Variable Selection for Classification of Multivariate Functional Data

Tomasz Górecki, Mirosław Krzyśko, and Waldemar Wołyński

Abstract New variable selection method is considered in the setting of classification with multivariate functional data (Ramsay and Silverman, Functional data analysis, 2005). The variable selection is a dimensionality reduction method which leads to replace the whole vector process, with a low-dimensional vector still giving a comparable classification error. The various classifiers appropriate for functional data are used. The proposed variable selection method is based on functional distance covariance (Székely et al. Ann Appl Stat 3(4):1236–1265, 2009; Stat Probab Lett 82(12):2278–2282, 2012). and is a modification of the procedure given by Kong et al. (Stat Med 34:1708–1720, 2015). The proposed methodology is illustrated on real data example.

1 Introduction

Much attention has been paid in recent years to methods for representing data as functions or curves. Such data are known in the literature as functional data (Ramsay and Silverman 2005; Horváth and Kokoszka 2012). Applications of functional data can be found in various fields, including medicine, economics, meteorology, and many others. In many applications there is a need to use statistical methods for objects characterized by multiple variables observed at many time points (doubly multivariate data). Such data are called multivariate functional data. In this paper we focus on the classification problem for multivariate functional data. In many cases, in the classification procedures, number of predictors p is much greater than the sample size n. It is thus natural to assume that only a small number of predictors are relevant to response Y.

Various basic classification methods have also been adapted to functional data, such as linear discriminant analysis (Hastie et al. 1995), logistic regression (Rossi

T. Górecki (✉) · M. Krzyśko · W. Wołyński
Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland
e-mail: tomasz.gorecki@amu.edu.pl; mkrzysko@amu.edu.pl; wolynski@amu.edu.pl

T. Imaizumi et al. (eds.), *Advanced Studies in Classification and Data Science*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-981-15-3311-2_17

et al. 2002), penalized optimal scoring (Ando 2009), kNN (Ferraty and Vieu 2003), SVM (Rossi and Villa 2006), and neural networks (Rossi et al. 2005). Moreover, the combining of classifiers has been extended to functional data (Ferraty and Vieu 2009). Górecki et al. (2016) adapted multivariate regression models to the classification of multivariate functional data.

Székely et al. (2007), Székely and Rizzo (2009), Székely and Rizzo (2012, 2013) defined the measures of dependence between random vectors: the distance covariance (dCov) coefficient and the distance correlation (dCor) coefficient. These authors showed that for all random variables with finite first moments, the dCor coefficient generalizes the idea of correlation in two ways. Firstly, this coefficient can be applied when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are of any dimensions and not only for the simple case where $p = q = 1$. Secondly, the dCor coefficient is equal to zero, if and only if there is independence between the random vectors. Indeed, a correlation coefficient measures linear relationships and can be equal to 0 even when the variables are related. Based on the idea of the distance covariance between two random vectors, we introduced the functional distance correlation between two random processes. We select a set of important predictors with large value of functional distance covariance. Our selection procedure is a modification of the procedure given by Kong et al. (2015). Entirely different approach to the variable selection in functional data classification is presented by Berrendero et al. (2016). It is clear that variable selection has, at least, an advantage when compared with other dimension reduction methods (functional principal component analysis (FPCA), see Górecki et al. 2014; Jacques and Preda 2014, functional partial least squares (FPLS) methodology, see Delaigle and Haal 2012, and other methods) based on general projections: the output of any variable selection method is always directly interpretable in terms of the original variables, provided that the required number d of selected variables is not too large.

The rest of this paper is organized as follows. In Sect. 2 we present the classification procedures used through the paper. In Sect. 3 we present the problem of representing functional data by orthonormal basis functions. In Sect. 4, we define a functional distance covariance and distance correlation. In Sect. 5 we propose a variable selection procedure based on the functional distance covariance. In Sect. 6 we illustrate the proposed methodology through a real data example. We conclude in Sect. 7.

2 Classifiers

The classification problem involves determining a procedure by which a given object can be assigned to one of q populations based on observation of p features of that object.

The object being classified can be described by a random pair $(\boldsymbol{X}, Y)$, where $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)' \in \mathbf{R}^p$ and $Y \in \{1, \ldots, q\}$. An automated classifier can be viewed as a method of estimating the posterior probability of membership in groups. For a given $\boldsymbol{X}$, a reasonable strategy is to assign $\boldsymbol{X}$ to that class with the highest posterior probability. This strategy is called the Bayes' rule classifier.

2.1 Linear and Quadratic Discriminant Classifiers

Now we make the Bayes' rule classifier more specific by the assumption that all multivariate probability densities are multivariate normal having arbitrary mean vectors and a common covariance matrix. We shall call this model the linear discriminant classifier (LDC). Assuming that class-covariance matrices are different, we obtain quadratic discriminant classifier (QDC).

2.2 Naive Bayes Classifier

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with independence assumptions. When dealing with continuous data, a typical assumption is that the continuous values associated with each class are distributed according to a one-dimensional normal distribution or we estimate density by kernel method.

2.3 k-Nearest Neighbor Classifier

Most often we do not have sufficient knowledge of the underlying distributions. One of the important nonparametric classifiers is a k-nearest neighbor classifier (kNN classifier). Objects are assigned to the class having the majority in the k nearest neighbors in the training set.

2.4 Multinomial Logistic Regression

It is a classification method that generalizes logistic regression to multiclass problem using one vs. all approach.

3 Functional Data

We now assume that the object being classified is described by a p-dimensional random process $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)' \in L_2^p(I)$, where $L_2(I)$ is the Hilbert space of square-integrable functions, and $\mathrm{E}(\boldsymbol{X}) = \mathbf{0}$.

Moreover, assume that the kth component of the vector $\boldsymbol{X}$ can be represented by a finite number of orthonormal basis functions $\{\varphi_b\}$

$$X_k(t) = \sum_{b=0}^{B_k} \alpha_{kb}\varphi_b(t),\ t \in I,\ k = 1, \ldots, p, \tag{1}$$

where $\alpha_{k0}, \alpha_{k1}, \ldots, \alpha_{kB_k}$ are the unknown coefficients.

Let $\boldsymbol{\alpha} = (\alpha_{10}, \ldots, \alpha_{1B_1}, \ldots, \alpha_{p0}, \ldots, \alpha_{pB_p})'$

and

$$\boldsymbol{\Phi}(t) = \begin{bmatrix} \boldsymbol{\varphi}_1'(t) & \mathbf{0}' & \ldots & \mathbf{0}' \\ \mathbf{0}' & \boldsymbol{\varphi}_2'(t) & \ldots & \mathbf{0}' \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{0}' & \mathbf{0}' & \ldots & \boldsymbol{\varphi}_p'(t) \end{bmatrix},$$

where $\boldsymbol{\varphi}_k(t) = (\varphi_0(t), \ldots, \varphi_{B_k}(t))', k = 1, \ldots, p$.

Using the above matrix notation, process $\boldsymbol{X}$ can be represented as:

$$\boldsymbol{X}(t) = \boldsymbol{\Phi}(t)\boldsymbol{\alpha}, \tag{2}$$

where $\mathrm{E}(\boldsymbol{\alpha}) = \mathbf{0}$. This means that the realizations of a process $\boldsymbol{X}$ are in finite-dimensional subspace of $L_2^p(I)$. We will denote this subspace by $\mathscr{L}_2^p(I)$.

We can estimate the vector $\boldsymbol{\alpha}$ on the basis of n independent realizations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ of the random process $\boldsymbol{X}$ (functional data). We will denote this estimator by $\hat{\boldsymbol{\alpha}}$.

Typically data are recorded at discrete moments in time. Let x_{kj} denote an observed value of the feature X_k, $k = 1, 2, \ldots, p$ at the jth time point t_j, where $j = 1, 2, \ldots, J$. Then our data consist of the pJ pairs (t_j, x_{kj}). These discrete data can be smoothed by continuous functions x_k and I is a compact set such that $t_j \in I$, for $j = 1, \ldots, J$.

Details of the process of transformation of discrete data to functional data can be found in Ramsay and Silverman (2005) or in Górecki et al. (2014).

4 Functional Distance Covariance and Distance Correlation

For jointly distributed random process $\boldsymbol{X} \in L_2^p(I)$ and random vector $\boldsymbol{Y} \in \mathbb{R}^q$, let

$$f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{l},\boldsymbol{m}) = \mathrm{E}\{\exp[i < \boldsymbol{l}, \boldsymbol{X} >_p + i < \boldsymbol{m}, \boldsymbol{Y} >_q]\}$$

be the joint characteristic function of $(\boldsymbol{X}, \boldsymbol{Y})$, where

$$< \boldsymbol{l}, \boldsymbol{X} >_p = \int_I \boldsymbol{l}'(t)\boldsymbol{X}(t)dt$$

and

$$< \boldsymbol{m}, \boldsymbol{Y} >_q = \boldsymbol{m}'\boldsymbol{Y}.$$

Moreover, we define the marginal characteristic functions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ as follows: $f_{\boldsymbol{X}}(\boldsymbol{l}) = f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{l}, \boldsymbol{0})$ and $f_{\boldsymbol{Y}}(\boldsymbol{m}) = f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{0}, \boldsymbol{m})$.

Here, for generality, we assume that $\boldsymbol{Y} \in \mathbb{R}^q$, although the label Y in the classification problem is a random variable, with values in $\{1, \ldots, q\}$. Label Y has to be transformed into the label vector $\boldsymbol{Y} = (Y_1, \ldots, Y_q)'$, where $Y_i = 1$ for $i = 1, \ldots, q$ if $\boldsymbol{X}$ belongs to class i, and 0 otherwise.

Now, let us assume that $\boldsymbol{X} \in \mathscr{L}_2^p(I)$. Then the process $\boldsymbol{X}$ can be represented as:

$$\boldsymbol{X}(t) = \boldsymbol{\Phi}(t)\boldsymbol{\alpha}, \tag{3}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{K+p}$ and $K = B_1 + \cdots + B_p$.

In this case, we may assume (Ramsay and Silverman 2005) that the vector weight function $\boldsymbol{l}$ and the process $\boldsymbol{X}$ are in the same space, i.e. the function $\boldsymbol{l}$ can be written in the form

$$\boldsymbol{l}(t) = \boldsymbol{\Phi}(t)\boldsymbol{\lambda}, \tag{4}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^{K+p}$.

Hence

$$< \boldsymbol{l}, \boldsymbol{X} >_p = \int_I \boldsymbol{l}'(t)\boldsymbol{X}(t)dt = \boldsymbol{\lambda}'[\int_I \boldsymbol{\Phi}'(t)\boldsymbol{\Phi}(t)dt]\boldsymbol{\alpha} = \boldsymbol{\lambda}'\boldsymbol{\alpha},$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ are vectors occurring in the representations (3) and (4) of process $\boldsymbol{X}$ and function $\boldsymbol{l}$, and

$$f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{l},\boldsymbol{m}) = \mathrm{E}\{\exp[i\boldsymbol{\lambda}'\boldsymbol{\alpha} + i\boldsymbol{m}'\boldsymbol{Y}]\} = f_{\boldsymbol{\alpha},\boldsymbol{Y}}(\boldsymbol{\lambda},\boldsymbol{m}),$$

where $f_{\boldsymbol{\alpha},\boldsymbol{Y}}(\boldsymbol{\lambda},\boldsymbol{m})$ is the joint characteristic function of the pair of random vectors $(\boldsymbol{\alpha}, \boldsymbol{Y})$.

On the basis of the idea of distance covariance between two random vectors (Székely et al. 2007), we can introduce functional distance covariance between random processes $\boldsymbol{X}$ and random vector $\boldsymbol{Y}$ as a nonnegative number $\nu_{\boldsymbol{X},\boldsymbol{Y}}$ defined by

$$\nu_{\boldsymbol{X},\boldsymbol{Y}} = \nu_{\boldsymbol{\alpha},\boldsymbol{Y}},$$

where

$$\nu^2_{\boldsymbol{\alpha},\boldsymbol{Y}} = \frac{1}{C_{K+p}C_q}\int_{\mathbb{R}^{K+p+q}} \frac{|f_{\boldsymbol{\alpha},\boldsymbol{Y}}(\boldsymbol{\lambda},\boldsymbol{m}) - f_{\boldsymbol{\alpha}}(\boldsymbol{\lambda})f_{\boldsymbol{Y}}(\boldsymbol{m})|^2}{\|\boldsymbol{\lambda}\|^{K+p+1}_{K+p}\|\boldsymbol{m}\|^{q+1}_q} d\boldsymbol{\lambda} d\boldsymbol{m},$$

and $|z|$ denotes the modulus of $z \in \mathbb{C}$, $\|\boldsymbol{\lambda}\|_{K+p}$, $\|\boldsymbol{m}\|_q$ the standard Euclidean norms on the corresponding spaces V chosen to produce scale free and rotation invariant measure that does not go to zero for dependent random vectors, and

$$C_r = \frac{\pi^{\frac{1}{2}(r+1)}}{\Gamma(\frac{1}{2}(r+1))}$$

is half the surface area of the unit sphere in $\mathbb{R}^{r+1}$.

The functional distance correlation between random vector process $\boldsymbol{X}$ and random vector $\boldsymbol{Y}$ is a nonnegative number defined by

$$\mathscr{R}_{\boldsymbol{X},\boldsymbol{Y}} = \frac{\nu_{\boldsymbol{X},\boldsymbol{Y}}}{\sqrt{\nu_{\boldsymbol{X},\boldsymbol{X}}\nu_{\boldsymbol{Y},\boldsymbol{Y}}}}$$

if both $\nu_{\boldsymbol{X},\boldsymbol{X}}$ and $\nu_{\boldsymbol{Y},\boldsymbol{Y}}$ are strictly positive, and defined to be zero otherwise.

We have $\mathscr{R}_{\boldsymbol{X},\boldsymbol{Y}} = \mathscr{R}_{\boldsymbol{\alpha},\boldsymbol{Y}}$ as $\nu_{\boldsymbol{X},\boldsymbol{Y}} = \nu_{\boldsymbol{\alpha},\boldsymbol{Y}}$.

For distributions with finite first moments, distance correlation characterizes independence in that $0 \leq \mathscr{R}_{\boldsymbol{X},\boldsymbol{Y}} \leq 1$ with $\mathscr{R}_{\boldsymbol{X},\boldsymbol{Y}} = 0$ if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent. We can estimate functional distance covariance using data $\{(\hat{\boldsymbol{\alpha}}_1, \boldsymbol{y}_1), \ldots, (\hat{\boldsymbol{\alpha}}_n, \boldsymbol{y}_n)\}$.

Let

$$\bar{\boldsymbol{\alpha}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\alpha}}_k, \quad \bar{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{k}}_k,$$

$$\tilde{\boldsymbol{\alpha}_k} = \hat{\boldsymbol{\alpha}_k} - \bar{\boldsymbol{\alpha}}, \quad \tilde{\boldsymbol{y}_k} = \boldsymbol{y}_k - \bar{\boldsymbol{y}}, \ k = 1, \ldots, n$$

and

$$\boldsymbol{A} = (a_{kl}), \quad \boldsymbol{B} = (b_{kl}),$$

$$\tilde{\boldsymbol{A}} = (A_{kl}), \quad \tilde{\boldsymbol{B}} = (B_{kl}),$$

where

$$a_{kl} = \|\hat{\boldsymbol{\alpha}}_k - \hat{\boldsymbol{\alpha}}_l\|_{K+p}, \quad b_{kl} = \|\boldsymbol{y}_k - \boldsymbol{y}_l\|_q,$$
$$A_{kl} = \|\tilde{\boldsymbol{\alpha}}_k - \tilde{\boldsymbol{\alpha}}_l\|_{K+p}, \quad B_{kl} = \|\tilde{\boldsymbol{y}}_k - \tilde{\boldsymbol{y}}_l\|_q, \quad k, l = 1, \ldots, n.$$

Hence

$$\tilde{\boldsymbol{A}} = \boldsymbol{HAH}, \quad \tilde{\boldsymbol{B}} = \boldsymbol{HBH},$$

where

$$\boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$$

is the centering matrix.

Let $\tilde{\boldsymbol{A}} \circ \tilde{\boldsymbol{B}} = (A_{kl}B_{kl})$ denote the Hadamard product of the matrices $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{B}}$. Then, on the basis of the result of Székely et al. (2007), we have

$$\hat{\nu}^2_{\boldsymbol{X},\boldsymbol{Y}} = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}.$$

The sample functional distance correlation is then defined by $\hat{\mathscr{R}}_{\boldsymbol{X},\boldsymbol{Y}} = \hat{\mathscr{R}}_{\boldsymbol{\alpha},\boldsymbol{Y}}$, where

$$\hat{\mathscr{R}}_{\boldsymbol{\alpha},\boldsymbol{Y}} = \frac{\hat{\nu}_{\boldsymbol{\alpha},\boldsymbol{Y}}}{\sqrt{\hat{\nu}_{\boldsymbol{\alpha},\boldsymbol{\alpha}}\hat{\nu}_{\boldsymbol{Y},\boldsymbol{Y}}}}$$

if both $\hat{\nu}_{\boldsymbol{\alpha},\boldsymbol{\alpha}}$ and $\hat{\nu}_{\boldsymbol{Y},\boldsymbol{Y}}$ are strictly positive, and zero otherwise.

5 Variable Selection Based on the Distance Covariance

In this section we propose the selection procedure built upon the distance covariance. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_q)'$ be the response vector, and $\boldsymbol{X} = (X_1, \ldots, X_p)'$ be the predictor p-dimensional process. Assume that only a small number of predictors are relevant to $\boldsymbol{Y}$. We select a set of important predictors with large $\hat{\mathscr{R}}_{\boldsymbol{X},\boldsymbol{Y}} = \hat{\mathscr{R}}_{\boldsymbol{\alpha},\boldsymbol{Y}}$. We utilize the functional distance covariance because it allows for arbitrary relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$, regardless of whether it is linear or nonlinear.

The functional distance covariance also permits univariate and multivariate response. Thus, this distance covariance procedure is completely model-free. Kong et al. (2015) prove the following theorem.

Theorem 1 *Suppose random vectors* $\boldsymbol{X}, \boldsymbol{Z} \in \mathbb{R}^p$ *and* $\boldsymbol{Y} \in \mathbb{R}^q$*, and assume* $\boldsymbol{Z}$ *is independent of* $(\boldsymbol{X}, \boldsymbol{Y})$*, then*

$$\nu^2_{(\boldsymbol{X},\boldsymbol{Z}),\boldsymbol{Y}} \leq \nu^2_{\boldsymbol{X},\boldsymbol{Y}}.$$

And a consequence of this theorem is the statement in the next corollary.

Corollary 1 *For the sample distance covariance, if n is large enough, we should have*

$$\hat{\nu}^2_{(\boldsymbol{X},\boldsymbol{Z}),\boldsymbol{Y}} \leq \hat{\nu}^2_{\boldsymbol{X},\boldsymbol{Y}},$$

under the assumption of independence between $(\boldsymbol{X}, \boldsymbol{Y})$ *and* $\boldsymbol{Z}$*.*

We implemented the above theorem as a stopping rule in the selections of responses. The procedure took the following steps:

1. Calculate marginal distance covariances for X_k, $k = 1, \ldots, p$ with the response $\boldsymbol{Y}$.
2. Rank the variables in decreasing order of the distance covariances. Denote the ordered predictors as $X_{(1)}, X_{(2)}, \ldots, X_{(p)}$. Start with $\boldsymbol{X}_S = \{X_{(1)}\}$.
3. For k from 2 to p, keep adding $X_{(k)}$ to $\boldsymbol{X}_S$ if $\hat{\nu}^2_{\boldsymbol{X}_S,\boldsymbol{Y}}$ does not decrease. Stop otherwise.

6 Real Example

As a real example we used `Japanese Vowels` data set which is available at UCI Machine Learning Repository (Lichman 2013). Nine male speakers uttered two Japanese vowels /ae/ successively. For each utterance, it was applied 12° linear prediction analysis to obtain a discrete-time series with 12 LPC cep-strum coefficients. This means that one utterance by a speaker forms a time series whose length is in the range 7–29 and each point of a time series is of 12 features (12 coefficients). The number of the time series is 640 in total. The samples in this data set are of different lengths. They were extended to the length of the longest sample in the data set (Górecki and Łuczak 2015).

During the smoothing process we used Fourier basis with five components. In the next step we applied the described earlier method of selecting variables (we stopped the procedure if the increase in covariance measure was less than 0.01). In such way we obtained four variables (Fig. 1).

Next, we applied described classifiers to reduced functional data and to full functional data. To estimate the error rate of the classifiers we used tenfold cross-validation method. The results are in Table 1.

We can observe that the error rate increases if we reduce our data set. This behavior is expected. However, the increase seems not too big. Particularly inter-

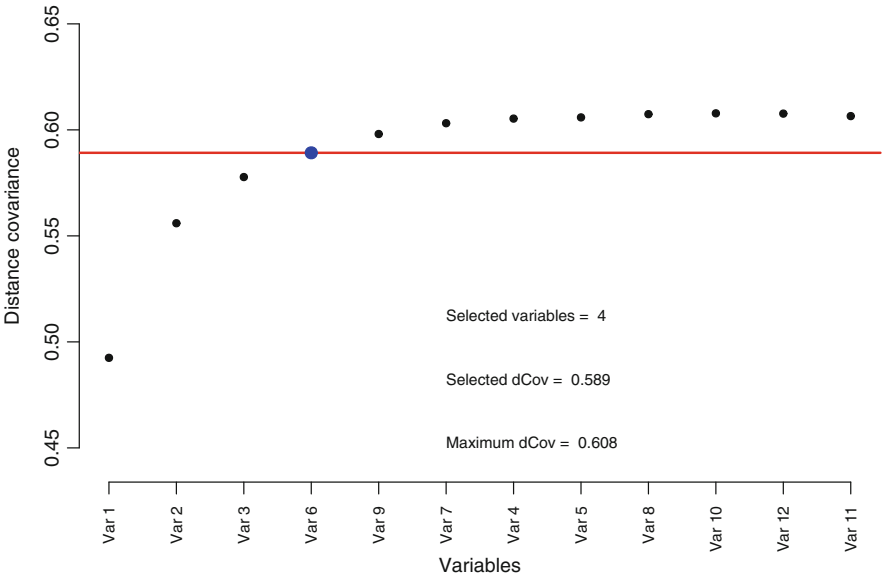

Fig. 1 Variables selection for `Japanese Vowels` data set

Table 1 Classification accuracy (in %) for `Japanese Vowels` data set

Classifier	Selected variables (4)	All variables (12)
LDC	93.60	99.37
Logistic regression	91.06	97.97
kNN ($k = 1, \ldots, 8$)	90.94	96.71
Naive Bayes (normal)	90.77	95.50
Naive Bayes (kernel)	90.15	94.34
QDC	89.85	Too small groups

esting is the case of QDC. For this method we do not have enough data to estimate covariance matrices for all groups for full data. When we select only four variables this procedure could be performed. We can also notice that the order of classifiers stays unchanged (the best classifier for full data is LDC, and the same is the best for reduced data).

During the calculations we used `R` (R Core Team 2017) software and `caret` (Kuhn 2017), `energy` (Rizzo and Székely 2016), and `fda` (Ramsay et al. 2014) packages.

7 Conclusion

The paper introduces variable selection for classification of multivariate functional data. Use of distance covariance as a tool to reduce dimensionality of data set suggests that the technique provides useful results for classification of multivariate functional data. For the analyzed data set only four from twelve variables were included in the final model. We can observe that classification accuracy could drop a little. However, we expect that this drop should be reasonable and in return we could gain a lot of computation time.

In practice, it is important not to depend entirely on variable selection criteria because none of them works well under all conditions. So our approach could be seen as a competitive to another variable selection methods. Additionally, model obtained by the proposed method of variable selection seems comparable with the full model (model without variables reduction). Finally, the researcher needs to evaluate the models using various diagnostic procedures.

References

Ando, T.: Penalized optimal scoring for the classification of multi-dimensional functional data. Stat. Methodol. 6, 565–576 (2009)

Berrendero, J.R., Cuevas, A., Torrecilla, J.L.: Variable selection in functional data classification: a maxima-hunting proposal. Stat. Sin. **26**(2), 619–638 (2016)

Delaigle, A., Haal, P.: Methodology and theory for partial least squares applied to functional data. Ann. Stat. **40**, 322–352 (2012)

Ferraty, F., Vieu, P.: Curve discrimination. A nonparametric functional approach. Comput. Stat. Data Anal. **44**, 161–173 (2003)

Ferraty, F., Vieu, P.: Additive prediction and boosting for functional data. Comput. Stat. Data Anal. **53**(4), 1400–1413 (2009)

Górecki, T., Krzyśko, M., Waszak, Ł., Wołyński, W.: Methods of reducing dimension for functional data. Stat. Transition New Series **15**, 231–242 (2014)

Górecki, T., Krzyśko, M., Wołyński, W.: Multivariate functional regression analysis with application to classification problems. In: Wilhelm Adalbert, F.X., Kestler Hans, A. (eds.) Analysis of Large and Complex Data, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 173–183. Springer, Berlin (2016)

Górecki, T., Łuczak, M.: Multivariate time series classification with parametric derivative dynamic time warping. Expert Syst. Appl. **42**(5), 2305–2312 (2015)

Hastie, T.J., Tibshirani, R.J., Buja, A.: Penalized discriminant analysis. Ann. Stat. **23**, 73–102 (1995)

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

Jacques, J., Preda, C.: Model-based clustering for multivariate functional data. Comput. Stat. Data Anal. **71**, 92–106 (2014)

Kong, J., Wang, S., Wahba G.: Using distance covariance for improved variable selection with application to learning genetic risk models. Stat. Med. **34**, 1708–1720 (2015)

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B.: caret: Classification and Regression Training. R package version 6.0-76 (2017). https://CRAN.R-project.org/package=caret

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

R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna (2017). https://www.R-project.org/

Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer, New York (2005)

Ramsay, J.O., Wickham, H. Graves, S., Hooker, G.: fda: Functional Data Analysis. R package Version 2.4.4 (2014). https://CRAN.R-project.org/package=fda

Rizzo, M.L., Székely, G.J.: Energy: E-Statistics: Multivariate Inference via the Energy of Data. R Package Version 1.7-0 (2016). https://CRAN.R-project.org/package=energy

Rossi, F., Delannayc, N., Conan-Gueza, B., Verleysenc, M.: Representation of functional data in neural networks. Neurocomputing **64**, 183–210 (2005)

Rossi, F., Villa, N.: Support vector machines for functional data classification. Neural Comput. **69**, 730–742 (2006)

Rossi, N., Wang, X., Ramsay, J.O.: Nonparametric item response function estimates with EM algorithm. J. Educ. Behav. Stat. **27**, 291–317 (2002)

Székely, G.J., Rizzo, M.L., Bakirov, N.K.: Measuring and testing dependence by correlation of distances. Ann. Stat. **35**(6), 2769–2794 (2007)

Székely, G.J., Rizzo, M.L.: Brownian distance covariance. Ann. Stat. **3**(4), 1236–1265 (2009)

Székely, G.J., Rizzo, M.L.: On the uniqueness of distance covariance. Stat. Probab. Lett. **82**(12), 2278–2282 (2012)

Székely, G.J., Rizzo, M.L.: The distance correlation t-test of independence in high dimension. J. Multivar. Anal. **117**, 193–213 (2013)