1

Introduction

Briefly stated, a *stochastic process* is an indexed collection of random variables all of which are defined on a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$. If we denote the index set by E, then this can be described mathematically as

$$\{X(t,\omega) : t \in E, \omega \in \Omega\},$$

where $X(t,\cdot)$ is a $\mathscr{F}$-measurable function on the sample space Ω. The ω argument will generally be suppressed and $X(t,\omega)$ will typically be shortened to just $X(t)$.

Once the $X(t)$ have been observed for every $t \in E$, the process has been realized and the resulting collection of real numbers is called a *sample path* for the process. Functional data analysis (fda), in the sense of this text, is concerned with the development of methodology for statistical analysis of data that represent sample paths of processes for which the index set is some (closed) interval of the real line; without loss, the interval can be taken as $[0,1]$. This translates into observations that are functions on $[0,1]$ and data sets that consist of a collection of such random curves.

From a practical perspective, one cannot actually observe a functional data set in its entirety; at some point, digitization must occur. Thus, analysis might be predicated on data of the form

$$x_i(j/r), j = 1,\ldots,r, i = 1,\ldots,n,$$

involving n sample paths $x_1(\cdot),\ldots,x_n(\cdot)$ for some stochastic process with each sample path only being evaluated at r points in $[0,1]$. When viewed from this perspective, the data is inherently finite dimensional and the temptation is to treat it as one would data in a multivariate analysis (mva) context.

Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators, First Edition. Tailen Hsing and Randall Eubank.

However, for truly functional data, there will be many more "variables" than observations; that is, $r \gg n$. This leads to drastic ill conditioning of the linear systems that are commonplace in mva which has consequences that can be quite profound. For example, Bickel and Levina (2004) showed that a naive application of multivariate discriminant analysis to functional data can result in a rule that always classifies by essentially flipping a fair coin regardless of the underlying population structure.

Rote application of mva methodology is simply not the avenue one should follow for fda. On the other hand, the basic mva techniques are still meaningful in a certain sense. Data analysis tools such as canonical correlation analysis, discriminant analysis, factor analysis, multivariate analysis of variance (MANOVA), and principal components analysis exist because they provide useful ways to summarize complex data sets as well as carry out inference about the underlying parent population. In that sense, they remain conceptually valid in the fda setting even if the specific details for extracting the relevant information from data require a bit of adjustment. With that in mind, it is useful to begin by cataloging some of the multivariate methods and their associated mathematical foundations, thereby providing a roadmap of interesting avenues for study. This is the subject of the following section.

1.1 Multivariate analysis in a nutshell

mva is a mature area of statistics with a rich history. As a result, we cannot (and will not attempt to) give an in-depth overview of mva in this text. Instead, this section contains a terse, mathematical sketch of a few of the methods that are commonly employed in mva. This will, hopefully, provide the reader with some intuition concerning the form and structure of analogs of mva techniques that are used in fda as well as an appreciation for both the similarities and the differences between the two fields of study. Introductions to the theory and practice of mva can be found in a myriad of texts including Anderson (2003), Gittins (1985), Izenman (2008), Jolliffe (2004), and Johnson and Wichern (2007).

Let us begin with the basic set up where we have a p-dimensional random vector $X = (X_1, \ldots, X_p)^T$ having (variance-)covariance matrix

$$\mathscr{K} = \mathbb{E}\left[(X - m)(X - m)^T\right] \tag{1.1}$$

with

$$m = \mathbb{E}X \tag{1.2}$$

the mean vector for X. Here, $\mathbb{E}$ corresponds to mathematical expectation and v^T indicates the transpose of a vector v. The matrix $\mathscr{K}$ admits an

eigenvalue–eigenvector decomposition of the form

$$\mathcal{K} = \sum_{j=1}^{p} \lambda_j e_j e_j^T \tag{1.3}$$

for eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ and associated orthonormal eigenvectors $e_j = \left(e_{1j}, \ldots, e_{pj}\right)^T, j = 1, \ldots, p$ that satisfy

$$e_i^T \mathcal{K} e_j = \lambda_j \delta_{ij},$$

where δ_{ij} is 1 or 0 depending on whether or not i and j coincide. This provides a basis for principal components analysis (pca).

We can use the eigenvectors in (1.3) to define new variables $Z_j = e_j^T(X - m)$, which are referred to as principal components. These are linear combinations of the original variables with the weight or loadings e_{ij} that is applied to X_i in the jth component indicating its importance to Z_j; more precisely,

$$\mathrm{Cov}(Z_j, X_i) = \lambda_j e_{ij}.$$

In fact,

$$X = m + \sum_{j=1}^{p} Z_j e_j \tag{1.4}$$

as, if $\mathcal{K}$ is full rank, $e_1, \ldots, e_p$ provide an orthonormal basis for $\mathbb{R}^p$; this is even true when $\mathcal{K}$ has less than full rank as $e_j^T X$ is zero with probability one when $\lambda_j = 0$. The implication of (1.4) is that X can be represented as a weighted sum of the eigenvectors of $\mathcal{K}$ with the weights/coefficients being uncorrelated random variables having variances that are the eigenvalues of $\mathcal{K}$.

In practice, one typically retains only some number $q < p$ of the components and views them as providing a summary of the (covariance) relationship between the variables in X. As with any type of summarization, this results in a loss of information. The extent of this loss can be gauged by the proportion of the total X variance $V := \mathrm{trace}(\mathcal{K})$ that is recovered by the principal components that are retained. In this regard, we know that

$$V = \sum_{j=1}^{p} \lambda_j$$

while the variance of the jth component is

$$\begin{aligned} \mathrm{Var}(Z_j) &= e_j^T \mathcal{K} e_j \\ &= \lambda_j. \end{aligned}$$

Thus, the jth component accounts for $100\lambda_j/V$ percentage of the total variance and $100\left(1-\sum_{k=1}^{j}\lambda_k/V\right)$ is the percentage of variability that is not accounted for by $Z_1,\ldots,Z_j$.

Principal components possess various optimality features such as the one catalogued in Theorem 1.1.1.

Theorem 1.1.1 $\mathrm{Var}(Z_j)=\max_{\{e^Te=1,e^T\mathscr{K}e_i=0,i=1,\ldots,j-1\}}\mathrm{Var}(e^TX)$.

The proof of this result is, e.g., a consequence of developments in Section 4.2. It can be interpreted as saying that the jth principle component is the linear combination of X that accounts for the maximum amount of the remaining total variance after removing the portion that was explained by $Z_1,\ldots,Z_{j-1}$.

The discussion to this point has been concerned with only the population aspects of pca. Given a random sample $x_1,\ldots,x_n$ of observations on X, we estimate $\mathscr{K}$ by the sample covariance matrix

$$\mathscr{K}_n=(n-1)^{-1}\sum_{i=1}^{n}\left(x_i-\overline{x}_n\right)\left(x_i-\overline{x}_n\right)^T \tag{1.5}$$

with

$$\overline{x}_n=n^{-1}\sum_{i=1}^{n}x_i \tag{1.6}$$

the sample mean vector. As $\mathscr{K}_n$ is positive semidefinite, it has the eigenvalue–eigenvector representation

$$\mathscr{K}_n=\sum_{j=1}^{p}\lambda_{jn}e_{jn}e_{jn}^T, \tag{1.7}$$

where the e_{in} are orthonormal and satisfy

$$e_{in}^T\mathscr{K}_ne_{jn}=\lambda_{jn}\delta_{ij}.$$

This produces the sample principle components $z_{jn}=e_{jn}^T(x-\overline{x}_n)$ for $j=1,\ldots,p$ with $x=(x_1,\ldots,x_p)^T$ and the associated scores $e_{jn}^T(x_i-\overline{x}_n), i=1,\ldots,n$ that provide sample information concerning the Z_j.

Theorems 9.1.1 and 9.1.2 of Chapter 9 can be used to deduce the large sample behavior of the sample eigenvalue–eigenvector pairs, $(\lambda_{jn},e_{jn}),j=1,\ldots,r$. The limiting distributions of $\sqrt{n}(\lambda_{jn}-\lambda_j)$ and $\sqrt{n}(e_{jn}-e_j)$ are found to be normal which provides a foundation for hypothesis testing and interval estimation.

The next step it to assume that X consists of two subsets of variables that we indicate by writing $X=(X_1^T,X_2^T)^T$, where $X_1=(X_{11},\ldots,X_{1p})^T$ and

$X_2 = (X_{21}, \ldots, X_{2q})^T$. Questions of interest now concern the relationships that may exist between X_1 and X_2. Our focus will be on those that are manifested in their covariance structure. For this purpose, we partition the covariance matrix $\mathscr{K}$ for X from (1.1) as

$$\mathscr{K} = \begin{bmatrix} \mathscr{K}_1 & \mathscr{K}_{12} \\ \mathscr{K}_{21} & \mathscr{K}_2 \end{bmatrix}. \tag{1.8}$$

Here, $\mathscr{K}_1, \mathscr{K}_2$ are the covariance matrices for X_1, X_2, respectively, and $\mathscr{K}_{12} = \mathscr{K}_{21}^T$ is sometimes called the cross-covariance matrix.

The goal is now to summarize the (cross-)covariance properties of X_1 and X_2. Analogous to the pca approach, this will be accomplished using linear combinations of the two random vectors. Specifically, we seek vectors $a_1 \in \mathbb{R}^p$ and $a_2 \in \mathbb{R}^q$ that maximize

$$\rho^2(a_1, a_2) = \frac{\mathrm{Cov}^2(a_1^T X_1, a_2^T X_2)}{\mathrm{Var}\left(a_1^T X_1\right) \mathrm{Var}\left(a_2^T X_2\right)}. \tag{1.9}$$

This optimization problem can be readily solved with the help of the singular value decomposition: e.g., Corollary 4.3.2. Assuming that X_1, X_2 contain no redundant variables, both $\mathscr{K}_1$ and $\mathscr{K}_2$ will be positive-definite with nonsingular square roots $\mathscr{K}_i^{1/2}, i = 1, 2$. This allows us to write

$$\rho^2(a_1, a_2) = \frac{\left(\tilde{a}_1^T \mathscr{R}_{12} \tilde{a}_2\right)^2}{\tilde{a}_1^T \tilde{a}_1 \tilde{a}_2^T \tilde{a}_2}, \tag{1.10}$$

where

$$\mathscr{R}_{12} = \mathscr{K}_1^{-1/2} \mathscr{K}_{12} \mathscr{K}_2^{-1/2}, \tag{1.11}$$

$\tilde{a}_1 = \mathscr{K}_1^{1/2} a_1$ and $\tilde{a}_2 = \mathscr{K}_2^{1/2} a_2$. The matrix $\mathscr{R}_{12}$ can be viewed as a multivariate analog of the linear correlation coefficient between two variables. Using the singular value decomposition in Corollary 4.3.2, we can see that (1.10) is maximized by choosing $\tilde{a}_1, \tilde{a}_2$ to be the pair of singular vectors $\tilde{a}_{11}, \tilde{a}_{21}$ that correspond to its largest singular value ρ_1. The optimal linear combinations of X_1 and X_2 are therefore provided by the vectors $a_{11} = \mathscr{K}_1^{-1/2} \tilde{a}_{11}$ and $a_{21} = \mathscr{K}_2^{-1/2} \tilde{a}_{21}$. The corresponding random variables $U_{11} = a_{11}^T X_1$ and $U_{21} = a_{21}^T X_2$ are called the first canonical variables of the X_1 and X_2 spaces, respectively. They each have unit variance and correlation ρ_1 that is referred to as the first canonical correlation.

The summarization process need not stop after the first canonical variables. If $\mathscr{K}_{12}$ has rank r, then there are actually $r - 1$ additional canonical variables that can be found: namely, for $j = 2, \ldots, r$, we have

$$U_{1j} = a_{1j}^T X_1 \tag{1.12}$$

and

$$U_{2j} = a_{2j}^T X_2, \tag{1.13}$$

where $a_{1j} = \mathscr{K}_1^{-1/2}\tilde{a}_{1j}$, $a_{2j} = \mathscr{K}_2^{-1/2}\tilde{a}_{2j}$ with $\tilde{a}_{1j}, \tilde{a}_{2j}$, the other singular vector pairs from $\mathscr{R}_{12}$ that correspond to its remaining nonzero singular values $\rho_2 \geq \cdots \geq \rho_r > 0$. For each choice of the index j, the random variable pair $\left(U_{1j}, U_{2j}\right)$ is uncorrelated with all the other canonical variable pairs and has corresponding canonical correlation ρ_j. When all this is put Together, it gives us

$$\mathscr{R}_{12} = \mathscr{K}_1^{1/2}\mathscr{A}_1\mathscr{D}\mathscr{A}_2^T\mathscr{K}_2^{1/2} \tag{1.14}$$

with

$$\mathscr{A}_i = \left[a_{i1}, \ldots, a_{ir}\right]$$

the matrix of canonical weight vectors for $X_i, i = 1, 2$, and

$$\mathscr{D} = \text{diag}\,(\rho_1, \ldots, \rho_r)$$

a diagonal matrix containing the corresponding canonical correlations.

There are various other ways of characterizing the canonical correlations and vectors. As they stem from the singular values and vectors of $\mathscr{R}_{12}$, they are the eigenvalue and eigenvectors obtained from $\mathscr{K}_1^{-1/2}\mathscr{K}_{12}\mathscr{K}_2^{-1}\mathscr{K}_{21}\mathscr{K}_1^{-1/2}$ and $\mathscr{K}_2^{-1/2}\mathscr{K}_{21}\mathscr{K}_1^{-1}\mathscr{K}_{12}\mathscr{K}_2^{-1/2}$. For example, the squared canonical correlations and canonical vectors of the X_1 space can be derived from the linear system

$$\mathscr{K}_1^{-1}\mathscr{K}_{12}\mathscr{K}_2^{-1}\mathscr{K}_{21}a_1 = \rho^2 a_1. \tag{1.15}$$

The population canonical correlations and associated canonical vectors can be estimated from the sample covariance matrix (1.5). For this purpose, one partitions $\mathscr{K}_n$ analogous to $\mathscr{K}$ and carries out the same form of singular value decomposition except using sample entities in place of $\mathscr{K}_1, \mathscr{K}_2$, and $\mathscr{K}_{12}$. The resulting sample canonical correlations have a limiting multivariate normal distribution under various conditions as detailed in Muirhead and Waternaux (1980).

The vectors X_1 and X_2 are linearly independent when $\mathscr{K}_{12}$ is a matrix of all zeros which we now recognize as being equivalent to $\rho_1 = \cdots = \rho_{\min(p,q)} = 0$. Test statistics for this and other related hypotheses can be developed from the sample canonical correlations.

Canonical correlation occupies a pervasive role in classical multivariate analysis. One place it arises naturally is in the (linear) prediction of X_1 from X_2. A best linear unbiased predictor(BLUP) is provided by the vector $\beta_0 + \beta_1 X_2$, where β_0, β_1 are, respectively, the $p \times 1$ vector and $p \times q$ matrix that minimize

$$\mathbb{E}(X_1 - b_0 - b_1 X_2)^T(X_1 - b_0 - b_1 X_2)$$

as a function of b_0, b_1. The minimizers are easily seen to be

$$\beta_0 = m_1 - \beta_1 m_2$$
$$\beta_1 = \mathscr{K}_{12}\mathscr{K}_2^{-1},$$

wherein $m_j = \mathbb{E}X_j, j = 1, 2$. From this, we recognize β_1 as the least-squares regression coefficients for the regression of X_1 on X_2. Now, premultiply (1.14) by $\mathscr{K}_1^{1/2}$ and postmultiply by $\mathscr{K}_2^{-1/2}$ to obtain

$$\beta_1 = \sum_{j=1}^{r} \rho_j \mathscr{K}_1 a_{1j} a_{2j}^T,$$

where, again, r is the rank of $\mathscr{K}_{12}$. This establishes a fundamental relationship between the best linear predictor and the canonical variables: namely,

$$\beta_0 + \beta_1 X_2 = \beta_0 + \sum_{j=1}^{r} \rho_j \mathscr{K}_1 a_{1j} U_{2j}. \tag{1.16}$$

Thus, canonical correlation lies at the heart of the linear prediction of X_1 from X_2. The converse is seen to be true as well by simply interchanging the roles of X_1 and X_2 in the above-mentioned discussion.

In finding our linear predictor for X_1, we chose a priori to restrict attention to only those that were linear functions of X_2. A related, but distinct, variation on this theme is to presume the existence of a linear model that relates X_1 to X_2 by an expression such as

$$X_1 = \beta_0 + \beta_1 X_2 + \varepsilon \tag{1.17}$$

with β_0 and β_1 of dimension $p \times 1$ and $p \times q$ as before and ε a $p \times 1$ random vector that is uncorrelated with X_2 while having zero mean and covariance matrix

$$\mathscr{K}_\varepsilon = \mathbb{E}\varepsilon\varepsilon^T.$$

In this event,

$$\mathscr{K}_{12} = \beta_1 \mathscr{K}_2, \tag{1.18}$$
$$\begin{aligned}\mathscr{K}_1 &= \beta_1 \mathscr{K}_2 \beta_1^T + \mathscr{K}_\varepsilon \\ &= \mathscr{K}_{12}\mathscr{K}_2^{-1}\mathscr{K}_{21} + \mathscr{K}_\varepsilon\end{aligned} \tag{1.19}$$

and the canonical correlations now satisfy the relationship

$$\mathscr{A}_1 \mathscr{D} \mathscr{A}_2^T = (\beta_1 \mathscr{K}_2 \beta_1^T + \mathscr{K}_\varepsilon)^{-1} \beta_1. \tag{1.20}$$

Weyl's inequality (e.g., Thompson and Freede, 1971) tells us that the eigenvalues of $\left(\beta_1 \mathscr{K}_2 \beta_1^T + \mathscr{K}_\varepsilon\right)$ are at least as large as those for $\mathscr{K}_\varepsilon$. Thus, β_1 is null if and only if $\mathscr{D}$ in (1.20) is the zero matrix.

Factor analysis is another multivariate analysis method that aims to examine the relationship between the two sets of variables. However, in this setting, only the values of the response variable X_1 are observed while X_2 is viewed as a collection of latent variables whose values represent the object of the analysis. The basic premise is that X_1 and X_2 are linearly related in that

$$X_1 = X_2 + \varepsilon \tag{1.21}$$

with ε a vector of zero mean random errors with variance–covariance matrix $\mathscr{K}_\varepsilon$ as before. The goal is now to use X_1 to predict the unobserved values of X_2.

Canonical correlation again provides the tool that allows us to make a modicum of progress toward solving the prediction problem posed by model (1.21). In this regard, we can look for a linear combination $a_1^T X_1$ that is maximally correlated with a linear combination $a_2^T X_2$ of X_2. There are some simplifications in this instance in that $\mathscr{K}_{12} = \mathscr{K}_2$ and $\mathscr{K}_1 = \mathscr{K}_2 + \mathscr{K}_\varepsilon$ lead us to consideration of the objective function

$$\begin{aligned}\mathrm{Corr}^2(a_1^T X_1, a_2^T X_2) &= \frac{\left(a_1^T \mathscr{K}_2 a_2\right)^2}{a_1^T \mathscr{K}_1 a_1 a_2^T \mathscr{K}_2 a_2} \\ &= \frac{\left(\tilde{a}_1^T \mathscr{K}_1^{-1/2} \mathscr{K}_2^{1/2} \tilde{a}_2\right)^2}{\tilde{a}_1^T \tilde{a}_1 \tilde{a}_2^T \tilde{a}_2}\end{aligned}$$

with $\tilde{a}_1 = \mathscr{K}_1^{1/2} a_1, \tilde{a}_2 = \mathscr{K}_2^{1/2} a_2$. Thus, for example, the optimal correlations ρ and choices for $\tilde{a}_1$ can be obtained as solutions of the eigenvalue problem

$$\mathscr{K}_1^{-1/2} \mathscr{K}_2 \mathscr{K}_1^{-1/2} \tilde{a}_1 = \rho^2 \tilde{a}_1.$$

A little algebra then reveals this to be equivalent to finding solutions of

$$\mathscr{K}_1 a_1 = \frac{1}{1-\rho^2} \mathscr{K}_\varepsilon a_1. \tag{1.22}$$

To proceed further, we need to impose some additional structure on X_2. The standard approach is to assume that

$$X_2 = \Phi Z \tag{1.23}$$

for some unknown $p \times r$ matrix

$$\Phi = \{\phi_{ij}\}_{i=1:p, j=1:r} = \left[\phi_1, \ldots, \phi_r\right] \tag{1.24}$$

and $Z = (Z_1, \ldots, Z_r)^T$ a vector of zero mean random variables with

$$\mathbb{E}\left[ZZ^T\right] = I.$$

The elements of Z are referred to as *factors* while the elements of Φ are called *factor loadings*. A typical identifiability constraint arising from maximum likelihood estimation is to have

$$\phi_i^T \mathscr{K}_\varepsilon^{-1} \phi_j = 0, i \neq j. \tag{1.25}$$

This has the consequence of making $\Phi^T \mathscr{K}_\varepsilon^{-1} \Phi$ a diagonal matrix, which we will subsequently presume to be the case.

When (1.23) holds

$$\begin{aligned} \mathscr{K}_1 &= \Phi\Phi^T + \mathscr{K}_\varepsilon \\ &= \mathscr{K}_\varepsilon^{1/2} \left(\mathscr{K}_\varepsilon^{-1/2} \Phi\Phi^T \mathscr{K}_\varepsilon^{-1/2} + I \right) \mathscr{K}_\varepsilon^{1/2}. \end{aligned}$$

It is readily verified that under (1.25) the matrix $\mathscr{K}_\varepsilon^{-1/2} \Phi\Phi^T \mathscr{K}_\varepsilon^{-1/2}$ has eigenvalues $\gamma_j = \phi_j^T \mathscr{K}_\varepsilon^{-1} \phi_j$ with associated eigenvectors $\mathscr{K}_\varepsilon^{-1/2} \phi_j$. However, this means that $\mathscr{K}_\varepsilon^{-1/2} \mathscr{K}_1 \mathscr{K}_\varepsilon^{-1/2}$ has eigenvalues $1 + \gamma_j$ associated with this same set of eigenvectors; i.e.,

$$\left[\mathscr{K}_\varepsilon^{-1/2} \mathscr{K}_1 \mathscr{K}_\varepsilon^{-1/2} - \left(1 + \gamma_j\right) I \right] \mathscr{K}_\varepsilon^{-1/2} \phi_j = 0 \tag{1.26}$$

or

$$\left[\mathscr{K}_1 - (1 + \gamma_j) \mathscr{K}_\varepsilon \right] \mathscr{K}_\varepsilon^{-1} \phi_j = 0. \tag{1.27}$$

Comparing this with (1.22) leads to the conclusion that $\mathscr{K}_\varepsilon^{-1} \phi_j$ is a canonical weight vectors for the X_1 space with $\gamma_j = \rho_j^2/(1 - \rho_j^2)$ obtained from its corresponding canonical correlation.

To see where these developments might take us consider the unrealistic scenario where we knew $\mathscr{K}_1$ and $\mathscr{K}_\varepsilon$ but not Φ. If that were the case, the coefficient matrix could be recovered from the canonical weight functions for the X_1 space as $\Phi = \mathscr{K}_\varepsilon \left[a_{11}, \ldots, a_{1r} \right]$. We could then predict Z via the best linear unbiased predictor: namely, the linear transformation of X_1 that minimizes the prediction error $\mathbb{E}(Z - \mathscr{L}X_1)^T (Z - \mathscr{L}X_1)$ over all possible choices for the $r \times p$ matrix $\mathscr{L}$. The minimum is attained with $\mathscr{L} = \mathrm{Cov}(Z, X_1) \mathscr{K}_1^{-1}$ giving

$$\hat{Z} = \Phi^T \left(\Phi\Phi^T + \mathscr{K}_\varepsilon \right)^{-1} X_1 \tag{1.28}$$

as the optimal predictor.

Of course, in practice, we will not know either of $\mathscr{K}_1$ or $\mathscr{K}_\varepsilon$. However, given a random sample of values from X_1, the first of these two quantities is easy to estimate using the sample variance–covariance matrix. While estimation of $\mathscr{K}_\varepsilon$ is more problematic, there are various ways this can be accomplished that produce at least acceptable initial estimators. Such estimators can be substituted into (1.27) to obtain an estimator of Φ. This, in turn, provides an update

of $\mathscr{K}_\varepsilon = \mathscr{K}_1 - \Phi\Phi^T$. The result is one possible iterative estimation algorithm that is employed in the factor analysis genre.

Our particular development of factor analysis is due to Rao (1955). A detailed treatments of this and many other factor analysis-related topics can be found in Basilevsky (1994).

A case of particular interest that can be treated with a linear model is MANOVA and its predictive analog known as discriminant analysis. To develop these ideas, we begin with the model

$$X_1 = \overline{m} + \left[m_1 - \overline{m}, \ldots, m_{q+1} - \overline{m}\right] \tilde{X}_2 + \varepsilon,$$

where $\tilde{X}_2 = \left(\tilde{X}_{21}, \ldots, \tilde{X}_{2(q+1)}\right)^T$ has a multinomial distribution with $\sum_{j=1}^{q+1} \tilde{X}_{2j} = 1$ and success probabilities $\pi_1, \ldots, \pi_{q+1}$, $m_1, \ldots, m_{q+1}$ are the X_1 mean vectors for the $q+1$ different populations, $\overline{m} = \sum_{j=1}^{q+1} \pi_j m_j$ is the grand mean and ε is a p-variate random vector with mean zero and covariance matrix $\mathscr{K}_\varepsilon$. As, $\pi_{q+1} = 1 - \sum_{j=1}^{q} \pi_j$ and $\tilde{X}_{2(q+1)} = 1 - \sum_{j=1}^{q} \tilde{X}_{2j}$, the previous model can be equivalently expressed as

$$X_1 = m_{q+1} + \left[m_1 - m_{q+1}, \ldots, m_q - m_{q+1}\right] X_2 + \varepsilon \tag{1.29}$$

with

$$X_2 := \left(\tilde{X}_{21}, \ldots, \tilde{X}_{2q}\right)^T. \tag{1.30}$$

This corresponds to (1.17) with $\beta_1 = \left[(m_1 - m_{q+1}), \ldots, (m_q - m_{q+1})\right]$ and $\beta_0 = m_{q+1}$.

To apply formula (1.20), we must calculate $\mathscr{K}_1, \mathscr{K}_{12}$ and $\mathscr{K}_2$. In this regard, first note that

$$\begin{aligned} \mathbb{E}X_2 &= (\pi_1, \ldots, \pi_q)^T \\ &=: \pi \end{aligned}$$

and

$$\begin{aligned} \mathscr{K}_2 &= \mathrm{Var}(X_2) \\ &= \mathrm{diag}\left(\pi_1, \ldots, \pi_q\right) - \pi\pi^T. \end{aligned}$$

Then, one may check that

$$\mathscr{K}_2^{-1} = \mathrm{diag}\left(\pi_1^{-1}, \ldots, \pi_q^{-1}\right) + \pi_{q+1}^{-1} 1 1^T$$

for a q-vector 1 of all unit elements. From these identities, we obtain

$$\mathscr{K}_{12} = \left[\pi_1(m_1 - \overline{m}), \ldots, \pi_q(m_q - \overline{m})\right]$$

and

$$\mathscr{K}_1 = \mathscr{K}_B + \mathscr{K}_\varepsilon$$

with

$$\mathscr{K}_B = \sum_{j=1}^{q+1} \pi_j (m_j - \overline{m})(m_j - \overline{m})^T = \mathscr{K}_{12}\mathscr{K}_2^{-1}\mathscr{K}_{21}.$$

Relations (1.18) and (1.15) can now be used to see that in this instance the canonical correlations and canonical vectors of the X_1 space are characterized by

$$\mathscr{K}_\varepsilon^{-1}\mathscr{K}_B a = \frac{\rho^2}{1-\rho^2} a. \tag{1.31}$$

Thus, all the canonical correlations are zero if and only if $\mathscr{K}_B$ is the zero matrix, which, in turn, is equivalent to the standard MANOVA null hypothesis that all of the $q+1$ populations have the same mean vector. Statistical tests for this null model can then be constructed from the sample canonical correlations.

If the mean vectors are the same for all of our $q+1$ populations, it will not generally be possible to distinguish between them on the basis of location. However, if the MANOVA null model is rejected, one can expect to have at least some success in categorizing incoming observations according to population membership. The process of doing so is often referred to as *discriminant analysis*. While there are many discrimination methods that appear in the literature, our focus here will be limited to Fisher's classical proposal. The idea is to find a linear combination, or discriminant function, $h^T X_1$ that provides the maximum separation between the populations in the sense of maximizing the ratio

$$\frac{h^T \mathscr{K}_B h}{h^T \mathscr{K}_\varepsilon h}. \tag{1.32}$$

However, this is just a variation of a problem we already encountered with pca. The solution is the largest eigenvalue λ_1 of $\mathscr{K}_\varepsilon^{-1/2}\mathscr{K}_B\mathscr{K}_\varepsilon^{-1/2}$ with the optimal discriminant function weight vector $h_1 = \mathscr{K}_\varepsilon^{-1/2}u_1$ obtained from the eigenvector u_1 that corresponds to λ_1. These quantities are characterized by

$$\mathscr{K}_\varepsilon^{-1/2}\mathscr{K}_B\mathscr{K}_\varepsilon^{-1/2}u_1 = \lambda_1 u_1$$

or, equivalently, by

$$\mathscr{K}_\varepsilon^{-1}\mathscr{K}_B h_1 = \lambda_1 h_1.$$

So, $\lambda_1 = \rho_1^2/(1-\rho_1^2)$ with ρ_1 the first canonical correlation.

Additional discriminant functions are provided by maximizing (1.32) conditional on the resulting linear combinations of variables being uncorrelated with the discriminant functions that have already been determined from this iterative process. The resulting eigenvalues will, of course, enjoy the

same relation to their corresponding canonical correlations. If we now opt to retain $r \leq \min(p, q)$ discriminant functions having weight vectors $h_1, \ldots h_r$, a new observation x is classified as being from the population whose index minimizes

$$\sum_{j=1}^{r} (h_j^T x - h_j^T m_i)^2 \tag{1.33}$$

over $i = 1, \ldots, q + 1$.

As one might expect, the relationship between Fisher's discriminant functions and canonical variables goes much deeper than just the eigenvalue connection. The equivalence of Fisher's discriminant analysis and canonical correlation is revealed by observing that

$$\begin{aligned} \delta_{ij} &= \frac{a_{1i}^T \mathscr{K}_B a_{1j}}{\rho_i^2} \\ &= \frac{h_i^T \mathscr{K}_B h_j}{\lambda_i}. \end{aligned}$$

This shows that the h_i and a_{1i} are eigenvector of $\mathscr{K}_B$ that have been adjusted to have norms λ_i and ρ_i^2. So, $a_{1i} = \rho_i h_i / \sqrt{\lambda_i}$. In particular, this means that (1.33) will produce the same classification as would be obtained using canonical variables with the criterion

$$\sum_{j=1}^{r} \frac{(a_j^T x - a_j^T m_i)^2}{1 - \rho_j^2}.$$

A typical situation with data would have us observing p dimensional random vectors

$$X_{ij}, i = 1, \ldots, q + 1, j = 1, \ldots, n_i$$

with $n = \sum_{i=1}^{q+1} n_i$. Then, estimators of $\mathscr{K}_\varepsilon, \mathscr{K}_B$ are

$$n^{-1} \sum_{i=1}^{q+1} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i \right) \left(X_{ij} - \overline{X}_i \right)^T$$

and

$$n^{-1} \sum_{i=1}^{q+1} n_i \left(\overline{X}_i - \overline{X} \right) \left(\overline{X}_i - \overline{X} \right)^T,$$

respectively, with $\overline{X}_i = n_i^{-1} \sum_{j=1}^{n_i} X_{ij}, \overline{X} = n^{-1} \sum_{i=1}^{q+1} n_i \overline{X}_i$. We then carry out classification by replacing $\mathscr{K}_\varepsilon, \mathscr{K}_B$ by their sample analogs in the previous formulation.

1.2 The path that lies ahead

The basic concepts described in Section 1.1 remain conceptually valid in the context of fda. The first challenge faced by researchers in the area lies in developing them fully and rigorously from a mathematical perspective. This is the essential precursor to the growth and maturation of inferential methodology in general and for fda in particular. One cannot estimate a "parameter" if it is undefined and it is easy to take a misstep in fda formulations that lead to exactly such a conundrum.

The theory of multivariate analysis is inextricably interwoven with matrix theory. If the observations in fda are viewed as vectors of infinite length, then we would anticipate that the infinite dimensional analog of matrices would represent the tools of the trade for advancing our understanding of this emerging field. Such entities are called linear operators and, in particular, compact operators give us the infinite dimensional extension of matrices that arise naturally in fda. Just as one cannot venture far into multivariate analysis without understanding matrix theory, one cannot expect to appreciate the mathematical aspects of fda without a thorough background in compact operators and their Hilbert–Schmidt and trace class variants.

After developing the necessary background on linear spaces in Chapter 2, we accumulate some of the essential ingredients of functional analysis and operator theory in Chapters 3–5. Chapter 3 provides a general overview that introduces linear operators and linear functionals as well as fundamental concepts such as the inverse and adjoint of an operator, nonnegative and projection operators. Compact operators are then treated in Chapter 4 where we develop both eigenvalue and singular value expansions and treat the special cases of Hilbert–Schmidt and trace class operators. This work plays an important role throughout the remainder of this text. Chapter 5 deals with perturbation theory for compact operators and provides key tools for the treatment of functional pca in Chapter 9.

Data smoothing methods tend to be a prominent aspect of most fda inferential methodology making a foray into the mathematical aspects of smoothing somewhat de rigueur for this particular treatise. Our treatment of the topic focuses on an abstract penalized smoothing problem that can be specialized to recover the spline smoothers that are most common in the fda literature. Penalized smoothing methods recur in Chapters 8 and 11 where we study their performance for estimation of certain functional parameters.

Classical statistics is concerned with inference about the distribution of a basic random variable that we are able to sample repeatedly. The same can be said for fda except that the meaning of the "random variable" phrase requires a bit of reinterpretation before it becomes relevant to that setting. What is needed is the concept of a random element of a Hilbert space. That idea along with its associated probabilistic machinery is developed in Chapter 7.

Here, we extend the concepts of a mean vector and covariance matrix to the infinite dimensional scenario and develop some asymptotic theory that is relevant for random samples of Hilbert space valued random elements.

Chapters 8–11 will provide extended, detailed illustrations of how the mathematical machinery in Chapters 2–6 can be used to address problems that arise in the fda environment. In Chapter 8, this takes the form of analysis of the large sample properties of three types of estimators of the mean element and covariance function: ones that are based on the sample mean element and covariance operator for completely observed functional data and local linear and penalized least-squares estimators for the discretely observed case. This is followed in Chapter 9 with an investigation of the asymptotic behavior of the principle components estimators that are produced by the covariance estimators introduced in Chapter 8.

Chapter 10 deals with the bivariate case where, for example, one has two stochastic processes and wishes to analyze their dependence structure. Somewhat more generally, it provides a development of abstract canonical correlation for two Hilbert space valued random elements. By specializing this theory to the fda stochastic processes context, we are then able to obtain parallels of results in Section 1.1 for functional analogs of linear prediction, regression, factor analysis, MANOVA, and discriminant analysis.

Finally, Chapter 11 deals with the important case of bivariate data having both a scalar and functional (i.e., stochastic process) response. The large sample properties of a particular penalized least-squares estimator of the regression coefficient function are investigated in this setting.