Multi-task Function-on-function Regression with Co-grouping Structured Sparsity

Pei Yang
South China University of Technology
Arizona State University
cs.pyang@gmail.com

Qi Tan
School of Computer Science and Engineering
South China Normal University
tanqi@scnu.edu.cn

Jingrui He
School of Computing, Informatics, Decision Systems Engineering
Arizona State University
jingrui.he@asu.edu

ABSTRACT

The growing importance of functional data has fueled the rapid development of functional data analysis, which treats the infinite-dimensional data as continuous functions rather than discrete, finite-dimensional vectors. On the other hand, heterogeneity is an intrinsic property of functional data due to the variety of sources to collect the data. In this paper, we propose a novel multi-task function-on-function regression approach to model both the functionality and heterogeneity of data. The basic idea is to simultaneously model the relatedness among tasks and correlations among basis functions by using the co-grouping structured sparsity to encourage similar tasks to behave similarly in shrinking the basis functions. The resulting optimization problem is challenging due to the non-smoothness and non-separability of the co-grouping structured sparsity. We present an efficient algorithm to solve the problem, and prove its separability, convexity, and global convergence. The proposed algorithm is applicable to a wide spectrum of structured sparsity regularized techniques, such as structured $\ell_{2,p}$ norm and structured Schatten p-norm. The effectiveness of the proposed approach is verified on benchmark functional data sets collected from various domains.

CCS CONCEPTS

•Computing methodologies → Multi-task learning; •Mathematics of computing → Regression analysis;

KEYWORDS

Function-on-function regression; multi-task learning; co-grouping structured sparsity; generalized Schatten norm

1 INTRODUCTION

A large amount of functional data is being produced in many areas, such as healthcare [18], biology [19], climatology [23], signal processing [17], etc. Their units of observation are functions defined on some continuous domains. In the Canadian weather prediction example [23], the temperature and precipitation profiles were collected daily over a period of 30 years in 35 weather stations. Another example is neuroimaging, where the activation levels observed at voxels in the brain are the responses over two or three dimensions of space and possibly time as well [16]. The distinctive characteristic of functional data analysis [23] is that it treats the infinite-dimensional data as real-valued functions rather than discrete, finite-dimensional vectors which are commonly used by machine learning algorithms. On the other hand, heterogeneity is an intrinsic property of functional data due to the fact that data is usually collected from multiple sources. For example, the relationship between temperature and precipitation in one area is likely to be different from its neighbor areas although they are closely correlated. Intuitively, leveraging such kind of data heterogeneity appropriately may help build a better regression model for the functional data.

In this paper, for the first time, we focus on heterogeneous functional data analysis, and propose a novel Multi-task Function-On-function REgression approach (Multi-Fore) to model both the functionality and heterogeneity of the data. The goal is to leverage the relatedness among tasks and the correlations among basis functions to improve the regression performance for all the tasks. We propose to use the co-grouping structured sparsity regularization to characterize the grouping of basis functions and the soft-clustering of tasks. It encourages the similar tasks to behave similarly in selecting the basis functions. The objective of Multi-Fore is to minimize the regression loss resulting from using functional predictor to recover the functional response for each task, while imposing the co-grouping structured sparsity regularization on the regression weights. The intuition of co-grouping is that the regression weights can be row-wise grouped according to the types of basis functions, and column-wise grouped according to the soft-clustering of tasks.

The resulting optimization problem is challenging due to the non-smoothness and non-separability of the objective function. To address this challenge, we first transform the non-smooth objective function into a smooth one by making use of an auxiliary function. Then, we prove the separability of the reformulated objective function, which leads to an efficient algorithm to solve the problem. Furthermore, we build the connections between the original problem and

KDD'17, August 13–17, 2017, Halifax, NS, Canada.

DOI: http://dx.doi.org/10.1145/3097983.3098133

its reformulation by showing that both of them are strictly convex and converge to the same global optimum. The proposed approach covers a wide range of sparsity regularizations. In particular, it provides a unified framework to solve both the structured $\ell_{2,p}$ norm and the structured Schatten p-norm regularized problems. The effectiveness of the proposed method is evaluated on functional data sets in comparison with various existing functional regression algorithms.

In summary, the main contributions of this paper are as follows:

- We introduce a new learning problem targeting heterogeneous functional data.
- We propose a novel multi-task functional regression approach based on co-grouping structured sparsity to model both the relatedness among tasks and correlations among basis functions.
- We develop an efficient algorithm to solve the co-grouping structured sparsity regularized problem, which enjoys the nice properties of separability, convexity, and global convergence. It covers a wide spectrum of structured sparsity regularizations such as structured $\ell_{2,p}$ norm and structured Schatten p-norm.
- We demonstrate the effectiveness of Multi-Fore on various functional data sets.

The rest of the paper is organized as follows. After a review of the related work in Section 2, we present the Multi-Fore approach in Section 3, and show the experimental results in Section 4. Finally, we conclude the paper in Section 5.

2 RELATED WORK

In this section, we review the related work on functional data analysis (FDA) [23]. In particular, we focus on functional regression, which is one of the most active areas in FDA.

Most of the current functional regression work involves either functional predictor or functional response. To name a few, Yao et al. [27] proposed the Principal Analysis by Conditional Expectation (PACE) model aiming at estimating functional principal component scores for sparse longitudinal data; Goldsmith et al. [4] proposed the Penalized Functional Regression (PFR) using truncated power series for basis functions with ℓ_2 penalties; Zhang et al. [30] presented a spline-based Functional Linear Model (FLM) with periodic spline bases via roughness regularization; Yuan and Cai [29] introduced a general Reproducing Kernel Hilbert Space (RKHS) approach to functional linear regression regularized by roughness penalties. In contrast, some work on the functional regression used sparsity constraints. Zhu and Cox [33] proposed a functional generalized linear model, in which functional principal components were used to reduce the model to multivariate logistic regression and a group Lasso penalty was applied to select useful functional covariates among multiple curves; James et al. [10] introduced the FLiRTI model which used ℓ_1 penality to encourage sparsity for interpretability; Lee and Park [14] developed a sparse functional linear regression model that imposed different sparsity penalties such as Lasso and adaptive Lasso on the regression coefficients; Fan et al. [3] proposed a Functional Additive Regression (FAR) model which used group-Lasso to perform variable selection across multiple functional predictors using orthogonal basis transform.

There has been comparatively little work for function-on-function regression, which involves both functional predictor and functional response [18]. Some function-on-function regression methods [19, 28] utilized Functional Principal Component Analysis (FPCA) to capture the variance of functional data. For example, FPCreg [28] is a nonparametric regression model based on functional principal component decomposition. Different from FPCreg, the Functional Additive Models (FAM) model [19] utilized the functional principal components in an additive way. Regularization is widely used in the function-on-function models. FdaLM [23] is a linear function-on-function regression model with roughness penalties (or ℓ_2 regularization). The regression model proposed in [8] considered a variety of regularization techniques for linear B-spline basis functions, including basis truncation, roughness penalty, and ℓ_1 sparsity penalty. The Penalized Function-on-Function Regression (PFFR) method [9] accommodated multiple functional predictors and response observed on dense or sparse grids. PFFR applied quadratic roughness penalties to avoid overfitting. A functional additive mixed model [24] was developed for the regression of functional response on both functional and scale predictors. Quadratic penality was used and the smoothing parameters were estimated by the restricted maximum likelihood method. A multi-level functional regression model proposed in [17] incorporated various basis expansions including principal components, spline-based and wavelet-based functional representations. The FarMost model [25] used mode-sparsity constraint to automatically filter out the irrelevant basis functions for both predictors and responses. In addition, some nonlinear functional regression models have also been proposed, such as the Triple-Basis Estimator (3BE) [21] and the functional RKHS approach [12].

Multi-task learning aims to improve the performance of each task by borrowing the knowledge learned from related tasks. Different assumptions on task relatedness lead to different multi-task learning models. Some typical work include: multi-task feature learning [1], clustered multi-task learning [31], low-dimensional subspace learning [11], robust multi-task learning [5], etc. Sparsity regularizations are widely used in multi-task learning, such as $\ell_{2,1}$-norm regularized method [1], weighted Lasso regularized method [15], group Lasso regularized method [26], sparse group Lasso regularized methods [32], tree-guided group Lasso regularized method [13], tree-guided fused lasso [7], elastic net regularized method [6], etc.

Different from existing work on either function-on-function regression or multi-task learning, in this paper, we aim to model both the functionality and heterogeneity of data. To the best of our knowledge, this is the first work on heterogeneous functional regression.

3 THE MULTI-FORE APPROACH

In this section, we introduce the Multi-Fore model, and then present an efficient algorithm to solve the problem.

3.1 Generalized Schatten Norm

We first introduce the proposed generalized Schatten norm. The entrywise $\ell_{p,q}$ norm of a matrix W is defined as

$$\|W\|_{p,q} = \left[\sum_i \left(\sum_j |W_{ij}|^p \right)^{q/p} \right]^{1/q}.$$

The case where $p = q = 2$ yields the Frobenius norm. The case where $p = 2, q = 1$ yields the $\ell_{2,1}$ norm.

The Schatten p-norm ($0 < p < \infty$) of a matrix $W \in \mathcal{R}^{n\times m}$ is defined as

$$\|W\|_{S_p} = \left(\sum_{i=1}^{min\{n,m\}} \sigma_i^p \right)^{\frac{1}{p}} = tr\left[\left((MM^T)^{\frac{p}{2}} \right) \right]^{\frac{1}{p}}$$

where σ_i is the i^{th} largest singular value of W. The case where $p = 2$ yields the Frobenius norm, while the case where $p = 1$ yields the trace norm (or nuclear norm).

We define a new matrix norm named generalized Schatten norm as follows.

*Definition 3.1 (**Generalized Schatten Norm**).* The generalized Schatten p-norm of a matrix W is defined as

$$\|W\|_{(\Delta,p)} = \begin{cases} \|W\|_{S_p}, & \Delta = 0 \\ \|W\|_{2,p}, & \Delta = 1 \end{cases}$$

The generalized Schatten p-norm covers both the $\ell_{2,p}$ norm and the Schatten p-norm. In particular, it subsumes some of the most popular sparsity regularizations such as $\ell_{2,1}$ norm and trace norm. Both of them result in sparse structures by forcing either the row-wise sparsity or the low-rank matrix.

3.2 The Multi-Fore Model

Suppose there are T functional regression tasks which are correlated with each other. For the i^{th} task, denote the functional predictor by $x_i(s) \in \mathcal{R}^{n_i\times 1}$ measured at grid point $s \in \mathcal{S}$, and functional response $y_i(t) \in \mathcal{R}^{n_i\times 1}$ measured at $t \in \mathcal{T}$, where n_i is the number of curves, $\mathcal{S}$ and $\mathcal{T}$ are two interval domains. Note that we do not require the predictor and response to be defined on the same domains. Our goal is to build a regression model of functional response y_i on functional predictor x_i by borrowing the information from other related regression tasks.

The basic functional regression model [23] can be represented as

$$y_i(t) = \int x_i(s)\beta_i(s,t)\,ds + \varepsilon_i(t) \tag{1}$$

where $\beta_i(s,t)$ is a bivariate regression coefficient function, and $\varepsilon_i(t)$ is the residual function. Intuitively, the regression function $\beta_i(s,t)$ for a fixed value of t can be interpreted as the related weight placed on $x_i(s)$ required for predicting $y_i(t)$. We consider the bivariate regression function $\beta_i(s,t)$ as a double expansion in terms of K_1 basis functions $\varphi_u(1 \le u \le K_1)$ and K_2 basis functions $\theta_v(1 \le v \le K_2)$ as follows.

$$\beta_i(s,t) = \sum_{u=1}^{K_1}\sum_{v=1}^{K_2} w_{uv}\varphi_u(s)\theta_v(t) = \varphi^T(s)W_i\theta(t) \tag{2}$$

where $\varphi(s) \in \mathcal{R}^{K_1\times 1}$, $\theta(t) \in \mathcal{R}^{K_2\times 1}$, and W_i is a $K_1 \times K_2$ matrix of coefficients w_{uv}. Denote $X_i = \int x_i(s)\varphi^T(s)\,ds$. The substitution of Eq. 2 into in Eq. 1 gives

$$y_i(t) = X_iW_i\theta(t) + \varepsilon_i(t). \tag{3}$$

Next, we introduce the proposed multi-task function-on-function regression model. It aims to leverage both the relatedness among tasks and the correlations among basis functions to build a better functional regression model. On one hand, the basis functions are naturally classified into groups according to their types, such as Fourier basis functions, Spline basis functions, Wavelet functions, functional PCA, etc. Different types of basis functions are suitable for different functional data since they have different properties. It is challenging to choose which kinds of basis functions to fit the given functional data. On the other hand, we may have prior knowledge on the grouping of tasks in many cases. By better modeling the relatedness among tasks, the performance of multiple regression tasks could be significantly improved.

The key idea of the proposed Multi-Fore approach is to build a multi-task functional regression model by using the co-grouping sparsity regularization, which encourages the tasks in the same cluster to behave similarly in selecting either the individual or the groups of basis functions. The intuition of co-grouping is that the regression weights can be row-wise grouped according to the types of basis functions, and column-wise grouped according to the soft-clustering of tasks.

We assume that the regression tasks are clustered into overlapping groups, i.e., each task belongs to one or more task clusters. Let G be the number of task clusters. Denote the set of task indices in the k^{th} cluster by $g(k)$, and the set of task clusters by $\Omega = \{g(k)|1 \le k \le G\}$. Denote $W = [W_1, \cdots, W_T]$, which is the concatenation of the weight matrices of all the tasks. Let $W_{(k)}$ be the block matrix corresponding to the k^{th} cluster, which is a horizontally concatenated matrix of all the weight matrices $W_i(1 \le i \le T)$ if $i \in g(k)$. Suppose the basis functions θ are divided into B groups. Then, W can be vertically partitioned into B sub-matrices. Let $W^{(b)}(1 \le b \le B)$ be the sub-matrix corresponding to the b^{th} group of basis functions.

We propose to use structured generalized Schatten norm to model both the soft-clustering of tasks and the grouping of basis functions. The objective of Multi-Fore is to minimize the loss resulting from regressing the functional response on the functional predictor for each task, while imposing the co-grouping structured sparsity on the regression weights:

$$\begin{aligned}\min_{W} \sum_{i=1}^{T} \int & \left\| y_i\left(t\right) - X_i W_i \theta\left(t\right) \right\|_F^2 dt \\ & + \sum_{k=1}^{G} \alpha_k \left\| W_{(k)} \right\|_{(\Delta,p)}^{p} + \sum_{b=1}^{B} \beta_b \left\| W^{(b)} \right\|_F \end{aligned} \tag{4}$$

where $p (1 \leq p \leq 2)$ controls the sparisty of the regression weights. The co-grouping structured sparsity regularization is imposed on the clusters of tasks and the grouping of basis functions. α_k and β_b are non-negative tradeoff parameters. The first regularization imposed on the clusters of tasks encourages the tasks in the same cluster to behave similarly in selecting basis functions, and ensures task-cluster-specific sparsity by shrinking irrelevent basis function. The second regularization imposed on the vertically partition of W ensures the group sparsity of basis functions by shrinking irrelevant groups. In such a way, we can simultaneously model the relatedness between tasks and the correlations among basis functions.

The proposed structured generalized Schatten norm covers a wide range of models with sparse regularizations, such as structured $\ell_{2,p}$ norm $\|W_{(k)}\|_{2,p}^{p}$ and structured Schatten p-norm $\|W_{(k)}\|_{S_p}$. For the former, it encourages the similar tasks to behave similarly in shrinking the basis functions. For the latter, it encourages the similar tasks to share the similar low-rank structures. Both of them result in the sparse structure of the weight matrix.

Furthermore, the structured generalized Schatten norm provides a flexible mechanism to characterize the task relatedness. It allows for the modeling of soft-clustering of regression tasks. For example, we are able to simultaneously learn the task-dependent relatedness and task-independent relatedness by using the proposed Multi-Fore model. The task-dependent relatedness captures the specific properties of each individual task, while the task-independent relatedness models the common properties shared among all the tasks. In Multi-Fore, this corresponds to the case where there are $G = T + 1$ task clusters, i.e., $\Omega = \{\{1\}, \cdots, \{T\}, \{1, \cdots, T\}\}$. Note that each task belongs to two clusters, i.e., the single-task cluster which contains the task only and the all-in-one cluster which contains all the tasks. Then, Eq. 4 can be instantialized as follows,

$$\begin{aligned}\min_{W} \sum_{i=1}^{T} \int & \left\| y_i\left(t\right) - X_i W_i \theta\left(t\right) \right\|_F^2 dt \\ & + \left[\alpha_G \|W\|_{\Delta,p}^{p} + \sum_{i=1}^{T} \alpha_i \left\| W_i \right\|_{\Delta,p}^{p} \right] + \sum_{b=1}^{B} \beta_b \left\| W^{(b)} \right\|_F . \end{aligned} \tag{5}$$

Here, the norms $\|W_i\|_{\Delta,p}^{p}$ and $\|W\|_{\Delta,p}^{p}$ ensure the task-specific sparsity and task-common sparsity, respectively, while $\|W_a^b\|_F$ models the group sparsity.

3.3 Optimization

The optimization of Multi-Fore is challenging due to the non-smoothness and non-separability of co-grouping structured sparsity. To address this problem, we first transform the objective function into a smooth one, and prove the separability of the reformulated objective, which allows us to solve the problem in parallel. The proposed algorithm is convex, convergence-guaranteed, and adaptable to both structured $\ell_{2,p}$ and structured Schatten p-norm regularized problems.

We first introduce the following inequality [20].

Lemma 3.2. *For any positive definite matrices A and A_t, the following inequality holds when $0 < p \leq 2$:*

$$tr\left(A^{\frac{p}{2}}\right) - \frac{p}{2} tr\left(A A_t^{\frac{p-2}{2}}\right) \leq tr\left(A_t^{\frac{p}{2}}\right) - \frac{p}{2} tr\left(A_t A_t^{\frac{p-2}{2}}\right).$$

The Hadamard product of two matrices, X and Y, is denoted by $X \odot Y$. Let $\tilde{W} = W^{(\tau)}$ where τ is the iteration index. I is the identity matrix. Denote

$$\Phi_W^{\Delta} = \begin{cases} WW^T, & \Delta = 0 \\ \left(WW^T\right) \odot I, & \Delta = 1 \end{cases}$$

$$\tilde{\Phi}_W^{\Delta} = \begin{cases} \tilde{W}\tilde{W}^T, & \Delta = 0 \\ \left(\tilde{W}\tilde{W}^T\right) \odot I, & \Delta = 1 \end{cases}$$

It is easy to verify that both Φ_W^{Δ} and $\tilde{\Phi}_W^{\Delta}$ are positive definite matrices.

In Theorem 3.3, we transform the non-smooth objective defined in Eq. 4 into a smooth one by making use of an auxiliary function, and derive an iterative update algorithm to solve the problem.

Theorem 3.3 (**Update Rule**). *The objective function defined in Eq.4 is non-increasing under the update*

$$\begin{aligned} W^{(\tau+1)} = \arg\min_{W} \sum_{i=1}^{T} \int & \left\| y_i\left(t\right) - X_i W_i \theta\left(t\right) \right\|_F^2 dt \\ & + \sum_{k=1}^{G} \alpha_k tr\left(W_{(k)}^T D_k^{(\tau)} W_{(k)} \right) \\ & + \sum_{b=1}^{B} \frac{\beta_b}{2\|\tilde{W}^{(b)}\|_F} tr\left(W^{(b)T} W^{(b)} \right) \end{aligned} \tag{6}$$

where τ is the iteration index, and $D_k^{(\tau)} (1 \leq k \leq G)$ is as follows,

$$D_k^{(\tau)} = \frac{p}{2} \left[\tilde{\Phi}_{W_{(k)}}^{\Delta} \right]^{\frac{p-2}{2}}. \tag{7}$$

Proof. Denote

$$L(W) = \sum_{i=1}^{T} \int \left\| y_i\left(t\right) - X_i W_i \theta\left(t\right) \right\|_F^2 dt.$$

The objective function in Eq. 4 is rewritten into

$$\begin{aligned} F\left(W\right) &= L(W) + \sum_{k=1}^{G} \alpha_k \left\| W_{(k)} \right\|_{(\Delta,p)}^{p} + \sum_{b=1}^{B} \beta_b \left\| W^{(b)} \right\|_F \\ &= L(W) + \sum_{k=1}^{G} \alpha_k tr\left[(\Phi_{W_{(k)}}^{\Delta})^{\frac{p}{2}} \right] + \sum_{b=1}^{B} \beta_b \left\| W^{(b)} \right\|_F . \end{aligned}$$

The key issue is to design an auxiliary function for $F(W)$. Define a new function,

$$J\left(W, W^{(\tau)}\right) = L(W) + \sum_{b=1}^{B}\beta_b \frac{\left\|W^{(b)}\right\|_F^2 + \left\|\tilde{W}^{(b)}\right\|_F^2}{2\left\|\tilde{W}^{(b)}\right\|_F} +$$
$$\sum_{k=1}^{G}\alpha_k\left[\frac{2-p}{2}tr\left((\tilde{\Phi}^{\triangle}_{W_{(k)}})^{\frac{p}{2}}\right) + \frac{p}{2}tr\left(\Phi^{\triangle}_{W_{(k)}}(\tilde{\Phi}^{\triangle}_{W_{(k)}})^{\frac{p-2}{2}}\right)\right].$$

Note that $J\left(W, W^{(\tau)}\right)$ is an auxiliary function of $F\left(W\right)$ due to the facts:

$$J\left(W, W\right) = F\left(W\right)$$

and

$$J\left(W, W^{(\tau)}\right) \geq F\left(W\right)$$

where the latter follows from: 1) $u^2+v^2 \geq 2uv$ for any scalars u and v; 2) Lemma 3.2 since both $\Phi^{\triangle}_{W}$ and $\tilde{\Phi}^{\triangle}_{W}$ are positive definite matrices.

Since $J\left(W, W^{(\tau)}\right)$ is an auxiliary function of $F\left(W\right)$, $F\left(W\right)$ is non-increasing under the update

$$W^{(\tau+1)} = \arg\min_{W} J\left(W, W^{(\tau)}\right).$$

We obtain the derivative of $J\left(W, W^{(\tau)}\right)$:

$$\begin{aligned}
&\frac{\partial}{\partial W}J\left(W, W^{(\tau)}\right)\\
=&\frac{\partial}{\partial W}L(W) + \frac{\partial}{\partial W}\sum_{k=1}^{G}\alpha_k\frac{p}{2}tr\left[(\Phi^{\triangle}_{W_{(k)}})(\tilde{\Phi}^{\triangle}_{W_{(k)}})^{\frac{p-2}{2}}\right]\\
&+\frac{\partial}{\partial W}\sum_{b=1}^{B}\beta_b\frac{\left\|W^{(b)}\right\|_F^2}{2\left\|\tilde{W}^{(b)}\right\|_F}\\
=&\frac{\partial}{\partial W}L(W) + \frac{\partial}{\partial W}\sum_{k=1}^{G}\alpha_k tr\left(\Phi^{\triangle}_{W_{(k)}}D_k^{(\tau)}\right)\\
&+\frac{\partial}{\partial W}\sum_{b=1}^{B}\beta_b\frac{\left\|W^{(b)}\right\|_F^2}{2\left\|\tilde{W}^{(b)}\right\|_F}\\
=&\frac{\partial}{\partial W}L(W) + \frac{\partial}{\partial W}\sum_{k=1}^{G}\alpha_k tr\left(W_{(k)}^T D_k^{(\tau)} W_{(k)}\right)\\
&+\frac{\partial}{\partial W}\sum_{b=1}^{B}\frac{\beta_b}{2\|\tilde{W}^{(b)}\|_F}tr\left(W^{(b)T}W^{(b)}\right)
\end{aligned}$$

where the last equality follows from the equivalence of the derivatives.

Since $F(W^{(\tau)}) = J(W^{(\tau)}, W^{(\tau)}) \geq \min_W J(W, W^{(\tau)}) = J(W^{(\tau+1)}, W^{(\tau)}) \geq F(W^{(\tau+1)})$, the objective function $F(W)$ is non-increasing under the update. □

Theorem 3.3 provides an iteratively update way to solve the optimization problem in Eq. 4. In Theorem 3.4, we further show that the reformulation of the co-grouping structured sparsity is separable, which leads to an efficient solution of the problem.

Theorem 3.4 (**Separability**). *The co-grouping structured sparsity is separable:*

$$\begin{aligned}
&\sum_{k=1}^{G}\alpha_k tr\left[W_{(k)}^T D_k^{(\tau)} W_{(k)}\right] + \sum_{b=1}^{B}\frac{\beta_b}{2\|\tilde{W}^{(b)}\|_F}tr\left(W^{(b)T}W^{(b)}\right)\\
=&\sum_{i=1}^{T}tr\left[W_i^T\Gamma_i^{(\tau)}W_i\right]
\end{aligned} \tag{8}$$

where

$$\Gamma_i^{(\tau)} = \Lambda^{(\tau)} + \sum_{1\leq k\leq G, i\in g(k)}\alpha_k D_k^{(\tau)} \tag{9}$$

for $1 \leq i \leq T$*, and* $\Lambda^{(\tau)}$ *is a block diagonal matrix,*

$$\left[\Lambda^{(\tau)}\right]_{bb} = \frac{\beta_b}{2\|\tilde{W}^{(b)}\|_F}I \tag{10}$$

for $1 \leq b \leq B$*.*

Proof.

$$\begin{aligned}
&\sum_{k=1}^{G}\alpha_k tr\left(W_{(k)}^T D_k^{(\tau)} W_{(k)}\right) + \sum_{b=1}^{B}\frac{\beta_b}{2\|\tilde{W}^{(b)}\|_F}tr\left(W^{(b)T}W^{(b)}\right)\\
=&\sum_{k=1}^{G}\left[\alpha_k\sum_{i\in g(k)}tr\left(W_i^T D_k^{(\tau)} W_i\right)\right] + tr\left[W^T\Lambda^{(\tau)}W\right]\\
=&\sum_{i=1}^{T}tr\left[W_i^T\left(\sum_{1\leq k\leq G, i\in g(k)}\alpha_k D_k^{(\tau)}\right)W_i\right]\\
&+tr\left[W^T\Lambda^{(\tau)}W\right]\\
=&\sum_{i=1}^{T}tr\left[W_i^T\left(\sum_{1\leq k\leq G, i\in g(k)}\alpha_k D_k^{(\tau)}\right)W_i\right]\\
&+\sum_{i=1}^{T}tr\left[W_i^T\Lambda^{(\tau)}W_i\right]\\
=&\sum_{i=1}^{T}tr\left[W_i^T\left(\Lambda^{(\tau)} + \sum_{1\leq k\leq G, i\in g(k)}\alpha_k D_k^{(\tau)}\right)W_i\right]\\
=&\sum_{i=1}^{T}tr\left[W_i^T\Gamma_i^{(\tau)}W_i\right].
\end{aligned}$$

□

Due to the separability property of co-grouping structured sparsity, the objective function in Eq.6 can be transformed into:

$$\min_{W}\sum_{i=1}^{T}\int\left\|y_i\left(t\right) - X_iW_i\theta\left(t\right)\right\|_F^2 dt + tr\left[W_i^T\Gamma_i^{(\tau)}W_i\right]. \tag{11}$$

The benefit of separability is that the reformulated problem can be split into multiple independent sub-problems, which could be solved in parallel.

Define $J_{uv^T} = \int u\left(t\right)v^T\left(t\right)dt$ for any functional data $u\left(t\right)$ and $v\left(t\right)$. The Kronecker product between two matrices U and V is denoted by $U \otimes V$. Let $w = vec\left(W\right)$ be the vectorization of the matrix W into a vector w. We derive the optimal solution of the problem in Theorem 3.5.

Theorem 3.5 (**Optimum**). *The optimal solution for Eq.11 is as follows,*

$$vec\left(W_i\right) = \left[J_{\theta\theta^T} \otimes \left(X_i^T X_i\right) + I_{K_2} \otimes \Gamma_i^{(\tau)}\right]^{-1} vec\left(X_i^T J_{y_i\theta^T}\right). \tag{12}$$

Proof. The zero gradient condition of Eq. 11 with respect to W_i gives

$$\int X_i^T X_i W_i \theta\left(t\right)\theta^T\left(t\right)dt + \Gamma_i^{(\tau)} W_i = \int X_i^T y_i(t)\theta^T(t)dt.$$

Based on the property of Kronecker product, $vec\left(UHV^T\right) = \left(V \otimes U\right) vec\left(H\right)$ for any matrices U, V, and H, we have

$$\int vec\left[X_i^T X_i W_i \theta\left(t\right)\theta^T\left(t\right)\right] dt + vec(\Gamma_i^{(\tau)} W_i)$$
$$= vec\left[\int X_i^T y_i(t)\theta^T(t)dt\right].$$

Therefore,

$$\left[J_{\theta\theta^T} \otimes \left(X_i^T X_i\right)\right] vec(W_i) + \left[I_{K_2} \otimes \Gamma_i^{(\tau)}\right] vec(W_i)$$
$$= vec\left(X_i^T J_{y_i\theta^T}\right)$$

which completes the proof. □

3.4 Convexity and Convergence

We conduct theoretical analysis on convexity and convergence of the proposed optimization algorithm.

First, we prove the convexity of the original objective and the reformulated objective functions.

Theorem 3.6 (**Convexity of Reformulated Objective Function**). *The reformulated objective function in Eq. 11 is strictly convex.*

Proof. Eq. 11 is a linear combination of multiple independent sub-problems, each of which is related to $W_i (1 \leq i \leq T)$, respectively. Consider the sub-problem:

$$J(W_i) = \int \|y_i(t) - X_i W_i \theta(t)\|_F^2 dt + tr(W_i^T \Gamma_i^{(\tau)} W_i).$$

We can obtain the derivative of $J(W_i)$,

$$\nabla J(W_i) = 2X_i^T X_i W_i J_{\theta\theta^T} + 2\Gamma_i^{(\tau)} W_i - 2X_i^T J_{y_i\theta^T},$$

and the Hessian matrix,

$$H\left[J(W_i)\right] = 2\left[J_{\theta\theta^T} \otimes (X_i^T X_i)\right] + 2I \otimes \Gamma_i^{(\tau)}.$$

Note that the Kronecker product of two positive definite matrices is positive definite. Since the matrices $J_{\theta\theta^T}$, $X_i^T X_i$, and $\Gamma_i^{(\tau)}$ are positive definite, the Hessian matrix $H\left[J(W_i)\right]$ is also positive definite. Therefore, the sub-problem $J(W_i)$ is strictly convex. To sum up, the reformulated objective function $J(W)$ is strictly convex. □

Likewise, we can prove the convexity of the original objective function.

Theorem 3.7 (**Convexity of Original Objective Function**). *The objective function in Eq. 4 is strictly convex when* $1 \leq p \leq 2$.

Since the reformulated problem is a combination of independent convex sub-problems, we adopt a divide-and-conquer strategy to solve it. Algorithm 1 shows the proposed Multi-Fore algorithm. It iteratively updates the regression weight matrix $W_i^{(\tau)}$ and the matrices $D_k^{(\tau)}$ and $\Gamma_i^{(\tau)}$ until convergence. To speed up, it updates $W_i^{(\tau)}$ in parallel.

Algorithm 1 Multi-Fore Algorithm

```
Input: functional predictors $x_i(s)$ and functional responses
    $y_i(t)$ $(1 \leq i \leq T)$; basis functions $\varphi$ and $\theta$; $\alpha_k (1 \leq k \leq G)$
    and $\beta_b (1 \leq b \leq B)$, maximum iteration $\tau_{max}$.
Output: regression weights $W_i (1 \leq i \leq T)$.
Set $\tau = 1$;
Initialize $\Gamma_i^{(\tau)} (1 \leq i \leq T)$ as identity matrix;
repeat
    Update $W_i^{(\tau)} (1 \leq i \leq T)$ for each task by Eq.12 in parallel;
    Update $D_k^{(\tau)} (1 \leq k \leq G)$ for each soft-cluster of tasks by Eq.7;
    Update $\Gamma_i^{(\tau)} (1 \leq i \leq T)$ for each task by Eq.9;
    $\tau \leftarrow \tau + 1$;
until convergence
```

Theorem 3.8 demonstrates the global convergence of Multi-Fore. Also, it builds the equivalence between the original objective function and the reformulated function in the sense that both of them converge to the same global optimum.

Theorem 3.8 (**Global Convergence**). *The proposed Multi-Fore algorithm is guaranteed to converge to the global optimum. Also, both of the objective function in Eq. 4 and the reformulated function in Eq. 11 converge to the same global optimum.*

Proof. In Theorem 3.3, we show that the objective in Eq. 4 is non-increasing under the update.

It is easy to verify that the reformulated function $J(W, W^{(\tau)})$ in Eq. 11 and the original objective function $F(W)$ in Eq. 4 have the same derivative at $W^{(\tau)}$, i.e.,

$$\nabla F(W)|_{W^{(\tau)}} = \nabla J(W, W^{(\tau)})|_{W^{(\tau)}}.$$

Thus, if $W^{(\tau)}$ is a local optimal solution of $J(W, W^{(\tau)})$, i.e., $\nabla J(W, W^{(\tau)})|_{W^{(\tau)}} = \mathbf{0}$, $W^{(\tau)}$ should also be the local optimal solution of $F(W)$ since

$$\nabla F(W)|_{W^{(\tau)}} = \nabla J(W, W^{(\tau)})|_{W^{(\tau)}} = \mathbf{0}.$$

If $W^{(\tau)}$ is not a local optimal solution of $J(W, W^{(\tau)})$, it means that $W^{(\tau)} \neq W^{(\tau+1)} = argmin_W J(W, W^{(\tau)})$. Therefore, we have $F(W^{(\tau)}) = J(W^{(\tau)}, W^{(\tau)}) > \min_W J(W, W^{(\tau)}) = J(W^{(\tau+1)}, W^{(\tau)}) \geq F(W^{(\tau+1)})$. Obviously, the objective function $F(W)$ is strictly decreasing.

In summary, the objective function $F(W)$ is strictly decreasing under the update until it reaches the local optimum, i.e., $\nabla F(W)|_{W^{(\tau)}} = \mathbf{0}$.

According to Theorems 3.6 and 3.7, both of the objective function in Eq. 4 and the reformulated function in Eq. 11 are strictly convex. For the convex function, the local optimum is also the global optimum. Therefore, the proposed Multi-Fore algorithm is guaranteed to converge to the global optimum.

Also, both the original objective function in Eq. 4 and the reformulated function in Eq. 11 will converge to the same global optimum, i.e.,

$$\nabla F(W)|_{W^{(*)}} = \nabla J(W, W^{(\tau)})|_{W^{(*)}} = \mathbf{0}.$$

□

In Algorithm 1, the most time and space consuming procedure is to update W_i in Step 4. Note that the algorithm complexity in Eq. 12 is related to K_1 and K_2, which are the numbers of basis functions. Even though the number of curves n or the number of points d is large, a far less number of basis functions is enough to well represent the data. In other words, we usually have $K_i \ll max(n, d)$ $(i = 1, 2)$. Also, owing to the separability of the co-grouping structured sparsity, we can adopt the divide-and-conquer strategy to solve the problem in parallel. Therefore, the proposed Multi-Fore algorithm is scalable to large data.

4 EXPERIMENTS

We evaluate the proposed method on the functional data sets in comparison with various functional regression algorithms.

4.1 Setup

Three benchmark functional data sets collected from various domains are used in our experiment. Table 1 shows the main statistics of the data sets, where $|\mathcal{S}|$ and $|\mathcal{T}|$ are numbers of points in the corresponding domains.

The Canadian Weather[1] data [23] consists of the temperature and precipitation profiles collected from 35 weather stations over a period of 30 years. There are 365 time points for either daily temperature or precipitation profile. The goal is to predict the daily precipitation profile of a weather station from the daily temperature profile.

The LipEMG data[2] consists of 32 records of the movement acceleration of the lower lip when a subject was required to repeatedly say the the sentence, 'say bob again', and the corresponding electromyographical (EMG) activities of the muscle [22]. Each of these measures has 501 time points. According to Newtons's second law, acceleration reflects the force applied to tissue by muscle contraction. The goal is to predict the lip acceleration from the EMG activities.

The diffusion tensor imaging (DTI2) data [2] involves 340 DTI scans from the subjects with multiple sclerosis. Multiple sclerosis is a demyelinating autoimmune-mediated disease that is associated with brain lesions and results in severe disability. From each DTI scan, fractional anisotropy (FA) profiles are obtained along the corpus callosum (FA-CCA) and right corticospinal (FA-RCTS) tracts. Each of the FA-CCA or FA-RCTS tracts has 93 or 55 grid points, respectively. To study the spatial association between demyelinations along the tracts, we try to build a functional regression model to predict FA-CCA from FA-RCTS.

Table 1: Statistics of different data sets.

Data set	n	x	$\|\mathcal{S}\|$	y	$\|\mathcal{T}\|$
Weather	35	Temperature	365	Precipitation	365
LipEMG	32	EMG	501	Acceleration	501
DTI2	340	FA-RCTS	55	FA-CCA	93

For each data set, we cluster the curves into T groups by using K-means algorithm, each group corresponding to one task. We empirically set $T = 5$. The dynamic time warping distance is used as the similarity metric to measure the discrepancy between two curves. Since the model defined in Eq. 5 is able to characterize both task-specific and task-common relatedness, we implement this model in the experiments. Note that in this case each task belongs to two clusters, i.e., the single-task cluster and the all-in-one cluster.

4.2 Evaluation Metrics and Comparison Methods

We adopt various evaluation metrics including MRSS and MRPE to evaluate the proposed method.

Suppose there are totally m curve pairs in the test data. Let $f_j(t)$ and $y_j(t)$ represent the predicted and true regression functions respectively for the $j^{th}(j = 1, \cdots, m)$ curve. RSS is the residual squares integrated over the domain, and MRSS is the mean of RSS over testing curves, defined as,

$$MRSS = \frac{1}{m}\sum\nolimits_{j=1}^{m} \int [y_j(t) - f_j(t)]^2 dt.$$

MRPE is the mean of the relative prediction errors over testing curves, which is defined as,

$$MRPE = \frac{1}{m}\sum\nolimits_{j=1}^{m} \frac{\int [y_j(t) - f_j(t)]^2\, dt}{\int y_j^2(t)dt}.$$

We compare our approach with various function-on-function regression methods including FdaLM [23] with ℓ_2 regularization, FPCreg [28], and FAM [19]. Both FPCreg and FAM are available online in the PACE[3] package.

We repeat the experiments 10 times and report the mean and standard deviation of the performance. Note that the smaller the value of either MRSS or MRPE on the test data, the better the performance of the algorithm. The parameters are tuned for each algorithm by cross-validation on training data. The basis functions used in Multi-Fore include both Fourier series and B-splines. The numbers of basis functions are empirically set to 120.

4.3 Comparison of Multi-Fore Variants

Since the proposed Multi-Fore algorithm covers both structured $\ell_{2,p}$ norm and structured Schatten p-norm, we first compare their performance differences. We implement two

[1] http://www.psych.mcgill.ca/misc/fda/

[2] http://www.stats.ox.ac.uk/~silverma/fdacasebook/lipemg.html

[3] http://www.stat.ucdavis.edu/PACE/

variants of Multi-Fore, i.e., Multi-Fore-L_p and Multi-Fore-S_p, which correspond to structured $\ell_{2,p}$ and structured Schatten p-norm, respectively. The sparsity parameter p varies from 0.25 to 2.0 with step size 0.25. Figures 1 - 3 show the performance MRPE varying with p on the three data sets.

Both Multi-Fore-L_p and Multi-Fore-S_p perform better when p decreases from 2 to 1 in most cases. As proved in Theorems 3.6 and 3.7, the objective function in Multi-Fore is strictly convex when $1 \leq p \leq 2$. Note that the case where $p = 1$ yields much sparser structures of the regression weight matrix than the case where $p = 2$. The results suggest that the regression performance could be improved by learning the low-rank structures and enhancing the sparsity of the regression weights. Also, Multi-Fore-L_p and Multi-Fore-S_p show the comparable performances on these data sets although they differ in the ways of enhancing the sparsity.

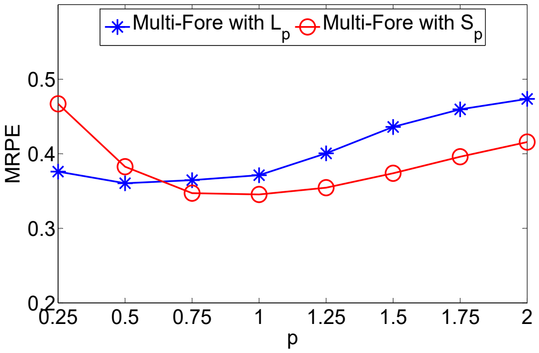

Figure 1: MRPE varies with p on Weather data.

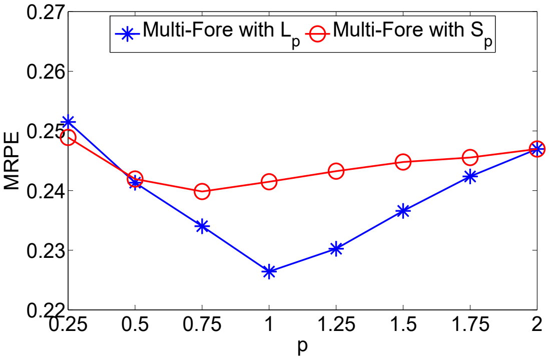

Figure 2: MRPE varies with p on LipEMG data.

In addition, we study the performance of Multi-Fore when $0.25 \leq p < 1$. It shows that the performances of both Multi-Fore-L_p and Multi-Fore-S_p in this case are not very stable across different data sets. This might be due to the non-convexity of objective function when $0 < p < 1$, which prevents the algorithm from reaching the global optimum. Based on the above results, we set $p = 1$ in the following experiments.

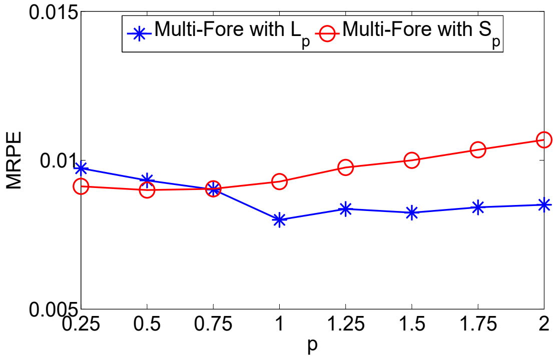

Figure 3: MRPE varies with p on DTI2 data.

4.4 Performance Evaluation

We first examine intuitively how well the predictions of different algorithms fit the ground-truth curves. For the LipEMG data, we take the first response curve in the test data as an example. Figure 4 shows the regression results. We can see that most of the algorithms are basically capable of catching the trend of the ground-truth curve. However, the proposed Multi-Fore fits the ground-truth curve better than the others. Next, we will quantitatively compare their performance.

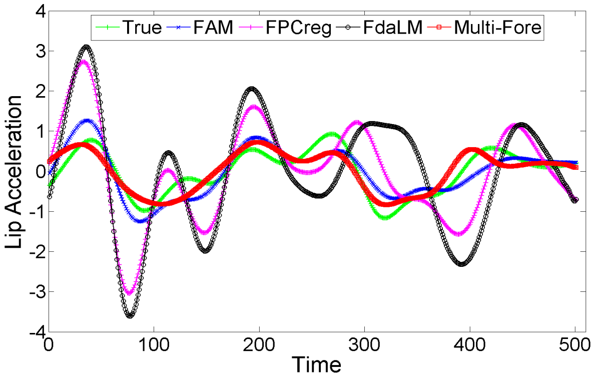

Figure 4: Regression curves on LipEMG data.

Table 2 shows the regression performances in terms of MRPE and MRSS respectively on the three data sets. The proposed Multi-Fore algorithms including both Multi-Fore-L_p and Multi-Fore-S_p perform better than the comparison algorithms. FPCreg and FAM are based on functional principal component analysis which keeps only the first few principal components that explain the most variability. The results show that the functional principal components explaining the largest amount of variability in predictor may not necessarily be the most discriminative in predicting response. FdaLM shows relatively poor performance compared to the other algorithms, indicating that using only the roughness penalty may not be adequate to build a robust regression model. In contrast, the competency of Multi-Fore is that it models the

Table 2: Regression performance of different algorithms on the three data sets.

Algorithm	Weather		LipEMG		DTI2	
	MRPE	MRSS	MRPE	MRSS	MRPE	MRSS
Multi-Fore-L_p	0.371±0.02	23.47±0.78	**0.226±0.04**	**85.71±0.86**	**0.008±0.00**	**0.85±0.03**
Multi-Fore-S_p	**0.346±0.03**	**21.22±0.53**	0.241±0.03	90.02±0.73	0.009±0.00	0.88±0.07
FdaLM	0.992±0.20	66.72±13.58	1.992±0.50	532.62±43.78	0.070±0.02	6.98±0.85
FPCreg	0.517±0.05	33.77±4.50	0.962±0.24	248.65±0.69	0.012±0.00	1.71±0.21
FAM	0.655±0.08	40.78±6.58	0.287±0.03	122.03±31.73	0.016±0.00	2.05±0.17

task relatedness by using the co-grouping structured sparsity regularization to encourage the similar tasks to share either the individual or groups of basis functions. Multi-Fore automatically picks out the most informative individual or groups of basis functions for the similar tasks in building the regression model, while truncating the irrelevant ones. As a consequence, its performance on the unseen responses can be improved.

In the three data sets, Multi-Fore-L_p performs better on LipEMG and DTI2 data, while Multi-Fore-S_p performs better on Weather data. Nevertheless, their performances are comparable with each other.

4.5 Parameter Sensitivity

We study the performance sensitivity of Multi-Fore on the regularization parameters α and β. Note that we have $G = T + 1$ task clusters (see Eq. 5). We empirically set $[\alpha_1, \cdots, \alpha_T, \alpha_G] = [\alpha, \cdots, \alpha, T\alpha]$, which is used to control the importance of the structured sparsity. The result on the Weather data is shown in Figure 5. Multi-Fore performs worse as α approaches 0 when no sparsity constraint is imposed on the regression weights. The optimal performance is achieved at $\alpha = 16$. The results show that the performance of Multi-Fore can be significantly improved by learning the task relatedness via the structured sparsity. Figure 5 shows the similar trend of the performance along with the changes of β, which controls group sparsity of the regression weights. We empirically set $[\beta_1, \cdots, \beta_B] = [\beta, \cdots, \beta]$. The results suggest that truncating the irrelevant groups of basis functions could lead to better performance.

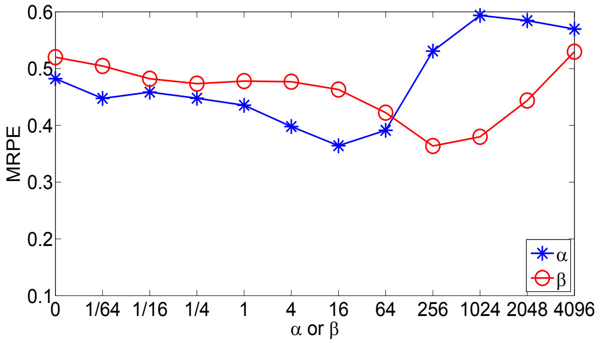

Figure 5: MRPE varies with α or β (log_4 scale).

4.6 Convergence

We study the convergence property of Multi-Fore. Figure 6 shows the performance varying with iterations on the Weather data. We can see that MRPE drops sharply in the first several iterations and becomes stable after about 5 iterations. The results show that Multi-Fore converges very fast.

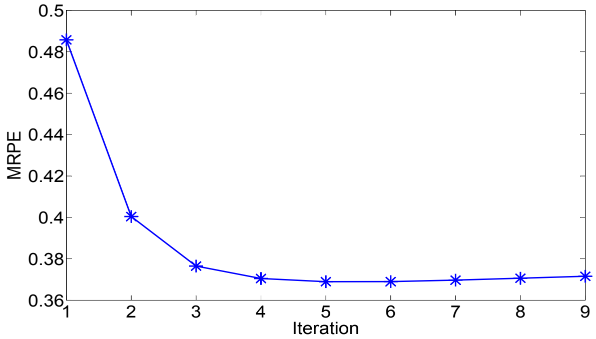

Figure 6: MRPE varies with iterations.

5 CONCLUSION

We propose a novel multi-task functional regression method, which uses the co-grouping structured sparsity to encourage similar tasks to share either the individual or groups of basis functions. An efficient algorithm is developed to solve the resulting non-smooth and non-separable problem. The proposed algorithm enjoys the advantages of separability, convexity, global convergence, and wide adaptability. The experimental results on various functional data demonstrate the effectiveness of the proposed approach.

ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China under Grant No. 61473123, National Science Foundation under Grant No. IIS-1552654, ONR under Grant No. N00014-15-1-2821, and an IBM Faculty Award. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

REFERENCES

[1] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. 2006. Multi-Task Feature Learning. In *NIPS*. 41–48.

[2] Peter J. Basser, Sinisa Pajevic, Carlo Pierpaoli, Jeffrey Duda, and Akram Aldroubi. 2000. In Vivo Fiber Tractography using DT-MRI Data. *Magnetic Resonance in Medicine* 44, 4 (2000), 625–632.

[3] Yingying Fan, Gareth M. James, and Peter Radchenko. 2015. Functional Additive Regression. *The Annals of Statistics* 43, 5 (2015), 2296–2325.

[4] Jeff Goldsmith, Jennifer Bobb, Ciprian M. Crainiceanu, Brian Caffo, and Daniel Reich. 2011. Penalized Functional Regression. *Journal of Computational and Graphical Statistics* 20, 4 (2011), 830–851.

[5] Pinghua Gong, Jieping Ye, and Changshui Zhang. 2012. Robust Multi-task Feature Learning. In *KDD*. 895–903.

[6] Pinghua Gong, Jiayu Zhou, Wei Fan, and Jieping Ye. 2014. Efficient Multi-task Feature Learning with Calibration. In *KDD*. 761–770.

[7] Lei Han and Yu Zhang. 2015. Learning Tree Structure in Multi-Task Learning. In *KDD*. 397–406.

[8] Jaroslaw Harezlak, Brent A. Coull, Nan M. Laird, Shannon R. Magari, and David Christiani. 2007. Penalized Solutions to Functional Regression Problems. *Computational Statistics and Data Analysis* 51, 10 (2007), 4911–4925.

[9] AndradaE. Ivanescu, Ana-Maria Staicu, Fabian Scheipl, and Sonja Greven. 2015. Penalized Function-on-function Regression. *Computational Statistics* 30, 2 (2015), 539–568.

[10] Gareth M. James, Jing Wang, and Ji Zhu. 2009. Functional Linear Regression That's Interpretable. *The Annals of Statistics* 37, 5A (2009), 2083–2108.

[11] Shuiwang Ji and Jieping Ye. 2009. An Accelerated Gradient Method for Trace Norm Minimization. In *ICML*. 457–464.

[12] Hachem Kadri, Emmanuel Duflos, Philippe Preux, Stéphane Canu, Alain Rakotomamonjy, and Julien Audiffren. 2015. Operator-valued Kernels for Learning from Functional Response Data. *Journal of Machine Learning Research* (2015).

[13] Seyoung Kim and Eric P. Xing. 2010. Tree-Guided Group Lasso for Multi-Task Regression with Structured Sparsity. In *ICML*. 543–550.

[14] Eun R. Lee and Byeong U. Park. 2012. Sparse Estimation in Functional Linear Regression. *Journal of Multivariate Analysis* 105, 1 (2012), 1–17.

[15] Giwoong Lee, Eunho Yang, and Sung Ju Hwang. 2016. Asymmetric Multi-task Learning Based on Task Relatedness and Loss. In *ICML*. 230–238.

[16] Daniel J. Levitin, Regina L. Nuzzo, Bradley W. Vines, and J. O. Ramsay. 2007. Introduction to Functional Data Analysis. *Canadian Psychology* 48, 3 (2007), 135–155.

[17] Mark J. Meyer, Brent A. Coull, Francesco Versace, Paul Cinciripini, and Jeffrey S. Morris. 2015. Bayesian Function-on-function Regression for Multilevel Functional Data. *Biometrics* 71, 3 (2015), 563–574.

[18] Jeffrey S. Morris. 2015. Functional Regression. *Annual Review of Statistics and Its Application* 2, 1 (2015), 321–359.

[19] Hans-Georg Müller and Fang Yao. 2008. Functional Additive Models. *J. Amer. Statist. Assoc.* 103, 484 (2008), 1534–1544.

[20] Feiping Nie, Heng Huang, and Chris H. Q. Ding. 2012. Low-Rank Matrix Recovery via Efficient Schatten p-Norm Minimization. In *AAAI*.

[21] Junier B. Oliva, Willie Neiswanger, Barnabás Póczos, Eric P. Xing, Hy Trac, Shirley Ho, and Jeff G. Schneider. 2015. Fast Function to Function Regression. In *AISTATS*.

[22] J.O. Ramsay and B.W. Silverman. 2002. *Applied Functional Data Analysis* (1st ed.). New York: Springer-Verlag.

[23] J.O. Ramsay and B.W. Silverman. 2005. *Functional Data Analysis* (2nd ed.). New York: Springer-Verlag.

[24] Fabian Scheipl, Ana-Maria Staicu, and Sonja Greven. 2015. Functional Additive Mixed Models. *Journal of Computational and Graphical Statistics* 24, 2 (2015), 477–501.

[25] Pei Yang and Jingrui He. 2016. Functional Regression with Mode-Sparsity Constraint. In *ICDM*. 1311–1316.

[26] Xiaolin Yang, Seyoung Kim, and Eric P. Xing. 2009. Heterogeneous Multitask Learning with Joint Sparsity Constraints. In *NIPS*. 2151–2159.

[27] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. 2005. Functional Data Analysis for Sparse Longitudinal Data. *J. Amer. Statist. Assoc.* 100, 470 (2005), 577–590.

[28] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. 2005. Functional Linear Regression Analysis for Longitudinal Data. *The Annals of Statistics* 33, 6 (2005), 2873–2903.

[29] Ming Yuan and T. T. Cai. 2010. A Reproducing Kernel Hilbert Space Approach to Functional Linear Regression. *The Annals of Statistics* 38, 6 (2010), 3412–3444.

[30] Daowen Zhang, Xihong Lin, and MaryFran Sowers. 2007. Two-Stage Functional Mixed Models for Evaluating the Effect of Longitudinal Covariate Profiles on a Scalar Outcome. *Biometrics* 63, 2 (2007), 351–362.

[31] Jiayu Zhou, Jianhui Chen, and Jieping Ye. 2011. Clustered Multi-Task Learning Via Alternating Structure Optimization. In *NIPS*. 702–710.

[32] Jiayu Zhou, Jun Liu, Vaibhav A. Narayan, and Jieping Ye. 2012. Modeling Disease Progression via Fused Sparse Group Lasso. In *KDD*. 1095–1103.

[33] Hongxiao Zhu and Dennis D. Cox. 2009. A Functional Generalized Linear Model with Curve Selection in Cervical Pre-cancer Diagnosis Using Fluorescence Spectroscopy. *Lecture Notes-Monograph Series* 57 (2009), 173–189.