Appendix

Some algebraic and functional techniques

This appendix covers various topics involving matrix decompositions, projections in vector and function spaces, and the constrained maximization of certain quadratic forms through the solution of appropriate eigenequations.

A.1 Inner products $\langle x, y \rangle$

An advance in mathematical notation occurs when we separate the name for an operation from explicit instructions on how to carry it out. Consider, for example, the operation $+$. Suppose one opens a mathematics book at a random page, and discovers the expression $x + y$. One might imagine that everyone would always mean the same by $x + y$, but a moment's thought shows that computing the sum can involve very different techniques depending on whether x and y are real numbers, complex numbers, vectors, matrices of the same dimensions or functions. What really counts is that any author who uses the symbol $+$ can be assumed to mean an operation that obeys the basic properties of addition, $x+y = y+x$ and $(x+y)+z = x+(y+z)$, and that this operation also interlocks with the multiplication operation $\times$ through $(x + y) \times z = x \times z + y \times z$ and $x \times (y + z) = x \times y + x \times z$. The author assumes that we can actually carry out the operation involved ourselves, or else in some exotic situations he or she furnishes us with detailed instructions. The notation $x+y$ allows the basic structure of addition to be assumed, almost subconsciously, leaving the

details to be supplied in any particular case if necessary. Hiding the details focusses our attention on what really matters.

A.1.1 Some specific examples

We now discuss a generic notation for inner products, extending the familiar idea of the inner product of two vectors $\mathbf{x}$ and $\mathbf{y}$. Consider the *Euclidean inner product* operation $\mathbf{x}'\mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ are vectors of the same length. The operation has the following simple properties:

Symmetry: $\mathbf{x}'\mathbf{y} = \mathbf{y}'\mathbf{x}$ for all $\mathbf{x}$ and $\mathbf{y}$,

Positivity: $\mathbf{x}'\mathbf{x} \geq 0$ for all $\mathbf{x}$, with $\mathbf{x}'\mathbf{x} = 0$ if and only if $\mathbf{x} = 0$, and

Bilinearity: for all real numbers a and b, $(a\mathbf{x} + b\mathbf{y})'\mathbf{z} = a\mathbf{x}'\mathbf{z} + b\mathbf{y}'\mathbf{z}$ for all vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$.

Of course, these properties follow from the instructions implied in the definition

$$\mathbf{x}'\mathbf{y} = \sum_i x_i y_i. \tag{A.1}$$

However, it is important to note that the Euclidean inner product operation, and the instructions defining it, are of critical importance in multivariate data analysis *because* of the properties of symmetry, positivity and bilinearity, which can therefore be considered of more fundamental significance than the definition (A.1) itself.

This basic role of symmetry, positivity and bilinearity is further emphasized when we realize that $\mathbf{x}'\mathbf{W}\mathbf{y}$, where $\mathbf{W}$ is a positive definite matrix of appropriate order, also has these properties and, indeed, can be used almost anywhere that we use $\mathbf{x}'\mathbf{y}$. So, for example, we use $\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{y}$, where $\mathbf{\Sigma}$ is a population covariance matrix, to define the multivariate normal distribution, to compute Mahalanobis distances, to define generalized least squares estimates instead of ordinary least squares, and many other useful things.

Now suppose that x and y are not vectors, but rather functions with values $x(t)$. The natural functional counterpart to $x'y$ is $\int x(t)y(t)\,dt$, replacing the sum in (A.1) by an integral. Again we have an operation on two functions x and y that is denoted by presenting the instructions for computing its value, but we know that this, too, is symmetric in x and y, linear in either function, and satisfies the positivity requirement. The same conclusions can be drawn for the operation $\int \omega(t)x(t)y(t)\,dt$, where ω is a strictly positive weight function, and indeed for the more general operation $\int\int \omega(s,t)x(s)y(t)\,ds\,dt$ if ω is strictly positive-definite, which simply means that the positivity requirement for the inner product is satisfied.

It should by now be clear that we can achieve a great leap forward in generality by using a common notation for these various real-valued

operations that is understood to imply symmetry, positivity and bilinearity, without bothering with the details of the computation. We call such an operation an *inner product*, and we use the generic notation $\langle x, y\rangle$ for the inner product of x and y. The fundamental properties of an inner product are:

Symmetry: $\langle x, y\rangle = \langle y, x\rangle$ for all x and y;

Positivity: $\langle x, x\rangle \geq 0$ for all x, with $\langle x, x\rangle = 0$ if and only if $x = 0$;

Bilinearity: for all real numbers a and b, $\langle ax + by, z\rangle = a\langle x, z\rangle + b\langle y, z\rangle$ for all vectors x, y and z.

Note that bilinearity in the second argument follows from symmetry and bilinearity in the first.

A.1.2 General properties: association, size, angle, distance

We can think of the inner product as defining a scalar measure of *association* between pairs of quantities x and y. The symmetric nature of the measure means that it is, as we would usually require, invariant with respect to the order of the quantities, and its bilinearity means that changing the scale of either argument and/or using a sum as either argument leaves the measure unchanged in its essential properties.

Positivity means that the inner product of any x with itself is essentially a measure of its *size*. The positive square root of this size measure is called the *norm* of x, written $\|x\|$, so that

$$\|x\|^2 = \langle x, x\rangle \tag{A.2}$$

with $\|x\| \geq 0$. In the special case where x is an n-vector, and the inner product is the Euclidean inner product (A.1), the norm of x is simply the length of the vector measured in n-dimensional space. In the case of a function f, a basic type of norm is $\|f\| = \sqrt{\int f^2}$, and is called its $\mathcal{L}^2$ norm.

Whatever inner product is used, the standard properties of inner products lead to the following properties of the norm:

1. $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$
2. $\|ax\| = |a|\|x\|$ for all real numbers a
3. $\|x + y\| \leq \|x\| + \|y\|$.

From the properties of the inner product also follows the *Cauchy-Schwarz inequality*:

$$|\langle x, y\rangle| \leq \|x\|\|y\| = \sqrt{\langle x, x\rangle\langle y, y\rangle}.$$

This inequality links the inner product with the derived size measure or norm, and also leads to the *cosine inequality*:

$$-1 \leq \langle x, y\rangle/(\|x\|\|y\|) \leq 1.$$

The cosine inequality links the inner product to the geometrical concept of *angle*; the angle between x and y can be defined to be the angle θ such that

$$\cos\theta = \frac{\langle x, y \rangle}{\|x\|\|y\|}.$$

Where x and y are n-vectors and the inner product is Euclidean inner product, θ is the angle between x and y in the usual geometric sense. Similarly, the cosine of the angle between two functions f and g can be defined as $\int fg / \sqrt{(\int f^2)(\int g^2)}$. The use of the cosine inequality to justify the idea of the angle between two vectors or functions further illuminates the notion that $\langle x, y \rangle$ is a association measure. Once we have obtained a scale invariant coefficient by dividing by $\|x\|\|y\|$, we have a useful index of the extent to which x and y are measuring the same thing.

The particular relation $\langle x, y \rangle = 0$, called *orthogonality*, implies that x and y can be considered as being at right angles to one another. Because of bilinearity, orthogonality remains unchanged under any rescaling of either quantity. Orthogonality plays a key role in the operation of *projection* that is discussed in Section A.2.1.

From the inner product, we also derive a measure of *distance* between x and y

$$d_{xy} = \|x - y\| = \sqrt{\langle x - y, x - y \rangle}$$

that has extremely wide applications; again, in the Euclidean case, distance corresponds to the usual geometric definition.

Thus, the simple algebraic properties of symmetry, positivity and bilinearity of the inner product lead easily to very useful definitions of the size of a quantity x, and of the angle and distance between x and y. We can be confident that, no matter how we define $\langle x, y \rangle$ in a particular application, the essential characteristics of these three measures remain unchanged.

The nature of the inner product depends on something more fundamental about x and y: They are elements of a *vector space* in which elements can be added, can be multiplied by real numbers to yield new vectors, and in which addition distributes with respect to scalar multiplication. The ensemble of a vector space and an associated inner product is called an *inner product* space.

Finally, of the three properties, only symmetry and bilinearity are really crucial. We can often get by with relaxing positivity to the weaker condition that $\langle x, x \rangle \geq 0$, so that $\langle x, x \rangle$ may be zero for some x's that are not themselves zero. Then the inner product is called a *semi-inner product* and the norm a *seminorm*. Most properties of inner products remain true for semi-inner products.

A.1.3 Descriptive statistics in inner product notation

As an example of how inner products can work for us, we consider how standard descriptive statistics can be expressed in inner product notation. Consider the space of possible univariate samples $x = (x_1, \ldots, x_N)$ of size N. Define the inner product to be the Euclidean inner product

$$\langle x, y\rangle = \sum_i x_i y_i = x'y.$$

Let 1 indicate the vector of size N all of whose elements are unity. Then some familiar univariate descriptive statistics become

Mean: $\bar{x} = N^{-1}\langle x, 1\rangle$. Note that $\bar{x}$, being a multiple of an inner product, is a scalar and not a vector. The vector of length N all of whose elements are $\bar{x}$ is $\bar{x}1$.

Variance: $s_x^2 = N^{-1}\langle x - \bar{x}1, x - \bar{x}1\rangle = N^{-1}\|x - \bar{x}1\|^2$

Covariance: $s_{xy} = N^{-1}\langle x - \bar{x}1, y - \bar{y}1\rangle$

Correlation: $r_{xy} = s_{xy}/(s_x s_y)$.

It is easy to show that the covariance s_{xy} is itself a semi-inner product between x and y. It is then an immediate consequence of the cosine inequality that the correlation coefficient satisfies the well-known *correlation inequality*

$$-1 \leq r_{xy} \leq 1.$$

Now suppose that we stop using the Euclidean inner product, but instead go for

$$\langle x, y\rangle = \sum_i w_i x_i y_i,$$

where w_i is a nonnegative weight to be applied to observation i. What difference would this make? None at all, except of course we must now divide by the constant $\sum_i w_i$ instead of N in defining $\bar{x}, s_x^2$, and s_{xy}. The essential characteristics of these statistics depend on the characteristics of the inner product, and not on precisely how the inner product is actually calculated. Of course, the precise weighting affects the values of the statistics, but the essential meanings of the various descriptive statistics, for example as measures of location, scale and dependence remain basically unchanged.

We can generalize this idea further: Suppose that the sequence of observations is known to be correlated, with covariance matrix $\mathbf{\Sigma}$. Then we can use $\langle x, y\rangle = x'\mathbf{\Sigma}^{-1}y$ to provide a basis for descriptive statistics that compensate for the known covariance structure on the observations.

Now consider these same statistics in the context of x being a function with values $x(t)$, where argument t takes values within some real interval

such as $[0, T]$. Thus the index i taking N possible values has been replaced by the index t taking an infinity of values. Define the inner product as

$$\langle x, y\rangle = \int_0^T x(t)y(t)\,dt,$$

where we assume that the functions are sufficiently well behaved that the integral is always defined and finite. Then the various descriptive statistics continue to be defined as above, except that we divide by $\int_0^T dt = T$ instead of N and the vector 1 is replaced by the function $1 = 1(t)$ which takes the value of unity for all t. In the functional case, $\bar{x}$ becomes the mean level of the function x, s_x^2 becomes a measure of its variation about its mean level, and s_{xy} and r_{xy} measure the correspondence between the variation of x and y. Moving to

$$\langle x, y\rangle = \int_0^T \omega(t)x(t)y(t)\,dt,$$

for some positive weight function ω, and dividing by $\int \omega(t)\,dt$ really wouldn't change these interpretations in any essential way, except that different parts of the range of t would be regarded as being of different importance.

Finally, we note that even the divisors in these statistics can be defined in inner product terms, meaning that our fundamental descriptive statistics can be written in the unifying form

$$\begin{aligned} \bar{x} &= \langle x, 1\rangle/\|1\|^2 \\ s_x^2 &= \|x - \bar{x}1\|^2/\|1\|^2 \\ s_{xy} &= \langle x - \bar{x}1, y - \bar{y}1\rangle/\|1\|^2. \end{aligned}$$

A.1.4 Some extended uses of inner product notation

In this book, we take the somewhat unorthodox step of using inner product notation to refer to certain *linear operations* that, strictly speaking, do not fall within the rubric of inner products.

So far in our discussion, the result of an inner product has always been a single real number. One way in which we extend our notation is the following. Let $x = (x_1, \ldots, x_m)'$ be a vector of length m, each element of which is an element of some vector space, whether finite dimensional or functional. Then the notation $\langle x, y\rangle$, where y is a single element of the same space, indicates the m-vector whose elements are $\langle x_1, y\rangle, \ldots, \langle x_m, y\rangle$. Furthermore, if y is similarly a vector of length, say, n, then the notation $\langle x, y'\rangle$ defines the *matrix* with m rows and n columns containing the values $\langle x_i, y_j\rangle, i = 1, \ldots, m; j = 1, \ldots, n$. We only use this convention in situations where the context should make clear whether x and/or y are vectors of elements of the space in question.

In the functional context, we sometimes write

$$\langle z, \beta \rangle = \int z(s)\beta(s)\, ds$$

even when the functions z and β are not in the same space. We hope that the context of this use of inner product notation will make clear that a true inner product is not involved in this case. The alternative would have been the use of different notation such as (z, β), but we considered that the possibilities of confusion justified avoiding this convention.

An important property is that $\langle z, \beta \rangle$ is always a *linear operator* when regarded as a function of either of its arguments; generally speaking a linear operator on a function space is a mapping A such that, for all f_1 and f_2 in the space, and for all scalars a_1 and a_2, we have $A(a_1 f_1 + a_2 f_2) = a_1 A f_1 + a_2 A f_2$.

A.2 Further aspects of inner product spaces

We briefly review two further aspects of inner product spaces that are useful in our later development.

A.2.1 Projections

Let $u_1, \ldots, u_n$ be any n elements of our space, and let $\mathcal{U}$ be the subspace consisting of all possible linear combinations of the u_i. We can characterize the subspace $\mathcal{U}$ by using suitable vector notation. Let u be the n-vector whose elements are the $u_1, \ldots, u_n$. Then every member of $\mathcal{U}$ is of the form $u'c$ for some real n-vector c.

Associated with the subspace $\mathcal{U}$ is the *orthogonal projection onto* $\mathcal{U}$, which is defined to be a linear operator P with the following properties:

1. For all z, the element Pz falls in $\mathcal{U}$, and so is a linear combination of the functions $u_1, \ldots, u_n$.

2. If y is in $\mathcal{U}$ already, then $Py = y$.

3. For all z, the *residual* $z - Pz$ is orthogonal to all elements v of $\mathcal{U}$.

From the first two of these properties, it follows at once that $PP = P^2 = P$. From the third property, it is easy to show that the operator P maps each element z to its *nearest* point in $\mathcal{U}$, distance being measured in terms of the norm. This makes projections very important in statistical contexts such as least squares estimation.

A.2.2 Quadratic optimization

Some of our functional data analysis methodology require the solution of a particular kind of constrained optimization problem. Suppose that A is a linear operator on a function space satisfying the condition

$$\langle x, Ay\rangle = \langle Ax, y\rangle \text{ for all } x \text{ and } y.$$

Such an operator is called a *self-adjoint* operator.

Now consider the problem of maximizing $\langle x, Ax\rangle$ subject to the constraint $\|x\| = 1$. In Section A.5, we set out results relating this optimization problem to the eigenfunction/eigenvalue problem $Au = \lambda u$. We then go on to consider the more general problem of maximizing $\langle x, Ax\rangle$ subject to a constraint on $\langle x, Bx\rangle$ for a second self-adjoint operator B.

A.3 Matrix decompositions and generalized inverses

We describe two important matrix decompositions, the singular value decomposition and the QR decomposition. Both of these are standard techniques in numerical linear algebra, and can be carried out within packages such as S-PLUS and MATLAB®. We do not give any details of the way the decompositions are computed; for these see, for example, Golub and Van Loan (1989) or the standard numerical linear algebra package LINPACK (Dongarra et al., 1979).

A.3.1 Singular value decompositions

Suppose $\mathbf{Z}$ is an $m \times n$ matrix. For many purposes it is useful to carry out a *singular value decomposition* (SVD) of $\mathbf{Z}$. This expresses $\mathbf{Z}$ as the product of three matrices

$$\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{V}' \tag{A.3}$$

where, for some integer $q \leq \min(m, n)$,

- $\mathbf{U}$ is $m \times q$ and $\mathbf{U}'\mathbf{U} = \mathbf{I}_q$, where $\mathbf{I}_q$ is the identity matrix of order q;
- $\mathbf{D}$ is a $q \times q$ diagonal matrix with strictly positive elements on the diagonal;
- $\mathbf{V}$ is $n \times q$ and $\mathbf{V}'\mathbf{V} = \mathbf{I}_q$.

The diagonal elements $d_1, d_2, \ldots, d_q$ of $\mathbf{D}$ are called the *singular values* of $\mathbf{Z}$, and the SVD can always be carried out in such a way that the diagonal elements $d_1, d_2, \ldots, d_q$ satisfy

$$d_1 \geq d_2 \geq \ldots \geq d_q > 0. \tag{A.4}$$

In this case, the number q is equal to the rank of the matrix $\mathbf{Z}$, i.e., the maximum number of linearly independent rows or columns of $\mathbf{Z}$.

In the special case where $\mathbf{Z}$ is square and symmetric, the requirement that the diagonal elements of $\mathbf{D}$ are necessarily positive is usually dropped, but the matrices $\mathbf{U}$ and $\mathbf{V}$ are chosen to be identical. Furthermore we may allow $q \geq$ rank $\mathbf{Z}$. The d_i then include all the nonzero eigenvalues of $\mathbf{Z}$, together with some or all of the zero eigenvalues if there are any. We have

$$\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{U}' \text{ with } \mathbf{U}'\mathbf{U} = \mathbf{I}. \tag{A.5}$$

If, in addition, $\mathbf{Z}$ is positive semi-definite, so that $x'\mathbf{Z}x \geq 0$ for all vectors x, then

$$d_1 \geq d_2 \geq \ldots \geq d_q \geq 0. \tag{A.6}$$

A.3.2 Generalized inverses

Given any $m \times n$ matrix $\mathbf{Z}$, we can define a *generalized inverse* or *g-inverse* of $\mathbf{Z}$ to be any $n \times m$ matrix $\mathbf{Z}^-$ such that

$$\mathbf{Z}\mathbf{Z}^-\mathbf{Z} = \mathbf{Z}. \tag{A.7}$$

If $m = n$ and $\mathbf{Z}$ is an invertible matrix, then it follows from (A.7) that $\mathbf{Z}^{-1}$ is a g-inverse of $\mathbf{Z}$. Furthermore, by pre and post multiplying (A.7) by $\mathbf{Z}^{-1}$, we see that $\mathbf{Z}^{-1}$ is the *unique* g-inverse of $\mathbf{Z}$ in this case.

In the more general case, the matrix $\mathbf{Z}^-$ is not generally unique, but a particular g-inverse, called the *Moore-Penrose g-inverse* $\mathbf{Z}^+$ can always be calculated using the singular value decomposition (A.3) of the matrix $\mathbf{Z}$. Set

$$\mathbf{Z}^+ = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}'. \tag{A.8}$$

It is easy to check that $\mathbf{Z}^+$ is a g-inverse of $\mathbf{Z}$ and also that

$$\mathbf{Z}^+\mathbf{Z}\mathbf{Z}^+ = \mathbf{Z}^+ \text{ and } \mathbf{Z}\mathbf{Z}^+ = \mathbf{U}\mathbf{U}'. \tag{A.9}$$

A.3.3 The QR decomposition

The *QR decomposition* of an $m \times n$ matrix $\mathbf{Z}$ is a different decomposition that yields the expression

$$\mathbf{Z} = \mathbf{Q}\mathbf{R},$$

where $\mathbf{Q}$ is an $m \times m$ orthogonal matrix (so that $\mathbf{Q}'\mathbf{Q} = \mathbf{Q}\mathbf{Q}' = \mathbf{I}$) and $\mathbf{R}$ is an $m \times n$ upper-triangular matrix (so that $\mathbf{R}_{ij} = 0$ if $i > j$).

If $m > n$ then the last $(m - n)$ rows of $\mathbf{R}$ will be zero, and each of the last $(m - n)$ columns x of $\mathbf{Q}$ will satisfy $x'\mathbf{Z} = 0$. Dropping these rows and columns will yield a *restricted QR decomposition* $\mathbf{Z} = \mathbf{Q}_1\mathbf{R}_1$ where $\mathbf{R}_1$ is an $n \times n$ upper-triangular matrix and $\mathbf{Q}$ is an $m \times n$ matrix of orthonormal columns.

A.4 Projections

In discussing the key concept of *projection*, we first consider projection matrices in m-dimensional spaces, and then go on to consider more general inner product spaces.

A.4.1 Projection matrices

Suppose that an $m \times m$ matrix $\mathbf{P}$ has the property that $\mathbf{P}^2 = \mathbf{P}$. Define $\mathcal{P}$ to be the subspace of R^m spanned by the columns of $\mathbf{P}$. The matrix $\mathbf{P}$ is then called a *projection matrix* onto the subspace $\mathcal{P}$. The following two properties, which are easily checked, give the reason for this definition:

1. Every m-vector $\mathbf{z}$ is mapped by $\mathbf{P}$ into the subspace $\mathcal{P}$.

2. If $\mathbf{z}$ is already a linear combination of columns of $\mathbf{P}$, so that $\mathbf{z} = \mathbf{P}\mathbf{u}$ for some vector $\mathbf{u}$, then $\mathbf{P}\mathbf{z} = \mathbf{z}$.

If $\mathbf{P}$ is a symmetric matrix, then $\mathbf{P}$ is called an *orthogonal* projection matrix, and will have several nice properties. For example, for any vector $\mathbf{z}$ we have

$$(\mathbf{P}\mathbf{z})'\{(\mathbf{I}-\mathbf{P})\mathbf{z}\} = \mathbf{z}'\mathbf{P}'(\mathbf{I}-\mathbf{P})\mathbf{z} = \mathbf{z}'(\mathbf{P}\mathbf{z}-\mathbf{P}^2\mathbf{z}) = 0.$$

This means that the projected vector $\mathbf{P}\mathbf{z}$ and the 'residual vector' $\mathbf{z}-\mathbf{P}\mathbf{z}$ are orthogonal to one another, in the usual Euclidean sense. Furthermore, suppose $\mathbf{v}$ is any vector in $\mathcal{P}$. Then, by a very similar argument,

$$\mathbf{v}'(\mathbf{z}-\mathbf{P}\mathbf{z}) = (\mathbf{P}\mathbf{v})'(\mathbf{I}-\mathbf{P})\mathbf{z} = \mathbf{v}'\mathbf{P}(\mathbf{I}-\mathbf{P})\mathbf{z} = 0,$$

so that the residual vector is orthogonal to all vectors in $\mathcal{P}$.

Suppose that $\mathbf{w}$ is any vector in $\mathcal{P}$ other than $\mathbf{P}\mathbf{z}$. Then $\mathbf{w}-\mathbf{P}\mathbf{z}$ is also in $\mathcal{P}$ and therefore is orthogonal to $\mathbf{z}-\mathbf{P}\mathbf{z}$. Defining $\langle x, y\rangle = x'y$ and $\|x\|$ to be the usual Euclidean inner product and norm, we then have

$$\begin{aligned} \|\mathbf{z}-\mathbf{w}\|^2 &= \|\mathbf{z}-\mathbf{P}\mathbf{z}\|^2 + \|\mathbf{P}\mathbf{z}-\mathbf{w}\|^2 + 2\langle \mathbf{z}-\mathbf{P}\mathbf{z}, \mathbf{P}\mathbf{z}-\mathbf{w}\rangle \\ &= \|\mathbf{z}-\mathbf{P}\mathbf{z}\|^2 + \|\mathbf{P}\mathbf{z}-\mathbf{w}\|^2 > \|\mathbf{z}-\mathbf{P}\mathbf{z}|^2. \end{aligned} \tag{A.10}$$

This means that $\mathbf{P}\mathbf{z}$ is the closest point to $\mathbf{z}$ in the subspace $\mathcal{P}$. Thus orthogonal projections onto a subspace have the property of mapping each vector to the nearest point in the subspace.

More generally, if the inner product is $\langle \mathbf{x}, \mathbf{y}\rangle = \mathbf{x}'\mathbf{W}\mathbf{y}$, and if $\mathbf{P}$ is a projection onto the space $\mathcal{P}$ such that $\mathbf{W}\mathbf{P}$ is symmetric, then $\mathbf{P}$ is orthogonal with respect to this inner product, meaning that $\langle \mathbf{P}\mathbf{z}, \mathbf{z}-\mathbf{P}\mathbf{z}\rangle = 0$ and $\langle \mathbf{v}, \mathbf{z}-\mathbf{P}\mathbf{z}\rangle = 0$ for all $\mathbf{v}$ in $\mathcal{P}$.

A.4.2 Finding an appropriate projection matrix

Now suppose we are not given a projection matrix, but instead we are given a subspace $\mathcal{U}$ of R^m, and we wish to find an orthogonal projection matrix $\mathbf{P}$ that projects onto $\mathcal{U}$.

Let $\mathbf{Z}$ be any matrix whose columns are m-vectors that span the subspace $\mathcal{U}$. There is no need for the columns to be linearly independent. Define $\mathbf{P}$ by

$$\mathbf{P} = \mathbf{Z}\mathbf{Z}^{-}.$$

It is straightforward to show that $\mathbf{P}$ is a projection onto the subspace $\mathcal{U}$ as required.

In order to get an *orthogonal* projection, define $\mathbf{P}$ using the Moore-Penrose g-inverse $\mathbf{Z}^{+}$. Then, in terms of the SVD of $\mathbf{Z}$, we have $\mathbf{P} = \mathbf{U}\mathbf{U}'$, so that $\mathbf{P}$ is a symmetric matrix and hence an orthogonal projection.

A.4.3 Projections in more general inner product spaces

We can extend these ideas to projections in more general inner product spaces as discussed in Section A.2.1. As in that section, let $u_1, \ldots, u_n$ be any n elements of our space, and let u be the n-vector whose elements are the $u_1, \ldots, u_n$. Let $\mathcal{U}$ be the subspace consisting of all possible linear combinations $c'u$ for real n-vectors c. Suppose that P is an orthogonal projection onto $\mathcal{U}$ as specified in Section A.2.1. The proof that P maps each element z to the nearest member $P\mathbf{z}$ of $\mathcal{U}$ is identical to the argument given in (A.10) because that depends only on the defining properties of an inner product and associated norm.

How are we to find an orthogonal projection of this kind? Extend our notation to define $\mathbf{K} = \langle u, u' \rangle$ to be the symmetric $n \times n$ matrix with elements $\langle u_i, u_j \rangle$. Given any real n-vector x, we have $x'\mathbf{K}x = \langle x'u, u'x \rangle = \|x'u\|^2 \geq 0$, so the matrix $\mathbf{K}$ is positive semi-definite.

Define the operator P by

$$Pz = u'\mathbf{K}^{+}\langle u, z \rangle$$

for all z. By definition Pz is a linear combination of the elements of u and hence is in $\mathcal{P}$. We shall show that P is an orthogonal projection onto $\mathcal{P}$.

If y is in $\mathcal{P}$, then $y = u'c$ for some real vector c, so that $Py = u'\mathbf{K}^{+}\mathbf{K}c$, and $y - Py = u'd$ where $d = (\mathbf{I} - \mathbf{K}^{+}\mathbf{K})c$. Therefore, since $\mathbf{K}\mathbf{K}^{+}\mathbf{K} = \mathbf{K}$,

$$\|y - Py\|^2 = d'\mathbf{K}d = d'(\mathbf{K} - \mathbf{K}\mathbf{K}^{+}\mathbf{K})c = 0,$$

implying that $\|y - Py\|^2 = 0$ and $Py = y$.

Finally, given any v in $\mathcal{P}$, and any z, use the property (A.9) to show that

$$\begin{aligned} \langle Pz - v, Pz \rangle &= \langle P(z - v), Pz \rangle = \langle z - v, u' \rangle \mathbf{K}^{+}\mathbf{K}\mathbf{K}^{+}\langle u, z \rangle \\ &= \langle z - v, u' \rangle \mathbf{K}^{+}\langle u, z \rangle = \langle Pz - v, z \rangle \end{aligned}$$

and therefore that $\langle Pz - v, z - Pz\rangle = 0$, completing the proof that P is the required orthogonal projection onto $\mathcal{P}$.

A.5 Constrained maximization of a quadratic function

A.5.1 The finite-dimensional case

Suppose that $\mathbf{A}$ is a symmetric $p \times p$ matrix. An important result in linear algebra concerns the constrained maximization problem

$$\max \mathbf{x}'\mathbf{A}\mathbf{x} \text{ for } p\text{-vectors } \mathbf{x} \text{ subject to } \mathbf{x}'\mathbf{x} = 1. \tag{A.11}$$

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ be the eigenvalues of $\mathbf{A}$, and let $\mathbf{u}_i$ be the corresponding eigenvectors, each normalized to have $\|\mathbf{u}_i\| = 1$. Let $\mathbf{U}$ be the matrix whose columns are the eigenvectors $\mathbf{u}_i$ and $\mathbf{D}$ be the diagonal matrix with diagonal elements λ_i. We then have $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}'$, and $\mathbf{U}\mathbf{U}' = \mathbf{U}'\mathbf{U} = \mathbf{I}$.

Set $\mathbf{y} = \mathbf{U}'\mathbf{x}$ in (A.11), so that $\mathbf{x} = \mathbf{U}\mathbf{y}$. We have $\mathbf{x}'\mathbf{x} = \mathbf{y}'\mathbf{U}'\mathbf{U}\mathbf{y} = \mathbf{y}'\mathbf{y}$, so the constraint $\mathbf{x}'\mathbf{x} = 1$ is equivalent to $\mathbf{y}'\mathbf{y} = 1$. Therefore, in terms of $\mathbf{y}$, the maximization problem (A.11) can be rewritten as

$$\max \mathbf{y}'\mathbf{D}\mathbf{y} \text{ for } p\text{-vectors } \mathbf{y} \text{ subject to } \mathbf{y}'\mathbf{y} = 1. \tag{A.12}$$

This is clearly solved by setting $\mathbf{y}$ to be the vector $(1, 0, \ldots, 0)'$, so that $\mathbf{x}$ is the first column of $\mathbf{U}$, in other words the leading normalized eigenvector $\mathbf{u}_1$ of $\mathbf{A}$.

By an extension of this argument, we can characterize all the eigenvectors of $\mathbf{A}$ as solutions of successive optimization problems. The jth normalized eigenvector $\mathbf{u}_j$ solves the problem (A.11) subject to the additional constraint of being orthogonal to all the solutions found so far:

$$\max \mathbf{x}'\mathbf{A}\mathbf{x} \text{ subject to } \mathbf{x}'\mathbf{x} = 1 \text{ and } \mathbf{x}'\mathbf{u}_1 = \mathbf{x}'\mathbf{u}_2 = \ldots = \mathbf{x}'\mathbf{u}_{j-1} = 0. \tag{A.13}$$

Setting $\mathbf{x} = \mathbf{u}_j$, we have $\mathbf{x}'\mathbf{A}\mathbf{x} = \lambda_j \mathbf{u}_j'\mathbf{u}_j = \lambda_j$, the jth eigenvalue.

A.5.2 The problem in a more general space

Now suppose we are working within a more general inner product space. The role of a symmetric matrix is now played by a self-adjoint linear operator A, that is, one satisfying the condition

$$\langle x, Ay\rangle = \langle Ax, y\rangle \text{ for all } x \text{ and } y.$$

We shall assume that A is a completely continuous (or compact) symmetric transformation on a Hilbert space; there is no need at all for the reader to understand what this means, but anyone interested is referred to Aubin (2000) or any other standard text on functional analysis. The reader can

always take it on trust that the assumptions are satisfied when we appeal to the results of this section.

The problem

$$\max\langle x, Ax\rangle \text{ subject to } \|x\| = 1 \tag{A.14}$$

corresponds to the maximization problem (A.11), and we can define a sequence u_j as the solutions to the succession of optimization problems

$$\max\langle x, Ax\rangle \text{ subject to } \|x\| = 1 \text{ and } \langle x, u_i\rangle = 0 \text{ for } i < j. \tag{A.15}$$

Under the conditions referred to above, these optimization problems can be solved by considering the eigenfunction problem

$$Au = \lambda u$$

and normalizing the eigenfunctions u to satisfy $\|u\| = 1$. Suppose the eigenvalues are $\lambda_1 \geq \lambda_2 \geq \ldots$ with eigenfunctions $u_1, u_2, \ldots$. Then the leading eigenfunction u_1 solves the optimization problem (A.14) and the value of the maximum is λ_1. The successive eigenfunctions u_j solve the constrained problem (A.15), and the maximum at the jth stage is $\langle u_j, Au_j\rangle = \lambda_j \|u_j\|^2 = \lambda_j$.

A.5.3 Generalized eigenproblems

We sometimes wish to modify the optimization problems we have considered by the introduction of a positive definite symmetric matrix $\mathbf{B}$ into the constraints, replacing the constraint $\|\mathbf{x}\| = 1$ by $\mathbf{x}'\mathbf{B}\mathbf{x} = 1$ or, more generally, $\langle \mathbf{x}, \mathbf{B}\mathbf{x}\rangle = 1$, and similarly defining orthogonality with respect to the matrix $\mathbf{B}$.

Consider the solutions of the *generalized eigenproblem*

$$Av = \rho Bv,$$

where v is either a function or a vector, and A and B are corresponding linear operators acting on V. We normalize the solutions to satisfy $\langle v, Bv\rangle = 1$. Suppose the solutions are $v_1, v_2, \ldots$, with corresponding generalized eigenvalues $\rho_1 \geq \rho_2 \geq \ldots$. Under suitable conditions, which are always satisfied in the finite-dimensional case, and are analogous to those noted above for more general spaces, the leading generalized eigenvector or function v_1 solves the problem

$$\max\langle v, Av\rangle \text{ subject to } \langle v, Bv\rangle = 1, \tag{A.16}$$

and the maximizing value is equal to ρ_1. The jth generalized eigenvector or function v_j solves the problem

$$\max\langle v, Av\rangle \text{ subject to } \langle v, Bv\rangle = 1 \text{ and } \langle v, Bv_i\rangle = 0 \text{ for } i < j$$

and the maximizing value is ρ_j.

Finally, we note that the problem of maximizing the ratio

$$\frac{\langle v, Av\rangle}{\langle v, Bv\rangle} \tag{A.17}$$

for $v \neq 0$ is equivalent to that of maximizing $\langle v, Av\rangle$ subject to the constraint $\langle v, Bv\rangle = 1$. To see this, note that scaling any v to satisfy the constraint does not affect the value of the ratio (A.17), and so the maximum of the ratio is unaffected by the imposition of the constraint. Once the constraint is imposed, the denominator of (A.17) is equal to 1, and so maximizing the ratio subject to $\langle v, Bv\rangle = 1$ is exactly the same as the original maximization problem (A.16).

A.6 Kronecker Products

Let $\mathbf{A}$ be an m by n matrix and let $\mathbf{B}$ be a p by q matrix. The Kronecker product $\mathbf{A}\otimes\mathbf{B}$ is the super or composite matrix of order mp by nq consisting of sub-matrices $a_{ij}\mathbf{B}$. That is,

$$\mathbf{A}\otimes\mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2n}\mathbf{B} \\ \vdots & \vdots & & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix}.$$

One of the most common applications of the Kronecker product is to express a linear equation of the form

$$\mathbf{AXB}' = \mathbf{C},$$

which cannot be solved for $\mathbf{X}$ by conventional matrix algebra, in the form

$$(\mathbf{B}\otimes\mathbf{A})\text{vec}\,(\mathbf{X}) = \text{vec}\,(\mathbf{C}),$$

where vec $(\mathbf{X})$ indicates the vector of length nq obtained by writing matrix $\mathbf{X}$ as a vector column-wise, and, in the same way, vec $(\mathbf{C})$ indicates the vector of length mp obtained by writing matrix $\mathbf{C}$ as a vector column-wise. Then we can express the solution directly as

$$\text{vec}\,(\mathbf{X}) = (\mathbf{B}\otimes\mathbf{A})^{-1}\text{vec}\,(\mathbf{C}),$$

provided that, of course, matrix $\mathbf{B}\otimes\mathbf{A}$ is nonsingular.

The Kronecker product is *bilinear* in the sense that

$$\text{vec}\,(\mathbf{A}_1\mathbf{XB}_1' + \mathbf{A}_2\mathbf{XB}_2') = (\mathbf{B}_1\otimes\mathbf{A}_1 + \mathbf{B}_2\otimes\mathbf{A}_2)\text{vec}\,(\mathbf{X}).$$

Other useful relations for simplifying expressions involving Kronecker products are

$$\mathbf{A}\otimes(\mathbf{B}\otimes\mathbf{C}) \quad = \quad (\mathbf{A}\otimes\mathbf{B})\otimes\mathbf{C}$$

$$\begin{aligned}
(\mathbf{A} \otimes \mathbf{B})' &= \mathbf{A}' \otimes \mathbf{B}' \\
(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) &= (\mathbf{AC}) \otimes (\mathbf{BD}) \\
(\mathbf{A} + \mathbf{B}) \otimes \mathbf{C} &= (\mathbf{A} \otimes \mathbf{C}) + (\mathbf{B} \otimes \mathbf{C}) \\
\mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) &= (\mathbf{A} \otimes \mathbf{B}) + (\mathbf{A} \otimes \mathbf{C}) \\
\operatorname{trace}(\mathbf{A} \otimes \mathbf{B}) &= (\operatorname{trace}\mathbf{A})(\operatorname{trace}\mathbf{B}),
\end{aligned}$$

Finally, if both $\mathbf{A}$ and $\mathbf{B}$ are nonsingular, then

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}.$$

A.7 The multivariate linear model

We now return to a more statistical topic. A review of the multivariate linear model may be helpful, both to fix ideas and notation, and because some of the essential concepts transfer without much more than a change of notation to functional contexts. But a slight change of perspective is helpful on what the design matrix means. Moreover, a notion that is used repeatedly for functional data is *regularization*, and we introduce regularization in Section A.8 within the multivariate context.

A.7.1 Linear models from a transformation perspective

Let $\mathbf{Y}$ be a $N \times p$ matrix of dependent variable observations, $\mathbf{Z}$ be a $N \times q$ matrix, and $\mathbf{B}$ be a $q \times p$ matrix. In classical terminology, $\mathbf{Z}$ is the *design matrix* and $\mathbf{B}$ is a matrix of *parameters*.

The multivariate linear model is

$$\mathbf{Y} = \mathbf{ZB} + \mathbf{E}. \tag{A.18}$$

The rows of the disturbance or residual matrix $\mathbf{E}$ are often thought of, at least at the population level, as independent samples from a common population of p-variate observations with mean 0 and finite covariance matrix $\mathbf{\Sigma}$.

Although in many contexts it is appropriate to think of the columns of $\mathbf{Z}$ as corresponding to variables, it is better for our purposes to take the more general view that $\mathbf{Z}$ represents a linear transformation that maps matrices $\mathbf{B}$ into matrices with the dimensions of $\mathbf{Y}$. This can be indicated by the notation

$$\mathbf{Z} : R^{q \times p} \to R^{N \times p}.$$

The space of all possible transformed values $\mathbf{ZB}$ then defines a subspace of $R^{N \times p}$, denoted by $R(\mathbf{Z})$, and is called the *range space* of $\mathbf{Z}$.

A.7.2 The least squares solution for $\mathbf{B}$

When it is assumed that the rows of the disturbance matrix $\mathbf{E}$ are independent, each with covariance matrix $\mathbf{\Sigma}$, the natural inner product to use in the observation space $R^{N\times p}$ is

$$\langle \mathbf{X}, \mathbf{Y}\rangle = \operatorname{trace} \mathbf{X}\mathbf{\Sigma}^{-1}\mathbf{Y}' = \operatorname{trace} \mathbf{Y}'\mathbf{X}\mathbf{\Sigma}^{-1} \tag{A.19}$$

for $\mathbf{X}$ and $\mathbf{Y}$ in $R^{N\times p}$. We then measure the goodness of fit of any parameter matrix $\mathbf{B}$ to the observed data $\mathbf{Y}$ making use of the corresponding norm

$$\texttt{LMSSE}(\mathbf{B}) = \|\mathbf{Y} - \mathbf{Z}\mathbf{B}\|^2 = \operatorname{trace}\,(\mathbf{Y} - \mathbf{Z}\mathbf{B})'\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{Z}\mathbf{B}). \tag{A.20}$$

Suppose, for the moment, that the matrix $\mathbf{Z}$ is of full column rank, or that $N \geq q$ and the columns of $\mathbf{Z}$ are independent. A central result on the multivariate linear model is that the matrix $\hat{\mathbf{B}}$ that minimizes $\texttt{LMSSE}(\mathbf{B})$ is given by

$$\hat{\mathbf{B}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}. \tag{A.21}$$

The corresponding predictor of $\mathbf{Y}$ is given by

$$\hat{\mathbf{Y}} = \mathbf{Z}\hat{\mathbf{B}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}. \tag{A.22}$$

The matrix $\hat{\mathbf{Y}}$ can be thought of as the matrix in the subspace $R(\mathbf{Z})$ that minimizes $\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2$ over all possible approximations $\hat{\mathbf{Y}} = \mathbf{Z}\mathbf{B}$ falling in $R(\mathbf{Z})$.

Note that the least squares estimator $\hat{\mathbf{B}}$ and the best linear predictor $\hat{\mathbf{Y}}$ do not depend on the variance matrix $\mathbf{\Sigma}$, even though the fitting criterion $\texttt{LMSSE}(\mathbf{B})$ does. It turns out that when the details of the minimization of $\texttt{LMSSE}(\mathbf{B})$ are carried through, the variance matrix $\mathbf{\Sigma}$ cancels out. But if there are covariances among errors or residuals *across* observations, contained in a variance-covariance matrix $\mathbf{\Gamma}$, say, then the inner product (A.19) becomes

$$\langle \mathbf{X}, \mathbf{Y}\rangle = \operatorname{trace}\, \mathbf{Y}'\mathbf{\Gamma}^{-1}\mathbf{X}\mathbf{\Sigma}^{-1}.$$

Using this inner product in the definition of goodness of fit, the estimator of $\mathbf{B}$ and the best predictor of $\mathbf{Y}$ becomes

$$\hat{\mathbf{B}} = (\mathbf{Z}'\mathbf{\Gamma}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{\Gamma}^{-1}\mathbf{Y}$$

and

$$\hat{\mathbf{Y}} = \mathbf{Z}(\mathbf{Z}'\mathbf{\Gamma}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{\Gamma}^{-1}\mathbf{Y}.$$

Thus, the optimal solution does depend on how one treats errors *across* observations.

A.8 Regularizing the multivariate linear model

One of the major themes of this book is *regularization*, and for readers familiar with multivariate analysis, it may be helpful to introduce this idea in the multivariate context first. Others, especially those who are familiar with curve estimation already, may prefer to omit this section.

Suppose now that we are dealing with an under-determined problem, where $q > N$ and the matrix $\mathbf{Z}$ is of full row rank N. This means that the range space $R(\mathbf{Z})$ is the whole of $R^{N\times p}$.

A.8.1 Definition of regularization

Regularization involves attaching a *penalty term* to the basic squared error fitting criterion:

$$\texttt{LMSSE}_\lambda(\mathbf{B}) = \|\mathbf{Y} - \mathbf{Z}\mathbf{B}\|^2 + \lambda \times \texttt{PEN}(\mathbf{B}). \tag{A.23}$$

The purpose of the penalty term $\texttt{PEN}(\mathbf{B})$ is to require that the estimated value of $\mathbf{B}$ not only yields a good fit in the sense of small $\|\mathbf{Y} - \mathbf{Z}\mathbf{B}\|^2$, but also that some aspect of $\mathbf{B}$ captured in the function PEN is kept under control. The positive penalty parameter λ quantifies the relative importance of these two aims. If λ is large, then we are particularly concerned with keeping $\texttt{PEN}(\mathbf{B})$ small, and getting a good fit to the data is only of secondary importance; if λ is small, then we are not so concerned about the value of $\texttt{PEN}(\mathbf{B})$.

One example of this type of regularization is the *ridge regression* technique, often used to stabilize regression coefficient estimates in the presence of highly collinear independent variables. In this case, what is penalized is the size of the regression coefficients themselves, in the sense that $\texttt{PEN}(\mathbf{B}) = \text{trace}(\mathbf{B}'\mathbf{B})$, the sum of squares of the entries of $\mathbf{B}$. The solution to the minimization of $\texttt{LMSSE}_\lambda(\mathbf{B})$ is then

$$\mathbf{B} = (\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{I})^{-1}\mathbf{Z}'\mathbf{Y}.$$

As λ approaches zero, $\mathbf{B}$ approaches the least squares solution described in Section A.7, but as λ grows, $\mathbf{B}$ approaches zero. Thus, ridge regression is said to shrink the solution towards zero.

A.8.2 Hard-edged constraints

One way to obtain a well-determined problem is to place constraints on the matrix $\mathbf{B}$. For example, consider the model where it is assumed that the coefficients in each column of $\mathbf{B}$ are a constant vector, so all we have to do is to estimate a single number for each column. If we define the $(q-1)\times q$ matrix $\mathbf{L}$ to have $L_{ii} = 1$ and $L_{i,i+1} = -1$ for each i, and all other entries

zero, then our assumption about $\mathbf{B}$ can be written as the constraint

$$\mathbf{LB} = 0. \tag{A.24}$$

In order for the elements of $\mathbf{B}$ to be identifiable on the basis of the observed data, the design matrix $\mathbf{Z}$ has to satisfy the condition

$$\mathbf{Z1} \neq 0, \tag{A.25}$$

where $\mathbf{1}$ is a vector of q unities.

The transformation $\mathbf{L}$ reduces multiples of the vector $\mathbf{1}$ exactly to zero. The identifiability condition (A.25) can be replaced by the condition that the zero vector is the only q-vector $\mathbf{b}$ such that both $\mathbf{Lb}$ and $\mathbf{Zb}$ are zero. Equivalently, the matrix $[\mathbf{Z}' \ \ \mathbf{L}']$ is nonsingular.

A.8.3 Soft-edged constraints

Instead of enforcing the hard-edged constraint $\mathbf{LB} = 0$, we may wish to let the coefficients in any column of $\mathbf{B}$ vary, but not more than really necessary, by exploring compromises between the rank-one extreme implied by (A.24) and a completely unconstrained underdetermined fit. We might consider this a soft-edged constraint, and it can be implemented by a suitable regularization procedure. If we define

$$\texttt{PEN}(\mathbf{B}) = \|\mathbf{LB}\|^2 = \text{trace}(\mathbf{B}'\mathbf{L}'\mathbf{LB}) \tag{A.26}$$

then the penalty $\texttt{PEN}(\mathbf{B})$ quantifies how far the matrix $\mathbf{B}$ is from satisfying the constraint $\mathbf{LB} = 0$.

The regularized estimate of $\mathbf{B}$, obtained by minimizing the criterion (A.23), now satisfies

$$(\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{L}'\mathbf{L})\mathbf{B} = \mathbf{Z}'\mathbf{Y}. \tag{A.27}$$

For any $\lambda > 0$, a unique solution for $\mathbf{B}$ requires the nonsingularity of the matrix $[\mathbf{Z}'\,\mathbf{L}']$, precisely the condition for identifiability of the model subject to the constraint (A.24).

In the limit as the parameter $\lambda \to \infty$, the penalized fitting criterion (A.23) automatically enforces on $\mathbf{B}$ the one-dimensional structure $\mathbf{LB} = 0$. On the other hand, in the limit $\lambda \to 0$, no penalty at all is applied, and $\mathbf{B}$ takes on whatever value results in minimizing the error sum of squares to zero, due to the underdetermined character of the problem. Thus, from the regularization perspective, the constrained estimation problem $\mathbf{LB} = 0$ that arises frequently in linear modelling designs is simply an extreme case of the regularization process where $\lambda \to \infty$.

We have concentrated on a one-dimensional constrained model, corresponding to a $(q-1)\times q$ matrix $\mathbf{L}$, but of course the ideas can be immediately extended to nonsingular $s \times q$ constraint matrices $\mathbf{L}$ that map a q-vector into a space of vectors of dimension $s \leq q$. In this case, the constrained model is of dimension $q - s$. Note also that the specification of the matrix $\mathbf{L}$

corresponding to any particular constrained model is not unique, and that if $\mathbf{L}$ is specified differently the regularized estimates are in general different.

Finally, we note in passing that Bayesian approaches to regression, in which a multivariate normal prior distribution is proposed for $\mathbf{B}$, can also be expressed in terms of a penalized least squares problem of the form (A.23). For further details see, for example, Kimeldorf and Wahba (1970), Wahba (1978) or Silverman (1985).

References

Abraham, C., Cornillion, P. A., Matzner–Lober, E. and Molinari, N. (2003) Unsupervised curve-clustering using b-splines. *Scandinavian Journal of Statistics,* **30,** 581–595.

Aguilera, A. M., Ocaña, F. A. and Valderrama, M. J. (1999) Forecasting with unequally spaced data by a functional principal component approach, *Test,* **8,** 233-253.

Anderson, T. W. (1984) *An Introduction to Multivariate Statistical Analysis.* Second edition. New York: Wiley.

Anselone, P. M. and Laurent, P. J. (1967) A general method for the construction of interpolating or smoothing spline-functions. *Numerische Mathematik*, **12**, 66–82.

Ansley, C. F. and Kohn, R. (1990) Filtering and smoothing in state space models with partially diffuse initial conditions. *Journal of Time Series Analysis,* **11,** 275–293.

Ansley, C. F., Kohn, R. and Wong, C-M. (1993) Nonparametric spline regression with prior information. *Biometrika,* **80,** 75–88.

Aronszajn, N. (1950) Theory of reproducing kernels. *Transactions of the American Mathematical Society,* **68,** 337–404.

Atteia, M. (1965) Spline-fonctions généralisées. *Comptes Rendus de l'Académie des Sciences Série I: Mathématique*, **261**, 2149–2152.

Aubin, J-P. (2000) *Applied Functional Analysis, Second Edition.* New York: Wiley-Interscience.

Basilevsky, A. (1994) *Statistical Factor Analysis and Related Methods.* New York: Wiley.

Besse, P. (1979) Etude descriptive des processus: Approximation et interpolation. Thèse de troisième cycle, Université Paul-Sabatier, Toulouse.

Besse, P. (1980) Deux exemples d'analyses en composantes principales filtrantes. *Statistique et Analyse des Données*, **3**, 5–15.

Besse, P. (1988) Spline functions and optimal metric in linear principal component analysis. In J. L. A. van Rijkevorsel and J. de Leeuw (eds.) *Component and Correspondence Analysis: Dimensional Reduction by Functional Approximation*. New York: Wiley.

Besse, P. (1991) Approximation spline de l'analyse en composantes principales d'une variable aléatoire hilbertienne. *Annales de la Faculté des Sciences de Toulouse,* **12,** 329–349.

Besse, P. and Ramsay, J. O. (1986) Principal components analysis of sampled functions. *Psychometrika*, **51**, 285–311.

Besse, P. and Cardot, H (1997) Approximation spline de la prévision d'un processus fonctionnel autorégressif d'ordre 1. *Canadian Journal of Statistics,* **24,** 467–487.

Besse, P., Cardot, H. and Ferraty, F. (1997). Simultaneous nonparametric regressions of unbalanced longitudinal data. *Computational Statistics and Data Analysis,* **24,** 255–270.

Besse, P., Cardot, H. and Stephenson, D. (2000) Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics,* **27,** 673–687.

Bock, R. D. and Thissen, D. (1980) Statistical problems of fitting individual growth curves. In F. E. Johnston, A. F. Roche and C. Susanne (eds.) *Human Physical Growth and Maturation: Methodologies and Factors*. New York: Plenum.

Bookstein, F. L. (1991) *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge: Cambridge University Press.

Bosq, D. (2000) *Linear processes in function spaces*. New York: Springer.

Brumback, B. and Rice, J. A. (1998) Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *Journal of the American Statistical Association,* **93,** 961–994.

Buckheit, J. B, Olshen, R. A., Blouch, K. and Myers, B. D. (1997) Modeling of progressive glomerular injury in humans with lupus nephritis. *American Journal of Physiology*, **237**, 158–169.

Buja, A., Hastie, T. and Tibshirani, R. (1989) Linear smoothers and additive models (with discussion). *Annals of Statistics*, **17**, 453–555.

Cai, Z., Fan, J. and Li, R. (2000) Efficient estimation and inferences for varying-coefficient models, *Journal of the American Statistical Association,* **95,** 888–902.

Cai, Z., Fan, J. and Yao, Q. (2000) Functional-coefficient fregression models for non–linear time series, *Journal of the American Statistical Association,* **95,** 941–956.

Cailliez, F. and Pagès, J. P. (1976) *Introduction à l'analyse des données*. Paris: SMASH (Société de Mathétiques Appliquées et de Sciences Humaines.

Cardot, H. (2002) Spatially adaptive splines for statistical linear inverse problems. *Journal of Multivariate Analysis,* **81,** 100–119.

Cardot, H. (2004) Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *Journal of Nonparametric Statistics,* to appear.

Cardot, H., Ferraty, F. and Sarda, P. (1999) Functional linear model, *Statistics & Probability Letters,* **45,**, 11–22.

Cardot, H., Ferraty, F. and Sarda, P. (2003) Spline estimators for the functional linear model. *Statistica Sinica,* **13,**, 571–591.

Cardot, H., Faivre, R. and Goulard, M. (2003). Functional approaches for predicting land use with the temporal evolution of coarse resolution remote sensing data. *Journal of Applied Statistics,* **30,** 1185–1199.

Cardot, H., Ferraty, F., Mas, A. and Sarda, P. (2003). Testing Hypotheses in the functional linear model. /em Scandinavian Journal of Statistics, **30,** 241–255.

Cardot, H., Faivre, R. and Maisongrande, P. (2004). Random Effects Varying Time Regression Models: Application to Remote Sensing. J. Antoch (ed.) /em Compstat 2004 proceedings, Physica-Verlag, 777–784.

Cardot, H., Crambes, C. and Sarda, P. (2004). Estimation spline de quantiles conditionnels pour variables explicatives fonctionnelles. *Comtes Rendu de l'Académie des Sciences. Paris, serie I,* **339,** 141–144.

Cardot H., Goia, A. and Sarda, P. (2004). Testing for no effect in functional linear regression models, some computational approaches. /em Communications in Statistics – Simulation and Computation, **33,** 179–199.

Cardot, H., Ferraty, F., Mas, A. and Sarda, P. (2004) Testing hypotheses in the functional linear model. *Scandinavian Journal of Statistics,* **30,** 241–255.

Castro, P. E., Lawton, W. H. and Sylvestre, E. A. (1986) Principal modes of variation for processes with continuous sample curves. *Technometrics*, **28**, 329–337.

Char, B. W., Geeddes, K. O., Gonnet, G. H., Leong, B. L., Monagan, M. B. and Watt, S. M. (1991) *MAPLE V Language Reference Manual*. New York: Springer.

Chiou, J.-M., Müller, H.-G. and Wang, J.-L. (2003) Functional quasi-likelihood regression models with smooth random effects. *Journal of the Royal Statistical Society, Series B.,* **65,** 405–423.

Chui, C. K. (1992) *An Introduction to Wavelets*. San Diego: Academic Press.

Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Coddington, E. A. (1989) *An Introduction to Ordinary Differential Equations*. New York: Dover.

Coddington, E. A. and Levinson, N. (1955) *Theory of Ordinary Differential Equations*. New York: McGraw-Hill.

Cook, R. D. and Weisberg, S. (1982) *Residuals and Influence in Regression*. New York: Chapman and Hall.

Cox, D. R. and Lewis, P. A. W. (1966) *The Statistical Analysis of Series of Events.* London: Methuen.

Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, *Numerische Mathematik*, **31**, 377–403.

Cressie, N. (1991) *Statistics for Spatial Data.* New York: Wiley.

Cueva, A., Febrero, M. and Fraiman, R. (2002) Linear functional regression: The case of fixed design and functional response. *The Canadian Journal of Statistics,* **30,** 285–300.

Dalzell, C. J. and Ramsay, J. O. (1993) Computing reproducing kernels with arbitrary boundary constraints. *SIAM Journal of Scientific Computing*, **14**, 511–518.

Daubechies, I. (1992) *Ten Lectures on Wavelets.* CBMS-NSF Series in Applied Mathematics, **61**. Philadelphia: Society for Industrial and Applied Mathematics.

Dauxois, J. and Pousse, A. (1976) Les analyses factorielles en calcul des probabilité et en statistique: Essai d'étude synthètique. Thèse d'état, Université Paul-Sabatier, Toulouse.

Dauxois, J., Pousse, A. and Romain, Y. (1982) Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, **12**, 136–154.

Dauxois, J. and Nkiet, G. M. (2002) Measure of association for Hilbertian subspaces and some applications. *Journal of Multivariate Analysis,* **82,** 263–297.

de Boor, C. (2001) *A Practical Guide to Splines.* Revised Edition. New York: Springer.

Delves, L. M. and Mohamed, J. L. (1985) *Computational Methods for Integral Equations.* Cambridge: Cambridge University Press.

Deville, J. C. (1974) Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'INSEE*, **15**, 7–97.

Diggle, P. J., Liang, K.-Y. and Zeger, S. L. (1994) *Analysis of Longitudinal Data.* New York: Oxford University Press.

Dongarra, J. J., Bunch, J. R., Moler, C. B. and Stewart, G. W. (1979) *LINPACK Users' Guide.* Philadelphia: Society for Industrial and Applied Mathematics.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995) Wavelet shrinkage: asymptopia? (with discussion). *Journal of the Royal Statistical Society*, Series B, **57**, 301–369.

Draper, N. R. and Smith, H. (1998) *Applied Regression Analysis. Third Edition.* New York: Wiley.

Dreesman, J. M. and Tutz, G. (2001) Non–stationary conditional models for spatial data based on varying coefficients, *The Statistician,* **50,** 1–15.

Eaton, M. L. (1983) *Multivariate Statistics: A Vector Space Approach.* New York: Wiley.

Eaton, M. L. and Perlman, M. D. (1973) The non-singularity of generalized sample covariance matrices. *Annals of Statistics*, **1**, 710–717.

Efron, B. and Tibshirani, R. J. (1992) *An Introduction to the Bootstrap.* London: Chapman & Hall.

Eilers, P. H. C. and Marx, B. D. (1996) Flexible smoothing with B-splines and penalties, with comments. *Statistical Science,* **11**, 89–121.

Engle, R. F., Granger, C. W. J., Rice, J. A. and Weiss, A. (1986) Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, **81**, 310–320.

Escabias, M., Aguilera, A. M. and Valderrama, M. J. (2004) Principal component estimation of functional logistic regression: Discussion of two different approaches. *Nonparametric Statistics,* In press.

Eubank, R. L. (1999) *Spline Smoothing and Nonparametric Regression, Second Edition.* New York: Marcel Dekker.

Eubank, R. L., Muñoz Maldonado, Y., Wang, N. and Wang, S. (2004) Smoothing spline estimation in varying coefficient models. *Journal of the Royal Statistical Society, Series B.,* **66,** 653–667.

Falkner, F. (ed.) (1960) *Child Development: An International Method of Study.* Basel: Karger.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and its Applications.* London: Chapman and Hall.

Fan, J. and Lin, S.-K. (1998) Tests of significance when data are curves. *Journal of the Americal Statistical Association,* **93,** 1007–1021.

Fan, J., Yao, Q. and Cai, Z. (2003) Adaptive varying-coefficient linear models, *Journal of the Royal Statistical Society, Series B,* **65,** 57–80.

Faraway, J. J. (1997) Regression analysis for a functional response, *Technometrics,* **39,** 254–261.

Ferraty, F. and Vieu, P. (2001) The functional nonparametric model and its applications to spectrometric data. *Computational Statistics,* **17,** 545–564.

Ferraty, F., Goia, A. and Vieu, P. (2002) Functional nonparametric model for time series: A fractal approach for dimension reduction. *Test,* **11,** 317–344.

Frank, I. E. and Friedman, J. H. (1993) A statistical view of some chemometrics regression tools. *Technometrics,* **35**, 109–135.

Friedman, J. H. and Silverman, B. W. (1989) Flexible parsimonious smoothing and additive modeling (with discussion). *Journal of the American Statistical Association*, **31**, 1–39.

Gasser, T. and Kneip, A. (1995). Searching for structure in curve samples. *Journal of the American Statistical Association*, **90**, 1179–1188.

Gasser, T., Kneip, A., Binding, A., Prader, A. and Molinari, L. (1991a) The dynamics of linear growth in distance, velocity and acceleration. *Annals of Human Biology*, **18**, 187–205.

Gasser, T., Kneip, A., Ziegler, P., Largo, R., Molinari, L., and Prader, A. (1991b) The dynamics of growth of height in distance, velocity and acceleration. *Annals of Human Biology*, **18**, 449–461.

Gasser, T., Kneip, A., Ziegler, P., Largo, R. and Prader, A. (1990) A method for determining the dynamics and intensity of average growth. *Annals of Human Biology*, **17**, 459–474.

Gasser, T. and Müller, H.-G., (1979) Kernel estimation of regression functions. In T. Gasser and M. Rosenblatt (eds.) *Smoothing Techniques for Curve Estimation*. Heidelberg: Springer, pp. 23–68.

Gasser, T. and Müller, H.-G., (1984) Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, **11**, 171–185.

Gasser, T. , Müller, H.-G., Köhler, W., Molinari, L. and Prader, A. (1984). Nonparametric regression analysis of growth curves. *Annals of Statistics*, **12**, 210–229.

Gauss, C. F. (1809) *Theoria motus corporum celestium*. Hamburg: Perthes et Besser.

Gelfand, A. E., Kim, H-J., Sirmans, C.F. and Banerjee, S. (2003) Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association,* **98,** 387–396.

Gervini, D. and Gasser, T. (2004) Self–modeling warping functions. *Journal of the Royal Statistical Society, Series B,* **66,** 959–971.

Golub, G. and Van Loan, C. F. (1989) *Matrix Computations*. Second edition. Baltimore: Johns Hopkins University Press.

Grambsch, P. M., Randall, B. L. Bostick, R. M., Potter, J. D. and Louis, T. A. (1995) Modeling the labeling index distribution: an application of functional data analysis. *Journal of the Americal Statistical Association*, **90**, 813–821.

Green, G. (1828) *An Essay on the Mathematical Analysis to the Theories of Electricity and Magnetism.* Privately printed.

Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman and Hall.

Greenhouse, S. W. and Geisser, S. (1959) On methods in the analysis of profile data. *Psychometrika*, **24**, 95–112.

Grenander, U. (1981) *Abstract Inference*. New York: Wiley.

Gu, C. (2002) *Smoothing Spline ANOVA Models*. New York: Springer.

Guo, W. (2002) Functional mixed effects models. *Biometrics,* **58,** 121–128.

Hall, P. and Heckman, N. E. (2002) Estimating and depicting the structure of a distribution of random runctions. *Biometrika,* **89,** 145–158.

Hanson, M. H. and Kooperberg, C. (2002) Spline adaptation in extended linear models, with discussion. *Statistical Science,* **17,** 2–51.

Härdle, W. (1990) *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. New York: Chapman and Hall.

Hastie, T. and Tibshirani, R. (1990) Varying-coefficient models. *Journal of the Royal Statistical Society, Series B,* **55,** 757–796.

Hastie, T., Buja, A. and Tibshirani, R. (1995) Penalized discriminant analysis. *Annals of Statistics*, **23**, 73–102.

He, G. Z., Müller, H-G. and Wang, J. I. (2003) Functional canonical analysis for square integrable stochastic processes. *Jurnal of Multivariate Analysis,* **85,** 54–77.

Heckman, N. E. and Ramsay, J. O. (2000). Penalized regression with model based penalties. *Canadian Journal of Statistics.* **28,** 241–258.

Hermanussen, M., Thiel, C., von Büren, E., de los Angeles Rol. de Lama, M., Pérez Romero, A., Ariznaverreta Ruiz, C., Burmeister, J., and Tresguerres, J. A. F. (1998). Micro and macro perspectives in auxology: Findings and considerations upon the variability of short term and individual growth and the stability of population derived parameters. *Annals of Human Biology*, **25**, 359–395.

Huynh, H. S. and Feldt, L. (1976) Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, **1**, 69–82.

Houghton, A. N., Flannery, J. and Viola, M. V. (1980) Malignant melanoma in Connecticut and Denmark. *International Journal of Cancer*, **25**, 95–104.

Hutchison, M. F. and de Hoog, F. R. (1985) Smoothing noisy data with spline functions. *Numerische Mathematik*, **47**, 99–106.

Ince, E. L. (1956) *Ordinary Differential Equations.* New York: Dover.

James, G. M. (2002) Generalized linear models with functional predictors, *Journal of the Royal Statistical Society, Series B.*, **64,** 411-432.

James, G. M. and Hastie, T. J. (2001) Functional linear discriminant analysis for irregularly sampled curves, *Journal of the Royal Statistical Society, Series B.*, **63,** 533–550.

James, G. M. and Sugar, C. A. (2003) Clustering sparsely sampled functional data. *Journal of the Americal Statistical Association,* **98,** 397–408.

James, G. M., Hastie, T. J. and Sugar, C. A. (2000) Principal component models for sparse functional data, *Biometrika*, **87,** 587–602.

Johnson, R. A. and Wichern, D. A. (1988) *Applied Multivariate Statistical Analysis.* Englewood Cliffs, N. J.: Prentice Hall.

Johnstone, I. M. and Silverman, B. W. (1997) Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B*, **59**, 319–351.

Jolicoeur, P., Pontier, J., Abidi, H. (1992) Asymptotic models for the longitudinal growth of human stature. *American Journal of Human Biology*, **4**, 461–468.

Karhunen, K. (1947) Über linear Methoden in der Warscheinlichkeitsrechnung. *Annales Academiae Scientiorum Fennicae*, **37**, 1–79.

Keselman, H. J. and Keselman, J. C. (1993) Analysis of repeated measurements. In L. K. Edwards (ed.) *Applied analysis of Variance in Behavioral Science*, New York: Marcel Dekker, 105–145.

Kim, Y-J. and Gu, G. (2004) Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *Journal of the Royal Statistical Association, Series B.,* **66,** 337–356.

Kimeldorf, G. S. and Wahba, G. (1970) A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, **41**, 495–502.

Kneip, A. and Engel, J. (1995) Model estimation in nonlinear regression under shape invariance. *Annals of Statistics*, **23**, 551–570.

Kneip, A. and Gasser, T. (1988) Convergence and consistency results for self-modeling nonlinear regression. *Annals of Statistics*, **16**, 82–112.

Kneip, A. and Gasser, T. (1992) Statistical tools to analyze data representing a sample of curves. *Annals of Statistics*, **20**, 1266–1305.

Kneip, A., Li, X., MacGibbon, K. B. and Ramsay, J. O. (2000) Curve registration by local regression. *The Canadian Journal of Statistics,* **28,** 19–29.

Kneip, A. and Utikal, K. J. (2001) Inference for density families using functional principal components analysis. *Journal of the Americal Statistical Association,* **96,** 519–542.

Koenker, R., Ng, P. and Portnoy, S. (1994) Quantile smoothing splines. *Biometrika,* **81,** 673–680.

Krommer, A. R. and Überhuber, C. W. (1998) *Computational Integration.* Philadelpha: Society for Industrial and Applied Mathematics.

Lancaster, H. O. (1969) *The Chi-squared Distribution.* New York: Wiley.

Largo, R. H., Gasser, T., Prader, A., Stützle, W. and Huber, P. J. (1978) Analysis of the adolescent growth spurt using smoothing spline functions. *Annals of Human Biology*, **5**, 421–434.

Legendre, A. M. (1805) *Nouvelles Méthodes pour la Détermination des Orbites des Comètes*. Paris: Courcier.

Leurgans, S. E., Moyeed, R. A. and Silverman, B. W. (1993) Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society*, Series B, **55**, 725–740.

Li, K-C. (1991) Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association,* **86**, 316–342.

Li, K-C., Aragon, Y., Shedden, K. and Thomas Agnan, C. (2003) Dimension reduction for multivariate response data. *Journal of the American Statistical Association,* **98**, 99–109.

Liggett, W., Cazares, L. and Semmes, O. J. (2003) A look at mass spectral measurement. *Chance,* **16**, 24-28.

Lindsey, J. K. (1993) *Models for Repeated Measurements*. New York: Oxford University Press.

Lindstrom, M. J. (2002) Bayesian estimation of free-knot splines using reversible jumps. *Computational Statistics and Data Analysis,* **41,** 255–269.

Lindstrom, M. J. and Kotz, S. (2004) Free-knot splines. In S. Kotz, N. L. Johnson and B. R. Campbell (eds.) *Encyclopedia of Statistics.* New York: Wiley.

Liu, X. and Müller, H-G. (2004) Functional convex averaging and synchronization for time-warped random curves. *Journal of the American Statistical Association,* **99,** 687–699.

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T. and Cohen, K. L. (1999) Robust principal component analysis for functional data. *Test,* **8,** 1–73.

Loève, M. (1945) Fonctions aléatoires de second ordre. *Comptes Rendus de l'Académie des Sciences, Série I: Mathématique*, **220**, 469.

Mao, W. and Zhao, L. H. (2003) Free–knot polynomial splines with confidence intervals. *Journal of the Royal Statistical Society, Series B.,* **65,** 901–919.

Maxwell, S. E. and Delaney, H. D. (2003) *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Belmont, CA: Wadsworth.

Moore, R. E. (1985) *Computational Functional Analysis*. Chichester: Wiley.

Mulaik, S. A. (1972) *The Foundations of Factor Analysis*. New York: McGraw-Hill.

Müller, H-G. and Stadtmüller, U. (2004) Generalized functional linear models. *Annals of Statistics,* to appear.

Muñoz Maldonado, Y., Staniswalis, J. G., Irwin, L. N. and Byers, D. (2002) A similarity analysis of curves. *The Canadian Journal of Statistics,* **30,** 373–381.

Nadaraya, E. A. (1964) On estimating regression. *Theory of Probability and its Applications*, **10**, 186–190.

Nason, G. P. and Silverman, B. W. (1994) The discrete wavelet transform in S. *Journal of Computational and Graphical Statistics*, **3**, 163–191.

Nielsen, H. A., Nielsen, T. S., Joensen, A. K., Madsen, H. and Holst, Ja. (2000) Tracking time-varying coefficient functions. *International Journal of Adaptive Control and Signal Processing,* **14**, 813–828.

Ocaña, F. A., Aguilera, A. M. and Valderrama, M. J. (1999) Functional principal components analysis by choice of norm. *Journal of Multivariate Analysis,* **71,** 262–276.

OECD (1995) *Quarterly National Accounts*, **3**.

Øksendal, B. (1995) *Stochastic Differential Equations: An Introduction with Applications*. New York: Springer.

Olshen, R. A., Biden, E. N., Wyatt, M. P. and Sutherland, D. H. (1989) Gait analysis and the bootstrap. *Annals of Statistics*, **17**, 1419–1440.

O'Sullivan, F. (1986) A statistical perspective on ill-posed linear inverse problems. *Statistical Science*, **1**, 502–527.

O'Sullivan, F., Yandell, B. and Raynor, W. (1986) Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association,* **81,** 441–455.

Parzen, E. (1961) An approach to time series analysis. *Annals of Mathematical Statistics*, **32**, 951–989.

Parzen, E. (1963) Probability density functionals and reproducing kernel Hilbert spaces, In M. Rosenblatt (ed.) *Proceedings of the Symposium on Time Series Analysis*, Providence, RI.: Brown University, 155–169.

Pezzulli, S. D. and Silverman, B. W. (1993) Some properties of smoothed principal components analysis for functional data. *Computational Statistics*, **8**, 1–16.

Pinheiro, J. C. and Bates, D. M. (2000) *Mixed-effects Models in S and S-PLUS.* New York: Springer.

Pousse, A. (1992) Etudes asymptotiques. In J.-J. Droesbeke, B. Fichet and P. Tassi (Eds.) *Modèles pour l'Analyse des Données Multidimensionelles.* Paris: Economica.

Press W. H., Teukolsky S. A., Vetterling W. T. and Flannery B. P. (1999) *Numerical recipes in C.* Second edition. Cambridge: Cambridge University Press.

Ramsay, J. O. (1982). When the data are functions. *Psychometrika*, **47**, 379–396.

Ramsay, J. O. (1989) The data analysis of vector-valued functions. In E. Diday (ed.) *Data Analysis, Learning Symbolic and Numeric Knowledge.* Commack, N. Y.: Nova Science Publishers.

Ramsay, J. O. (1996a) Principal differential analysis: data reduction by differential operators. *Journal of the Royal Statistical Society*, Series B, **58**, 495–508.

Ramsay, J. O. (1996b) Estimating smooth monotone functions. *Journal of the Royal Statistical Society, Series B.*, **60**, 365–375.

Ramsay, J. O. (1996c) Pspline: An Splus module for polynomial spline smoothing. Computer software in the `statlib` archive.

Ramsay, J. O. (2000). Functional components of variation in handwriting. *Journal of the American Statistical Association*, **95**, 9–15.

Ramsay, J. O., Altman, N. and Bock, R. D. (1994) Variation in height acceleration in the Fels growth data. *Canadian Journal of Statistics*, **22**, 89–102.

Ramsay, J. O., Bock. R. D. and Gasser, T. (1995) Comparison of height acceleration curves in the Fels, Zurich, and Berkeley growth data. *Annals of Human Biology*, **22**, 413–426.

Ramsay, J. O. and Dalzell, C. J. (1991) Some tools for functional data analysis (with Discussion). *Journal of the Royal Statistical Society*, Series B, **53**, 539–572.

Ramsay, J. O, Heckman, N. and Silverman, B. W. (1997) Some general theory for spline smoothing. *Behavioral Research: Instrumentation, Methods, and Computing*, **29,** 99–106.

Ramsay, J. O. and Li, X. (1996) Curve registration. *Journal of the Royal Statistical Society, Series B.*, **60,** 351–363.

Ramsay, J. O., Munhall, K., Gracco, V. L. and Ostry, D. J. (1996) Functional data analyses of lip motion. *Journal of the Acoustical Society of America*, **99**, 3718–3727.

Ramsay, J. O. and Silverman, B. W. (2002) *Applied Functional Data Analysis.* New York: Springer.

Ramsay, J. O., Wang, X. and Flanagan, R. (1995) A functional data analysis of the pinch force of human fingers. *Applied Statistics*, **44**, 17–30.

Rao, C. R. (1958) Some statistical methods for comparison of growth curves. *Biometrics*, **14**, 1–17.

Rao, C. R. (1987) Prediction in growth curve models (with discussion). *Statistical Science*, **2**, 434–471.

Ratcliffe, S. J., Leader, L. R. and Heller, G. Z. (2002) Functional data analysis with application to periodically stimulated foetal heart rate data. I: Functional regression. *Statistics in Medicine,* **21,** 1103–1114.

Ratcliffe, S. J., Heller, G. Z. and Leader, L. R. (2002) Functional data analysis with application to periodically stimulated foetal heart rate data. II: Functional logistic regression. *Statistics in Medicine,* **21,** 1115–1127.

Reinsch, C. (1967) Smoothing by spline functions. *Numerische Mathematik*, **10**, 177–183.

Reinsch, C. (1970) Smoothing by spline functions II. *Numerische Mathematik*, **16**, 451–454.

Rice, J. A. and Wo, C. O. (2001) Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics,* **57,** 253–259.

Rice, J. A. and Silverman, B. W. (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society*, Series B, **53**, 233–243.

Ripley, B. D. (1991) *Statistical Inference for Spatial Processes.* Cambridge: Cambridge University Press.

Roach, G. F. (1982) *Green's Functions.* Second edition. Cambridge: Cambridge University Press.

Roche, A. (1992) *Growth, Maturation and Body Composition: The Fels Longitudinal Study 1929–1991.* Cambridge: Cambridge Press.

Rønn, B. B. (2001) Nonparametric maximum likelihood estimation for shifted curves. *Journal of the Royal Statistical Society, Series B.,* **63,** 243–259.

Rossi, N., Wang, X. and Ramsay, J. O. (2002) Nonparametric item response function estimates with the EM algorithm. *Journal of the Behavioral and Educational Sciences,* **27,** 291–317.

Sakoe, H. and Chiba, S. (1978) Dynamic programming algorithm for optimizating for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal processing,* **26**, 43–49.

Scott, D. W. (1992) *Multivariate Density Estimation.* New York: Wiley.

Schumaker, L. (1981) *Spline Functions: Basic Theory.* New York: Wiley.

Scott, D. W. (1992) *Multivariate Density Estimation.* New York: Wiley.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992) *Variance Components.* New York: Wiley.

Seber, G. A. F. (1984) *Multivariate Observations.* New York: Wiley.

Silverman, B. W. (1982) On the estimation of a probability density function by the maximum penalized likelihood method. *Annals of Statistics*, **10**, 795–810.

Silverman, B. W. (1985) Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society*, Series B, **47**, 1–52.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

Silverman, B. W. (1994) Function estimation and functional data analysis. In A. Joseph, F. Mignot, F. Murat, B. Prum and R. Rentschler (eds.) *First European Congress of Mathematics.* Basel: Birkhäuser. vol II, pp. 407–427.

Silverman, B. W. (1995). Incorporating parametric effects into functional principal components analysis. *Journal of the Royal Statistical Society*, Series B, **57**, 673–689.

Silverman, B. W. (1996) Smoothed functional principal components analysis by choice of norm. *Annals of Statistics*, **24**, 1–24.

Silverman, B. W. (1999) Wavelets in statistics: Beyond the standard assumptions. *Philosophical Transactions of the Royal Society of London, Series A.* **357,** 2459–2473.

Silverman, B. W. (2000) Wavelets for regression and other statistical problems. In M. G. Schimek (Ed.) *Smoothing and Regression: Approaches, Computation and Application.* New York: Wiley.

Silverman, B. W. and Vassilicos, J. C. (1999) Wavelets: The key to intermittent information? *Philosophical Transactions of the Royal Society of London, Series A.* **357,** 2393–2395.

Snyder, D. L. and Miller, M. I. (1991) *Random Point Processes in Time and Space.* New York: Springer.

Simonoff, J. S. (1996) *Smoothing Methods in Statistics.* New York: Springer.

Spitzner, D. J., Marron, J. S. and Essick, G. K. (2003) Mixed–model functional ANOVA for studying human tactile perception. *The Journal of the American Statistical Association,* **98,** 263–272.

Statistical Sciences (1995) *S-PLUS Guide to Statistical and Mathematical Analysis, Version 3.3.* Seattle: StatSci, a division of MathSoft, Inc.

Stoer, J. and Bulirsch, R. (2002) *Introduction to Numerical Analysis.* Third Edition. New York: Springer.

Stone, M. (1987) *Coordinate-Free Multivariable Statistics.* Oxford: Clarendon Press.

Tarpey, T. and Kinateder, K. K. J. (2003) Clustering functional data. *Journal of Classification,*, **20,** 93–114.

Tenenbaum, M. and Pollard, H. (1963) *Ordinary Differential Equations.* New York: Harper and Row.

Tucker, L. R. (1958) Determination of parameters of a functional relationship by factor analysis. *Psychometrika*, **23**, 19–23.

Tuddenham, R. D. and Snyder, M. M. (1954) Physical growth of California boys and girls from birth to eighteen years. *University of California Publications in Child Development* **1**, 183–364.

Valderrama, M. J., Aguilera, A. M. and Ocaña, F. A. (2000) *Predicción Dinámica Mediante Análisis de Datos Funcionales.* Madrid: Hespérides.

Viele, K. (2001) Evaluating fit in functional data analysis using model embeddings. *The Canadian Journal of Statistics,* **29,** 51–66.

Vinod, H. D. (1976) Canonical ridge and econometrics of joint production. *Journal of Econometrics*, **4**, 147–166.

Viviani, R., Grön, G. and Spitzer, M. (2005) *Human Brain Mapping,* **24,** 109–129.

Wahba, G. (1978) Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society*, Series B, **40**, 364–372.

Wahba, G. (1990) *Spline Models for Observational Data.* Philadelphia: Society for Industrial and Applied Mathematics.

Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing.* London: Chapman and Hall.

Wang, K. and Gasser, T. (1997) The alignment of curves by dynamic time warping. *The Annals of Statistics,* **25,** 1251–1276.

Wang, K. and Gasser, T. (1998) Asymptotic and boostrap confidence bounds for the structural average of curves. *The Annals of Statistics,* **26,** 972–991.

Wang, K. and Gasser, T. (1999) Synchronizing sample curves nonparametrically. *The Annals of Statistics,* **27,** 439–460.

Watson, G. S. (1964) Smooth regression analysis. *Sankhyā*, Series A, **26**, 101–116.

Weinert, H. L., Byrd, R. H. and Sidhu, G. S. (1980) A stochastic framework for recursive computation of spline functions: Part II, smoothing splines. *Journal of Optimization Theory and Applications,* **2,** 255–268.

West, M. and Harrison, P. J. (1989) *Bayesian Forecasting and Dynamic Models.* New York: Springer.

West, M., Harrison, P. J. and Migon, H. S. (1985) Dynamic generalized linear models and Bayesian forecasting (with discussion). *Journal of the American Statistical Association,* **80,** 73–97.

Wilson, A. M., Seelig, T. J., Shield, R. A. and Silverman, B. W. (1996) The effect of imposed foot imbalance on point of force application in the equine. Technical report, Department of Veterinary Basic Sciences, Royal Veterinary College.

Wolfram, S. (1991) *The Mathematica Book, fifth edition.* Champaign, Illinois: Wolfram Research Inc.

Wu, C. O., Chiang, C.-T. and Hoover, D. R. (1998) Asymptotic confidence regions for kernel smoothing of a varying coefficient model with longitudinal data. *Journal of the American Statistical Association,* **93,** 1388–1418.

Yao, F., Müller, H.-G., Clifford, A. J., Dueker, S. R., Follet, J., Lin, Y., Bucholz, B. A. and Vogel, J. S. (2003) Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics,* **59,** 676–685.

Yao, F., Müller, H.-G. and Wang, J.-L. (2004) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association,* to appear.

Zhang, C. (2003) Calibrating the degrees of freedom for automatic data smoothing and effective curve checking. *The Journal of the American Statistical Association,* **98,** 609–628.

Zhang, W. and Lee, S.-K. (2000) Variable bandwidth selecton in varying-coefficient models, *Journal of Multivariate Analysis,* **74**, 116–134.

Zhang, W., Lee, S.-K. and Song, X. (2002) Semiparametric smooth coefficient models, *Journal of Business and Economic Statistics,* **20,** 412–422.

Index