Neurocomputing 422 (2021) 235–244

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Region of interest selection for functional features

Qiyue Wang [a,*], Yao Lu [a], Xiaoke Zhang [b], James Hahn [a]

[a] Department of Computer Science, The George Washington University, USA
[b] Department of Statistic, The George Washington University, USA

ARTICLE INFO

Article history:
Received 11 July 2019
Revised 14 June 2020
Accepted 3 October 2020
Available online 14 October 2020
Communicated by Steven Hoi

Keywords:
Feature selection
Functional data
Machine learning

ABSTRACT

Feature selection is a critical component in supervised learning to improve model performance. Searching for the optimal feature candidates can be NP-hard. With limited data, cross-validation is widely used to alleviate overfitting, which unfortunately suffers from high computational cost. We propose a highly innovative strategy in feature selection to reduce the overfitting risk but without cross-validation. Our method selects the optimal sub-interval, i.e., region of interest (ROI), of a functional feature for functional linear regression where the response is a scalar and the predictor is a function. For each candidate sub-interval, we evaluate the overfitting risk by calculating a necessary sample size to achieve a pre-specified statistical power. Combining with a model accuracy measure, we rank these sub-intervals and select the ROI. The proposed method has been compared with other state-of-the-art feature selection methods on several reference datasets. The results show that our proposed method achieves an excellent performance in prediction accuracy and reduces computational cost substantially.

1. Introduction

In supervised learning, we attempt to make predictions based on what we learn from limited existing data. However, the inherently high-dimensional real data can be a curse for learning. To generalize a learning model well, the amount of data needed is expected to grow exponentially with data dimensionalities (i.e., features). Feature selection (FS) process is extremely important to reduce data dimensionality [1]. In general, most FS methods can be categorized into three types: the filter methods, the wrapper methods and the embedded methods [2]. Filter methods evaluate features based on their individual mapping potency to the response [3]. The selection process ignores the relationships between the features and is irrelevant to the choice of the learning model. Therefore, a filter method tends to choose features with high redundancy and the model trained accordingly tends to perform poorly. On the contrary, wrapper or embedded methods fully consider a learning model during FS [4]. Wrapper methods consider the FS as a searching problem, which evaluates the subsets of features based on their performance under the learning model. Therefore, wrapper methods typically result in a better performance. However, with insufficient but high-dimensional training samples, wrapper methods often suffer from overfitting and the resulting learning model cannot be generalized well with the selected features [2]. To deal with the overfitting problem, cross-validation is typically introduced to wrapper models to evaluate the potential risk of overfitting and to select the optimal feature subset that achieves the best trade-off between the variance and bias [5]. However, the cross-validation process either significantly increases the computational workload, e.g., leave-one-out-cross-validation (LOOCV), or suffers from result uncertainties due to random data partitioning, e.g., K-fold cross-validation. Unlike the wrapper methods, the embedded methods alleviate overfitting during FS with a penalty against complexity [6,7]. However, this regularization to drop redundant features may perform poorly when a dataset contains highly intra-correlated relevant features. In this paper, we propose a novel method that can evaluate the overfitting risk without cross-validation.

Many real-world features, such as sound and images, are in essence of a continuous or functional form. Whilst the recording process inevitably discretizes a feature, the intrinsic order and continuity (i.e., smoothness and dependency) of these discretized measurements carry important information on this feature. Thus, treating these measurements naively as multivariate features in the modeling is very inefficient and often computationally unstable. Functional data analysis (FDA) [8–12], an increasingly important area in statistics, has shown its superiority in dealing with this type of data, called "functional data," which considers the feature as a function varying over a continuum. The functionalization process turns the high-dimensionality curse into a blessing and shows robustness in dealing with data with different sampling rate [13]. In this paper, we focus on functional linear regression

* Corresponding author.
E-mail address: wangqiyue@gwu.edu (Q. Wang).

https://doi.org/10.1016/j.neucom.2020.10.009
0925-2312/

with a scalar response and a functional feature and we aim to locate the optimal interval for the domain of the functional feature, i.e., the region of interest (ROI).

The main contribution of this paper is threefold. First, we propose a novel measure to evaluate the risk of overfitting based on a statistical modeling framework, i.e., functional linear regression. Second, our framework trades off the model accuracy and overfitting risk without the need for splitting data as in cross-validation, which effectively reduces the computational cost. Third, our method is highly applicable and effective for moderate datasets.

2. Related works

We will first review the functional FS methods and then list representative methods for multivariate FS which can be adapted to functional ROI through some preprocessing steps.

Functional ROI selection. Our aim here is to select the ROI for a functional feature within functional linear regression. Although the functional feature is always discretely measured over its domain in practice, the classical FS methods, which are originally designed for multivariate features, cannot be applied directly since they cannot take into account the intrinsic order and dependencies between these measurements. Prior work on this topic is limited, which only include [14,15] to the best of our knowledge. They both transformed the functional linear regression model to a classical linear regression model by approximating the functional feature using B-spline basis functions and then applied a LASSO-type method [20] to select the B-spline coefficients. Since each B-spline basis is defined on a local region, the ROI of the functional feature can be selected accordingly.

Extended Multivariate FS methods. Multivariate FS methods are usually categorized into three types: filter methods, wrapper methods and embedded methods [3]. A filter method typically evaluates the informativeness of each feature individually using a criterion score regardless of the learning model [16–18]. Without a training model, filter methods are typically fast and robust. In contrast a wrapper method involves a learning model during FS and ranks the candidate feature subsets according to the learning performance [5,19]. Wrapper methods usually perform better performance but suffer from higher computation cost. Embedded methods are similar to wrapper methods in that they both search for a feature subset that fits the model best, but they do not separate the feature selection from model training which increase the efficiency of FS. LASSO [20] is a commonly used embedded method, which introduces the L1 regularization to penalize model complexity and excludes ineffective features simultaneously. Similarly, ridge regression adopts the L2 regularization and Elastic-Net [21] combines the L1 and L2 regularizations. LASSO has been further developed for feature selection purposes in many recent works [22,23,2]. Embedded methods take advantage of both filter methods and wrapper methods. Compared to the wrapper methods, embedded methods are typically faster and suffer less from overfitting. To extend the general multivariate FS methods to functional data, the functional feature can be transformed with some basis functions such as the eigenfunctions obtained from the functional principal component analysis, Fourier functions and B-spline bases. The transformed features are usually very similar to classical multivariate features and multivariate FS methods can be accordingly applied.

3. Methodology

3.1. Functional linear models

In scalar-on-functional regression, we want to map a smooth (e.g., continuous) functional feature $X(t) \in L^2(I)$ defined on domain I to the scalar response $Y \in \mathbb{R}$. For simplicity we focus on a functional linear model as in Eq. (1). $E(X(t)) = 0, t \in I$. For example, in practice, we could center the functional feature at its cross-sectional mean. The $\beta(t) \in L^2(I)$ is the coefficient function and $\beta_0 \in \mathbb{R}$ is the intercept.

$$Y \approx \beta_0 + \int_I \beta(t)X(t)dt. \tag{1}$$

It is difficult to fit Eq. (1) directly since both $X(t)$ and $\beta(t)$ are in essence infinite-dimensional. However, they can be transformed to a classical linear model as in Eq. (2) in terms of a set of basis functions as in Eq. (3), where $\omega_k(t)$ denotes a basis function; ξ_k and θ_k are the transformed parameters of $X(t)$ and $\beta(t)$, respectively. Commonly used basis functions include the eigenfunctions from the functional principal component analysis (FPCA) of $X(t)$, B-splines and Fourier functions [24–26]. In this paper, we adopt the FPCA of $X(t)$ to transform the original functional feature.

$$Y \approx \alpha_0 + \sum_{k=1}^{\infty} \theta_k \xi_k. \tag{2}$$

$$\begin{cases} \beta(t) = \sum_{k=1}^{\infty} \theta_k \omega_k(t), \quad \theta_k = \int_I \beta(t)\omega_k(t)dt; \\ X(t) = \sum_{k=1}^{\infty} \xi_k \omega_k(t), \quad \xi_k = \int_I X(t)\omega_k(t)dt. \end{cases} \tag{3}$$

In practice, the functional feature $X(t)$ is discretely measured at N points over the domain I. We collect data from n subjects, so the upper limit of the summation is at most $\min(N, n)$. Moreover, the first C_n principal components (PCs) often suffice to approximate $X(t)$ well if they cumulatively explain at least η proportion of the variance of $X(t)$, e.g., 95%. Therefore, the linear model in Eq. (2) is almost equivalent to Eq. (4),

$$\begin{cases} Y = \alpha_0 + \sum_{k=1}^{C_n} \theta_k \xi_k + \delta; \\ C_n = \underset{S \leqslant \min(N,n)}{\arg\min} \left(\sum_{i=1}^{S} \lambda_i / \sum_{i=1}^{\min(N,n)} \lambda_i \geqslant \eta \right). \end{cases} \tag{4}$$

where $\lambda_i,\ i = 1, \ldots, C_n$ are positive eigenvalues in a decreasing order and δ is the noise term.

3.2. ROI selection for a functional feature

We define the original functional feature domain as Γ, which, without loss of generality, is assumed to be $[0, 1]$. Our goal here is to find the optimal sub-interval of Γ such that the best trade-off between the model accuracy and overfitting risk of the functional linear model in Eq. (1) is achieved. Fig. 1 illustrates our proposed method. First, for each sub-interval Γ_j of Γ, a candidate ROI, we let $I = \Gamma_j$ in Eq. (1) and then transform Eq. (1) to Eq. (4). Then we propose two measures to evaluate the model accuracy and overfitting risk in terms of Eq. (4). Finally, we define a metric that balances the two measures and select the optimal sub-interval, i.e., ROI, that achieves the best trade-off.

Model accuracy. For each sub-interval Γ_j, we solve Eq. (4) by the least square method and obtain the coefficient estimates $\hat{\theta}_i^j$ and residual $\boldsymbol{\tau_j}$ in Eq. (5), where $\mathbf{Y}$ is the vector of all subjects' response vector, $\hat{\mathbf{Y}}_j$ is the fitted response and $\hat{\boldsymbol{\xi}}_i^j$ is the vector that consists of the i^{th} PC of all subjects.

$$\boldsymbol{\tau_j} = \mathbf{Y} - \hat{\mathbf{Y}}_j, \hat{\mathbf{Y}}_j = \hat{\alpha_0} + \sum_{i=1}^{C_n^j} \hat{\theta}_i^j \hat{\boldsymbol{\xi}}_i^j. \tag{5}$$

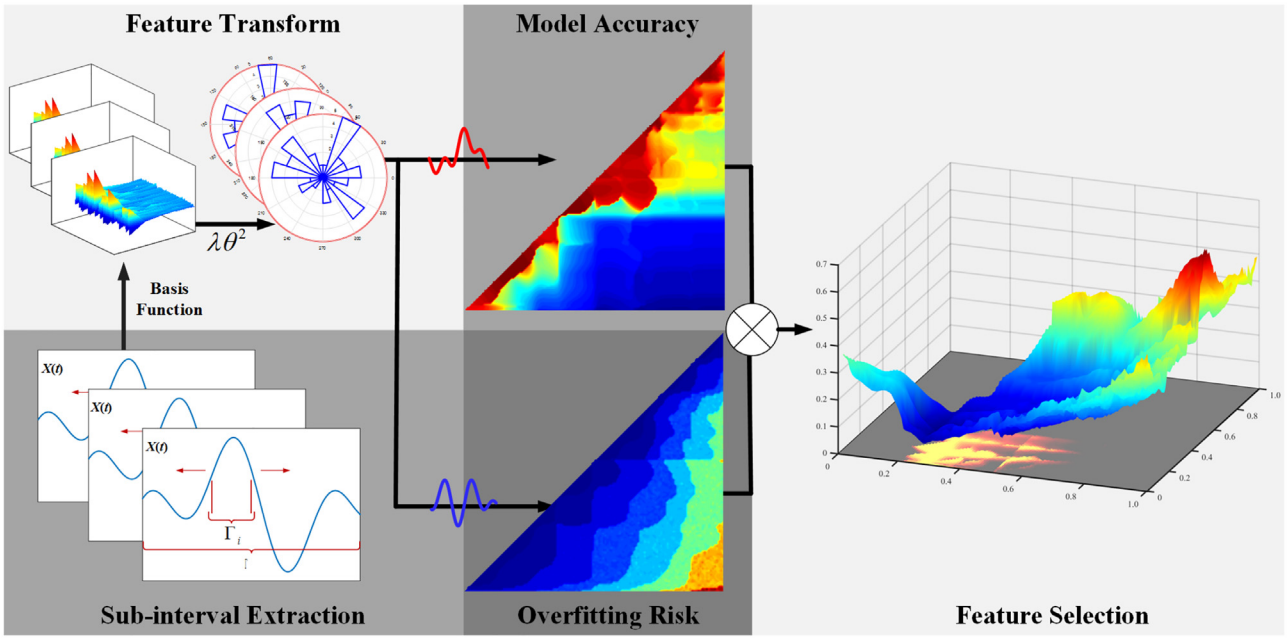


Fig. 1. An illustration of our approach. Sub-Interval Extraction: sub-intervals are defined as candidate ROIs. Feature Transform: the functional feature of each candidate ROI is transformed using basis functions, e.g., eigenfunctions via FPCA. Two measures are proposed to evaluate the model accuracy and overfitting risk. Feature Selection: the ROI is selected as the subinterval that achieves the best trade-off between the two measures.

The model accuracy is measured by $\hat{v}_j^2 = \frac{\tau_j^T \tau}{n-1-C_n^j}$, the estimated noise variance in terms of the residual $\boldsymbol{\tau}_j$, which is commonly used in classical linear regression.

Overfitting risk. A learning model trained with limited data is prone to suffer from overfitting, which means that it cannot be generalized well to unseen data. The larger the training dataset we use, the lower the overfitting risk. In theory, as the training data size increases, the training error converges to the generalization error and the overfitting risk converges to zero. We propose a novel measure to quantify the overfitting risk by evaluating the necessary sample size for the training model to achieve a sufficiently large statistical power. A small necessary sample size indicates a low risk of overfitting, whereas a large necessary sample size indicates a high risk. We will illustrate this relation using a real dataset in Section 4.3.

We follow [26] to obtain the sample size based on statistical power. First, we define

$$\hat{V}_i^j = \hat{\lambda}_i^j \left(\hat{\theta}_i^j\right)^2, i = 1, 2, \ldots, C_n^j. \tag{6}$$

The $\hat{V}_i^j$ is not only closely related to the coefficients of determination (i.e., R^2) of the training model, but also takes the variation of PCs, i.e., $\hat{\lambda}_i^j$, into account. We denote their order statistics by $\hat{V}_{(i)}^j$ in a decreasing order, with the concomitant coefficient estimate $\hat{\theta}_{(i)}^j$ and PCs $\hat{\boldsymbol{\xi}}_{(i)}^j$.

Then we define a test statistic for the null hypothesis $H_0 : \beta_j(t) = 0$ for all $t \in [0, 1]$ in Eq. (7) where γ is a pre-specified fraction, e.g., 0.95.

$$\begin{cases} T_j = \sum_{i=1}^{K_n^j} \frac{\left(\hat{\theta}_{(i)}^j\right)^2}{v\hat{a}r\left(\hat{\theta}_{(i)}^j\right)} = \frac{1}{\hat{v}_j^2} \sum_{i=1}^{K_n^j} \frac{\left(\mathbf{Y}^T \hat{\boldsymbol{\xi}}_{(i)}^j\right)^2}{\hat{\boldsymbol{\xi}}_{(i)}^{jT} \hat{\boldsymbol{\xi}}_{(i)}^j}; \\ K_n^j = \underset{S \leqslant C_n^j}{arg\,min} \left(\sum_{i=1}^{S} \hat{V}_{(i)}^j / \sum_{i=1}^{C_n^j} \hat{V}_{(i)}^j \geqslant \gamma \right). \end{cases} \tag{7}$$

By Theorem 1 of [26], under the null hypothesis, the test function T_j in Eq. (7) is asymptotically equivalent to the sum of the first K_n^j order statistics of C_n^j independent chi-square random variables with degree of freedom 1. With the significance level α, let $q_{(\alpha, K_n^j, C_n^j)}$ denote the $100(1-\alpha)\%$ quantile of T_j under the null hypothesis, which can be found accurately by 10,000 Monte Carlo simulations.

By Theorem 2 of [26], under the significance level α, the necessary sample size h_j to achieve the statistical power p is determined by Eq. (8), where Φ is the cumulative distribution function of the standard normal distribution.

$$\begin{cases} 1 - \Phi\left(\frac{q_{(\alpha, K_n^j, C_n^j)} - \left(K_n^j + \frac{h_j}{\hat{v}_j^2} Z\right)}{\sqrt{2\left(K_n^j + 2\frac{h_j}{\hat{v}_j^2} Z\right)}} \right) \geqslant p; \\ Z = \sum_{i=1}^{K_n^j} \hat{V}_{(i)}^j. \end{cases} \tag{8}$$

Once h_j is obtained, the overfitting risk is measured by

$$\psi_j = \log\left(\frac{h_j}{n} + 1\right). \tag{9}$$

The risk ψ_j is always positive as in Eq. (9).

ROI selection. We define a metric f which linearly trades off the normalized model accuracy and overfitting risk with the weight w as in Eq. (10).

$$\begin{cases} f\left(\psi_j, \hat{v}_j^2\right) = (1-w)P(\psi_j) + wP\left(\hat{v}_j^2\right); \\ P(x) = \frac{x - \min(x)}{\max(x) - \min(x)}. \end{cases} \tag{10}$$

For each sub-interval Γ_j, we compute a value of a $f\left(\psi_j, \hat{v}_j^2\right)$ following Eq. (10) and select the ROI of which f value is the smallest.

4. Discussion

The ROI is determined jointly by both model accuracy and overfitting risk via a weight w by Eq. (10). In this section we discuss how this weight influences the ROI selection in detail using a real dataset.

Dataset. Some studies indicated the human body shape descriptors, such as the circumference and waist-hip ratio, can predict the visceral adipose tissue (VAT) value. We have collected a 3D human body shape data using a depth vision sensor (DVS) based 3D human body scan and reconstruct system [27]. The VAT value is measured by the CoreScan®(GE Healthcare, Madison, WI), a Dual-energy X-ray Absorptiometry (DXA) based VAT assessment software. The accuracy has been validated by multiple studies [28,29]. Given the high cost of data collection, this is a very representative moderate-sized medical dataset, which contains 60 male and 87 female subjects. The body level circumference, one type of the anthropometric data, has been verified as a very useful shape descriptor to summarize body shape related information [30,36]. Therefore, we extract $N = 128$ equal-spaced level circumferences of the 3D body shape from the neck to the ankle (Fig. 2a). These level circumferences can be viewed as discrete measurements of a functional feature defined on the domain $\Gamma = [0, 1]$. The original functional feature derived from the level circumference is shown in Fig. 2b. We use the 60 male subjects to analyze the influence of the weight on the ROI selection.

Implementation. To study the effect of the weight w, we generate a real sequence ranging from 0.1 to 0.9 with the increment 0.1. A large w value implies an emphasis on model accuracy, while a small w implies an emphasis on suppressing overfitting. As the level circumference is measured at $N = 128$ points, for each value of the weight w, we traverse all 8128 (C^2_{128}) sub-intervals to locate the ROI following the procedures in Section 3.

We set the power $p = 0.8$, significance level $\alpha = 0.05$, threshold $\eta = 0.99$ and threshold $\gamma = 0.95$, all of which are commonly used statistical parameter settings in hypothesis testing and functional data analysis [26].

4.1. Mix weight

The heatmaps of the model accuracy and overfitting risk are illustrated in Fig. 3. Unsurprisingly, their trends are reversed: a larger sub-interval corresponds to a more complex model and consequently, the model training error decreases while the overfitting risk increases. The heatmaps of the f function in Eq. (10) with different weights are shown in Fig. 4. We also highlight the top candidate ROIs with the lowest 500 f values and their ranks have been color-coded.

As shown in Fig. 4, a large weight attaches more importance to model accuracy and thus a wider ROI will be selected; however, a larger sub-interval results in a higher risk of overfitting. This is the reason why a weighted combination is desirable to achieve a balance. Nonetheless, there is no explicit reference on how to choose the best weight, since it is highly dependent on the data and training model. For our dataset, in Fig. 5, we demonstrate the generalization error distribution of the models trained with functional features corresponding to different ROIs selected by different weight w. The generalization error is derived from the LOOCV. The difference between the best and the worst median absolute errors is over 30%. We find 0.5 is a reasonable weight. In general, if the sample size is large enough and the risk of overfitting is low, a larger weight works better and vice versa. Unless such knowledge is available, we recommend the weight 0.5 to practitioners.

4.2. Computational cost analysis

Our approach does not use the cross-validation process for ROI selection. Here we want to compare the computing time of our method with a cross-validation based method. In order to keep the results consistent, we choose the original cross-validation embedded in our functional linear regression model in Eq. (1). Meantime, we choose the mix weight $w = 0.5$ for our method. The K = 5,6,7,8,9,10 folds cross-validation and LOOCV are compared in this section.

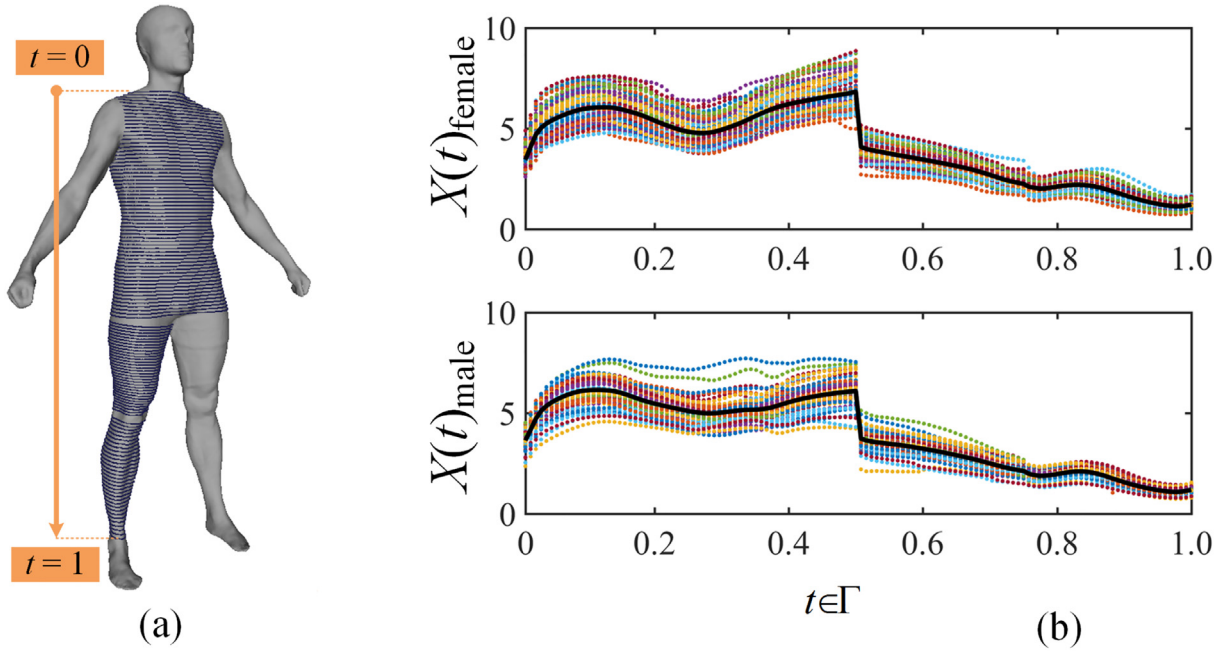


Fig. 2. Functional features of the dataset. (a) Level circumference extraction from 3D body shape. (b) The original functional feature derived from the level circumference and the functional mean (the bold line) for male and female datasets respectively.

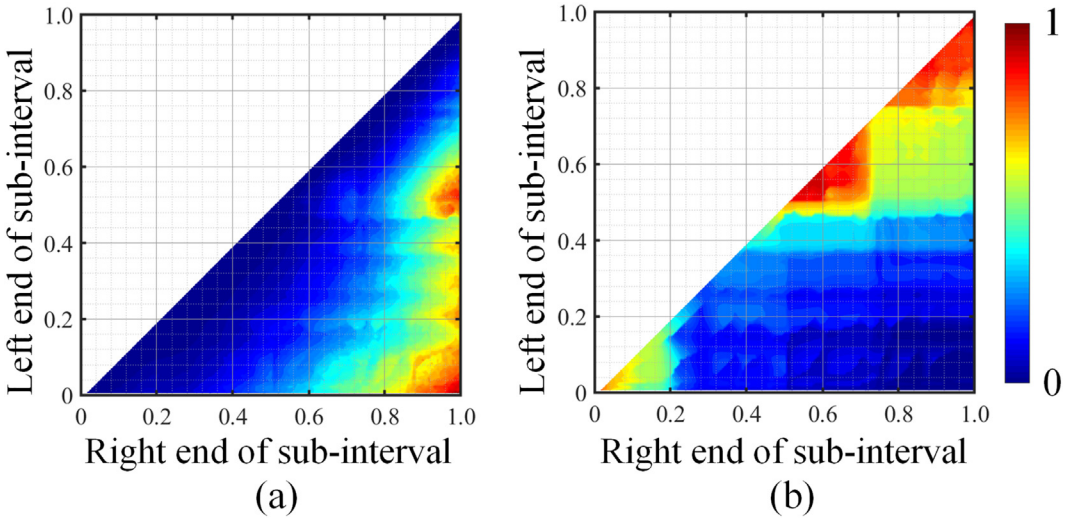


Fig. 3. Heatmaps of the model accuracy (a) and the overfitting risk (b).

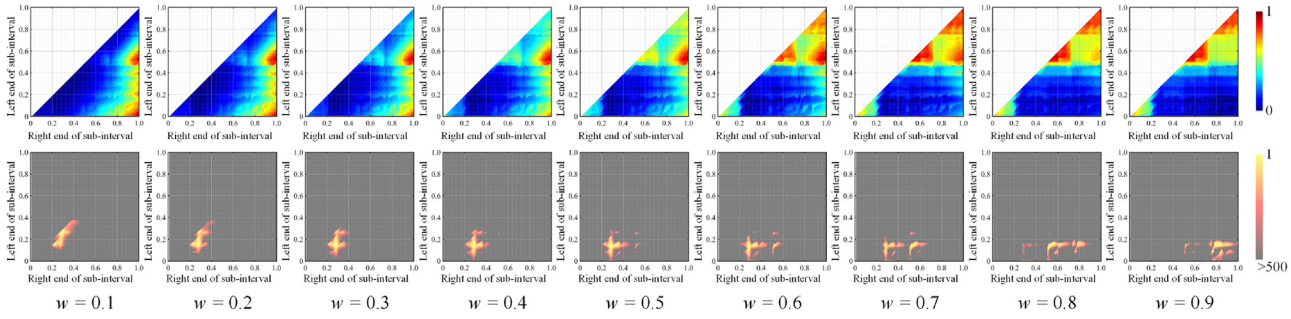


Fig. 4. ROI selection with different weights. (Top row) The f function heatmaps with different weights. (Bottom row) The distribution of the top 500 candidate ROIs corresponding to each heatmap.

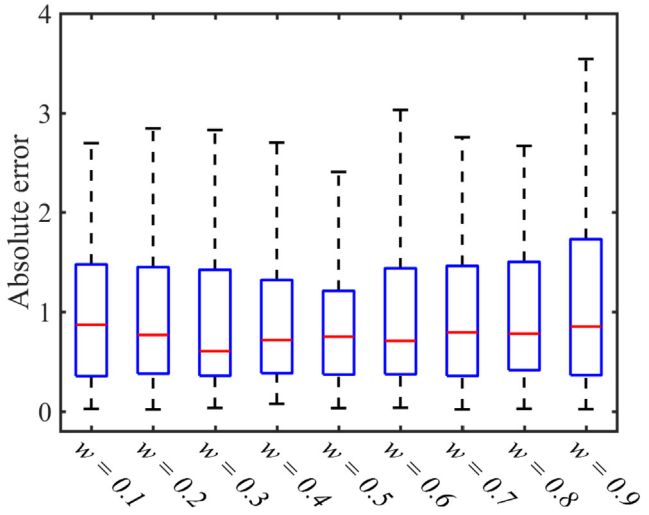


Fig. 5. Boxplots for absolute errors of models with different weights. The red horizontal bar in each boxplot refers to the median. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The results are shown in Table 1, which include the selected ROI, Root Mean Squared Error (RMSE), R- Squared value and execution time of each method in comparison. Table 1 shows that the execution time of the k-fold cross-validation is comparable with that of our method, but the ROI selected by the K-fold cross-validation is unstable for this small dataset as the partition to multiple folds is random. Thus one usually adopts LOOCV instead of K-folds cross validation for a small dataset. However, the LOOCV is unsurprisingly computationally very intensive, which as shown in Table 1 consumes the most execution time among all methods, more than six times than our proposed

Table 1
The computation complexity analysis.

Case	ROI (t)	RMSE (in^3)	R^2	Exec.Time
5-folds	[0.102, 0.920]	1.530	0.720	129
6-folds	[0.086, 0.920]	1.523	0.721	144
7-folds	[0.141, 0.531]	1.449	0.749	161
8-folds	[0.156, 0.516]	1.420	0.758	184
9-folds	[0.165, 0.820]	1.559	0.709	203
10-folds	[0.117, 0.508]	1.549	0.713	216
LOOCV	[0.078, 0.508]	1.503	0.722	1168
Proposed	[0.148, 0.281]	1.348	0.782	178

Table 2
The illustration of the relation between the overfitting risk and necessary sample size to achieve a particular statistical power using the male VAT data.

ROI	h_j	Training MSE	Test MSE	OR
[0.148, 0.281]	92	0.458	1.715	1.257
[0, 1]	233	0.412	2.381	1.969

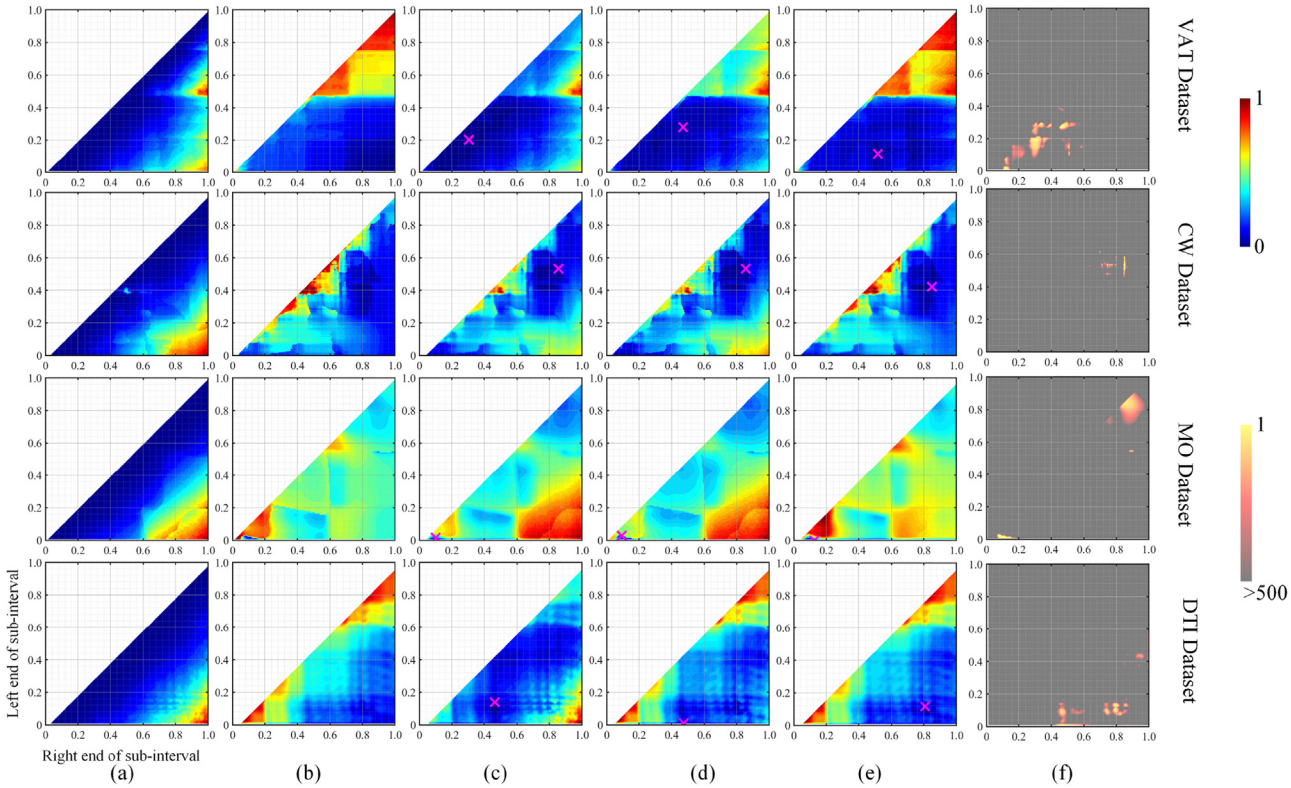


Fig. 6. The results of our approach applied to the VAT dataset, Canadian Weather (CW) dataset and Moisture (MO) dataset. (a) The overfitting risk heatmap. (b) The model accuracy heatmap. (c) (e) The f function heatmap and the selected ROI (the magenta cross) with $w = 0.3, 0.5, 0.7$. (f) The top 500 ranked candidate ROIs with $w = 0.5$.

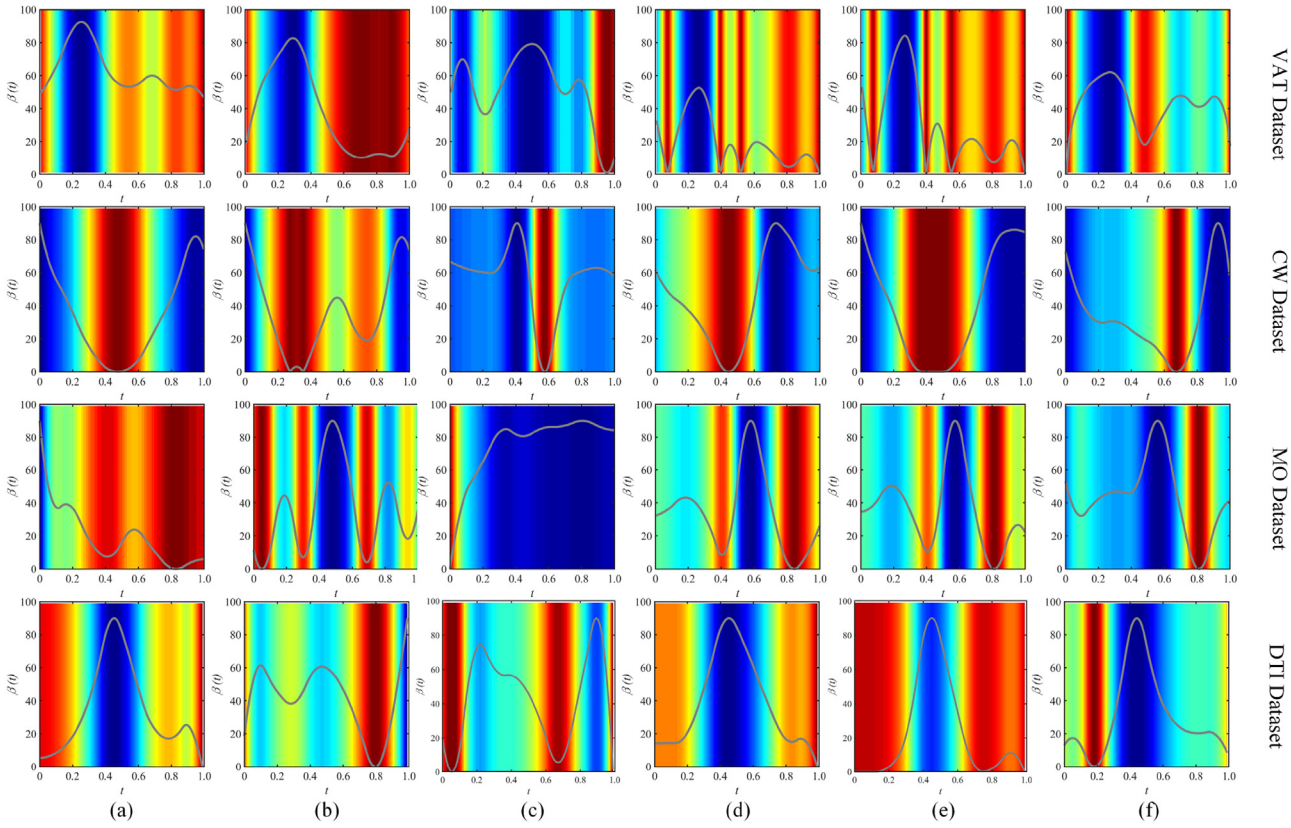


Fig. 7. ROIs selected by the reference methods: (a) DCS, (b) RReliefF, (c) SLS, (d) B-spline LASSO, (e) Elastic-Net and (f) Step-wise, according to the coefficient function $\beta(t)$. The blue region corresponds to high coefficient values; the red region corresponds to the low coefficient values; the gray curve corresponds to the coefficient function.

method. Without using cross-validation, our method is not sensitive to random partition as is the case for the K-fold cross-validation and does not suffer from intensive computation as is the case for the LOOCV.

4.3. Overfitting risk, ROI and h_j

To verify the relation between the overfitting risk and necessary sample size to achieve a particular statistical power as in Section 3.2, we use the male VAT data for illustration. We randomly partition the data into a training set (80%) and a test set (20%). We consider two ROIs, $[0.148, 0.281]$, the optimal ROI selected by our method, and the entire domain $[0, 1]$. Intuitively the second ROI $[0, 1]$ is expected to overfit the data since it includes an additional but redundant domain compared to the first ROI. For each ROI, we fit a functional linear regression, calculate the training mean squared error (MSE) and test MSE, together with the overfitting risk (OR) OR = test MSE − training MSE. We also obtain the corresponding h_j following Eq. (8) using the training set, the necessary sample size to achieve power 0.8 with the level of significance 0.05. The results are given in Table 2.

Table 2 shows that as expected the ROI $[0, 1]$ leads to a larger overfitting risk since its corresponding OR value is larger, with a correspondingly larger h_j. This confirms the relation between the overfitting risk and necessary sample size to achieve a statistical power, that is, a small necessary sample size indicates a low risk of overfitting, whereas a large necessary sample size indicates a high risk.

5. Experiment

We compare our approach with other state-of-the-art FS methods. We choose six reference methods: Distance Correlation Selection (DCS) [16], RReliefF [17], Supervised Laplacian Score (SLS) [18], B-spline LASSO [14], Elastic-Net [31] and Step-wise feature selection [19]. The DCS, RReliefF and SLS are filter methods, the Step-wise feature selection is a wrapper method. The B-spline LASSO and Elastic-Net are both embedded methods. Among these six methods, only the B-spline Lasso is originally designed for functional ROI selection. The other methods are adapted to functional ROI selection, as discussed in Section 2. For each method, we will obtain the selected ROI and assess its prediction performance.

In addition to the female VAT data in our medical dataset in Section 3, we also compare the performance of these methods when applied to the following three classical functional datasets.

Canadian Weather. The dataset includes daily temperature and precipitation at 35 different locations in Canada averaged over 1960 to 1994 [8]. The daily temperature (annually 365 days) is continuously recorded and can be taken as the functional feature. The corresponding annual rainfall is the response. The temperatures in different days play different roles in regard to the annual rainfall. Therefore, we want to obtain the most predictive duration in a year (ROI of time) for the annual rainfall.

Moisture. This data set consists of near-infrared reflectance spectra of 100 wheat samples, measured in 2 nm intervals from 1100 nm to 2500 nm and associated response variables, the

Table 3
The comparison of prediction performance. Note: The Canadian Weather Datasets is cyclic recorded, thus it allow the boundary larger than 1.

Dataset	Methods	ROI	RMSE	R-Squared
Medical	DCS	[0.172, 0.336]	1.356	0.667
	RReliefF	[0.156, 0.383]	1.367	0.662
	SLS	[0.406, 0.578]	1.419	0.635
	LASSO	[0.180, 0.328]	1.356	0.665
	Elastic-Net	[0.188, 0.328]	1.359	0.656
	Step-wise	[0.141, 0.352]	1.384	0.653
	Proposed($w = 0.3$)	**[0.223, 0.305]**	**1.006**	**0.770**
	Proposed($w = 0.5$)	**[0.289, 0.469]**	**1.105**	**0.746**
	Proposed($w = 0.7$)	**[0.117, 0.516]**	**1.326**	**0.677**
Canadian Weather	DCS	[0.835, 1.016]	4.314	0.729
	RReliefF	[0.853, 1.022]	4.402	0.718
	SLS	[0.359, 0.444]	5.272	0.596
	LASSO	[0.674, 0.805]	5.081	0.625
	Elastic-Net	[0.833, 1.041]	4.633	0.688
	Step-wise	[0.882, 0.978]	4.709	0.678
	Proposed($w = 0.3$)	**[0.533, 0.854]**	**4.474**	**0.710**
	Proposed($w = 0.5$)	**[0.574, 0.859]**	**4.415**	**0.709**
	Proposed($w = 0.7$)	**[0.423, 0.849]**	**4.736**	**0.679**
Moisture	DCS	[0.000, 0.024]	0.921	0.554
	RReliefF	[0.417, 0.549]	0.821	0.646
	SLS	[0.291, 1.000]	0.823	0.645
	LASSO	[0.528, 0.629]	0.852	0.619
	Elastic-Net	[0.531, 0.618]	0.853	0.618
	Step-wise	[0.514, 0.606]	0.845	0.625
	Proposed($w = 0.3$)	**[0.071, 0.109]**	**0.789**	**0.673**
	Proposed($w = 0.5$)	**[0.023, 0.086]**	**0.733**	**0.717**
	Proposed($w = 0.7$)	**[0.032, 0.124]**	**0.877**	**0.610**
DTI	DCS	[0.398, 0.505]	6.014	0.732
	RReliefF	[0.957, 1.000]	7.004	0.637
	SLS	[0.860, 0.925]	6.561	0.681
	LASSO	[0.333, 0.516]	6.412	0.695
	Elastic-Net	[0.323, 0.538]	6.428	0.694
	Step-wise	[0.376, 0.505]	6.328	0.7.0
	Proposed($w = 0.3$)	**[0.140, 0.452]**	**5.647**	**0.764**
	Proposed($w = 0.5$)	**[0.033, 0.482]**	**5.906**	**0.742**
	Proposed($w = 0.7$)	**[0.108, 0.796]**	**6.804**	**0.657**

samples' moisture content [32]. The spectrum is very wide and varies continuously, which is considered as functional feature as well. We aim to find the most predictive spectrum band for the moisture.

DTI The diffusion tensor imaging data [33] consists of 100 subjects of Fractional anisotropy (FA) tract profiles for the corpus callosum (cca) and a score of the Paced Auditory Serial Addition Test (pasat). The dataset consists of 93 contiguous locations and we want to find the best area to associate the pasat score.

5.1. ROI selection

For simplicity, the domain of the functional feature of each dataset above is transformed to [0, 1]. Without loss of generality, for our approach, we adopt the same statistical parameter settings as in Section 4 and let the mix weight $w = 0.3, 0.5, 0.7$ for comparison. We apply our method to the four datasets above and obtain the heatmaps of model accuracy, overfitting risk and f function, as shown in Fig. 6. The ROI selected by our method with different weighted values ($w = 0.3, 0.5, 0.7$) for these four datasets and corresponding p-values are also included in Fig. 6. Thus all functional features defined on their corresponding ROIs in the four datasets have significant predictabilities at the significance level 0.05.

Meanwhile, we also apply the six reference methods to these datasets. Fig. 7 illustrates the ROIs selected by the six reference methods according to the coefficient function $\beta(t)$. We draw the weights of the $\beta(t)$ and color encodes the significance of the whole domain. Obviously these selected ROIs in each dataset are very different. For the VAT dataset experiment, only our method and SLS are able to pick up the sub-interval $[0.406, 0.469]$, which is abdomen to hip, a critical region for the VAT prediction [34]. For the Canadian Weather dataset, according to Ramsay and Silverman [8], the temperature between October to November($[0.8, 0.9]$) is the most significant factor in influencing the annual precipitation. Our method selects the period between late summer and fall. The spectrum analysis often focuses on a narrow band and our method is able to extract the ROI ($[0.023, 0.086]$) in the Moisture data, which is a relatively narrow region. For the DTI dataset, it seems the region ($[0.2, 0.4]$) is more effective [35], and the reference methods usually include region ($[0.5, 1]$). It is worth mentioning that for these four datasets, the reference methods cannot guarantee localization of the estimated coefficient function $\beta(t)$ to zero-out irrelevant regions effectively, whereas our method overcomes this limitation.

5.2. Prediction

For each dataset, it is very difficult to identify the ground truth ROI without any prior knowledge. Therefore, we evaluate the performance of the selected ROIs in terms of predicting the response using the functional linear model in Eq. (1). For example, for the ROI selected by each method, we set $I = I_{select}$ in Eq. (1). We then calculate the R-Squared and Root Mean Squared Error (RMSE) of prediction based on LOOCV. The extracted ROIs of all datasets above by different approaches and evaluated reults are shown in Table 3. The prediction absolute errors are illustrated in Fig. 8. We can conclude that our proposed method is almost always supe-

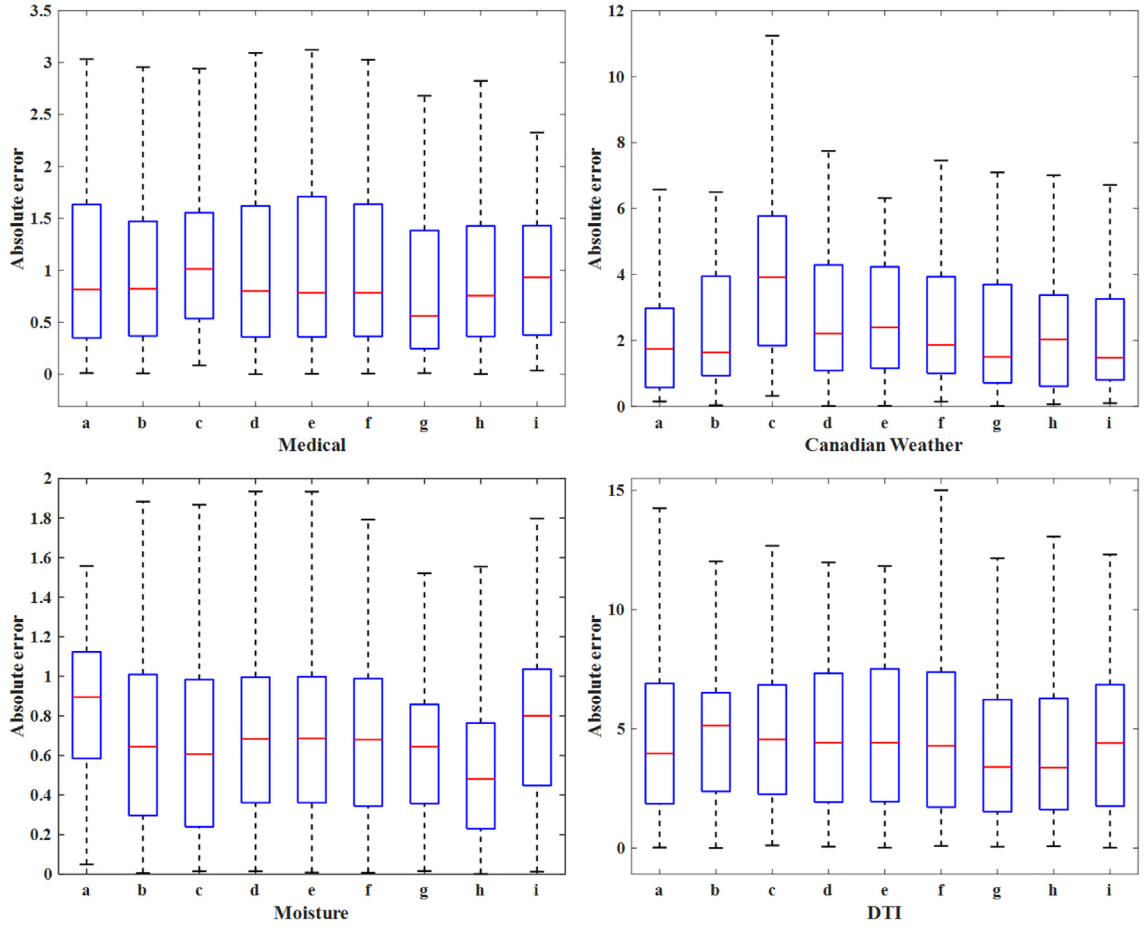


Fig. 8. The absolute errors based on LOOCV of all methods in comparison. The a - i are DCS, RReliefF, SLS, LASSO, Elastic-Net, Step-wise and Proposed method ($w = 0.3$, $w = 0.5$, $w = 0.7$).

rior compared with the reference methods. The performance of our method with different weight values $w = 0.3, 0.5, 0.7$, is slightly different. In comparison, $w = 0.7$ works the worst in all datasets, $w = 0.5$ performs best in Moisture dataset and $w = 0.3$ is the best in the other three datasets. For the Canadian Weather data, in terms of RMSE and R Squared, our proposed method is only slightly worse than DCS and RReliefF, but better than the others. With respect to the Medical, Moisture and DTI datasets, our method is better than all other methods, with both the smallest RMSE and largest R-Squared values. The four experiments show that our approach can effectively locate the ROI of a the functional feature in the context of functional linear regression.

6. Conclusion

This paper proposes an effective ROI selection method for functional features. Our proposed method provides a novel metric to balance model accuracy and overfitting risk. Under the framework of functional linear regression, the model accuracy is measured by the residual variance, while the overfitting risk is quantified through the necessary sample size to achieve a certain statistical power. We have evaluated the performance of our proposed method on four representative moderate sized datasets and compared it with six state-of-the-art reference methods. The proposed method almost always outperforms other reference methods.

The proposed framework may be generalized to other scenarios. First, it may be extended to ROI selection for multi-dimensional functional features, although in this paper we only illustrate its application to one-dimensional functional data. Similar to Theorem 2 of [26], the relationship between the sample size and statistical power for multi-dimensional functional linear regression may be attainable, which makes our method generalizable. One caveat is that it is impractical to exhaust all possible sub-intervals to select the optimal ROI for multi-dimensional functional features. Thus a more efficient search algorithm is needed, which is an interesting future research direction. Moreover, our proposed method can be adapted to nonlinear functional regression models as long as corresponding valid sample size estimation methods can be developed. Such extension requires substantial theoretical analyses, which are beyond the scope of this paper.

Our framework assumes independently and identically distributed (i.i.d.) data. This assumption is valid most of the time if the data are collected from a random sample from a population, so our proposed method is generally applicable to such data. When feature selection tasks arise from non i.i.d. data, such as time series, our proposed method may not be applicable or perform poorly.

CRediT authorship contribution statement

Qiyue Wang: Methodology, Software, Writing - original draft. **Yao Lu:** Conceptualization, Visualization. **Xiaoke Zhang:** Validation, Formal analysis. **James Hahn:** Funding acquisition, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

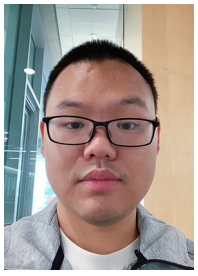
Acknowledgements

This study is supported by the USA NIH grants R21HL124443 and R01HD091179 and USA NSF grants CNS-1337722 and DMS-1832046 and George Washington University Cross Disciplinary Research Fund.

References

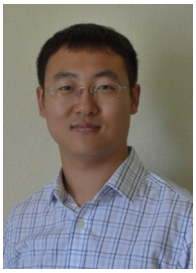
[1] L. Jian, J. Li, K. Shu, H. Liu, Multi-label informed feature selection, in: IJCAI, 2016, pp. 1627–1633.
[2] S. Hara, T. Maehara, Enumerate lasso solutions for feature selection, in: AAAI, 2017, pp. 1985–1991.
[3] G. Roffo, S. Melzi, U. Castellani, A. Vinciarelli, Infinite latent feature selection: a probabilistic latent graph-based ranking approach, in: Computer Vision and Pattern Recognition, 2017.
[4] X. Chang, F. Nie, Y. Yang, H. Huang, A convex formulation for semi-supervised multi-label feature selection, in: AAAI, 2014, pp. 1171–1177.
[5] D. Bertsimas, A. King, R. Mazumder, et al., Best subset selection via a modern optimization lens, The Annals of Statistics 44 (2) (2016) 813–852.
[6] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Ijcai, vol. 14, Montreal, Canada, 1995, pp. 1137–1145..
[7] R.B. Rao, G. Fung, R. Rosales, On the dangers of cross-validation. an experimental evaluation, in: Proceedings of the 2008 SIAM International Conference on Data Mining SIAM, SIAM, 2008, pp. 588–596.
[8] J.Ramsay,B.Silverman,Functionaldataanalysis,SpringerSeriesinStatistics(2005).
[9] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis: Theory and Practice, Springer Science & Business Media, 2006.
[10] L. Horváth, P. Kokoszka, Inference for functional data with applications, vol. 200, Springer Science & Business Media, 2012..
[11] T. Hsing, R. Eubank, Theoretical Foundations of Functional Data Analysis, with An Introduction to Linear Operators, John Wiley & Sons, 2015.
[12] J.-L. Wang, J.-M. Chiou, H.-G. Müller, Functional data analysis, Annual Review of Statistics and its Application 3 (2016) 257–295.
[13] X. Zhang, J.-L. Wang, From sparse to dense functional data and beyond, The Annals of Statistics 44 (5) (2016) 2281–2321.
[14] G.M. James, J. Wang, J. Zhu, et al., Functional linear regression that's interpretable, The Annals of Statistics 37 (5A) (2009) 2083–2108.
[15] J. Zhou, N.-Y. Wang, N. Wang, Functional linear model with zero-value coefficient function at sub-regions, Statistica Sinica 23 (1) (2013) 25.
[16] R. Li, W. Zhong, L. Zhu, Feature screening via distance correlation learning, Journal of the American Statistical Association 107 (499) (2012) 1129–1139.
[17] M. Robnik-Šikonja, I. Kononenko, Theoretical and empirical analysis of relieff and rrelieff, Machine Learning 53 (1–2) (2003) 23–69.
[18] X. Chen, G. Yuan, F. Nie, J.Z. Huang, Semi-supervised feature selection via rescaled linear regression., in: IJCAI, vol. 2017, 2017, pp. 1525–1531..
[19] N.R. Draper, H. Smith, Applied Regression Analysis, vol. 326, John Wiley & Sons, 2014.
[20] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Methodological) (1996) 267–288.
[21] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2) (2005) 301–320.
[22] Z. Xu, G. Huang, K.Q. Weinberger, A.X. Zheng, Gradient boosted feature selection, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2014, pp. 522–531.
[23] S.-B. Chen, Y. Zhang, C.H. Ding, Z.-L. Zhou, B. Luo, A discriminative multi-class feature selection method via weighted l2, 1-norm and extended elastic net, Neurocomputing 275 (2018) 1140–1149.
[24] O. Litany, T. Remez, E. Rodolà, A. Bronstein, M. Bronstein, Deep functional maps: structured prediction for dense shape correspondence, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 5660–5668.
[25] G. Ghiasi, C.C. Fowlkes, Laplacian pyramid reconstruction and refinement for semantic segmentation, in: European Conference on Computer Vision, Springer, 2016, pp. 519–534.
[26] Y.-R. Su, C.-Z. Di, L. Hsu, Hypothesis testing in functional linear models, Biometrics 73 (2) (2017) 551–561.
[27] Y. Lu, S. Zhao, N. Younes, J.K. Hahn, Accurate nonrigid 3d human body surface reconstruction using commodity depth sensors, Computer Animation and Virtual Worlds 29 (5) (2018) e1807.
[28] A. Mohammad, E.D.L. Rolfe, A. Sleigh, T. Kivisild, K. Behbehani, N.J. Wareham, S. Brage, T. Mohammad, Validity of visceral adiposity estimates from dxa against mri in kuwaiti men and women, Nutrition & Diabetes 7 (1) (2017) e238.
[29] I. Neeland, S. Grundy, X. Li, B. Adams-Huet, G. Vega, Comparison of visceral fat mass measurement by dual-x-ray absorptiometry and magnetic resonance imaging in a multiethnic cohort: the dallas heart study, Nutrition & Diabetes 6 (7) (2016) e221.
[30] Y. Lu, S. McQuade, J.K. Hahn, 3d shape-based body composition prediction model using machine learning, in: 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2018, pp. 3999–4002.
[31] J. Wang, J. Shen, P. Li, Provable variable selection for streaming features, in: International Conference on Machine Learning, 2018, pp. 5158–5166..
[32] P.T. Reiss, R.T. Ogden, Functional principal component regression and functional partial least squares, Journal of the American Statistical Association 102 (479) (2007) 984–996.
[33] J. Goldsmith, C.M. Crainiceanu, B. Caffo, D. Reich, Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements,

Journal of the Royal Statistical Society: Series C (Applied Statistics) 61 (3) (2012) 453–469.
[34] M. Snijder, R. Van Dam, M. Visser, J. Seidell, What aspects of body fat are particularly hazardous and how do we measure them?, International Journal of Epidemiology 35 (1) (2005) 83–92
[35] J. Goldsmith, J. Bobb, C.M. Crainiceanu, B. Caffo, D. Reich, Penalized functional regression, Journal of Computational and Graphical Statistics 20 (4) (2011) 830–851.
[36] Q. Wang, Y. Lu, X. Zhang, J.K. Hahn, A Novel Hybrid Model for Visceral Adipose Tissue Prediction using Shape Descriptors, in: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 1729–1732.



Qiyue Wang is pursuing his Ph.D. at the George Washington University, department of Computer Science. He received her B.S. degree in Electrical Engineering from the University of science and technology Beijing, China, in 2014 and M.S. degree from the Beihang University, China, in 2017. His research focuses on Medical image processing, 3D vision, and machine learning.

Yao Lu received her Ph.D. degree from the George Washington University, department of Computer Science in 2019. She received her B.S. degree in Electrical Engineering from Communication University of China, China, in 2012 and M.S. degree in Computer Science from the George Washington University, USA, in 2014. Her research focuses on 3D surface reconstruction, non-rigid registration, and machine learning.



Xiaoke Zhang received the B.S. degree in Statistics from Peking University, China in 2009. He received the Ph.D. degree in Statistics from the University of California, Davis, USA in 2014. He was an Assistant Professor of Statistics in the University of Delaware from August 2014 to July 2017. Since August in 2017, he has been an Assistant Professor of Statistics in George Washington University. His current research interests are functional data analysis, nonparametric statistics, statistical learning, and applied statistics.

James K. Hahn is currently a Professor in the Department of Computer Science and a Professor of Pediatrics in the School of Medicine and Health Sciences at the George Washington University where he has been a faculty since 1989. He founded the GW Institute for Biomedical Engineering and is the founding director of the Institute for Computer Graphics. His areas of interests are: medical simulation, image-guided surgery, medical informatics, visualization, and computer animation. He received his Ph.D. in Computer and Informations Science from the Ohio State University.