STATISTICS IN TRANSITION new series, September 2017
Vol. 18, No. 3, pp. 433–442, DOI 10. 21307

AN APPLICATION OF FUNCTIONAL MULTIVARIATE REGRESSION MODEL TO MULTICLASS CLASSIFICATION

Mirosław Krzyśko[1], **Łukasz Smaga**[2]

ABSTRACT

In this paper, the scale response functional multivariate regression model is considered. By using the basis functions representation of functional predictors and regression coefficients, this model is rewritten as a multivariate regression model. This representation of the functional multivariate regression model is used for multiclass classification for multivariate functional data. Computational experiments performed on real labelled data sets demonstrate the effectiveness of the proposed method for classification for functional data.

Key words: functional data analysis, multi-label classification problem, multivariate functional data, regression model.

1. Introduction

In recent decades, the analysis of data given as functions or curves has become a very popular branch of statistics. In the literature, such data are called functional data and have a broad perspective of applications, for example, in economics and medicine. The aim of functional data analysis (FDA) is to develop methods for analysing functional data. For instance, the books Ramsay and Silverman (2005), Ferraty and Vieu (2006), Horváth and Kokoszka (2012) and Zhang (2013), and the references therein, offer a broad perspective of such methods.

Methods for analysing multivariate functional data (e.g. vectors of functions) are of particular interest. Some solutions of such problems as analysis of variance, canonical correlation analysis, classification, cluster analysis, linear regression and prediction, or principal component analysis are known in the literature. For example, we refer to the following papers by Górecki and Smaga (2017), Górecki et al. (2016), Górecki et al. (2015), Jacques and Preda (2014), Collazos et al. (2016) and Berrendero et al. (2011), respectively, and the references therein.

[1]Inter-Faculty Department of Mathematics and Statistics, The President Stanisław Wojciechowski State University of Applied Sciences in Kalisz, Poland. E-mail: mkrzysko@amu.edu.pl.

[2]Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland. E-mail: ls@amu.edu.pl.

This paper discusses the multiclass classification problem for multivariate functional data. The classifiers are constructed based on the scale response functional multivariate regression model and basis functions representation of functional predictors and coefficients. The presented results may be seen as extensions of those given in Górecki et al. (2015) from binary to multi-label case.

The rest of the paper is organized as follows. We first (Section 2) construct and rewrite (using the basis functions representation of predictors and coefficients) the scale response functional multivariate regression model. We consider two versions of this model, i.e. with and without intercepts. In Section 3, we apply these results to the multi-label classification problem for multivariate functional data. Section 4 contains the description of computational experiments for comparison of the proposed classifiers and a discussion of their results. We conclude in Section 5 with discussion of possible improvement of performance of the proposed method.

2. Functional multivariate regression model

In this Section, we consider the scalar response functional multivariate regression model, which can be seen as an extension of the one-dimensional model studied, for example, in Horváth and Kokoszka (2012).

Let $L_2(T)$ denote the Hilbert space of square integrable functions over $T=[a,b]$. Assume that we have measured p (scalar) responses $Y_1,\ldots,Y_p$ and the same set of k (functional) predictors $x_1(t),\ldots,x_k(t)$ belonging to $L_2(T)$ on each sample unit. Moreover, suppose that the responses follow the scalar regression models, i.e.

$$Y_j=\sum_{i=1}^{k}\int_T x_i(t)\xi_{ji}(t)dt+e_j,\qquad j=1,\ldots,p,$$

where $\xi_{ji}\in L_2(T)$ are the unknown functional coefficients and e_j are the random errors such that $\mathbf{e}^\top=[e_1,\ldots,e_p]$ has zero expectation and covariance matrix $\mathbf{\Sigma}$. When we have a sample of N independent observations $\mathbf{Y}_1,\ldots,\mathbf{Y}_N$ of the vector $[Y_1,\ldots,Y_p]^\top$, the scalar response functional multivariate regression model is formulated as follows:

$$\mathbf{Y}=\int_T \mathbf{X}(t)\mathbf{\Xi}(t)dt+\mathbf{E}, \tag{1}$$

where

$$\mathbf{Y}=\begin{bmatrix}\mathbf{Y}_1^\top\\ \vdots\\ \mathbf{Y}_N^\top\end{bmatrix},\ \mathbf{X}(t)=\begin{bmatrix}\mathbf{x}_1^\top(t)\\ \vdots\\ \mathbf{x}_N^\top(t)\end{bmatrix},\ \mathbf{\Xi}(t)=\left[\boldsymbol{\xi}_1(t),\ldots,\boldsymbol{\xi}_p(t)\right],\ \mathbf{E}=\begin{bmatrix}\mathbf{e}_1^\top\\ \vdots\\ \mathbf{e}_N^\top\end{bmatrix}, \tag{2}$$

and $\mathbf{x}_i^\top(t) = [x_{i1}(t), \ldots, x_{ik}(t)]$, $i = 1, \ldots, N$, $\boldsymbol{\xi}_j^\top(t) = \left[\xi_{j1}(t), \ldots, \xi_{jk}(t)\right]$, $j = 1, \ldots, p$.

To handle the model (1), we assume that the predictors and functional coefficients can be represented by a finite number of orthonormal basis functions $(\varphi_{mn}(t))_{n=0}^{\infty}$, $m = 1, \ldots, k$ in $L_2(T)$, i.e. for $i = 1, \ldots, N$ and $j = 1, \ldots, p$

$$x_{im}(t) = \sum_{n=0}^{B_m} c_{imn}\varphi_{mn}(t), \quad \xi_{jm}(t) = \sum_{n=0}^{B_m} d_{jmn}\varphi_{mn}(t), \tag{3}$$

where c_{imn} and d_{jmn} are the unknown coefficients. More precisely, c_{imn} are the random variables with finite variance (see Ramsay and Silverman, 2005). To estimate the coefficients c_{imn} (for each predictor separately), the least squares method can be used (see, for instance, Krzyśko and Waszak, 2013). The selection method of the values B_m may depend on the aim of the research. For example, when we want to obtain the best fit, the Bayesian information criterion should perhaps be used (see Shmueli, 2010). Different bases can be used for different predictors.

For easier presentation of our results, we represent the equations (3) in matrix notation. Let

$$\boldsymbol{\Phi}(t) = \begin{bmatrix} \boldsymbol{\varphi}_1^\top(t) & \mathbf{0}_{B_2+1}^\top & \cdots & \mathbf{0}_{B_k+1}^\top \\ \mathbf{0}_{B_1+1}^\top & \boldsymbol{\varphi}_2^\top(t) & \cdots & \mathbf{0}_{B_k+1}^\top \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{B_1+1}^\top & \mathbf{0}_{B_2+1}^\top & \cdots & \boldsymbol{\varphi}_k^\top(t) \end{bmatrix},$$

where $\boldsymbol{\varphi}_l^\top(t) = [\varphi_{l0}(t), \ldots, \varphi_{lB_l}(t)]$ for $l = 1, \ldots, k$ and $\mathbf{0}_n$ is an $n \times 1$ vector of zeros. Then, the equations given in (3) can be rewritten as follows:

$$\mathbf{x}_i(t) = \boldsymbol{\Phi}(t)\mathbf{c}_i, \qquad \boldsymbol{\xi}_j(t) = \boldsymbol{\Phi}(t)\mathbf{d}_j \tag{4}$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, p$, where $\mathbf{c}_i^\top = [c_{i10}, \ldots, c_{i1B_1}, \ldots, c_{ik0}, \ldots, c_{ikB_k}]$ and $\mathbf{d}_j^\top = [d_{j10}, \ldots, d_{j1B_1}, \ldots, d_{jk0}, \ldots, d_{jkB_k}]$.

By (4), for $i = 1, \ldots, N$ and $j = 1, \ldots, p$, we have

$$\begin{aligned} \int_T \mathbf{x}_i^\top(t)\boldsymbol{\xi}_j(t)dt &= \int_T \mathbf{c}_i^\top \boldsymbol{\Phi}^\top(t)\boldsymbol{\Phi}(t)\mathbf{d}_j dt \\ &= \mathbf{c}_i^\top \int_T \boldsymbol{\Phi}^\top(t)\boldsymbol{\Phi}(t)dt\ \mathbf{d}_j \\ &= \mathbf{c}_i^\top \mathbf{d}_j, \end{aligned} \tag{5}$$

since the bases $(\varphi_{mn}(t))_{n=0}^{\infty}$, $m = 1, \ldots, k$, are orthonormal, i.e. $\int_T \boldsymbol{\Phi}^\top(t)\boldsymbol{\Phi}(t)dt$ is the identity matrix of size $\sum_{l=1}^{k} B_l + k$. From (2), it follows that

$$\int_T \mathbf{X}(t)\boldsymbol{\Xi}(t)dt = \begin{bmatrix} \int_T \mathbf{x}_1^\top(t)\boldsymbol{\xi}_1(t)dt & \dots & \int_T \mathbf{x}_1^\top(t)\boldsymbol{\xi}_p(t)dt \\ \vdots & \ddots & \vdots \\ \int_T \mathbf{x}_N^\top(t)\boldsymbol{\xi}_1(t)dt & \dots & \int_T \mathbf{x}_N^\top(t)\boldsymbol{\xi}_p(t)dt \end{bmatrix}.$$

Thus, by (5), we obtain

$$\begin{aligned} \int_T \mathbf{X}(t)\boldsymbol{\Xi}(t)dt &= \begin{bmatrix} \mathbf{c}_1^\top\mathbf{d}_1 & \dots & \mathbf{c}_1^\top\mathbf{d}_p \\ \vdots & \ddots & \vdots \\ \mathbf{c}_N^\top\mathbf{d}_1 & \dots & \mathbf{c}_N^\top\mathbf{d}_p \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{c}_1^\top \\ \vdots \\ \mathbf{c}_N^\top \end{bmatrix} [\mathbf{d}_1, \dots, \mathbf{d}_p] \\ &= \mathbf{CD}. \end{aligned}$$

Hence, the model (1) can be rewritten as

$$\mathbf{Y} = \mathbf{CD} + \mathbf{E}, \tag{6}$$

which is the multivariate regression model with the parameter matrix $\mathbf{D}$. Therefore, the problems connected with the functional multivariate regression model (1) (e.g. estimation of $\boldsymbol{\Xi}(t)$) can be replaced by the ones in the multivariate regression model (6). In the next Section, this relation is used for multiclass classification for multivariate functional data. Other results of such type and their usage are presented, for instance, in Kayano and Konishi (2009), Matsui and Konishi (2011), Matsui (2014), Górecki et al. (2015) and Collazos et al. (2016).

In the model (1), the intercepts were not considered. However, adding them to the model may improve the classification procedure based on it as we will see in Section 4. Thus, we extend the above results to the functional multivariate regression model with intercepts. Now, the scalar responses Y_j are modelled by the following regression models

$$Y_j = \xi_{j0} + \sum_{i=1}^{k} \int_T x_i(t)\xi_{ji}(t)dt + e_j, \qquad j = 1, \dots, p,$$

where ξ_{j0} are the (unknown) intercepts, and further the model (1) is replaced by

$$\mathbf{Y} = \boldsymbol{\Xi}_0 + \int_T \mathbf{X}(t)\boldsymbol{\Xi}(t)dt + \mathbf{E}, \tag{7}$$

where $\boldsymbol{\Xi}_0 = [\xi_{10}\mathbf{1}_N, \ldots, \xi_{p0}\mathbf{1}_N]$ and $\mathbf{1}_N$ is the $N \times 1$ vector of ones. Using the basis functions representation of predictors and functional coefficients given in (4), the model (7) can be rewritten as

$$\mathbf{Y} = [\mathbf{1}_N, \mathbf{C}] \begin{bmatrix} \boldsymbol{\xi}_0^\top \\ \mathbf{D} \end{bmatrix} + \mathbf{E} = \mathbf{C}_* \mathbf{D}_* + \mathbf{E}, \tag{8}$$

where $\boldsymbol{\xi}_0^\top = [\xi_{10}, \ldots, \xi_{p0}]$. Thus, the parameter matrix has one row more than in the earlier model.

3. Multiclass classification for functional data

In this Section, we investigate the multi-label classification problem for multivariate functional data by using the functional multivariate regression model considered in Section 2. More general information and results on classification problems based on regression models can be found in Krzyśko et al. (2008).

Assume that there are $K \geq 2$ populations and the objects are characterized by k features, which are given as functions in the space $L_2(T)$. Let

$$\mathbf{x}_i^\top(t) = [x_{i1}(t), \ldots, x_{ik}(t)], \quad i = 1, \ldots, N$$

be a sample from these populations. Each vector of functions $\mathbf{x}_i(t)$ is accompanied by the group label given by the $K \times 1$ vector

$$\mathbf{Y}_i^\top = [0, \ldots, 0, 1, 0, \ldots, 0]$$

with 1 in the lth place when the ith observation belongs to lth population.

In a classification problem, one wants to determine a procedure by which a given object can be assigned to one of K populations. For this purpose, the relation between vectors $\mathbf{x}_i(t)$ and $\mathbf{Y}_i$, $i = 1, \ldots, N$ is described by the scalar response functional multivariate regression model (1) or (7). Here we use the rewritten form (6) or (8) of it. The parameter matrices $\mathbf{D}$ and $\mathbf{D}_*$ in the models (6) and (8) can be estimated by the least squares method. The obtained estimators are of the form

$$\hat{\mathbf{D}} = (\mathbf{C}^\top \mathbf{C})^+ \mathbf{C}^\top \mathbf{Y}, \quad \hat{\mathbf{D}}_* = (\mathbf{C}_*^\top \mathbf{C}_*)^+ \mathbf{C}_*^\top \mathbf{Y},$$

where $\mathbf{M}^+$ is the Moore-Penrose pseudoinverse of the matrix $\mathbf{M}$. Then, the predicted

matrix is given by the formula

$$\hat{\mathbf{Y}} = \begin{cases} \mathbf{C}\hat{\mathbf{D}} = \mathbf{C}(\mathbf{C}^\top\mathbf{C})^+\mathbf{C}^\top\mathbf{Y}, & \text{for model (1),} \\ \mathbf{C}_*\hat{\mathbf{D}}_* = \mathbf{C}_*(\mathbf{C}_*^\top\mathbf{C}_*)^+\mathbf{C}_*^\top\mathbf{Y}, & \text{for model (7).} \end{cases}$$

To obtain the prediction for a new observation $\mathbf{x}_{\text{new}}(t)$, first its components have to be represented by a finite number of orthonormal basis functions, as it was described in Section 2, i.e.

$$\mathbf{x}_{\text{new}}(t) = \mathbf{\Phi}(t)\mathbf{c}_{\text{new}}.$$

Hence, the predicted vector $\hat{\mathbf{Y}}(\mathbf{x}_{\text{new}})$ for the new observation is of the form

$$\hat{\mathbf{Y}}(\mathbf{x}_{\text{new}})^\top = [\hat{Y}_1(\mathbf{x}_{\text{new}}), \ldots, \hat{Y}_K(\mathbf{x}_{\text{new}})] = \begin{cases} \mathbf{c}_{\text{new}}^\top\hat{\mathbf{D}}, & \text{for model (1),} \\ \left[1, \mathbf{c}_{\text{new}}^\top\right]\hat{\mathbf{D}}_*, & \text{for model (7).} \end{cases}$$

The lth component of the vector $\hat{\mathbf{Y}}(\mathbf{x}_{\text{new}})$ is the estimated value of the posterior probability of belonging to the lth population. Unfortunately, the components of this vector may not belong to the interval $[0,1]$. However, this may not matter if we get good predictions. Moreover, it can be shown that the sum of the components of $\hat{\mathbf{Y}}(\mathbf{x}_{\text{new}})$ is equal to one (see, for example, Krzyśko et al., 2008). Therefore, the classifier is given by the following formula

$$\hat{d}(\mathbf{x}_{\text{new}}) = \arg\max_{l=1,\ldots,K} \hat{Y}_l(\mathbf{x}_{\text{new}}). \tag{9}$$

In practice, the performance of this simple classifier may be satisfactory, as indicated by the real data examples of the next Section.

4. Computational experiments

In this Section, the accuracy of the proposed classifiers is examined using six real labelled data sets. All computational experiments were performed with R environment (R Development Core Team, 2015), and the codes are available from the authors.

The experiments were carried out on the following data sets: Arabic digits, Australian language, Character trajectories, Japanese vowels, ECG and Wafer. Table 1 shows the information on them. The first four data sets originate from Bache and Lichman (2013), and the remaining ones from Olszewski (2001). The discrete functional samples in each data set are of different lengths (see Table 1). For this reason, all discrete functional variables in a given data set were extended to the same length of the longest one by the method described and used, for example, in Górecki et al.

(2015) (see also Rodriguez et al., 2005).

Table 1. Summary of data sets

Data sets	k	N	K	Max length	Min length
Arabic digits	13	8800	10	93	4
Australian language	22	2565	95	136	45
Character trajectories	3	2858	20	205	109
ECG	2	200	2	152	39
Japanese vowels	12	640	9	29	7
Wafer	6	1194	2	198	104

To obtain the basis functions representation (3) of the observations, the orthonormal Fourier basis and the least squares method of estimating the coefficients were used (see Krzyśko and Waszak, 2013). As we noted in Section 2, the quantities B_m, $m = 1,\ldots,k$ in (3) can be chosen depending on the problem at hand. In our classification problem, we choose these quantities which minimize the classification error. In computational experiments, since we used the Fourier basis, we took into account $B_1 = \cdots = B_k = B$ and $B \in \{3,5,\ldots,I\}$, where I is the greatest odd number less than or equal to the number of design time points of a given data set, i.e. points on which functions are observed in practice.

The classifiers (9) based on models (1) and (7) were used for the classification process. The classification error rates are calculated by 10-fold cross-validation method. Figure 1 and Table 2 present the results. Observe that both classifiers give very good classification results for the data sets Arabic digits, Character trajectories, Japanese vowels and Wafer. However, the classification error rates are not so satisfactory for the data sets Australian language and ECG. This suggests that they are difficult to recognize.

Table 2. The smallest 10-fold cross-validation error rates (as percentages) and B's for which they are achieved by using classifiers (9) based on models (1) and (7)

	Model (1)		Model (7)	
Data sets	10CV error	B	10CV error	B
Arabic digits	4.35	15	4.01	27 or 33
Australian language	13.1	11	13.3	11
Character trajectories	1.23	175	1.19	127 or 171
ECG	11.5	31	11.5	31
Japanese vowels	1.88	5	1.41	5
Wafer	0.50	25 or 27 or 39	0.50	25 or 27 or 39

It seems that the classifier based on model (7) with intercepts performs at least as good as or even better than that based on model (1) without intercepts in most

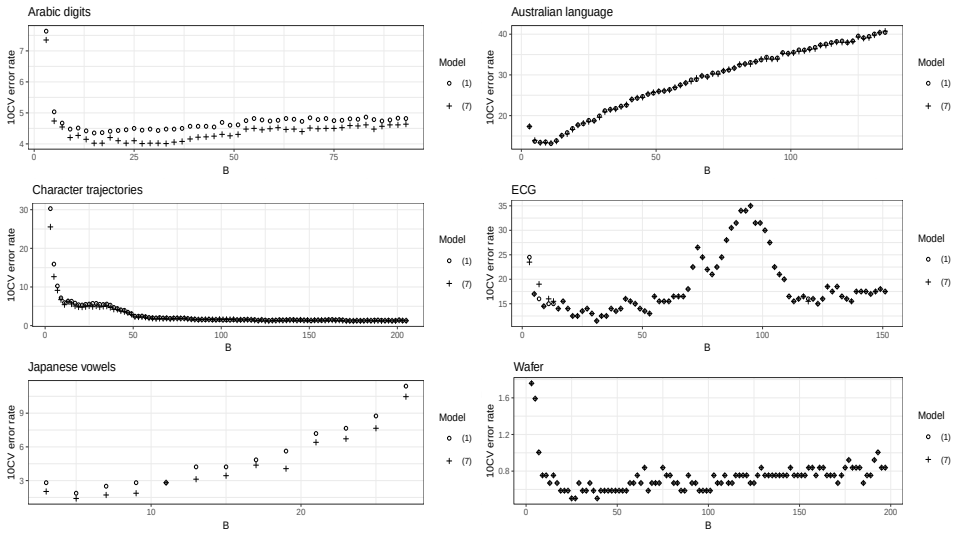

Figure 1: 10-fold cross-validation error rates (as percentages) for different values of B by using classifiers (9) based on models (1) and (7)

situations. However, for the data set Australian language, the smallest classification error rate of the method based on model (1) is slightly smaller than these of the second one (see Table 2). Therefore, for a given practical problem, both models may be examined, and we choose the one which minimizes the classification error.

From Figure 1 and Table 2, we see that the 10-fold cross-validation error rates behave differently for different values of B. In some cases, the best classification results are obtained for small values of B (e.g. for Japanese vowels) while in others for greater ones (e.g. for Character trajectories). Moreover, the values of B, for which the smallest classification error rates were achieved, may not be the same for classifiers based on models (1) and (7).

5. Conclusions

This paper discusses the construction of the scale response functional multivariate regression model and its application to multiclass classification problem for multivariate functional data. The computational experiments based on real labelled data sets suggest good performance of the proposed classification methods. From models with and without intercepts, the first one seems to be preferable.

For simplicity, in our real data examples, we used the orthonormal Fourier basis and equal lengths of basis functions representation of the observations, i.e. equal B_m's in (3). However, in practice, the performance of the considered classifiers may

be improved by using more appropriate orthonormal bases to different features and more varied values of B_m in (3).

REFERENCES

BACHE, K., LICHMAN, M., (2013). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science (http://archive.ics.uci.edu/ml).

BERRENDERO, J. R., JUSTEL, A., SVARC, M., (2011). Principal Components for Multivariate Functional Data. Computational Statistics & Data Analysis, 55, 2619–2634.

COLLAZOS, J. A. A., DIAS, R., ZAMBOM, A. Z., (2016). Consistent Variable Selection for Functional Regression Models. Journal of Multivariate Analysis, 146, 63–71.

FERRATY, F., VIEU, P., (2006). Nonparametric Functional Data Analysis: Theory and Practice, New York: Springer.

GÓRECKI, T., KRZYŚKO, M., WASZAK, Ł., WOŁYŃSKI, W., (2016). Selected Statistical Methods of Data Analysis for Multivariate Functional Data. Statistical Papers (Accepted) doi:10.1007/s00362-016-0757-8.

GÓRECKI, T., KRZYŚKO, M., WOŁYŃSKI, W., (2015). Classification Problem Based on Regression Models for Multidimensional Functional Data. Statistics in Transition new series, 16, 97–110.

GÓRECKI, T., SMAGA, Ł., (2017). Multivariate Analysis of Variance for Functional Data. Journal of Applied Statistics, 44, 2172–2189.

HORVÁTH, L., KOKOSZKA, P., (2012). Inference for Functional Data with Applications, New York: Springer.

JACQUES, J., PREDA, C., (2014). Model-Based Clustering for Multivariate Functional Data. Computational Statistics & Data Analysis, 71, 92–106.

KAYANO, M., KONISHI, S., (2009). Functional Principal Component Analysis via Regularized Gaussian Basis Expansions and its Application to Unbalanced Data. Journal of Statistical Planning and Inference, 139, 2388–2398.

KRZYŚKO, M., WASZAK, Ł., (2013). Canonical Correlation Analysis for Functional Data. Biometrical Letters, 50, 95–105.

KRZYŚKO, M., WOŁYŃSKI, W., GÓRECKI, T., SKORZYBUT, M., (2008). Learning Systems, Warsaw: WNT (in Polish).

MATSUI, H., (2014). Variable and Boundary Selection for Functional Data via Multiclass Logistic Regression Modeling. Computational Statistics & Data Analysis, 78, 176–185.

MATSUI, H., KONISHI, S., (2011). Variable Selection for Functional Regression Models via the L_1 Regularization. Computational Statistics & Data Analysis, 55, 3304–3310.

OLSZEWSKI, R. T., (2001). Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA (http://www.cs.cmu.edu/bobski).

RAMSAY, J. O., SILVERMAN, B. W., (2005). Functional Data Analysis, Second Edition, New York: Springer.

R DEVELOPMENT CORE TEAM, (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (https://www.R-project.org/).

RODRIGUEZ, J. J., ALONSO, C. J., MAESTRO, J. A., (2005). Support Vector Machines of Interval Based Features for Time Series Classification. Knowledge-Based Systems, 18, 171–178.

SHMUELI, G., (2010). To Explain or to Predict? Statistical Science, 25, 289–310.

ZHANG, J.-T., (2013). Analysis of Variance for Functional Data, London: Chapman & Hall.