{\nu=1}^{M} d_{i\nu}\phi_\nu$$

up to the degree of approximation involved in any choice of the number M of basis functions considered. Write $\mathbf{a}$ and $\mathbf{b}$ for the vectors of coefficients of the basis expansions of the functions $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$.

Define $M \times M$ covariance matrices $\tilde{\mathbf{V}}_{11}$, $\tilde{\mathbf{V}}_{12}$ and $\tilde{\mathbf{V}}_{22}$ to be the matrices with (ν, ρ) entries

$$N^{-1}\sum_i c_{i\nu}c_{i\rho}, \;\; N^{-1}\sum_i c_{i\nu}d_{i\rho}, \;\text{ and }\; N^{-1}\sum_i d_{i\nu}d_{i\rho},$$

respectively, the sample variance and covariance matrices corresponding to the basis expansions of the data. It can be shown that, in the basis expansion domain, we carry out the smoothed CCA of the given data by solving the generalized eigenvalue problem

$$\begin{bmatrix} \mathbf{0} & \mathbf{J}\tilde{\mathbf{V}}_{12}\mathbf{J} \\ \mathbf{J}\tilde{\mathbf{V}}_{21}\mathbf{J} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \rho \begin{bmatrix} \mathbf{J}\tilde{\mathbf{V}}_{11}\mathbf{J} + \lambda_1\mathbf{K} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}\tilde{\mathbf{V}}_{22}\mathbf{J} + \lambda_2\mathbf{K} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}.$$

As in Chapter 14, we should choose the number of basis functions M large enough to ensure that the regularization is controlled by the choice of the smoothing parameter(s) λ rather than that of dimensionality M. Values of M of around 20 should give good results without imposing an excessive computational burden.

11.6.2 The roughness of the canonical variates

A third algorithmic possibility is related to the idea of quantifying of the roughness of a variate, as discussed in Chapter 5. Just as in the case of smoothing data, this idea is of both conceptual and algorithmic value,

and can be used to elucidate the regularization method we propose for functional canonical correlation analysis.

Suppose $z_i = \int \boldsymbol{\xi} X_i$ is a possible canonical variate value, and let $\mathbf{z}$ be the N-vector containing these values. Let $\mathbf{R}_X$ be the matrix $\mathbf{R}$ as derived in Section 15.7.3, implying that $\mathbf{z}'\mathbf{R}_X\mathbf{z}$ is the roughness of the smoothest function ξ such that $\int \xi X_i = z_i$ for all i. It may be that $\mathbf{z}'\mathbf{R}_X\mathbf{z}$ is equal to $\|D^2\boldsymbol{\xi}\|^2$, or it may be that z_i can be obtained by integrating a smoother function against the X_i. In any case, we can consider $\mathbf{z}'\mathbf{R}_X\mathbf{z}$ in its own right as a measure of the roughness of z_i as a variate based on the X_i.

Similarly, let $\mathbf{R}_Y$ be a matrix such that the roughness of any vector of canonical variate values $\mathbf{w}$ relative to the observed covariate functions $\{Y_i\}$ is $\mathbf{w}'\mathbf{R}_Y\mathbf{w}$. Our smoothed canonical correlation method can then be recast as the determination of vectors $\mathbf{z}$ and $\mathbf{w}$ to maximize the sample covariance of z_i and w_i subject to

$$\mathtt{var}\{z_i\} + \lambda_1\mathbf{z}'\mathbf{R}_X\mathbf{z} = \mathtt{var}\{w_i\} + \lambda_2\mathbf{w}'\mathbf{R}_Y\mathbf{w} = 1. \tag{11.4}$$

Once we have found in this way a pair of canonical variates, the corresponding weight functions are defined as the smoothest functions $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ satisfying $z_i = \int \boldsymbol{\xi} X_i$ and $w_i = \int \boldsymbol{\eta} Y_i$ for all i.

We can maximize the sample covariance of $\{z_i\}$ and $\{w_i\}$ subject to the constraints (11.4) by solving an eigenvalue problem. Some care is necessary to deal with a slight complication caused by the presence of the sample mean in the formula for variance and covariance.

Assuming without loss of generality that the canonical variates have sample mean zero, write the constrained maximization problem as that of finding the maximum of $\mathbf{z}'\mathbf{w}$ subject to the constraints

$$\mathbf{z}'\mathbf{z} + \lambda_1\mathbf{z}'\mathbf{R}_X\mathbf{z} = \mathbf{w}'\mathbf{w} + \lambda_2\mathbf{w}'\mathbf{R}_Y\mathbf{w} = 1 \tag{11.5}$$

and the additional constraints

$$1'\mathbf{z} = 1'\mathbf{w} = 0. \tag{11.6}$$

For the moment, neglect the constraint (11.6) and consider the maximization of $\mathbf{z}'\mathbf{w}$ subject only to the constraints (11.5). This corresponds to the eigenvalue problem

$$\begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{z} \\ \mathbf{w} \end{bmatrix} = \rho \begin{bmatrix} \mathbf{I} + \lambda_1\mathbf{R}_X & \mathbf{0} \\ \mathbf{0} & \mathbf{I} + \lambda_2\mathbf{R}_Y \end{bmatrix} \begin{bmatrix} \mathbf{z} \\ \mathbf{w} \end{bmatrix}. \tag{11.7}$$

By premultiplying (11.7) by $[\mathbf{z}' \;\; \mathbf{w}']$ and taking the product of the two expressions for $\mathbf{z}'\mathbf{w}$ thus obtained, any solution of (11.7) satisfies

$$(\mathbf{z}'\mathbf{w})^2 = \rho^2(\mathbf{z}'\mathbf{z} + \lambda_1\mathbf{z}'\mathbf{R}_X\mathbf{z})(\mathbf{w}'\mathbf{w} + \lambda_2\mathbf{w}'\mathbf{R}_Y\mathbf{w}) \geq \rho^2(\mathbf{z}'\mathbf{z})(\mathbf{w}'\mathbf{w})$$

and so it is necessarily the case that $|\rho| \leq 1$. Since the smoothest functional interpolant of the constant vector has roughness zero, $\mathbf{R}_X 1 = \mathbf{R}_Y 1 = 0$, and so the condition $\mathbf{z} = \mathbf{w} = 1$ yields the leading solution of (11.7), with eigenvalue $\rho = 1$.

The solution of (11.7) with the *second* largest eigenvalue maximizes $\mathbf{z}'\mathbf{w}$ subject to the constraint (11.5) and the additional constraint

$$1'(\mathbf{I} + \lambda_1 \mathbf{R}_X)\mathbf{z} = 1'(\mathbf{I} + \lambda_2 \mathbf{R}_Y)\mathbf{w} = 0. \tag{11.8}$$

But since $\mathbf{R}_X 1 = \mathbf{R}_Y 1 = 0$, the constraint (11.8) is precisely equivalent to the constraint (11.6) that we temporarily neglected. It follows that the second and subsequent eigensolutions of (11.7) are the canonical variates we require, and automatically have sample mean zero; the leading solution is a constant and should be ignored.

11.7 Penalized optimal scoring and discriminant analysis

Hastie, Buja and Tibshirani (1995) consider functional forms of the multivariate techniques of optimal scoring and linear discriminant analysis, making use of ideas closely related to the functional canonical correlation analysis approach discussed in this chapter. We present a brief overview of their work; see the original paper for further details.

11.7.1 The optimal scoring problem

Assume that we have N paired observations (X_i, y_i) where each X_i is a function, and each y_i is a category or class taking values in the set $\{1, 2, \ldots, J\}$. For notational convenience, we code each y_i as a J-vector $\mathbf{y}_i$ with value 1 in position j if $y_i = j$, and 0 elsewhere.

We aim to obtain a function β and a J-vector $\boldsymbol{\theta}$ minimizing the criterion

$$\texttt{OSERR}(\boldsymbol{\theta}, \beta) = N^{-1} \sum_{i=1}^{N} (\int \beta X_i - \boldsymbol{\theta}' \mathbf{y}_i)^2$$

subject to the normalization constraint $N^{-1} \sum_i (\boldsymbol{\theta}' \mathbf{y}_i)^2 = 1$. The idea is to turn the categorical variable coded by the y-vectors into a quantitative variable taking the values θ_j. The θ_j are the scores for the various categories, chosen to give the best available prediction of a linear property $\int \beta X$ of the observed functional data.

For any given $\boldsymbol{\theta}$, the problem of finding the functions β is that of finding a function which satisfies a finite number of linear constraints. Because there are infinitely many degrees of freedom in the choice of a function, it is usually possible to choose β to give perfect prediction of any specified values $\boldsymbol{\theta}' \mathbf{y}_i$. This means that we cannot choose an optimal score vector $\boldsymbol{\theta}$ uniquely on the basis of the observed data. To deal with this difficulty, Hastie et al. (1995) introduced the *penalized* optimal scoring criterion

$$\texttt{OSERR}_\lambda(\boldsymbol{\theta}, \beta) = \texttt{OSERR}(\boldsymbol{\theta}, \beta) + \lambda \times \texttt{PEN}(\beta),$$

where λ is a smoothing parameter and $\texttt{PEN}(\beta)$ a roughness penalty.

11.7.2 The discriminant problem

The discriminant problem is similar to the optimal scoring problem. Again, we have functional observations X_i, each allocated to a category in $\{1, 2, \ldots, J\}$. For any proposed linear discriminant functional $\int \beta X_i$, define θ_j to be the average of the $\int \beta X_i$ for all X_i falling in category j. For each fixed β, this value of $\boldsymbol{\theta}$ minimizes the quantity $\texttt{OSERR}(\boldsymbol{\theta}, \beta)$, which can then be re-interpreted as the *within-class variance* of the $\int \beta X_i$. The *between-class variance* is simply the variance of the discriminant class means $\boldsymbol{\theta}'\mathbf{y}_i$, defining the J-vectors $\mathbf{y}_i$ by the same coding as above. Discriminant analysis aims to maximize the between-class variance subject to a constraint on the within-class variance.

The roles of objective function and constraint are exchanged in passing from optimal scoring to discriminant analysis, and minimization is replaced by maximization. Also, primary attention shifts from the score vector $\boldsymbol{\theta}$ in optimal scoring to the discriminant functional defined by the function β in discriminant analysis. Hastie et al. make the correspondence complete by proposing *penalized discriminant analysis* where we maximize the raw between-class variance subject to a penalized constraint on the within-class variance

$$\texttt{OSERR}(\boldsymbol{\theta}, \beta) + \lambda \times \texttt{PEN}(\beta) = 1.$$

11.7.3 The relationship with CCA

Simple modifications of arguments from multivariate analysis show that the penalized optimal scoring and the penalized discriminant analysis problems are both equivalent to the mixed functional-multivariate canonical correlation analysis problem of maximizing the covariance of $\int \xi X_i$ and $\boldsymbol{\eta}'\mathbf{y}_i$ subject to the constraints

$$\texttt{var}(\int \xi X_i) + \lambda \times \texttt{PEN}(\xi) = \texttt{var}(\boldsymbol{\eta}'\mathbf{y}_i) = 1. \tag{11.9}$$

In the notation we have used for CCA, the weight corresponding to the functional part X_i of the data is itself a function ξ, whereas the vector part $\mathbf{y}_i$ is mapped to its canonical variate by a weight *vector* $\boldsymbol{\eta}$. Only the functional part ξ is penalized for roughness in the constraints (11.9). The numerical approaches we have set out for CCA carry over to this case, with appropriate modifications because only the X_i are functions.

To obtain the solutions $(\beta, \boldsymbol{\theta})$ of the discriminant and optimal scoring problems, it is only necessary to rescale the estimated function ξ and vector $\boldsymbol{\eta}$ appropriately. The subsidiary variates are also interesting for these problems because they yield estimates of vector-valued scores θ_j and discriminants $\int \beta X_i$.

11.7.4 Applications

Hastie et al. present two fascinating applications of these techniques. For speech recognition, the frequency spectra of digitized recordings of various phonemes are used as data. A roughness penalty of the form $\texttt{PEN}(\beta) = \int\{D^2\beta(\omega)\}^2 w(\omega)d\omega$ is used, with the weight function $w(\omega)$ chosen to place different emphasis on different frequencies ω.

Their other application is the recognition of digits in handwritten postal addresses and zip codes. In this case, the observations X_i are functions of a bivariate argument t, defined in practice on a 16×16 pixel grid. The roughness penalty used is a discrete version of the Laplacian penalty $\int\int[\nabla^2\beta(t)]^2\,dt$.

11.8 Further readings and notes

The idea of canonical correlation between two function spaces has a rather substantial history. Lancaster (1969) is considered an early statement of the problem, considered in the context of a treatment of the chi-squared distribution. Caillez, F. and Pagès, J. P. (1976) and Dauxois and Pousse (1976) are two explorations in French of functional canonical correlation, the first being directed to applied statisticians, and the second being a severely abstract treatise that is yet to be published in the conventional sense. A recent contribution on the theoretical side is He, Müller and Wang (2003). Dauxois and Nkiet (2002) discuss some generalizations of canonical correlation analysis within a Hilbert space framework.