ABSTRACT - DOCTORAL DISSERTATION 'SELECTED MULTIDIMENSIONAL STATISTICAL METHODS FOR MULTIVARIATE FUNCTIONAL DATA'

ŁUKASZ WASZAK

In classical statistical methods, the studied objects are described by means of features observed at a fixed point in time. In the present work we shall consider objects described by means of functional variables. What does this mean? A functional variable X is a random variable that takes values in a certain functional space E. A functional data set is a sample $\{X_1, \ldots, X_n\}$ (also denoted by $\{X_1(t), \ldots, X_n(t)\}$, where convenient) taken from the distribution of a functional variable X. The term "functional data" was first used in the work of Ramsay and Dalzell (1991). In what follows we shall take E to be the Hilbert space of all square-integrable functions on a certain interval $[a, b]$, namely the space $L_2([a, b])$.

In this case, the functional data can be represented in the form

$$X(t) = \sum_{b=0}^{\infty} c_b \varphi_b(t),$$

where $\varphi_b(t)$ are known, fixed orthonormal functions, or in other words, the elements of an orthonormal basis $\{\varphi_0, \varphi_1, ...\}$. We note that the representation of functions using an infinite orthonormal series requires knowledge of an infinite number of coefficients c_b. Unfortunately, handling an infinite number of coefficients is not practically feasible. Therefore the function $X(t)$ is approximated using a truncated (finite) orthonormal series, otherwise called a partial sum, in the form

$$X_B(t) := \sum_{b=0}^{B} c_b \varphi_b(t) = \boldsymbol{c}'\boldsymbol{\varphi}(t) = \boldsymbol{\varphi}'(t)\boldsymbol{c}.$$

The parameter B, which is a whole number, is called the truncation point. Usually only a few coefficients of the expansion are significant, the remainder being of little importance. This leads to a significant compression of the data.

In general terms, the main statistical problems are the optimum choice of truncation point B and optimum estimation of the coefficients c_b. These issues are dealt with in Chapter 1 and 2.

The question naturally arises here as to whether functional data exist in reality. This is a valid question, since in practice the values of an observed random process $X(t)$ are always recorded at discrete time points $t_1, t_2, ..., t_J$, distributed more or less densely over the range of variation of the argument t. Hence we always end up with a time series $\{x(t_1), x(t_2), ..., x(t_J)\}$, or in other words, with a high-dimensional vector of observations. Nonetheless, there are numerous reasons to model such series as elements of a functional space, because functional data have many advantages compared with other methods of representing time series.

Firstly, they make it easy to deal with missing observations, which are an inevitable problem in many areas of research. Most methods of data analysis require complete time series. One solution is simply to remove a time series with missing values from the data set, but this operation may, and in general does, lead to a loss of information. Another possibility is to use one of the many statistical methods for prediction of missing data, but then the results will be dependent on the chosen method of interpolation. In the case of functional data, by contrast, the problem of missing observations is solved by the expression of time series in the form of a set of continuous curves.

Secondly, functional data preserve the structure of the observations in a natural way, that is, they preserve the time dependence of the observations and take account of information about every measurement.

Thirdly, the observation time points do not have to be evenly distributed in particular time series.

Fourthly, functional data allow one to avoid the "curse" of excessive dimensionality. When the total number of time points at which observations are made exceeds the number of time series being considered, most statistical methods fail to give satisfactory results, due to overparameterization. This problem is usually solved using dimension reduction techniques, such as principal component analysis. In that case, however, certain information about the space and time structure of the data may be lost. In the case of functional data the problem can be avoided, because the time series are replaced with a set of continuous curves which are not dependent on the total number of observation time points.

Although there was some earlier work relating to functional data, the symbolic starting point for statistical methods for functional data is taken to be the monograph of Ramsay and Silverman (1997), which is addressed to a wide range of readers, and in which practical considerations dominate over theory. This fine paper was supplemented by

a book, also by Ramsay and Silverman (2002), which addressed further practical aspects of the subject. In 2005 a second edition of the 1997 monograph was published. The appearance of this publication led to a flood of works on functional data analysis. Further significant publications include the books by Clarkson et al. (2005), Ferraty and Vieu (2006) (a theoretically oriented work), Bosq and Blanke (2007), Dabo-Niang and Ferraty (2008), Ramsay et al. (2009), and Ferraty and Romain (Eds.) (2011). The latest addition to the functional data literature is the book by Horvath and Kokoszka (2012), which contains a mixture of theoretical and practical aspects of functional data analysis.

Among review papers in this subject area, the following are worthy of mention: Rice (2004), Muller (2005), Gonzalez-Manteiga and Vieu (2011) (this contains a rich bibliography), Delsol et al. (2011), Febrero-Bande and Oviedo de la Fuente (2012), and Cuevas (2014).

Among the many statistical methods constructed for functional data, a prominent place is taken by three methods known jointly as dimension reduction methods. These are principal component analysis, discriminant analysis, and canonical correlation and variable analysis. The classical versions of these methods assume the studied objects to be described by multiple features. In the case of functional data, however, the work done to date has assumed that the objects are described by one-dimensional functional data. There is a visible divergence here between the assumptions made in the case of the classical methods and in the case of the methods used for functional data. In the present work, to bridge this gap, dimension reduction methods are constructed for multidimensional functional data. These methods are described in Chapters 2–4. Chapter 5 contains concrete examples of the application of the methods.