A note on variable selection in functional regression via random subspace method

Łukasz Smaga¹ · Hidetoshi Matsui²

Accepted: 12 January 2018 / Published online: 25 January 2018
© The Author(s) 2018. This article is an open access publication

Abstract Variable selection problem is one of the most important tasks in regression analysis, especially in a high-dimensional setting. In this paper, we study this problem in the context of scalar response functional regression model, which is a linear model with scalar response and functional regressors. The functional model can be represented by certain multiple linear regression model via basis expansions of functional variables. Based on this model and random subspace method of Mielniczuk and Teisseyre (Comput Stat Data Anal 71:725–742, 2014), two simple variable selection procedures for scalar response functional regression model are proposed. The final functional model is selected by using generalized information criteria. Monte Carlo simulation studies conducted and a real data example show very satisfactory performance of new variable selection methods under finite samples. Moreover, they suggest that considered procedures outperform solutions found in the literature in terms of correctly selected model, false discovery rate control and prediction error.

Keywords Basis functions representation · Functional regression analysis · Information criterion · Random subspace method · Variable selection

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10260-018-0421-7>) contains supplementary material, which is available to authorized users.

✉ Łukasz Smaga
ls@amu.edu.pl; lukaszsmaga17@gmail.com

Hidetoshi Matsui
hmatsui@biwako.shiga-u.ac.jp; hidetoshi.matsui@gmail.com

¹ Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Umultowska 87, 61-614 Poznań, Poland

² Faculty of Data Science, Shiga University, 1-1-1, Banba, Hikone, Shiga 522-8522, Japan

Mathematics Subject Classification 62J99 · 62M99**1 Introduction**

In functional data analysis (FDA), observations, called functional data, are viewed as functions defined over some set T . The data observed on a grid of space or time points can be represented by functions, and then methods of FDA draw information from the collection of these functions instead of discrete data. They include extensions of traditional statistical methods such as analysis of variance, cluster analysis, discriminant analysis, outlier detection, principal component analysis and regression analysis (see Ferraty and Vieu 2006; Horváth and Kokoszka 2012; Hubert et al. 2015; Ramsay and Silverman 2002, 2005; Zhang 2013 and the references therein). New developments in functional statistics can be found, for example, in recent special issues in FDA on *Econometrics and Statistics* (Kokoszka et al. 2017) and *Journal of Multivariate Analysis* (Aneiros et al. 2017b) as well as in Aneiros et al. (2017a) and Bongiorno et al. (2014).

Functional regression analysis is one of the most extensively studied branches of FDA. It is used to describe the relationship between response and explanatory variables when at least one of them contains a random function. By Horváth and Kokoszka (2012), the following cases are considered: the functional response model (the responses are curves, but the regressors are known scalars), the scalar response model (the responses are scalars and the regressors are curves), the fully functional model (the responses and regressors are curves). The functional response model was considered by Chiou et al. (2004) and Faraway (1997). The scalar response functional regression model has been intensively investigated mostly by using functional principal component analysis, penalized regularization or reproducing kernel Hilbert space approach (Hall and Horowitz 2007; Li and Hsing 2007; Yuan and Cai 2010). We also refer to the survey paper by Cardot and Sarda (2011) for excellent overview. This model was extended to nonlinear and semiparametric functional regression models, e.g., generalized functional regression models with known and unknown link (Chen et al. 2011; James 2002; Müller and Stadtmüller 2005) or additive functional regression model (Febrero-Bande and Gonzalez-Manteiga 2013; Goia and Vieu 2015). The fully functional model with one or more functional responses and its generalizations were investigated, for instance, in Chiou et al. (2016), Fan et al. (2014) and Radchenko et al. (2015).

One of important problems in functional regression analysis is variable selection. To motivate this statement, we consider a real data example from chemometrics, where functional data analysis is widely applicable. The data set contains the measurements of laboratory determinations of the quality of sugar, as, for example, ash content. For each sample of sugar, the emission spectra were measured at seven excitation wavelengths, and they can be seen as functional variables. Thus, the functional regression models can be used to describe the relationship between laboratory determinations of the quality of sugar and the fluorescence spectra. Such relationship is important, since it is easier and cheaper to use the spectra than chemical analysis in the laboratory. However, maybe not all seven excitation wavelengths have to be used in the regression model,

and a proper selection of them may result in even easier analysis. Variable selection methods in functional regression model may suggest excitation wavelengths, which should be used. The detailed description of the sugar data set and its analysis are presented in Sect. 5.

Although variable selection is very important problem in regression analysis, for functional regression models, it is more rarely considered in the literature. Most authors considered methods based on regularization techniques, where penalization is used for simultaneously shrinking parameters and selecting variables. For the scalar response model, Matsui and Konishi (2011) used the group SCAD regularization. The lasso was generalized by Hong and Lian (2011) for the functional response model. Matsui (2014) also used regularization techniques for variable selection in the multiclass logistic regression model, while Gertheiss et al. (2013) in the generalized functional linear model. Collazos et al. (2016) proposed different approach. They adapted the methodology of Bunea et al. (2006) to the scalar response model, and offered variable selection procedures based on significance testing of functional regressors. Aneiros and Vieu (2014) proposed another variable selection method based on the continuous structure of the functional regressors.

In this paper, we consider the problem of using random subspace method (RSM) to variable selection for the scalar response functional regression model. Main steps of our variable selection methods are as follows. First, the functional regressors and coefficient functions are represented by basis expansions. This representation is used to rewrite the functional regression model as a multiple linear regression one with design matrix consisting of basis expansion coefficients. Next, variables in this model are selected by using the random subspace method of Mielniczuk and Teisseyre (2014). Finally, the information from earlier steps is aggregated based on information criteria in two different ways, obtaining the final subset of variables. The random subspace method behaves promisingly in comparison to penalty-based ones in multiple linear regression model. A good behavior of the RSM procedure results in very satisfactory performance of new variable selection methods for the scalar response functional regression model. In fact, simulation studies show that our variable selection methods perform comparable to or even better than existing competitors. Apart from good finite sample properties, our procedures need less restrictive assumptions on the number of observations than certain known methods, which indicates that they have wider range of application. Although novel methods may be time consuming for large data sets, parallel implementation of the random subspace method results in reduction of computational cost of our procedures.

The remainder of the paper is organized as follows. Section 2 outlines the scalar response functional regression model and its representation via basis expansions. In Sect. 3, variable selection procedures based on random subspace method are explained. In Sect. 4, a finite sample behavior of new variable selection methods is investigated through simulation studies. A comparison with other variable selection procedures for the scalar response model is also presented there. Section 5 shows an application of the variable selection methods to sugar spectra data set. Finally, Sect. 6 draws some conclusions.

2 The scalar response functional regression model

In this section, the scalar response functional regression model is introduced and rewritten as a multiple linear regression model by using the basis functions representation of functional regressors and coefficient functions (For simplicity, we will refer to this model as the functional regression model).

Let us introduce the functional regression model. Assume that $y_i, i = 1, \dots, n$ are the scalar responses, and $x_{ij}(t), j = 1, \dots, p$ are the functional regressors belonging to the Hilbert space of square integrable functions over $[a_j, b_j], a_j, b_j \in R$, denoted by $\mathcal{L}_2([a_j, b_j])$. Let observations follow the functional regression model:

$$y_i = \beta_0 + \sum_{j=1}^p \int_{a_j}^{b_j} x_{ij}(t) \beta_j(t) dt + \varepsilon_i, \quad (1)$$

where β_0 is a constant term, $\beta_j(t) \in \mathcal{L}_2([a_j, b_j]), j = 1, \dots, p$ are the unknown coefficient functions, and $\varepsilon_1, \dots, \varepsilon_n$ are the unobservable independent random errors, assumed to have normal distribution with zero expectation and unknown variance σ^2 .

2.1 Basis representation of functional regression model

Since the functional regressors and coefficient functions of the model (1) belong to the spaces $\mathcal{L}_2([a_j, b_j]), j = 1, \dots, p$, they can be approximated arbitrarily well by taking a linear combination of a sufficiently large number of basis functions $\{\varphi_{jl}\}_{l=1}^\infty$ of these spaces (see Ramsay and Silverman 2005). Thus, we can assume that the functions $x_{ij}(t)$ and $\beta_j(t) (i = 1, \dots, n, j = 1, \dots, p)$ can be represented as a linear combination of a finite number of basis functions, i.e.,

$$x_{ij}(t) = \sum_{l=1}^{k_j} c_{ijl} \varphi_{jl}(t) = \mathbf{c}_{ij}^\top \boldsymbol{\varphi}_j(t), \quad \beta_j(t) = \sum_{l=1}^{k_j} d_{jl} \varphi_{jl}(t) = \mathbf{d}_j^\top \boldsymbol{\varphi}_j(t) \quad (2)$$

for $t \in [a_j, b_j]$, where $\mathbf{c}_{ij} = (c_{ij1}, \dots, c_{ijk_j})^\top$ and $\mathbf{d}_j = (d_{j1}, \dots, d_{jk_j})^\top$ are the vectors of unknown coefficients, and $\boldsymbol{\varphi}_j(t) = (\varphi_{j1}(t), \dots, \varphi_{jk_j}(t))^\top$ are the vectors of basis functions. As $x_{ij}(t)$ are random functions, c_{ijl} are random variables, supposed to have finite variance. For each regressor separately, they can be estimated, for example, by the least squares method (see Krzyśko and Waszak 2013) or by the regularized maximum likelihood method (see Matsui et al. 2008) or by the roughness penalty approach (see Ramsay and Silverman 2005, Chapter 5). The basis functions φ_{jl} as well as the values k_j may be selected depending on the data. In general, the spline and Gaussian radial bases are appropriate for time course data sets with a certain smoothness. For data sets with periodicity (such as daily or yearly data), the Fourier basis can be used. On the other hand, if we want to transform the data with some non-smooth “spikes” into functions, the wavelet basis is used in order to extract these peaks. To select the number of basis functions in (2), model selection criterion (such as Akaike and Bayesian information criteria) may be applied. Moreover, the bases and

the values k_j may be chosen based on the problem at hand. For example, they may be chosen to minimize prediction error of functional regression model.

By the basis functions representation of functional regressors and coefficient functions given in (2), for $i = 1, \dots, n$, we have

$$\begin{aligned} \sum_{j=1}^p \int_{a_j}^{b_j} x_{ij}(t) \beta_j(t) dt &= \sum_{j=1}^p \int_{a_j}^{b_j} \mathbf{c}_{ij}^\top \boldsymbol{\varphi}_j(t) \boldsymbol{\varphi}_j^\top(t) \mathbf{d}_j dt \\ &= \sum_{j=1}^p \mathbf{c}_{ij}^\top \int_{a_j}^{b_j} \boldsymbol{\varphi}_j(t) \boldsymbol{\varphi}_j^\top(t) dt \mathbf{d}_j. \end{aligned} \quad (3)$$

The matrices $\mathbf{J}_{\boldsymbol{\varphi}_j} := \int_{a_j}^{b_j} \boldsymbol{\varphi}_j(t) \boldsymbol{\varphi}_j^\top(t) dt$, $j = 1, \dots, p$ are the $k_j \times k_j$ cross product matrices corresponding to bases $\{\varphi_{jl}\}_{l=1}^\infty$. For orthonormal basis as the Fourier basis, the matrix $\mathbf{J}_{\boldsymbol{\varphi}_j}$ is equal to the identity matrix. However, for given data or aim of the research, some non-orthonormal basis may be more appropriate as, for instance, the Gaussian basis or the B-spline basis. The formula of cross product matrix for Gaussian basis is presented for example in Matsui and Konishi (2011). For computing the cross product matrix for the B-spline basis, the procedure in Kayano and Konishi (2009) can be used. The cross product matrix for these as well as other bases can be approximated by using the function `inprod` from the R package `fda` (R Core Team 2017; Ramsay et al. 2009, 2017).

From (3), it follows that the functional regression model given in (1) can be rewritten as

$$y_i = \beta_0 + \sum_{j=1}^p \mathbf{c}_{ij}^\top \mathbf{J}_{\boldsymbol{\varphi}_j} \mathbf{d}_j + \varepsilon_i = \mathbf{c}_i^\top \mathbf{d} + \varepsilon_i, \quad (4)$$

where $\mathbf{c}_i = (1, \mathbf{c}_{i1}^\top \mathbf{J}_{\boldsymbol{\varphi}_1}, \dots, \mathbf{c}_{ip}^\top \mathbf{J}_{\boldsymbol{\varphi}_p})^\top$, $i = 1, \dots, n$, and $\mathbf{d} = (\beta_0, \mathbf{d}_1^\top, \dots, \mathbf{d}_p^\top)^\top$. The above model can be seen as a basis representation of the model (1). In matrix notation, the model (4) is of the form

$$\mathbf{y} = \mathbf{C} \mathbf{d} + \boldsymbol{\varepsilon}, \quad (5)$$

where $\mathbf{y} = (y_1, \dots, y_n)^\top$, $\mathbf{C} = (\mathbf{c}_1, \dots, \mathbf{c}_n)^\top$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$. The vector $\mathbf{d}$ may be estimated by the least squares method or by the roughness penalty approach (see Ramsay and Silverman 2005, Chapter 15). The second method is more flexible and preferred. Therefore, the functional regression model (1) is re-expressed as the multiple linear regression model (5) with the $n \times (1 + \sum_{j=1}^p k_j)$ design matrix $\mathbf{C}$ and the $(1 + \sum_{j=1}^p k_j) \times 1$ vector $\mathbf{d}$ of unknown parameters. This relationship may be used to solve certain problems in functional regression by using results for multiple linear regression. For example, Collazos et al. (2016) tested the null hypotheses $H_0^j : \mathbf{d}_j = \mathbf{0}_{k_j}$ against $H_1^j : \mathbf{d}_j \neq \mathbf{0}_{k_j}$, $j = 1, \dots, p$ to test significance of the functional regressors and select variables in the model (1). In the following, we also use this relationship and propose two simple variable selection procedures for the functional regression model (1).

3 Methods

In this section, novel variable selection methods for functional regression model are proposed.

3.1 Variable selection via random subspace method

New variable selection methods for the functional regression model (1) use the random subspace method for linear regression proposed by Mielniczuk and Teisseyre (2014). Its extensions as well as implementation in the R package `regRSM` are presented in Teisseyre et al. (2016). For the convenience of the Reader, we briefly describe the RSM procedure. For more details, we refer the Reader to mentioned articles.

Consider the usual linear regression model $\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\epsilon}$, where $\mathbf{y}$ is an $n \times 1$ response vector, $\mathbf{X}$ is an $n \times p$ design matrix, $\mathbf{b}$ is a $p \times 1$ vector of unknown parameters, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of independent random errors with mean zero and unknown variance σ^2 . The number of parameters $p \geq n$ is allowed. Let $\mathbf{X}_m$ denote submatrix of $\mathbf{X}$ with columns corresponding to set of variables $m \subset \{1, \dots, p\}$. The pseudo code of the RSM procedure by Mielniczuk and Teisseyre (2014) is as follows:

Algorithm (RSM procedure, Mielniczuk and Teisseyre 2014; Teisseyre et al. 2016)

1. *Input: observed data $(\mathbf{y}, \mathbf{X})$, a number of subset draws B , a size of the subspace $|m| < \min(n, p)$, a cut off level $h \leq \min(n, p)$, a final model selection method.*
2. *Repeat the following procedure for $r = 1, \dots, B$ starting with the counter $C_{j,0} = 0$ for any variable j .*
 - *Randomly draw a subset of variables m_* (without replacement) from the original variable space with the same probability for each variable.*
 - *Fit model to data $(\mathbf{y}, \mathbf{X}_{m_*})$ and compute weight*

$$w(j, m_*) = T_{j,m_*}^2 = \hat{\sigma}_{j,m_*}^2 [\hat{\sigma}_{m_*}^2 (\mathbf{X}_{m_*}^\top \mathbf{X}_{m_*})^{-1}]^{-1}$$

for each $j \in m_$, where T_{j,m_*} is the t -statistic corresponding to the variable j when the model m_* is fitted to the data, $\hat{b}_{j,m_*}$ is the j th coordinate of least squares estimator $\hat{\mathbf{b}}_{m_*}$ based on model m_* , and $\hat{\sigma}_{m_*}^2$ is noise variance estimator based on that model. Set $w(j, m_*) = 0$ if $j \notin m_*$.*

- *Update the counter $C_{j,r} = C_{j,r-1} + I\{j \in m_*\}$.*
3. *For each variable j compute the final score $W_j^* = C_{j,B}^{-1} \sum_{m_*: j \in m_*} w(j, m_*)$.*
 4. *Sort the list of variables according to scores W_j^* : $W_{j_1}^* \geq W_{j_2}^* \dots \geq W_{j_p}^*$, obtaining the ordered list of variables $\{j_1, \dots, j_p\}$ and further the nested list of models $\{\emptyset, \{j_1\}, \{j_1, j_2\}, \dots, \{j_1, \dots, j_p\}\}$.*
 5. *Cut off the nested list of models at level h : $\{\emptyset, \{j_1\}, \{j_1, j_2\}, \dots, \{j_1, \dots, j_h\}\}$.*
 6. *Output: Final set of variables chosen from the list of models given in Step 5 by using a final model selection method.*

Originally, the RSM procedure was described in a little other way. However, the above description will result in easier presentation of our methods. Note that smaller

subspace size $|m|$ may result in missing informative variables or missing dependences between features. On the other hand, many spurious variables can be included in final model when $|m|$ is too large. By the empirical results of Mielniczuk and Teisseyre (2014) and Teisseyre et al. (2016), the reasonable choice is $|m| = \min(n, p)/2$. They also propose to use two final model selection methods based on minimizing the prediction error on validation set and on information criteria (In this case, $h = \min(n/2, p)$ should be set, since information criteria may select models which are close to the saturated one.). When the model includes an intercept (the first column of design matrix $\mathbf{X}$ consists of ones), a subset m_* is randomly drawn from the set of genuine regressors only, and then a model pertaining to regressors from m_* and an intercept is fitted.

3.2 Variable selection in functional regression model

Now, we can present our variable selection methods for functional regression model (1). Both methods have similar first steps. First, the functional regressors are expressed via basis expansions as in (2). Then, we compute the design matrix $\mathbf{C}$ of the multiple linear regression model (5), which represents the model (1). Next, both methods use the RSM procedure, but in different ways.

3.2.1 The first RSM based variable selection

In the first method, we apply the RSM procedure to the model with $\mathbf{X} = \mathbf{C}$ treating the basis representation coefficients of functional regressors as new variables, called further on basis representation variables. Then, the basis representation variables selected by the RSM procedure indicate the functional regressors which should be included in the final functional regression model. The detailed pseudo code of our first procedure is outlined below.

- Algorithm 1**
1. *Input: observed data $y_i, x_{ij}(t), t \in [a_j, b_j], i = 1, \dots, n, j = 1, \dots, p$, bases $\{\varphi_{jl}\}_{l=1}^\infty$ in $\mathcal{L}_2([a_j, b_j])$, truncation parameters k_j , a number of subset draws B , a size of the subspace $|m| < \min(n, \sum_{j=1}^p k_j)$, a cut off level $h \leq \min(n, \sum_{j=1}^p k_j)$, a final model selection method.*
 2. *Represent each functional regressor as a linear combination of a finite number of basis functions as in (2).*
 3. *Compute the $n \times (1 + \sum_{j=1}^p k_j)$ design matrix $\mathbf{C}$ of the multiple linear regression model (5).*
 4. *Perform the RSM procedure for $\mathbf{y} = (y_1, \dots, y_n)^\top$, the design matrix $\mathbf{X} = \mathbf{C}$, a number of subset draws B , a size of the subspace $|m|$, a cut off level h and a final model selection method.*
 5. *Output: Final set of functional variables consisting of those functional variables whose basis representation variables were selected by the RSM procedure in Step 4.*

In Algorithm 1, the information obtained from the RSM procedure is used in simple, but effective way (see Sects. 4 and 5). It is also worth noting that Algorithm 1 can be used for any number of functional regressors and almost any values of truncation parameters $k_j, j = 1, \dots, p$. This distinguishes our method from variable selection

methods for model (1) known in the literature, e.g., Collazos et al. (2016) have to assume that $1 + \sum_{j=1}^p k_j < n$. This property of Algorithm 1 follows from the fact that in the RSM procedure variables are ranked based on fitting small linear models, and hence this procedure does not impose any conditions on the number of candidate variables. Some standard restrictions about truncation parameters may have to be taken into account. For example, as the functional samples are not continuously observed in practice, i.e., each function is usually observed on a grid of design time points, truncation parameters should not be greater than a number of those points. Particular basis may also require specific values of k_j , e.g., for the Fourier basis they should be odd (see the implementation of this basis in the R package `fd`).

3.2.2 The second RSM based variable selection

Now we describe the second method, which uses the scores of basis representation variables obtained by the RSM procedure in other way than in Algorithm 1. In this method, we only use the first three steps of the RSM procedure to the model with $\mathbf{X} = \mathbf{C}$. The obtained scores of basis representation variables corresponding to different functional regressors are summed. In such a simple way, we aggregate the information about functional regressors from basis representation variables. Sorting these sums, we construct a nested list of models. From this list, we select the final model by using some model selection method. The detailed pseudo code of our second method is shown below.

- Algorithm 2** 1. *Input: observed data $y_i, x_{ij}(t), t \in [a_j, b_j], i = 1, \dots, n, j = 1, \dots, p$, bases $\{\varphi_{jl}\}_{l=1}^\infty$ in $\mathcal{L}_2([a_j, b_j])$, truncation parameters k_j , a number of subset draws B , a size of the subspace $|m| < \min(n, \sum_{j=1}^p k_j)$, a cut off level $h_* \leq \min(n, p)$, a final model selection method.*
2. *This step is the same as Step 2 of Algorithm 1.*
 3. *This step is the same as Step 3 of Algorithm 1.*
 4. *Perform Steps 1-3 of the RSM procedure for $\mathbf{y} = (y_1, \dots, y_n)^\top$, the design matrix $\mathbf{X} = \mathbf{C}$, a number of subset draws B and a size of the subspace $|m|$, obtaining the final scores of basis representation variables.*
 5. *For each functional variable j , compute the sum S_j of the scores of basis representation variables corresponding to it, obtained in Step 4.*
 6. *Sort the list of functional variables according to sums $S_j: S_{j_1} \geq S_{j_2} \dots \geq S_{j_p}$, obtaining the ordered list of functional variables $\{j_1, \dots, j_p\}$ and further the nested list of models $\{\emptyset, \{j_1\}, \{j_1, j_2\}, \dots, \{j_1, \dots, j_p\}\}$.*
 7. *Cut off the nested list of models at level h_* : $\{\emptyset, \{j_1\}, \{j_1, j_2\}, \dots, \{j_1, \dots, j_{h_*}\}\}$.*
 8. *Output: Final set of functional variables chosen from the list of models given in Step 7 by using a final model selection method.*

A final model selection method can be based on the model (5) and information criteria. As we assume normality of random errors, the generalized information criterion (GIC) can be written as $\text{GIC}(m) := n \log(\text{RSS}(m)) + |m|a_n$, where $\text{RSS}(m) = \|\mathbf{y} - \mathbf{C}_m \hat{\mathbf{d}}_m\|^2$ is the residual sum of squares for the model m , $\hat{\mathbf{d}}_m$ is the least squares estimator based on this model, and a_n is penalty. In simulations, we use the Bayesian Information Criterion ($a_n = \log(n)$), but other criteria are also possible, e.g., the Akaike Information

Criterion ($a_n = 2$). To compute the values of RSS for all models from the nested list in Step 7 of Algorithm 2, we can use the QR decomposition of design matrix similarly as in Teisseyre et al. (2016). Let $\mathbf{C}_{\{j_1, \dots, j_{h_*}\}}$ denote submatrix of $\mathbf{C}$ with columns corresponding to intercept and set of basis representation variables which represent the functional regressors from the ordered list $\{j_1, \dots, j_{h_*}\}$ obtained in Step 7 of Algorithm 2. If it is necessary, we have to order the columns (more precisely, blocks of columns corresponding to functional variables) of $\mathbf{C}_{\{j_1, \dots, j_{h_*}\}}$ according to $j_1, \dots, j_{h_*}$. Assume that the QR decomposition of the matrix $\mathbf{C}_{\{j_1, \dots, j_{h_*}\}}$ is as follows: $\mathbf{C}_{\{j_1, \dots, j_{h_*}\}} = \mathbf{Q}_{\{j_1, \dots, j_{h_*}\}} \mathbf{R}_{\{j_1, \dots, j_{h_*}\}}$, where $\mathbf{Q}_{\{j_1, \dots, j_{h_*}\}}$ is an $n \times (1 + \sum_{j \in \{j_1, \dots, j_{h_*}\}} k_j)$ orthogonal matrix and $\mathbf{R}_{\{j_1, \dots, j_{h_*}\}}$ is a $(1 + \sum_{j \in \{j_1, \dots, j_{h_*}\}} k_j) \times (1 + \sum_{j \in \{j_1, \dots, j_{h_*}\}} k_j)$ upper triangular matrix. First, we compute $\text{RSS}(\{j_1, \dots, j_{h_*}\})$ and then the values of RSS for other models from the nested list in Step 7 of Algorithm 2 are computed by using the following equality:

$$\begin{aligned} \text{RSS}(\{j_1, \dots, j_{h_*-i}\}) &= \text{RSS}(\{j_1, \dots, j_{h_*}\}) \\ &+ \sum_{r=0}^{\sum_{j \in \{j_{h_*-i+1}, \dots, j_{h_*}\}} k_j - 1} \left(\mathbf{y}^\top \mathbf{Q}_{(1 + \sum_{j \in \{j_1, \dots, j_{h_*}\}} k_j) - r} \right)^2 \end{aligned}$$

for $i = 1, \dots, h_* - 1$, where $\mathbf{Q}_s$ is the s th column of the matrix $\mathbf{Q}_{\{j_1, \dots, j_{h_*}\}}$, $s = 1, \dots, 1 + \sum_{j \in \{j_1, \dots, j_{h_*}\}} k_j$. This procedure is computationally faster than computing the values of RSS by fitting all nested models in Step 7 of Algorithm 2.

Similarly to Teisseyre et al. (2016), we propose to use $h_* = \min(n/2, p)$, but h_* has to satisfy the following additional condition: $1 + \sum_{r=1}^{h_*} k_{j_r} \leq n - 2$, i.e., a number of basis representation variables of model $\{j_1, \dots, j_{h_*}\}$ (see Step 7 of Algorithm 2) is less than $n - 2$. It is necessary to perform properly the final model selection method based on information criteria. This condition is a more stringent limitation for truncation parameters k_j used in Algorithm 2 as compared to Algorithm 1. However, in general, it is still less stringent than that for the methods of Collazos et al. (2016). Observe that in Step 5 of Algorithm 2, “simple” sum of scores is used. In this step, other connected functions may be used to obtain better results. It is also possible to choose other scores in the RSM procedure. So Algorithm 2 may be further modified in case of unsatisfactory performance of the present one.

The difference between considered algorithms is that whether we treat the basis representation variables as individual variables (Algorithm 1) or as grouped (functional) variables (Algorithm 2). We will compare the results of these algorithms in artificial and real data analysis. Simulation and real data experiments of the next sections indicate that the proposed variable selection procedures behave promisingly under finite samples. A disadvantage of our methods is that they may be time consuming for large data sets. However, Teisseyre et al. (2016) developed parallel implementations of the RSM procedure (the most time consuming step of new methods) which enable to reduce the computation cost of it, and thereby also of our methods, significantly. Experiments conducted by Teisseyre et al. (2016) showed the practical efficiency of the parallelization of the RSM procedure.

4 Simulations

In this section, the performance of the proposed variable selection methods as well as the procedures of Collazos et al. (2016) and the group lasso (Glasso, Matsui and Konishi 2011) is investigated through simulation studies. We also considered the group SCAD of Matsui and Konishi (2011), but the results were similar to or slightly worse than those for the Glasso, so they are not presented.

Simulation experiments were performed using the R programming language (R Core Team 2017). The basis functions representation of functional regressors was estimated by using the R package *fda* (Ramsay et al. 2009, 2017). To perform the RSM procedure, its implementation in the R package *regRSM* (Teisseyre et al. 2016) was used.

4.1 Simulation models

We generated $n = 50, 75, 100$ observations from model (1) of three and six functional regressors ($p = 3$ or $p = 6$). The functional regressors $x_{ij}(t)$, $t \in [a_j, b_j]$ are represented by their values in an equally spaced grid of 50 points in $[a_j, b_j]$, $i = 1, \dots, n$, $j = 1, \dots, p$. We consider independent and correlated functional predictors. In independent case, the functional regressors are generated according to the following rule: $x_{ij}(t_{js}) = z_{ij}(t_{js}) + \varepsilon_{ijs}$, where $\varepsilon_{ijs} \sim N(0, 0.025r_{x_{ij}})$, $r_{x_{ij}} = \max_s(z_{ij}(t_{js})) - \min_s(z_{ij}(t_{js}))$, $i = 1, \dots, n$, $j = 1, \dots, p$ and $s = 1, \dots, 50$. When $p = 3$, the functions $z_{ij}(t)$ and coefficient functions $\beta_j(t)$ are as follows:

$$\begin{aligned} z_{i1}(t) &= \cos(2\pi(t - c_1)) + c_2t, \quad t \in [0, 1], \quad c_1 \sim N(-5, 3), \quad c_2 \sim N(7, 1), \\ z_{i2}(t) &= d_1 \sin(2t) + d_2, \quad t \in [0, \pi/3], \quad d_1 \sim U(3, 7), \quad d_2 \sim N(0, 1), \\ z_{i3}(t) &= e_1t^3 + e_2t^2 + e_3t + e_4, \quad t \in [-1, 1], \quad e_1 \sim N(-3, 1.2), \quad e_2 \sim N(2, 0.5), \\ &\quad e_3 \sim N(-2, 1), \quad e_4 \sim N(2, 1.5), \\ \beta_1(t) &= \sin(2\pi t), \quad \beta_2(t) = \sin(\pi t), \quad \beta_3(t) = 0. \end{aligned}$$

In the case of $p = 6$, the following functions $z_{ij}(t)$ and coefficient functions $\beta_j(t)$ are considered:

$$\begin{aligned} z_{i1}(t) &= \cos(2\pi(t - c_1)) + c_2, \quad t \in [0, 1], \quad c_1 \sim N(-4, 3), \quad c_2 \sim N(7, 1.5), \\ z_{i2}(t) &= d_1 \sin(\pi t) + d_2, \quad t \in [0, \pi/3], \quad d_1 \sim U(3, 7), \quad d_2 \sim N(0, 1), \\ z_{i3}(t) &= e_1t^3 + e_2t^2 + e_3t, \quad t \in [-1, 1], \quad e_1 \sim N(-3, 1.2), \quad e_2 \sim N(2, 0.5), \\ &\quad e_3 \sim N(-2, 1), \\ z_{i4}(t) &= \sin(2(t - f_1)) + f_2t, \quad t \in [0, \pi/3], \quad f_1 \sim N(-2, 1), \quad f_2 \sim N(3, 1.5), \\ z_{i5}(t) &= g_1 \cos(2t) + g_2t, \quad t \in [-2, 1], \quad g_1 \sim U(2, 7), \quad g_2 \sim N(2, 0.4), \\ z_{i6}(t) &= h_1e^{-t/3} + h_2t + h_3, \quad t \in [-1, 1], \quad h_1 \sim N(4, 2), \quad h_2 \sim N(-3, 0.5), \\ &\quad h_3 \sim N(1, 1), \end{aligned}$$

$$\begin{aligned}\beta_1(t) &= \sin(t), \quad \beta_2(t) = \sin(2t), \quad \beta_3(t) = 0, \\ \beta_4(t) &= d \sin(2t), \quad \beta_5(t) = e \sin(\pi t), \quad \beta_6(t) = f t^2.\end{aligned}$$

In correlated case, the six ($p = 6$) functional predictors are generated in the following way: $x_{ij}(t_s) = z_{ij}(t_s) = \alpha_{ij}^\top \phi(t_s)$, where $\alpha_{ij}^\top = (\alpha_{ij1}, \dots, \alpha_{ijs})$ are subvectors of the random vector $(\alpha_{i1}^\top, \dots, \alpha_{ip}^\top)^\top \sim N_{5p}(\mathbf{0}_{5p}, \Sigma)$ with $\Sigma = (0.9^{|k-l|})$, $k = 1, \dots, 5p$, $l = 1, \dots, 5p$, and the vector $\phi(t) = (\phi_1(t), \dots, \phi_5(t))^\top$ contains the Fourier basis functions, $i = 1, \dots, n$, $j = 1, \dots, p$ and $s = 1, \dots, 50$. The functional canonical correlation analysis of functional predictors generated in such way indicates that the correlation between i -th and j -th predictors decreases as $|i - j|$ increases, similarly as in the autoregressive process with autocorrelation coefficient $\rho = 0.9$ (see, for example, Krzyśko and Waszak 2013). The following coefficient functions $\beta_j(t)$ are considered:

$$\begin{aligned}\beta_1(t) &= \xi_1 \sin(t), \quad \beta_2(t) = \xi_2 \sin(2t), \quad \beta_3(t) = \xi_3 t^2, \\ \beta_4(t) &= \xi_4 \sin(2t), \quad \beta_5(t) = \xi_5 \sin(\pi t), \quad \beta_6(t) = \xi_6 t^2,\end{aligned}$$

where $t \in [0, 1]$ and $\xi_j \in \{0, 1\}$, $j = 1, \dots, 6$ will be specified later.

The scalar responses $y_1, \dots, y_n$ are generated as $y_i = g(z_{i1}, \dots, z_{ip}) + \epsilon_i$, where $g(z_{i1}, \dots, z_{ip}) = \sum_{j=1}^p \int_{a_j}^{b_j} z_{ij}(t) \beta_j(t) dt$, $\epsilon_i = c r_y (Z_i - E(Z_i)) / (\text{Var}(Z_i))^{1/2}$, $c = 0.05, 0.1$ sets the amount of noise, $r_y = \max_i (g(z_{i1}, \dots, z_{ip})) - \min_i (g(z_{i1}, \dots, z_{ip}))$ and the distributions of random variables Z_i are taken from normal distribution, t_3 -distribution and χ_3^2 -distribution in all the models specified as follows: In independent case, we shall refer to the generated observations for $p = 3$ as Model 1, and to the observations generated for $p = 6$ and $\{d = e = f = 0\}$, $\{d = 1, e = f = 0\}$, $\{d = 1, e = 0.8, f = 0\}$, $\{d = 1, e = f = 0.8\}$ as Models 2, 3, 4 and 5, respectively. Model 1 was considered in Matsui and Konishi (2011), and Model 3 in Collazos et al. (2016). In Models 2, 4 and 5, slightly smaller or larger number of significant functional variables is considered than in Model 3. In correlated case, we consider the following three models specified by ξ_j , $j = 1, \dots, 6$: Model 6 with $\xi_1 = \xi_4 = 1$ and $\xi_2 = \xi_3 = \xi_5 = \xi_6 = 0$; Model 7 with $\xi_1 = \xi_4 = \xi_6 = 1$ and $\xi_2 = \xi_3 = \xi_5 = 0$; Model 8 with $\xi_1 = \dots = \xi_4 = 1$ and $\xi_5 = \xi_6 = 0$.

For simplicity, in the representation (2) of all functional regressors and coefficient functions, we used five functions of the same basis. In independent case, the Fourier and B-spline bases were considered. Since the Fourier basis is used to generate the data in correlated case, the variable selection methods are only performed with the B-spline basis in this case. We used the default values of the parameters of the RSM procedure implemented in R package `regRSM`, i.e., $B = 1000$, $|m| = \lfloor \min(n-1, \sum_{j=1}^p k_j)/2 \rfloor$, $h = \min(\lfloor n/2 \rfloor, \sum_{j=1}^p k_j)$ and the Bayesian Information Criterion as the final model selection method. The same criterion was used in Algorithm 2. In this algorithm, $h_* = \min(n/2, p) = p$ was also chosen as $1 + \sum_{r=1}^{h_*} k_{j_r} \leq n - 2$ holds for all k_j and n under consideration. The methods of Collazos et al. (2016) are referred to as the T_L^{BC} and T_L^{FDR} procedures, when the p values are corrected by the Bonferroni correction and the false discovery rate procedure of Benjamini and Yekutieli (2001),

respectively. They were performed with significance level $q = 0.01$ of their testing procedure. This level delivered the best results among levels considered by Collazos et al. (2016). The optimal value of the regularization parameter in the Glasso was chosen by minimizing the model selection criterion GIC (Matsui and Konishi 2011). In computing the prediction error for Algorithms 1 and 2 and the procedures of Collazos et al. (2016), we used the roughness penalty approach to estimate the vector $\mathbf{d}$ in model (5). The regularization parameter was chosen by cross-validation method implemented in the R package `fda`.

4.2 Simulation results

With 500 simulation experiments described in Sect. 4.1, the percentage of the correctly selected models (PCSM), the average of the false discovery rates (AFDR; $FDR = |\hat{M} \setminus M|/|\hat{M}|$, where $\hat{M}$ is the model selected by a given procedure and M is the set of relevant variables, i.e., coefficients corresponding to them in the model are nonzero) and the average of the mean square errors (AMSE) were computed and the results are given in Tables 1, 2, 3 and 4 and Tables 1–8 in the Supplementary Materials. The false discovery rate measures a fraction of chosen spurious functional regressors (false positives) with respect to all selected functional variables. When this rate is equal to zero, the final model does not contain any spurious variables.

First of all, we observe that each variable selection method performs similarly under all distributions considered. Thus, both algorithms and the procedures of Collazos et al. (2016) and the Glasso seem to be robust to non-normal distribution of random errors. However, all methods are usually sensitive to amount of noise in the data. More precisely, the simulation results for $c = 0.1$ are often worse than for $c = 0.05$. This is perhaps the most evident for prediction error (AMSE), but this seems to be natural.

Now, we discuss the simulation results obtained in independent case (Models 1–5). The values of PCSM are presented in Table 1 and Tables 1–2 in the Supplementary Materials. Algorithms 1 and 2 outperform the methods of Collazos et al. (2016) with respect to PCSM in most cases for Models 2–4 and Model 5 with $c = 0.05$. For Model 5 with $c = 0.1$, Algorithm 1 with the Fourier basis performs best, but the methods of Collazos et al. (2016) have at least slightly larger PCSM than the other new methods. For Model 1, our procedures work at least as good as their competitors by Collazos et al. (2016). The Glasso is outperformed by both algorithms in almost all cases. When the number of relevant variables is small (Models 2–3), the values of PCSM for Glasso are quite large and greater than those for the methods of Collazos et al. (2016), but in the other cases the opposite holds usually true. Table 2 and Tables 3–4 in the Supplementary Materials show the value of AFDR. For new procedures, the values of AFDR are usually quite small and decrease very quickly as number of observations increases. It is seen that the T_L^{BC} and T_L^{FDR} procedures select more (Models 1 and 5) or even much more (Models 2–4) spurious variables than our methods. For Model 1, the Glasso chooses much more spurious variables than the other procedures. In remaining models and $c = 0.05$ (resp. $c = 0.1$), the values of AFDR for Glasso are comparable to (resp. may be considerably greater than) those for Algorithms 1 and 2. The superiority of Algorithms 1 and 2 over their competitors in terms of PCSM and AFDR is usually

Table 1 Percentage of the correctly selected models (PCSM) for all variable selection methods obtained in Models 1–5 under normal distribution

M	c	n	Fourier basis					B-spline basis				
			Alg1	Alg2	T_L^{BC}	T_L^{FDR}	Glasso	Alg1	Alg2	T_L^{BC}	T_L^{FDR}	Glasso
1	0.05	50	96.8	99.4	96.0	94.0	69.2	99.2	98.8	96.4	95.4	65.2
		75	98.8	99.6	98.8	97.4	65.2	99.8	99.6	98.6	98.4	63.2
		100	99.0	99.8	98.0	97.6	68.2	100.0	99.8	98.6	97.4	64.8
	0.1	50	93.8	95.2	94.8	92.4	60.4	94.0	95.6	94.8	94.0	55.0
		75	98.6	97.8	98.8	97.2	58.0	97.6	97.8	97.6	97.4	55.0
		100	98.2	99.4	97.8	97.4	54.8	99.2	99.4	98.6	97.4	52.0
	0.05	50	78.0	95.2	42.6	40.4	86.2	97.6	94.4	42.8	41.4	80.4
		75	89.2	98.8	78.4	75.4	95.0	99.8	98.8	79.0	77.2	88.0
		100	93.8	99.6	89.0	87.2	97.0	100.0	99.4	89.4	88.2	97.0
2	0.1	50	73.4	94.2	42.2	40.0	68.6	95.8	93.8	42.6	40.6	72.0
		75	85.8	99.0	78.8	76.2	82.2	99.0	98.8	79.2	77.6	81.0
		100	89.6	99.6	89.0	87.4	88.2	100.0	99.6	89.2	87.8	88.2
	0.05	50	93.0	90.4	48.8	46.0	84.6	97.2	90.2	48.0	44.4	49.4
		75	98.4	98.2	82.4	78.0	87.0	100.0	98.0	83.0	80.8	68.2
		100	97.8	99.6	91.2	88.8	92.0	100.0	99.6	91.4	87.6	84.0
	0.1	50	88.4	85.8	49.0	44.6	72.8	90.0	85.6	46.8	42.8	53.6
		75	96.2	98.0	82.6	78.2	80.6	97.0	96.8	83.4	80.6	67.0
		100	97.4	99.4	91.2	88.8	90.0	98.8	99.2	91.4	87.4	81.4
3	0.05	50	96.0	85.8	57.2	52.4	53.2	99.0	85.2	57.8	52.0	63.2
		75	99.0	97.6	88.6	84.0	59.6	99.8	97.2	89.2	85.0	90.2
		100	99.8	99.4	94.4	90.4	68.6	100.0	99.6	93.4	91.0	94.0
	0.1	50	89.6	74.2	51.6	48.2	43.4	80.4	75.2	53.2	48.4	50.4
		75	96.4	93.0	88.0	83.8	46.8	90.4	94.4	88.6	84.8	62.6
		100	99.2	98.2	93.6	91.0	52.4	97.4	98.4	93.8	90.4	66.6
	0.05	50	98.6	80.8	73.0	68.2	48.8	97.6	83.0	72.0	67.2	53.4
		75	100.0	98.2	94.6	89.6	50.2	99.8	97.8	94.6	92.4	54.4
		100	100.0	99.4	97.2	94.0	54.6	100.0	99.2	96.2	95.0	57.4
4	0.1	50	81.8	55.8	57.6	54.6	42.4	58.8	56.2	58.4	57.4	52.0
		75	92.4	79.4	87.4	85.0	47.2	68.8	80.0	86.8	86.4	57.4
		100	97.4	87.6	90.6	90.4	53.8	78.2	87.0	92.2	91.4	59.0

The column “M” refers to different simulation models

especially evident for small number of observations, i.e., $n = 50$. This may follow from the fact that the RSM procedure used in the algorithms considers variables individually (the weights and scores are computed for each variable separately) in small subspaces of variables, while the other methods group variables, when they test the significance of corresponding parameters (Collazos et al. 2016) or compute the penalty (Glasso). It seems that individual approach needs less observations than grouping one. The values of AMSE are depicted in Table 3 and Tables 5–6 in the Supplementary Materials. For

Table 2 Average of the false discovery rates (AFDR) for all variable selection methods obtained in Models 1–5 under normal distribution

M	c	n	Fourier basis					B-spline basis				
			Alg1	Alg2	T_L^{BC}	T_L^{FDR}	Glasso	Alg1	Alg2	T_L^{BC}	T_L^{FDR}	Glasso
1	0.05	50	1.07	0.20	1.33	2.00	10.00	0.27	0.40	1.20	1.53	11.53
		75	0.40	0.13	0.40	0.87	11.53	0.07	0.13	0.47	0.53	12.27
		100	0.33	0.07	0.67	0.80	10.60	0.00	0.07	0.47	0.87	11.73
	0.1	50	1.97	0.27	1.20	2.00	12.60	0.27	0.33	1.33	1.60	14.87
		75	0.47	0.13	0.33	0.87	13.90	0.07	0.13	0.53	0.60	15.00
		100	0.60	0.07	0.67	0.80	15.07	0.00	0.07	0.47	0.87	16.00
	0.05	50	8.72	1.80	25.95	27.79	2.43	1.04	2.11	25.49	27.45	0.27
		75	3.84	0.40	7.79	9.11	0.33	0.07	0.40	7.59	8.45	0.00
		100	2.21	0.13	3.82	4.57	0.27	0.00	0.20	3.75	4.37	0.00
2	0.05	50	10.79	2.19	25.93	27.87	10.54	1.61	2.21	25.61	27.68	5.94
		75	5.27	0.33	7.72	8.82	5.62	0.33	0.40	7.59	8.33	3.80
		100	3.85	0.13	3.89	4.55	3.33	0.00	0.13	3.84	4.47	2.97
	0.1	50	2.07	2.89	17.02	18.61	1.66	1.10	2.98	16.97	18.92	0.10
		75	0.46	0.45	4.66	6.17	0.50	0.00	0.50	4.63	5.44	0.05
		100	0.55	0.10	2.29	2.97	0.25	0.00	0.10	2.29	3.38	0.00
	0.05	50	3.43	3.68	16.57	18.81	4.49	2.26	3.79	16.99	19.22	2.01
		75	1.10	0.50	4.64	6.03	2.18	0.71	0.78	4.56	5.49	0.76
		100	0.65	0.10	2.29	3.03	0.93	0.20	0.15	2.31	3.43	0.30
3	0.05	50	0.88	3.51	10.24	11.65	4.73	0.28	3.57	10.41	11.95	2.51
		75	0.20	0.48	2.39	3.47	2.49	0.04	0.56	2.24	3.19	1.67
		100	0.04	0.12	1.15	2.05	1.09	0.00	0.11	1.43	1.93	1.15
	0.1	50	1.86	3.46	10.38	11.91	9.66	2.18	4.01	10.37	11.86	9.43
		75	0.75	0.72	2.39	3.48	8.77	0.48	0.68	2.33	3.15	8.16
		100	0.16	0.12	1.25	1.87	7.85	0.08	0.15	1.27	2.01	7.39
	0.05	50	0.23	3.17	4.50	5.30	1.56	0.17	2.83	4.67	5.47	0.17
		75	0.00	0.30	0.90	1.73	0.99	0.00	0.37	0.90	1.27	0.03
		100	0.00	0.10	0.47	1.00	0.21	0.00	0.13	0.63	0.83	0.00
4	0.05	50	1.48	2.91	4.63	5.54	2.71	1.99	3.80	4.64	5.54	0.92
		75	0.41	0.33	0.88	1.51	2.09	0.79	0.57	0.87	1.24	0.40
		100	0.03	0.13	0.50	0.93	0.83	0.37	0.17	0.57	0.87	0.30

The column “M” refers to different simulation models

Model 1, all methods under consideration have similar AMSE, although the values of AMSE for Glasso may be slightly greater than for the other methods. For the rest of models with the exception of Models 4–5 with $c = 0.1$, the values of AMSE for the procedures of Collazos et al. (2016) and the Glasso are usually at least slightly greater than these for Algorithms 1 and 2. For Model 5 with $c = 0.1$, the Glasso performs better than the other procedures.

Table 3 Average of the mean square errors (AMSE) for all variable selection methods obtained in Models 1–5 under normal distribution

M	c	n	Fourier basis					B-spline basis				
			Alg1	Alg2	T_L^{BC}	T_L^{FDR}	Glasso	Alg1	Alg2	T_L^{BC}	T_L^{FDR}	Glasso
1	0.05	50	0.65	0.63	0.65	0.66	0.91	0.61	0.61	0.62	0.62	0.78
		75	0.49	0.48	0.49	0.49	0.68	0.50	0.50	0.50	0.50	0.59
		100	0.42	0.42	0.43	0.43	0.58	0.45	0.45	0.46	0.46	0.51
	0.1	50	2.55	2.81	2.62	2.66	2.96	2.42	2.32	2.11	2.12	3.08
		75	1.81	1.98	1.83	1.86	2.17	1.78	1.71	1.61	1.61	2.32
		100	1.56	1.56	1.58	1.59	1.82	1.37	1.36	1.34	1.36	2.04
2	0.05	50	1.19	1.06	1.61	1.67	2.87	0.80	0.83	1.19	1.24	1.76
		75	0.84	0.78	0.93	0.96	1.99	0.60	0.61	0.73	0.75	1.25
		100	0.68	0.64	0.72	0.73	1.52	0.50	0.51	0.57	0.58	1.07
	0.1	50	5.08	4.32	6.43	6.66	10.27	3.16	3.30	4.72	4.89	6.59
		75	3.50	3.06	3.67	3.74	7.02	2.34	2.33	2.82	2.86	4.45
		100	2.83	2.51	2.82	2.86	5.32	1.90	1.90	2.16	2.20	3.69
3	0.05	50	2.23	2.32	2.94	3.02	4.57	1.66	1.75	2.14	2.21	2.88
		75	1.69	1.70	1.89	1.94	3.23	1.27	1.29	1.43	1.46	2.10
		100	1.40	1.39	1.48	1.50	2.51	1.04	1.05	1.12	1.16	1.74
	0.1	50	9.40	9.84	12.10	12.54	15.49	7.31	7.87	8.97	9.03	9.84
		75	6.82	6.72	7.46	7.66	10.71	5.35	5.19	5.56	5.67	6.99
		100	5.56	5.53	5.82	5.91	8.29	4.36	4.11	4.33	4.48	5.77
4	0.05	50	3.48	3.73	4.09	4.18	4.92	2.68	2.87	3.12	3.18	3.55
		75	2.70	2.72	2.85	2.90	3.44	2.17	2.20	2.28	2.31	2.62
		100	2.25	2.25	2.32	2.36	2.78	1.83	1.84	1.90	1.92	2.16
	0.1	50	14.52	18.79	18.50	18.34	16.45	15.00	15.47	13.77	13.57	12.26
		75	10.68	12.15	11.26	11.40	12.01	10.66	9.38	8.64	8.74	8.85
		100	8.69	9.08	9.09	9.14	9.71	7.71	7.04	7.00	7.07	7.49
5	0.05	50	6.59	7.01	7.11	7.18	8.21	5.20	5.17	5.28	5.33	5.36
		75	5.18	5.23	5.29	5.35	5.93	3.98	4.02	4.06	4.08	3.97
		100	4.44	4.46	4.50	4.55	4.74	3.43	3.45	3.50	3.51	3.32
	0.1	50	29.22	43.50	34.58	34.25	26.71	34.12	40.65	26.78	25.66	17.51
		75	22.15	30.17	23.35	22.99	19.23	27.03	25.91	18.33	17.82	12.72
		100	18.06	22.81	19.69	19.03	15.30	21.70	18.72	14.70	14.36	10.76

The column “M” refers to different simulation models

The simulation results obtained in correlated case (Models 6–8) are depicted in Table 4 and Tables 7–8 in the Supplementary Materials. Similar results to those for Models 6–8 were obtained respectively for the following cases: $\xi_1 = \xi_3 = 1$ and $\xi_2 = \xi_4 = \xi_5 = \xi_6 = 0$, $\xi_1 = \xi_5 = 1$ and $\xi_2 = \xi_3 = \xi_4 = \xi_6 = 0$, $\xi_1 = \xi_6 = 1$ and $\xi_2 = \xi_3 = \xi_4 = \xi_5 = 0$; $\xi_1 = \xi_3 = \xi_5 = 1$ and $\xi_2 = \xi_4 = \xi_6 = 0$; $\xi_1 = \xi_2 = \xi_5 = \xi_6 = 1$ and $\xi_3 = \xi_4 = 0$. Therefore, the results for these cases are omitted to save space. Algorithms 1 and 2 outperform the methods of Collazos

Table 4 Percentage of the correctly selected models (PCSM), average of the false discovery rates (AFDR) and average of the mean square errors (AMSE) for all variable selection methods obtained in Models 6–8 under normal distribution

M	c	n	PCSM			AFDR			AMSE								
			Alg1	Alg2	T_L^{BC}	T_L^{FDR}	Glasso	Alg1	Alg2	T_L^{BC}	T_L^{FDR}	Glasso					
6	0.05	50	88.2	95.0	46.0	45.0	78.3	4.97	1.97	24.21	25.85	2.74	0.26	0.24	0.32	0.33	0.81
		75	96.0	99.2	80.6	79.0	86.4	1.63	0.27	7.33	8.44	0.58	0.20	0.19	0.22	0.22	0.58
		100	96.4	99.8	90.4	88.8	95.6	1.34	0.07	3.33	4.04	0.41	0.17	0.16	0.17	0.17	0.49
	0.1	50	81.0	92.8	43.2	42.0	69.7	7.55	2.39	24.51	26.15	10.91	0.98	0.89	1.63	1.62	2.83
		75	94.0	99.0	80.0	78.4	78.9	2.07	0.33	7.33	8.44	6.11	0.70	0.66	0.83	0.84	1.83
		100	96.4	99.8	90.4	88.8	86.4	1.30	0.07	3.33	4.04	3.41	0.54	0.53	0.58	0.58	1.48
7	0.05	50	79.2	89.4	48.0	45.6	51.8	6.45	3.31	15.83	17.62	2.89	0.50	0.49	0.76	0.76	1.04
		75	92.2	99.6	82.2	78.8	66.5	2.15	0.10	4.48	5.96	1.31	0.37	0.36	0.46	0.45	0.74
		100	95.8	99.6	93.4	91.6	85.2	1.14	0.10	1.51	2.23	0.87	0.31	0.30	0.33	0.33	0.63
	0.1	50	50.2	53.6	22.0	22.0	27.8	10.26	4.36	17.24	18.59	6.72	2.35	3.00	4.14	4.18	3.57
		75	67.8	75.2	40.4	39.4	34.6	5.32	1.41	4.97	6.32	2.93	1.66	1.89	3.43	3.37	2.36
		100	76.2	85.0	52.6	52.8	49.1	3.75	0.73	1.68	2.32	1.35	1.41	1.57	2.64	2.55	1.92
8	0.05	50	66.4	43.8	44.0	41.6	42.9	1.72	2.51	9.65	11.13	4.61	1.56	2.02	1.83	1.79	2.25
		75	72.0	53.2	67.8	65.4	66.3	0.37	0.16	2.48	3.57	2.55	1.39	1.62	1.36	1.32	1.86
		100	72.0	62.8	75.4	74.8	76.1	0.12	0.04	1.39	1.97	1.27	1.17	1.39	1.03	0.99	1.31
	0.1	50	33.2	6.2	12.4	13.4	10.6	4.37	1.84	11.28	12.13	9.49	6.07	10.04	16.80	17.95	8.76
		75	39.6	5.2	13.6	13.0	11.2	1.83	0.39	3.07	3.94	8.83	4.96	7.99	12.3	13.07	7.31
		100	42.8	5.0	17.0	19.0	14.9	0.87	0.17	1.72	2.20	7.94	3.95	6.70	8.67	9.29	6.21

The column “M” refers to different simulation models

et al. (2016) and the Glasso in almost all situations. Only when number of highly correlated relevant variables is greater (Model 8), the values of PCSM for Algorithm 2 are significantly smaller than for the Glasso and the T_L^{BC} and T_L^{FDR} procedures. However, the corresponding values of AFDR and AMSE for these methods indicate that Algorithm 2 selects almost only relevant variables, which have predictive power comparable to or better than relevant and spurious variables chosen by the methods of Collazos et al. (2016) and the Glasso. When the number of (more or less correlated) relevant variables is small or moderate (Models 6–7), Algorithm 2 performs better than Algorithm 1 in terms of correctly selected models and prediction error. However, the opposite holds true, when the number of relevant variables is greater (Model 8). The number of spurious variables in the final model is usually smaller for Algorithm 2 than in the case of Algorithm 1.

Summarizing, Algorithms 1 and 2 outperform the methods of Collazos et al. (2016) and the Glasso in terms of the correctly selected model, the false discovery rate and the mean square error, possibly with the exception of the case of noisy data with large number of relevant variables. The values of PCSM for Algorithms 1 and 2 are usually quite large, which indicates that the methods based on RSM select most relevant functional regressors. The number of spurious variables in the final models of both algorithms is small in most cases, which is important in the case of costly screening of such variables. Algorithm 1 and 2 select functional variables of very satisfactory predictive power.

The new procedures usually give similar results regardless of basis used. However, sometimes greater differences are observed. For example, for Model 2 (resp. Model 5 and $c = 0.1$) the values of AFDR (resp. PCSM) for Algorithm 1 and Fourier basis are evidently greater than those for the other new methods. In simulations and real data example of Sect. 5, we use default value of the subspace size $|m|$ of the RSM procedure. Reasonability of this choice is confirmed by our simulation results for Algorithms 1 and 2 and different values of $|m|$. Exemplary results are presented in Figures 1–3 in the Supplementary Materials, where $|m| = 5, 10, 15, 20, 25$ are considered and 15 is the default value used in our simulations for Model 3. However, we observe that smaller values of $|m|$ than the default one may result in at least slightly better performance of both algorithms. On the other hand, both algorithms behave rather worse for larger values of $|m|$ than for the default one. These all observations suggest that it may be needed to choose the algorithm, the basis, etc., carefully in practice. Such a choice may be based on prediction error obtained for selected functional regressors (See also the suggestions for selecting the basis given in Sect. 2.1.). More precisely, we perhaps should choose that variable selection method and those values of its parameters, which minimize prediction error of chosen functional regression model (see Sect. 5 for real data example). However, the smallest prediction error may not go hand in hand with choosing relevant variables.

4.3 Statistical comparison of variable selection methods

To see differences between the variable selection methods more exactly, we present a detailed statistical comparison (see Górecki and Smaga 2015, for a similar compari-

son). We test the null hypothesis that all methods perform the same and the observed differences are merely random. We used the Iman and Davenport (1980) test, which is a nonparametric equivalent of ANOVA. We perform this test separately for the results of the simulations for PCSM, AFDR and AMSE and for Models 1–5 and 6–8. Since in each case the differences between methods were significant (the p -values are given in Tables 5 and 6), we proceed with the Nemenyi post hoc test to detect significant pairwise differences among all of the methods (Nemenyi 1963). The Nemenyi test, recommended by Demšar (2006), is similar to the Tukey test for ANOVA and is used when all methods are compared with each other. Let R_{ij} be the rank of the j th of K methods on the i th of N settings, and $R_j = \frac{1}{N} \sum_{i=1}^N R_{ij}$. The performance of two methods is significantly different at the experimentwise error rate α if

$$|R_i - R_j| > q(\alpha, K, \infty) \sqrt{\frac{K(K+1)}{12N}}, \quad i = 1, \dots, K-1, \quad j = i+1, \dots, K,$$

where the values of $q(\alpha, K, \infty)$ are based on the Studentized range statistic.

The results of multiple comparisons of simulation results obtained under normal distribution (Tables 1, 2, 3 and 4) are given in Tables 5 and 6. Under t_3 -distribution and χ_3^2 -distribution, the results are similar, so they are omitted. Those methods that are connected by a sequence of letters have average ranks that are not significantly different from one another. For PCSM, the best procedures are in the first group (denoted by “a”), while for AFDR and AMSE, the best methods are in the last group. The results of multiple comparisons confirm the conclusions about simulation results given in Sect. 4.2.

5 Real data application

In chemometrics, the content of certain ingredients is usually determined by such functions as for example absorbance or emission spectra, which are usually much cheaper than chemical analysis. We investigate a sugar data set considered in Section 5.2 of Gertheiss et al. (2013), which contains the measurements of such spectra. This data set is described in Munck et al. (1998) and Bro (1999) and can be downloaded from the following address: http://www.models.life.ku.dk/Sugar_Process. The sample size is 268. Sugar was dissolved in un-buffered water, and the solution was measured spectrofluorometrically. For each sample of sugar, the emission spectra from 275–560 nm were measured in 0.5 nm intervals (i.e., at 571 wavelengths) at seven excitation wavelengths 230, 240, 255, 290, 305, 325 and 340 nm. Laboratory determinations of the quality of the sugar are also contained in the data set. One of them is ash content measuring the amount of inorganic impurities in the refined sugar. As Gertheiss et al. (2013), we would like to study the association between the ash content and the spectra measured at seven excitation wavelengths. More precisely, our aim is to determine the most useful excitation wavelengths to predict the ash content. If the results of our study indicate that not all excitation wavelengths have to be used in the analysis, it could become even easier and cheaper than chemical analysis.

Table 5 Results of the Nemenyi post hoc test for the PCSM, AFDR and AMSE of the variable selection methods under consideration obtained in Models 1–5 under normal distribution

PCSM			AFDR			AMSE		
Meth.	Ranks mean	Homogen. groups	Meth.	Ranks mean	Homogen. groups	Meth.	Ranks mean	Homogen. groups
Alg1B	8.67	a	T_L^{FDRF}	9.12	a	GlasseF	9.13	a
Alg1F	8.03	a	T_L^{FDRB}	8.72	a	T_L^{FDRF}	7.88	a
Alg2F	8.03	a	T_L^{BCF}	6.80	a	T_L^{BCF}	7.13	a
Alg2B	7.72	a	T_L^{BCB}	6.67	a	GlasseB	6.32	b
T_L^{BCB}	5.45	b	GlasseF	6.65	a	Alg2F	6.23	b
T_L^{BCF}	5.15	c	GlasseB	4.63	b	Alg1F	5.38	c
T_L^{FDRB}	3.25	c	Alg1F	4.00	c	T_L^{FDRB}	3.78	d
GlasseF	3.08	c	Alg2B	3.73	c	T_L^{BCB}	3.12	c
GlasseB	2.97	d	Alg2F	2.92	c	Alg2B	3.02	c
T_L^{FDRF}	2.65	e	Alg1B	1.77	d	Alg1B	3.00	e

The critical value of this test is equal to 2.473165. The p values of the Iman and Davenport (1980) test are less than 2.2e–16. After indicators of the methods, the letters F and B denote that the Fourier and B-spline basis was used, respectively

Table 6 Results of the Nemenyi post hoc test for the PCSM, AFDR and AMSE of the variable selection methods under consideration obtained in Models 6-8 under normal distribution

PCSM (3.376e-05)			AFDR (4.069e-11)			AMSE (3.406e-06)		
Meth.	Ranks mean	Homogen. groups	Meth.	Ranks mean	Homogen. groups	Meth.	Ranks mean	Homogen. groups
Alg1B	4.17	a	T_L^{FDRB}	4.83	a	GlassoB	4.11	a
Alg2B	3.78	a	T_L^{BCB}	3.67	a	T_L^{BCB}	3.56	a
T_L^{BCB}	2.81	a	GlassoB	2.78	b	T_L^{FDRB}	3.50	a
GlassoB	2.28	b	Alg1B	2.61	b	Alg2B	2.06	b
T_L^{FDRB}	1.97	b	Alg2B	1.11	c	Alg1B	1.78	b

The critical value of this test is equal to 1.437663. The p -values of the Iman and Davenport (1980) test are given in parentheses. After indicators of the methods, the letter B denotes that the B-spline basis was used

Table 7 Selected predictors (excitation wavelengths) for the sugar data set and results for 5-fold cross-validation

Method	q	Final model	5-fold Cross-validation									
			Prediction error	Mean size	Number of selection of variable							
					230	240	255	290	305	325	340	
Alg1		{240, 290, 305, 325, 340}	2.48	4.0	0	4	0	5	1	5	5	
Alg2		{290, 325, 340}	2.55	2.8	0	0	0	5	0	4	5	
T_L^{BC}	0.01	{255, 290, 340}	3.18	2.4	0	1	3	3	0	0	5	
T_L^{FDR}	0.01	{255, 290, 340}	5.88	2.0	0	1	3	2	0	0	4	
T_L^{BC}	0.05	{240, 255, 290, 340}	2.75	3.0	0	2	4	4	0	0	5	
T_L^{FDR}	0.05	{240, 255, 290, 340}	2.75	3.0	0	2	4	4	0	0	5	
T_L^{BC}	0.10	{240, 255, 290, 340}	2.74	3.2	0	3	4	4	0	0	5	
T_L^{FDR}	0.10	{240, 255, 290, 340}	2.75	3.4	1	3	4	4	0	0	5	
Glasso		{230, 240, 255, 290, 305}	2.91	3.4	2	3	2	5	2	1	2	

Since for each excitation wavelength, the emission spectra were measured at 571 wavelengths, the measurements can be treated as the values of a function at 571 design time points, i.e., discretized functional data. Then, the excitation wavelengths are seen as functional variables. The response variable is ash content. The relationship between these variables can be modeled by the scalar response functional regression model (1) with seven functional regressors. We apply Algorithms 1 and 2 as well as the methods of Collazos et al. (2016) and the Glasso for variable selection in this model. Since the spectra curves do not seem to be periodical, we use five B-spline basis functions only in our analysis. The other parameters of the methods are chosen in the same way as in simulations of Sect. 4, but additionally the procedures of Collazos et al. (2016) are also applied with significance levels $q = 0.05$ and $q = 0.1$. The selected variables by each procedure are given in Table 7. Moreover, this table shows the prediction error and the size of selected models averaged over 5 cross-validation splits as well as the number of times each variable was selected. We observe that the excitation wavelengths 230 and 305 nm are selected very rarely. Only the Glasso chooses these excitation wavelengths noticeably more often than other procedures. On the other hand, all methods (except possibly the Glasso) select the excitation wavelengths 290 and 340 nm. Additionally to these variables, Algorithms 1 and 2 choose the excitation wavelengths 325 nm and possibly 240 nm, while the methods of Collazos et al. (2016) (resp. the Glasso) select 240 and 255 nm (resp. 230, 240, 255 and 305 nm) ones. As the models obtained by both algorithms have the prediction errors smaller than the T_L^{BC} , T_L^{FDR} and Glasso procedures, the excitation wavelengths selected by Algorithms 1 and 2 should be preferred. From economic point of view, the model suggested by Algorithm 2 should perhaps be used, since it reduces the number of excitation wavelengths from seven to three.

6 Conclusions

In this paper, we have introduced and studied new variable selection methods for the scalar response functional regression model. The representation of this model by certain multiple linear regression model has allowed to appropriately use the very promising random subspace method of Mielniczuk and Teisseyre (2014) in variable selection problem under consideration. Our simulation study and real data example showed that the proposed methods lead to very good results, usually better than those of existing ones. Moreover, our methods are able to consider models with larger number of functional variables than some recent solutions. The computation time of novel algorithms can be reduced significantly by using parallel implementation of the RSM procedure in the R package `regRSM` (Teisseyre et al. 2016). The procedures were used with default values of parameters and two “standard” bases, therefore there might be a potential for their better behavior. Moreover, the new methods select functional variables individually. Perhaps appropriate use of interactions of variables could lead to a better performance of the variable selection process. This may be an interesting direction of the future research.

Acknowledgements The authors are thankful to three anonymous referees for helpful suggestions and comments for the improvement of this manuscript.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Aneiros G, Bongiorno EG, Cao R, Vieu P (eds) (2017a) Functional statistics and related fields. Springer, Contributions to Statistics
- Aneiros G, Cao R, Fraiman R, Vieu P (2017b) Special issue on functional data analysis and related topics. *J Multivariate Anal* (to appear)
- Aneiros G, Vieu P (2014) Variable selection in infinite-dimensional problems. *Statist Probab Lett* 94:12–20
- Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *Ann Statist* 29:1165–1188
- Bongiorno EG, Salinelli E, Goia A, Vieu P (2014) Contributions in infinite-dimensional statistics and related topics. Societa Editrice Esculapio
- Bro R (1999) Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis. *Chemom Intell Lab Syst* 46:133–147
- Bunea F, Wegkamp MH, Auguste A (2006) Consistent variable selection in high dimensional regression via multiple testing. *J Statist Plann Inference* 136:4349–4364
- Cardot H, Sarda P (2011) Functional linear regression. In: Ferraty F, Romain Y (eds) *The Oxford handbook of functional data analysis*. Oxford University Press, Oxford, pp 21–46
- Chen D, Hall P, Müller HG (2011) Single and multiple index functional regression models with nonparametric link. *Ann Statist* 39:1720–1747
- Chiou JM, Müller HG, Wang JL (2004) Functional response models. *Stat Sin* 14:675–693
- Chiou JM, Yang YF, Chen YT (2016) Multivariate functional linear regression and prediction. *J Multivariate Anal* 146:301–312
- Collazos JAA, Dias R, Zambom AZ (2016) Consistent variable selection for functional regression models. *J Multivariate Anal* 146:63–71
- Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 7:1–30

- Fan YY, Foutz N, James GM, Jank W (2014) Functional response additive model estimation with online virtual stock markets. *Ann Appl Stat* 8:2435–2460
- Faraway J (1997) Regression analysis for a functional response. *Technometrics* 39:254–261
- Febrero-Bande M, Gonzalez-Manteiga W (2013) Generalized additive models for functional data. *TEST* 22:278–292
- Ferraty F, Vieu P (2006) *Nonparametric functional data analysis: theory and practice*. Springer, New York
- Gertheiss J, Maity A, Staicu AM (2013) Variable selection in generalized functional linear models. *Stat* 2:86–101
- Goia A, Vieu P (2015) A partitioned single functional index model. *Comput Statist* 30:673–692
- Górecki T, Smaga Ł (2015) A comparison of tests for the one-way ANOVA problem for functional data. *Comput Statist* 30:987–1010
- Hall P, Horowitz JL (2007) Methodology and convergence rates for functional linear regression. *Ann Statist* 35:70–91
- Hong Z, Lian H (2011) Inference of genetic networks from time course expression data using functional regression with lasso penalty. *Commun Stat Theory Methods* 40:1768–1779
- Horváth L, Kokoszka P (2012) *Inference for functional data with applications*. Springer, New York
- Hubert M, Rousseeuw PJ, Segal P (2015) Multivariate functional outlier detection. *Stat Methods Appl* 24:177–202
- Iman RL, Davenport JM (1980) Approximations of the critical region of the Friedman statistic. *Commun Statist Theory Methods* 9:571–595
- James GM (2002) Generalized linear models with functional predictors. *J R Stat Soc Ser B Stat Methodol* 64:411–432
- Kayano M, Konishi S (2009) Functional principal component analysis via regularized Gaussian basis expansions and its application to unbalanced data. *J Stat Plann Inference* 139:2388–2398
- Kokoszka P, Oja H, Park B, Sangalli L (2017) Special issue on functional data analysis. *Econom Stat* 1:99–100
- Krzyśko M, Waszak Ł (2013) Canonical correlation analysis for functional data. *Biom Lett* 50:95–105
- Li Y, Hsing T (2007) On rates of convergence in functional linear regression. *J Multivariate Anal* 98:1782–1804
- Matsui H (2014) Variable and boundary selection for functional data via multiclass logistic regression modeling. *Comput Statist Data Anal* 78:176–185
- Matsui H, Araki Y, Konishi S (2008) Multivariate regression modeling for functional data. *J Data Sci* 6:313–331
- Matsui H, Konishi K (2011) Variable selection for functional regression models via the L_1 regularization. *Comput Statist Data Anal* 55:3304–3310
- Mielniczuk J, Teisseyre P (2014) Using random subspace method for prediction and variable importance assessment in regression. *Comput Statist Data Anal* 71:725–742
- Munck L, Nørgaard L, Engelsen SB, Bro R, Andersson CS (1998) Chemometrics in food science—a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance. *Chemom Intell Lab Syst* 44:31–60
- Müller HG, Stadtmüller U (2005) Generalized functional linear regression. *Ann Stat* 33:774–805
- Nemenyi PB (1963) *Distribution-free multiple comparisons*. Dissertation, Princeton University
- Radchenko P, Qiao X, James GM (2015) Index models for sparsely sampled functional data. *J Am Stat Assoc* 110:824–836
- Ramsay JO, Hooker G, Graves G (2009) *Functional Data Analysis with R and MATLAB*. Springer, Berlin
- Ramsay JO, Silverman BW (2002) *Applied functional data analysis. Methods and case studies*. Springer, New York
- Ramsay JO, Silverman BW (2005) *Functional data analysis*, 2nd edn. Springer, New York
- Ramsay JO, Wickham H, Graves S, Hooker G (2017) *fda: Functional Data Analysis*. R package version 2.4.7 <http://CRAN.R-project.org/package=fda>
- R Core Team (2017) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- Teisseyre P, Kłopotek RA, Mielniczuk J (2016) Random subspace method for high-dimensional regression with the R package *regRSM*. *Comput Statist* 31:943–972
- Yuan M, Cai T (2010) A reproducing kernel Hilbert space approach to functional linear regression. *Ann Statist* 38:3412–3444
- Zhang JT (2013) *Analysis of variance for functional data*. Chapman and Hall, London