Journal of Machine Learning Research 26 (2025) 1-25 Submitted 2/18; Revised 11/24; Published 1/25

Optimal Functional Bilinear Regression with Two-dimensional Functional Covariates via Reproducing Kernel Hilbert Space

Dan Yang DYANGHKU@HKU.HK
Jianlong Shao SJLPKU@CONNECT.HKU.HK
Haipeng Shen HAIPENG@HKU.HK
Faculty of Business and Economics, The University of Hong Kong

Hongtu Zhu HTZHU@EMAIL.UNC.EDU
Department of Biostatistics, University of North Carolina at Chapel Hill

Editor: Jie Peng

Abstract

Traditional functional linear regression usually takes a one-dimensional functional predictor as input and estimates the continuous coefficient function. Modern applications often generate two-dimensional covariates, which become matrices when observed at grid points. To avoid the inefficiency of the classical method involving estimation of a two-dimensional coefficient function, we propose a functional bilinear regression model, and introduce an innovative three-term penalty to impose roughness penalty in the estimation. The proposed estimator exhibits minimax optimal property for prediction under the framework of reproducing kernel Hilbert space. An iterative generalized cross-validation approach is developed to choose tuning parameters, which significantly improves the computational efficiency over the traditional cross-validation approach. The statistical and computational advantages of the proposed method over existing methods are further demonstrated via simulated experiments, the Canadian weather data, and a biochemical long-range infrared light detection and ranging data.

Keywords: functional principal component analysis; functional linear regression; tensor regression; scalar-on-image regression; Canadian weather data.

1. Introduction

The functional linear regression (FLR) is a powerful approach for predicting a scalar response from a one-dimensional functional predictor. It was first introduced by Ramsay and Dalzell (1991), and has been widely used in functional data analysis since then (Ramsay and Silverman, 2002, 2005a; Wang et al., 2016; Reiss et al., 2017). Consider a scalar response Y and a square integrable random function $X(\cdot)$ with mean 0 defined over the domain $\mathcal{T}$, the FLR model adopts the following form,

$$Y = \mu_0 + \int_{\mathcal{T}} X(t)\beta_0(t)\ dt + \epsilon, \tag{1}$$

where μ_0 is the intercept, $\beta_0(\cdot)$ is the unknown coefficient function, and ϵ is zero-mean noise.

The predictor $X(\cdot)$ in Model (1) is often one-dimensional, and is therefore sometimes referred to as one-way functional input. However, in recent years, it has been increasingly common to collect data in a two-way fashion. To be more specific, the random predictor

$X(\cdot)$ in these cases is a bivariate function defined in the domain $\mathcal{T}_1 \times \mathcal{T}_2$. For instance, the well-known Canadian weather data traditionally uses one-dimensional function of daily temperature to predict precipitation. In Section 5, it is shown that the hour-of-the-day and day-of-the-year information constitutes a two-dimensional functional predictor, which predicts precipitation more accurately and reveals more meteorological phenomena. Besides meteorology, such two-way functional data are now frequently encountered in fields like finance, economics, social science, neuroimaging, and so on. See Section G in Appendix for another real data example, where the two domains correspond to wavelength and range, respectively.

To generalize Model (1) in order to deal with two-way covariate, we propose the following functional bilinear regression (FBLR) model,

$$Y = \mu_0 + \int_{\mathcal{T}_1 \times \mathcal{T}_2} \alpha_0(s) X(s,t) \beta_0(t) \; dsdt + \epsilon. \tag{2}$$

Here, Y, μ_0 and ϵ are again the scalar response, intercept, and error terms, respectively, $X(\cdot,\cdot)$ is a square integrable bivariate zero-mean function defined on $\mathcal{T}_1 \times \mathcal{T}_2$, and $\alpha_0(\cdot)$ in the domain $\mathcal{T}_1$ and $\beta_0(\cdot)$ on $\mathcal{T}_2$ are two unknown coefficient functions. The error term ϵ is assumed to have mean zero and finite variance. We study the random design case where $X(\cdot,\cdot)$ is a stochastic process and independent of ϵ. The goal of the FBLR is to recover the two coefficient functions from n independent and identically distributed training sample $\{x_i(\cdot,\cdot), y_i\}_{i=1}^n$ and to make predictions for testing data.

Compared with Model (1), the two coefficient functions in Model (2) preserve the two-way functional structural information through a bilinear form, $\alpha_0(s)X(s,t)\beta_0(t)$. In the literature, this type of bilinear/multilinear combination is becoming increasingly common when dealing with two-way/multi-way/tensor data (e.g., Dyrholm et al., 2007; Zhou et al., 2013; Bi et al., 2018, 2021; Chen et al., 2022).

A naive and straightforward approach to deal with the two-way functional covariate is to convert the two-dimensional predictor into one-dimensional through stacking the data along one direction, then followed by implementing the traditional FLR Model (1). However, this conversion would destroy the two-way functional structure of the covariate, and the resulting one-way predictor is typically no longer a smooth function, which violates the underlying assumption of FLR. Therefore, adopting Model (1) is not a good choice for a two-dimensional functional predictor. More discussions on the vectorization methods to make the resulting long vector smoother can be found in Appendix E.1. But even though the resulting long vector is smoother, applying FLR on it still leads to worse performance than keeping the two-dimensional structure.

Note that $\alpha_0(\cdot)$ and $\beta_0(\cdot)$ are only identifiable up to a scalar, but their product $\alpha_0(\cdot)\beta_0(\cdot)$ is identifiable. That is, for any $c \neq 0$, $\alpha_0(\cdot)/c$ and $c\beta_0(\cdot)$ will lead to an equivalent model. For our primary focus on prediction accuracy, the un-identifiability is not a concern: as equivalent models will lead to identical predictions. To make the coefficient functions completely identifiable, one can adopt three choices of common practices to control the scaling issue: (1) assume $\|\alpha_0\| = \|\beta_0\| = 1$ and introduce one extra scaling scalar parameter; (2) assume either $\|\alpha_0\| = 1$ or $\|\beta_0\| = 1$ and absorb the scalar into the other; (3) assume $\|\alpha_0\| = \|\beta_0\|$. To further make the sign identifiable, one could assume the integral $\int_{\mathcal{T}_1} \alpha_0(s)ds$ to be positive and adjust the signs of α_0 and β_0 accordingly, or one could use domain expertise to

determine the signs. Please refer to Section 2.3 for detailed discussion of the impact of this un-identifiability property on the choice of the penalty involved.

A few articles have examined the problem of regression with two-way predictor, sometimes referred to scalar-on-image regression, as summarized in Happ et al. (2018). Reiss and Ogden (2010) studied this problem with 2D images as predictors by regressing on the principal component (PC) scores obtained from two-dimensional principal component analysis (PCA) of the observed images. Sangalli et al. (2013) proposed to penalize with the integral of the square of the Laplacian of the two-dimensional coefficient function. Guillas and Lai (2010) approached the problem through fixed bivariate spline. Wang et al. (2014) and Reiss et al. (2015) transformed the multi-dimensional functional problem into the estimation of wavelet coefficients via wavelet transformation. These aforementioned papers considered the following model

$$Y = \mu_0 + \int_{\mathcal{T}_1 \times \mathcal{T}_2} X(s,t)\beta_0(s,t)\ dsdt + \epsilon, \tag{3}$$

and focused on the estimation of the coefficient function $\beta_0(\cdot,\cdot)$.

The key difference between Model (2) and Model (3) is that our Model (2) adopts two one-dimensional coefficient functions compared to a single two-dimensional coefficient function in Model (3). Model (2) is a special case of Model (3) with restriction. The primary motivation for proposing this seemingly restrictive Model (2) is that estimation of Model (2) can lead to estimating the following model,

$$Y = \mu_0 + \sum_{r=1}^{R} \int_{\mathcal{T}_1 \times \mathcal{T}_2} \alpha_0^{[r]}(s) X(s,t) \beta_0^{[r]}(t)\ dsdt + \epsilon. \tag{4}$$

Model (4) is also a special case of Model (3) with extra assumption that the true two-dimensional function is approximated by the summation of a few terms of the products of two one-dimensional functions, i.e., $\beta_0(s,t) = \sum_{r=1}^{R} \alpha_0^{[r]}(s)\beta_0^{[r]}(t)$.

If Model (2) can be estimated via some approach, then Model (4) can be estimated in an iterative fashion via deflation: apply the approach to the original data $\{x_i(\cdot,\cdot), y_i\}_{i=1}^n$, obtain the estimate $\hat{\alpha}_0^{[1]}(s)\hat{\beta}_0^{[1]}(t)$ and retain the residuals $\{e_i\}_{i=1}^n$; re-apply the approach to the predictors and residuals $\{x_i(\cdot,\cdot), e_i\}_{i=1}^n$, obtain the estimate $\hat{\alpha}_0^{[2]}(s)\hat{\beta}_0^{[2]}(t)$ and retain the updated residuals; and repeat. Such deflation approach is commonly adopted in the literature of PCA.

There are a few reasons why Model (4) is preferred over Model (3) for two-dimensional functional regression. The first reason comes from the necessity of a low-rank assumption. Suppose the two-dimensional functional predictors are observed on a dense grid of equidistant points (s_i, t_j) for $i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$ and denote the observed data matrix as $x_{ij} = X(s_i, t_j)$, correspondingly $b_{ij} = \beta_0(s_i, t_j)$. Then Model (3) becomes $Y = \mu_0 + \sum_{ij} x_{ij} b_{ij} + \epsilon = \mu_0 + \langle \mathbf{X}, \mathbf{B} \rangle + \epsilon$. Such a scalar-on-matrix regression is one special type of tensor regression. For tensor regression, since the coefficient tensor/matrix, has multiple modes and high dimensions, it is necessary to assume the coefficient has a low-rank structure (Zhou et al., 2013; Lock, 2018; Raskutti et al., 2019; Chen et al., 2019). Suppose $\mathbf{B}$ has low-rank singular value decomposition (SVD) $\mathbf{B} = \mathbf{UDV}^T = \sum_{r=1}^{R} d_r \mathbf{u}_r \mathbf{v}_r$, where $\mathbf{u}_r, \mathbf{v}_r$ are the left and right singular vectors of unit length. Then on the observed grid, the low-rank model is $Y = \mu_0 + \sum_{r=1}^{R} \sum_{ij} x_{ij} d_r u_{ir} v_{jr} + \epsilon$. Bringing this approximation from the discrete case back to the continuous case, it boils down

to $Y = \mu_0 + \sum_{r=1}^{R} \int_{\mathcal{T}_1 \times \mathcal{T}_2} d_r u_r(s) X(s,t) v_r(t)\ dsdt + \epsilon$. Note that the constants d_r can be absorbed into $u_r(\cdot)$ or $v_r(\cdot)$, which becomes Model (4).

Second, adopting a framework of reproducing kernel Hilbert space (RKHS) to solve Model (3) will eventually lead to Model (4) as well. To solve Model (3) via RKHS, a four-dimensional kernel needs to be specified. To the best of our knowledge, there is no literature that studies the theories of Model (3) via RKHS. One natural choice for the four-dimensional kernel is the kernel associated with tensor product RKHSs in the context of smoothing spline (Gu, 2013). Model (3) with the tensor product kernel will be referred to as FLR+TPK from now on. By the representer theorem, the derivations in Section 4 show that the solution of FLR+TPK will be of the form $\sum_{jk} c_{jk} \phi_j^1(s) \phi_k^2(t) + \ldots$, where $\ldots$ represents some less important terms, and $\phi_j^1(s)$ and $\phi_k^2(t)$ are the basis of the two kernels from two domains. Plugging the representation back into the objective function and solving for the coefficient matrix $\mathbf{C} = (c_{jk})$ is again a scalar-on-matrix regression problem. Extending such a framework to even higher dimensional problem is scalar-on-tensor regression. It is well known in the tensor regression literature that it is imperative to assume low-rankness of $\mathbf{C}$, which is *exactly* equivalent to assuming Model (4) after simplifications.

The previous two reasons are rooted in the modeling perspective, and the next two reasons show the theoretical and numerical advantages of Model (4) over Model (3). Third, our theory shows that the two-dimensional FBLR has the same convergence rate as one-dimensional FLR. There is no literature that studies Model (3) via RKHS, but our conjecture is that the convergence rate should have another factor of "2" in the shoulder of the rate, just as in the non-parametric regression, which is slower than ours.

Fourth, extensive simulation studies and two real data examples demonstrate the superior statistical and computational performances of FBLR over FLR+TPK under all three two-dimensional models mentioned above, even when the data are generated according to Model (3). Furthermore, visualization and interpretation of the two one-dimensional functions from Model (4) are more straightforward and meaningful than the two-dimensional function from Model (3); see the real data applications for the details. Moreover, for many applications, the two domains are very different. In FLR+TPK, although two distinct kernels can be adopted, only one hyperparameter can be used. However, it is very likely that the levels of smoothness in the two domains are very different. Such a difference requires not only two kernels, but also two hyperparameters, which FBLR can accommodate but FLR+TPK cannot.

Lastly, the advantage of Model (4) over Model (3) is magnified when extension to even higher dimension d is made. Model (4) can be extended straightforwardly via multilinear form and Model (3) can be extended via multivariate integral. However, when Model (3) is extended, issues like infeasible computation, curse of dimensionality, and slow theoretical convergence rate will inevitably occur. These issues were mentioned in the two-dimensional functional PCA literature as well (Chen et al., 2017). On the other hand, when Model (4) is extended, we expect the same computational cost (by a factor of d, not exponentially in d) and identical theoretical convergence to remain.

Although the extension from linear to bilinear seems natural, the generalization is in fact non-trivial. Due to the interplay of the two coefficient functions in a product form, several challenges arise from the design of the penalty, the development of the algorithm, and the theoretical analysis. Therefore, the five-fold main contributions are elaborated below.

First, there are mainly two categories of approaches for one-way FLR, including functional PCA regression (FPCR) (e.g., Ramsay and Silverman, 2002, 2005a; Cai and Hall, 2006) and smoothness penalization under the framework of RKHS (e.g., Yuan and Cai, 2010; Cai and Yuan, 2012; Balasubramanian et al., 2022). Cai and Yuan (2012) showed that the penalty approach has an advantage over FPCR, because the penalty approach does not require the alignment of the reproducing kernel and the covariance kernel of the predictor, while the FPCR approach does. Hence, we take the penalty approach for the FBLR problem when extending from 1D to 2D. Our key innovation is the proposal of a three-term penalty that involves the Hilbert norm based on the reproducing kernel and the norm associated with the covariance kernel. This three-term penalty enjoys the invariant property and successfully separates the effects from the smoothness levels of the two coefficient functions $\alpha(\cdot)$ and $\beta(\cdot)$, which leads to a minimax rate optimal solution.

Second, upon the proposal of the penalty function, an iterative algorithm is developed to optimize the objective function because of the bi-convexity of the function. The main idea of the block descent algorithm is to reduce the two-way problem into a one-way FLR problem in each iteration, which can then be solved by the representer theorem (Wahba, 1990), with slight complications since the one-way problem involves updated 1D stochastic processes and reproducing kernels. Furthermore, there are two tuning parameters in the penalty corresponding to the different degrees of smoothness of the two coefficient functions. Naive cross-validation (CV) over a two-dimensional grid is outrageously time-consuming. A novel iterative generalized cross-validation (iGCV) approach is proposed whose computational time is only slightly more than the iterative FBLR algorithm with *fixed* tuning parameters, while achieving similar performance as the computationally expensive CV.

Third, one interesting finding is that the minimax convergence rate for the FBLR is identical to that of the FLR if we assume the domains, kernels and covariances are the same for the two dimensions of the two-way functional data. Unlike the FLR problem, whose solution is explicit and can be directly analyzed theoretically, the FBLR problem does not have an explicit solution since the two coefficient functions interact with each other. Therefore, the techniques from Cai and Yuan (2012) cannot be applied. To prove the minimax property, we need to combine 2D linearization, two-dimensional Gâteaux derivatives with sophisticated expressions, block matrix inversion, and RKHS, among others.

Fourth, this article is the first attempt to study the theoretical properties of scalar-on-matrix functional regression with a low-rank assumption, where the two-dimensional coefficient function has low rank of products of one-dimensional coefficient functions. This assumption is helpful to extend to functional tensor predictor of even higher order in the future. Tensor regression has become increasingly important recently. There are different combinations of response types and predictor types motivated by various applications, such as tensor-on-vector regression (e.g., Sun and Li, 2017; Zhou et al., 2021), scalar-on-tensor regression (e.g., Zhou et al., 2013; Liu et al., 2019), and tensor-on-tensor regression (e.g., Hoff, 2015; Chen et al., 2021). Most of the existing work focus on the low tensor rank assumption on the coefficient tensor with or without sparsity assumption. To the best of our knowledge, work on functional matrix or functional tensor predictor is very rare.

Lastly, we apply FBLR to two real data examples. One is the Canadian weather data, a well-known example in the functional data analysis (FDA) literature. The goal is to predict precipitation at different weather stations with temperature information. Existing

FDA studies (e.g., Ramsay and Dalzell, 1991; Ramsay and Silverman, 2005b; Cai and Yuan, 2012) focus on 1D PCA in Model (1), where each observation is a vector of 365 daily average temperatures. Besides daily variation, we introduce 2D predictors with the second domain reflecting the hourly temperature variation in Model (3). The extra domain not only boosts the prediction accuracy but also echoes some meteorological phenomena. The other real data example is the Light Detection and Ranging (LIDAR) data from the biochemistry field (Xun et al., 2013). The data generating process naturally produces smooth data and calls for FDA method. For both data sets, FBLR and its iterative variant have higher prediction accuracy and better interpretability compared to existing 2D FLR and 1D FLR methods.

The rest of the paper is organized as follows. A brief review of FLR and the methodology for FBLR are provided in Section 2. The optimal theoretical property of the proposed method is discussed in Section 3. Simulation and the Canadian data analysis are presented in Sections 4 and 5. Section 6 concludes with discussion. All proofs, more simulation results, and the biochemical data application are provided in the supplementary materials.

2. Methodology

2.1 Notation and Definitions

Suppose that $\mathcal{T}$ is a compact set. We denote by $\mathcal{H}(K)$ an RKHS associated with the reproducing kernel K, $\langle\cdot,\cdot\rangle_K$ the associated inner product, and $\|\cdot\|_K$ the induced norm. Then, we have $K(s,\cdot)\in\mathcal{H}(K)$ for all $s\in\mathcal{T}$ and $f(t)=\langle K(t,\cdot),f\rangle_K$ for all $f\in\mathcal{H}(K)$. We refer the readers to Wahba (1990), Gu (2013) and the references therein for more details. Let $K_1(\cdot,\cdot):\mathcal{T}_1\times\mathcal{T}_1\mapsto\mathbb{R}$ and $K_2(\cdot,\cdot):\mathcal{T}_2\times\mathcal{T}_2\mapsto\mathbb{R}$ be two reproducing kernels, and $\mathcal{H}(K_1)$ and $\mathcal{H}(K_2)$ be the corresponding RKHS's. The coefficients $\alpha_0(\cdot)$ and $\beta_0(\cdot)$ of Model (2) reside in $\mathcal{H}(K_1)$ and $\mathcal{H}(K_2)$, respectively.

The covariance function of the mean zero bivariate random function $X(\cdot,\cdot)$ plays another important role in developing both methodology and theory of FBLR. We define it as $C(s_1,t_1,s_2,t_2)=\mathbb{E}[X(s_1,t_1)X(s_2,t_2)]$, for any $s_1,s_2\in\mathcal{T}_1$ and $t_1,t_2\in\mathcal{T}_2$. As mentioned earlier, the two dimensions of $X(\cdot,\cdot)$ usually correspond to different domains, and hence the covariance can be reasonably assumed to have a decomposable or separable structure, that is, $C(s_1,t_1,s_2,t_2)=C_\alpha(s_1,s_2)C_\beta(t_1,t_2)$, where $C_\alpha(\cdot,\cdot)$ and $C_\beta(\cdot,\cdot)$ are two real bivariate functions that characterize the covariance structures along the first and second dimensions respectively. We note that this type of decomposable covariance structure has been widely used in the literature recently when dealing with two-way data (e.g., Zhou, 2014; Volfovsky and Hoff, 2015; Hafner et al., 2020; Aston et al., 2017; Chen et al., 2021, 2023). Similar to the reproducing kernels K_1 and K_2, the two covariance functions C_α and C_β are also symmetric and nonnegative definite. The subscripts α and β will be used frequently to differentiate between the two dimensions.

Lastly, for any $f\in\mathcal{H}(K_1)$ and $g\in\mathcal{H}(K_2)$, we define two semi-norms as follows,

$$\begin{aligned}\|f\|_{0\alpha}&=\left(\int_{\mathcal{T}_1\times T_1}f(s_1)C_\alpha(s_1,s_2)f(s_2)ds_1ds_2\right)^{1/2},\\ \|g\|_{0\beta}&=\left(\int_{\mathcal{T}_2\times T_2}g(t_1)C_\beta(t_1,t_2)g(t_2)dt_1dt_2\right)^{1/2}.\end{aligned}\tag{5}$$

They will appear in the penalty term of the regularization approach and play a crucial role in the development of the asymptotic theory. It can be verified that the variance of the integral of the bilinear form in Model (2) is equal to the square of the product of the two norms defined above,

$$\mathbb{E}\left(\int_{\mathcal{T}_1\times\mathcal{T}_2} f(s)X(s,t)g(t)dsdt\right)^2 = \|f\|_{0\alpha}^2\|g\|_{0\beta}^2. \tag{6}$$

Throughout this paper, following Yuan and Cai (2010), we assume that $\|f\|_{0\alpha}^2 \neq 0$ holds for any $f \neq 0$ that belongs to the null space of K_1, and similarly $\|f\|_{0\beta}^2 \neq 0$ holds for any $f \neq 0$ that belongs to the null space of K_2. This assumption is necessary to ensure that even if the estimation of the coefficient functions is constrained to the null spaces of K_1 and K_2, the objective function to be proposed in Section 2.3 can still be uniquely optimized.

2.2 Review of the Smoothness Regularization Approach for One-way FLR

We provide a brief review of the smoothness regularization approach for the FLR model (1) in this section to facilitate discussion of FBLR. For more details, please see Yuan and Cai (2010), Cai and Yuan (2012), and the references therein.

Consider the reproducing kernel $K(\cdot,\cdot)$ and the corresponding RKHS $\mathcal{H}(K)$. Assume that the coefficient function in Model (1) belongs to $\mathcal{H}(K)$. The smoothness regularized estimator can be obtained via minimizing the following objective with loss and penalty,

$$\widehat{\beta} = \underset{\beta\in\mathcal{H}(K)}{\arg\min}\, \ell_{n\lambda}(\beta) \overset{def}{=} \underset{\beta\in\mathcal{H}(K)}{\arg\min}\, \{\ell_n(\beta) + \lambda J(\beta)\}, \tag{7}$$

where $\ell_n(\beta) = n^{-1}\sum_{i=1}^n \left(y_i - \int_{\mathcal{T}} x_i(t)\beta(t)dt\right)^2$ is the normalized residual sum of squares measuring the goodness-of-fit, $J(\beta) = \|\beta\|_K^2$ is the squared RKHS norm measuring the smoothness, and λ is a tuning parameter that balances the trade-off between them.

Although the optimization in (7) is taken over an infinite-dimensional space, it can be solved by the representer theorem, i.e., Theorem 1 in Yuan and Cai (2010). This representer theorem is a generalization of the well-known representer lemma for smoothing splines (Wahba, 1990). The optimizer of (7) then has the following expression,

$$\widehat{\beta}(t) = \sum_{i=1}^n c_i \int_{\mathcal{T}} K(s,t)x_i(s)ds, \tag{8}$$

where the unknown scalars $c_1, c_2, \ldots, c_n \in \mathbb{R}$ can be readily computed once (8) is plugged back into (7), which leads to a quadratic function of $c_1, ..., c_n$. Please refer to Section 2 of Yuan and Cai (2010) for details on the explicit expression and implementation.

2.3 Objective Function of Two-way FBLR

The smoothness regularization approach of the one-way FLR can be extended to the two-way FBLR. Assume that the coefficient function α_0 resides in $\mathcal{H}(K_1)$ and β_0 in $\mathcal{H}(K_2)$, it is natural to estimate them by minimizing an objective function,

$$(\widehat{\alpha},\widehat{\beta}) = \underset{\alpha\in\mathcal{H}(K_1),\beta\in\mathcal{H}(K_2)}{\arg\min}\, \ell_{n\lambda}(\alpha,\beta) \overset{def}{=} \underset{\alpha\in\mathcal{H}(K_1),\beta\in\mathcal{H}(K_2)}{\arg\min}\, \{\ell_n(\alpha,\beta) + J(\alpha,\beta)\}. \tag{9}$$

This is a direct analogy of (7). $\ell_n(\alpha,\beta) = n^{-1}\sum_{i=1}^n \left(y_i - \int_{\mathcal{T}_1\times\mathcal{T}_2} \alpha(s)x_i(s,t)\beta(t)dsdt\right)^2$, the first part, is again the data fidelity term. However, the second part of the objective function

(9), which is the penalty $J(\alpha, \beta)$, cannot be trivially extended from the one-way case and will be discussed in detail below.

For the choice of $J(\alpha, \beta)$, we point out four properties that one wishes to consider. (1) As mentioned in the introduction, the functions α and β are only identifiable up to a scalar. We intentionally do not enforce any identifiability constraint so that the design of the penalty could be more convenient. Since $\alpha(s)\beta(t)$ is scale-invariant, the penalty $J(\alpha, \beta)$ should be scale-invariant as well; see Huang et al. (2009) for a similar requirement on two-way functional SVD. (2) The loss term $\ell_n(\alpha, \beta)$ in (9) is bi-quadratic in (α, β). Hence, the penalty part should ideally be bi-quadratic as well. (3) The norms $\|\alpha\|_{K_1}^2$ and $\|\beta\|_{K_2}^2$ should encourage smoothness. (4) The two functions α and β are from two domains and can have quite different levels of smoothness, which requires two tuning parameters $\lambda_\alpha, \lambda_\beta$ to control their smoothness respectively in the penalization.

These considerations suggest three potential candidates for the penalty $J(\alpha, \beta)$:

$$J(\alpha,\beta) = \begin{cases} \text{candidate 1: } \lambda_\alpha\lambda_\beta\|\alpha\|_{K_1}^2\|\beta\|_{K_2}^2, \\ \text{candidate 2: } \lambda_\alpha\|\alpha\|_{K_1}^2\|\beta\|_{0\beta}^2 + \lambda_\beta\|\alpha\|_{0\alpha}^2\|\beta\|_{K_2}^2, \\ \text{candidate 3: } \lambda_\alpha\|\alpha\|_{K_1}^2\|\beta\|_{0\beta}^2 + \lambda_\beta\|\alpha\|_{0\alpha}^2\|\beta\|_{K_2}^2 + \lambda_\alpha\lambda_\beta\|\alpha\|_{K_1}^2\|\beta\|_{K_2}^2, \end{cases}$$

where the norms $\|\cdot\|_{0\alpha}$ and $\|\cdot\|_{0\beta}$ involving covariance structure of the input are in (5).

A careful study of these three candidates reveals the following insights. Candidate 1 is simply the product of two one-way penalties, $\lambda_\alpha\|\alpha\|_{K_1}^2$ and $\lambda_\beta\|\beta\|_{K_2}^2$, which is scale-invariant and bi-quadratic. However, it is deficient because it cannot specialize in one-way penalty of FLR by setting one of λ_α and λ_β to zero when it is desirable to only penalize one dimension. Candidates 2 and 3 are both scale-invariant, bi-quadratic, and can both specialize to a form of one-way FLR. Nevertheless, Candidate 3 is the optimal choice since it ensures that the smoothness levels of α and β are detached as shown below, whereas Candidate 2 entangles the smoothness levels of α and β, which is undesirable.

The advantage of Candidate 3 can be seen as follows. After completing the squares of the data fidelity ℓ_n, the objective function $\ell_{n\lambda}$ with Candidate 3 has three types of terms as functions of α and β: bi-quadratic, bi-linear $n^{-1}\sum_{i=1}^n y_i \int \alpha(s)x_i(s,t)\beta(t)dsdt$, and constant $n^{-1}\sum_{i=1}^n y_i^2$. When Candidate 3 is adopted, all the bi-quadratic terms become

$$n^{-1}\sum_{i=1}^n \left(\int \alpha(s)x_i(s,t)\beta(t)dsdt\right)^2 + \lambda_\alpha\|\alpha\|_{K_1}^2\|\beta\|_{0\beta}^2 + \lambda_\beta\|\alpha\|_{0\alpha}^2\|\beta\|_{K_2}^2 + \lambda_\alpha\lambda_\beta\|\alpha\|_{K_1}^2\|\beta\|_{K_2}^2.$$

Because of (6), the population version of the bi-quadratic terms then becomes

$$\|\alpha\|_{0\alpha}^2\|\beta\|_{0\beta}^2 + \lambda_\alpha\|\alpha\|_{K_1}^2\|\beta\|_{0\beta}^2 + \lambda_\beta\|\alpha\|_{0\alpha}^2\|\beta\|_{K_2}^2 + \lambda_\alpha\lambda_\beta\|\alpha\|_{K_1}^2\|\beta\|_{K_2}^2.$$

These four bi-quadratic terms can be written as the product of terms related to α and β separately as

$$\left(\|\alpha\|_{0\alpha}^2 + \lambda_\alpha\|\alpha\|_{K_1}^2\right)\left(\|\beta\|_{0\beta}^2 + \lambda_\beta\|\beta\|_{K_2}^2\right).$$

Here, the bi-quadratic terms of α and β are completely decoupled, which has the following benefits. Imagine that α is known, then optimizing the objective function with respect to β will degenerate to a one-way FLR problem, where the quantities in front of $\|\beta\|_{0\beta}^2$ and $\lambda_\beta\|\beta\|_{K_2}^2$ are always proportional to each other no matter what value α takes, so that the level of smoothness of α does not affect the optimization over β or the smoothness of β.

However, if Candidate 2 is chosen, the bi-quadratic term cannot be written as a product of terms related to α and β separately, and the quantities in front of $\|\beta\|_{0\beta}^2$ and $\lambda_\beta\|\beta\|_{K_2}^2$ are $\lambda_\alpha\|\alpha\|_{K_1}^2$ and $\|\alpha\|_{0\alpha}^2$ respectively. This implies that when α is smoother, so that $\lambda_\alpha\|\alpha\|_{K_1}^2$

is smaller, $\lambda_\beta\|\beta\|^2_{K_2}$ will have a larger impact and β will also tend to be smoother. In summary, Candidate 2 will make the levels of smoothness of the two coefficient functions depend on each other.

This decoupling property of Candidate 3 is not only necessary to separate the smoothness of two coefficient functions in the objective function, but also crucial in Theorem 2 and its proof, where the measurements of the smoothness of α or β appear separately. Otherwise, some kind of measurement of the joint level of smoothness will be necessary to understand the theoretical property.

Note that the phenomenon of decoupling only occurs because of the choice of the norm $\|\alpha\|^2_{0\alpha}, \|\beta\|^2_{0\beta}$ defined in (5), which appears in Candidate 3. Adopting other norms to replace (5) in Candidate 3 will not simultaneously satisfy the invariant, bi-quadratic, specialization to one-way FLR, and decoupling requirements. Therefore, we will use Candidate 3 from now on.

We comment that other three-term penalties have been used to address the problem of SVD of two-way functional data (Huang et al., 2009) and the problem of bivariate smoothing (Xiao et al., 2013). However, the penalty in Huang et al. (2009) only involves the standard l_2 norm of a vector and the norm of second-order differences, which is a discrete version of $\left(\int (f'')^2\right)$, and the penalty in Xiao et al. (2013) involves spline basis and differencing matrix, while our penalty involves the relatively more complicated norm defined in (5) and the Hilbert norm in a more general framework.

Due to the un-identifiability issue of Model (2) and the design of the penalty in the objective function (9), the optimization does not have a unique solution. Specifically, given $(\widehat{\alpha}, \widehat{\beta})$ as the optimizer of (9), for any nonzero constant c, $(\widehat{\alpha}/c, c\widehat{\beta})$ is also the optimizer. We consider these as an equivalent set of solutions by varying c, since they will lead to identical model fit and prediction. As mentioned in the introduction right after Model (2), there are various ways to make the final optimization solution unique if needed, which can be appended to the algorithm in Section 2.4.

2.4 Optimization Algorithm of Two-way FBLR

Recall that the objective function is defined as follows. The bi-quadratic property of the function naturally calls for the block-descent algorithm to optimize the function iteratively. Given a starting point $\alpha^{(0)}$, one can iterate between minimizing one of α and β while holding the other fixed until convergence. For the rest of this section, the focus will be on the discussion of how to update β given α at each iteration. The updating rule for α given β can be obtained analogously.

For any integer $k \geq 1$, suppose the estimation of α in the $(k-1)$th step is $\alpha^{(k-1)}$. Denote a new 1D random input function

$$\tilde{x}_i(t) = \int_{\mathcal{T}_1} \alpha^{(k-1)}(s) x_i(s,t) ds, \tag{10}$$

and for any $f \in \mathcal{H}(K_2)$, define,

$$\|f\|^2_{\tilde{K}_2} = \left(\lambda_\alpha \|\alpha^{(k-1)}\|^2_{K_1}\right) \|f\|^2_{0\beta} + \left(\lambda_\beta \|\alpha^{(k-1)}\|^2_{0\alpha} + \lambda_\alpha\lambda_\beta \|\alpha^{(k-1)}\|^2_{K_1}\right) \|f\|^2_{K_2}, \tag{11}$$

where, given $\alpha^{(k-1)}$, the terms $\lambda_\alpha\|\alpha^{(k-1)}\|^2_{K_1}$ and $\lambda_\beta\|\alpha^{(k-1)}\|^2_{0\alpha} + \lambda_\alpha\lambda_\beta\|\alpha^{(k-1)}\|^2_{K_1}$ are both known constants. Recall the assumption on the relationship between C_β and K_2, which we make at the end of Section 2.1. It is seen that $\|\cdot\|_{\tilde{K}_2}$ is a norm since $\|f\|^2_{\tilde{K}_2} = 0$ if and only

if $f = 0$ and $\|f\|^2_{\tilde{K}_2}$ is quadratic. We further let $\tilde{K}_2(\cdot,\cdot)$ be the reproducing kernel associated with the norm $\|\cdot\|_{\tilde{K}_2}$.

Given $\alpha^{(k-1)}$, a new one-dimensional predictor (10) and a new kernel (11), the objective function $\ell_{n\lambda}(\alpha,\beta)$ for the k-th step becomes a functional of β alone, denoted by $\ell_{n\lambda}(\beta;\alpha^{(k-1)})$, and can be re-expressed in a compact form,

$$\ell_{n\lambda}(\beta;\alpha^{(k-1)}) = n^{-1/2}\sum_{i=1}^{n}\left(y_i - \int_{\mathcal{T}_2}\tilde{x}_i(t)\beta(t)dt\right)^2 + 1\times\|\beta\|^2_{\tilde{K}_2}. \tag{12}$$

The above objective function (12) is the same as the 1D FLR objective function (7) with inputs $\{(\tilde{x}_i(\cdot),y_i)\}_{i=1}^n$, kernel $\tilde{K}_2(\cdot,\cdot)$, and tuning parameter $\lambda = 1$. In other words, by fixing $\alpha^{(k-1)}$, the FBLR problem degenerates to an FLR problem with respect to β. Hence, the intermediate $\beta^{(k)}$ can be obtained from the 1D FLR via the representer theorem (8).

To summarize, the complete approach to obtain the estimators of FBLR is schematically presented in Algorithm 1 with given initialization, known covariance structure of the predictor, and fixed tuning parameters. Some details of the initialization, covariance structure, and tuning parameter selection are provided below.

Input:
1. The observations $(x_i(\cdot,\cdot),y_i), i = 1,...,n$; initial estimator $\alpha^{(0)}$;
2. Penalty parameters λ_α and λ_β; reproducing kernels K_1 and K_2;
3. Pre-specified norms $\|\cdot\|_{0\alpha}$ and $\|\cdot\|_{0\beta}$ as in (5);

repeat

1 To obtain $\beta^{(k)}$ while fixing $\alpha^{(k-1)}$

begin

2 Compute $\tilde{x}_i(\cdot)$ according to (10)

3 Evaluate $\|\alpha^{(k-1)}\|^2_{K_1}$ and $\|\alpha^{(k-1)}\|^2_{0\alpha}$

4 Derive the reproducing kernel $\tilde{K}(\cdot,\cdot)$ associated with the norm defined in (11)

5 Solve (12) for $\beta^{(k)}$ via the one-way FLR approach

end

6 To obtain $\alpha^{(k)}$ while fixing $\beta^{(k)}$

begin

7 Switch the role of α and β in Steps 2-5

end

until *Convergence*;

Output: Estimators $\widehat{\alpha}$ and $\widehat{\beta}$

Algorithm 1: The smoothness regularization approach to the FBLR problem (2).

Initialization. Based on the understanding that if the initial point is close to the truth, then the local contraction property will make the convergent point to be the global optimal one with appealing theoretical property. Hence, it is often true that one only needs a consistent estimator to begin with, and the iterative procedure will produce an optimal solution. There are two possible choices for initialization. The first one is to regress on the 2D functional PCs (Chen et al., 2017), because these estimated PCs have desirable theoretical properties. The second one is to initialize our algorithm with the estimator obtained from Ridge regression after naive vectorization of the two-way covariates. We recommend

the latter one for two reasons: (i) when the reproducing kernel and the covariance kernel align, the former one is computationally much more expensive than the latter while leading to almost identical results; (ii) when these two kernels are misaligned, the former one is inconsistent. These facts are revealed further in the simulation.

Estimation of the covariance function. To implement Algorithm 1, the input of $\|\cdot\|_{0\alpha}$ and $\|\cdot\|_{0\beta}$ as in (5) is required and it depends upon the separable covariance structures C_α and C_β of the covariate. However, for most applications, the two covariance functions are unknown. Hence, we adopt an iterative algorithm introduced in Werner et al. (2008) to estimate them. It basically fixes one of C_α and C_β and estimates the other.

Tuning parameter selection. The selection of tuning parameters plays a crucial role in determining the eventual performance of the algorithm. The most straightforward way is to use CV. However, since there are two tuning parameters, λ_α and λ_β, the search grid will be two-dimensional and the computational cost will be extremely high. So we propose to use the following iGCV approach. As described in Algorithm 1, α or β is updated via one-way FLR given the other. Fixing one of them, say β and λ_β, the selection of λ_α can be done via generalized cross-validation and the α can be updated once λ_α is chosen, and vice versa. The iGCV algorithm terminates when the choices of λ_α and λ_β in the current iteration remain the same as the previous one, and the distances of α and β between the current iteration and the previous one are less than some pre-determined tolerance level. We emphasize that once the iGCV algorithm stops, the tuning parameters λ_α and λ_β are selected, and the estimation of α and β is completed as well. Hence, there is no need to perform another round of iteration for the finally chosen tuning parameters.

3. Theoretical Results: Optimal Rate of Convergence

3.1 Preliminary

In this section, we introduce some notations and assumptions on the reproducing kernels and covariances that will be used in the development of the minimax rate of convergence.

From the methodology point of view, it is implicitly assumed that the two domains, the covariance structures of the two dimensions, and the levels of smoothness of the two coefficient functions may be different and require different penalties thereby. For notational simplicity, throughout the theoretical section, we assume that they are the same, that is, $\mathcal{T}_1 = \mathcal{T}_2 = \mathcal{T}$, $K_1 = K_2 = K$, $C_\alpha = C_\beta = C$ and $\lambda_\alpha = \lambda_\beta = \lambda$. It follows that the $\|\cdot\|_{0\alpha}$ and $\|\cdot\|_{0\beta}$ norms are identical, and hence we write them as $\|\cdot\|_0$. The theoretical results and proofs follow similar logic for the distinct version, with more complicated notation. In particular, the theorems below will hold with the slower rate produced by the two dimensions. See Section D in the Appendix for the statement of the theorems and a brief sketch of the proofs for the version with distinct domains.

The eigen-structures of the kernel K and the covariance C, and their alignment jointly determine the performance of FLR (Yuan and Cai, 2010; Cai and Yuan, 2012). Similarly, they play an important role in the theoretical property of FBLR as well. Suppose that both the reproducing kernel K and the covariance C are continuous and square integrable. By Mercer's Theorem, K and C have the following spectral decompositions, $K(s,t) = \sum_{k=1}^{\infty} s_k^K \phi_k^K(s)\phi_k^K(t)$ and $C(s,t) = \sum_{k=1}^{\infty} s_k^C \phi_k^C(s)\phi_k^C(t)$, where $s_1^K \geq s_2^K \geq \cdots \geq 0$ and

$s_1^C \geq s_2^C \geq \cdots \geq 0$ are the eigenvalues of K and C in descending order, and $\{\phi_1^K, \phi_2^K, ...\}$ and $\{\phi_1^C, \phi_2^C, \ldots\}$ are the corresponding orthonormal eigenfunctions of K and C, respectively.

Given the norms $\|\cdot\|_0$ and $\|\cdot\|_K$, for any function $f \in \mathcal{H}(K)$, define a new norm $\|\cdot\|_R$ that combines the two as $\|f\|_R^2 = \|f\|_0^2 + \|f\|_K^2$. Note that $\|\cdot\|_R$ is indeed a norm as discussed earlier in Section 2.4, since $\|f\|_0^2 \neq 0$ holds for any $\|f\|_K^2 = 0$ and $f \neq 0$. Let R be the corresponding kernel associated with the $\|\cdot\|_R$ norm. Since R is also continuous and square integrable, it follows from Mercer's Theorem that R has the following spectral decomposition, $R(s,t) = \sum_{k=1}^{\infty} s_k^R \phi_k^R(s)\phi_k^R(t)$, where $s_1^R \geq s_2^R \geq \cdots \geq 0$ and $\{\phi_k^R : k = 1, 2, \ldots\}$ are eigenvalues and eigenfunctions respectively.

In general, for a square integrable, symmetric, and non-negative definite function $R : \mathcal{T} \times \mathcal{T} \mapsto \mathbb{R}$ (similarly for K and C), its corresponding linear operator can be defined as $\mathcal{L}_R(f)(\cdot) = \int_{\mathcal{T}} R(t,\cdot)f(t)dt$. It follows from the definitions of the eigenvalues and eigenfunctions of the spectral decomposition of R that $\mathcal{L}_R(\phi_k^R) = s_k^R \phi_k^R$, for $k = 1, 2, \ldots$. Define the square root of the linear operator as $\mathcal{L}_{R^{1/2}}(\phi_k^R) = (s_k^R)^{1/2}\phi_k^R$, for $k = 1, 2, \ldots$. Now consider a new linear operator $\mathcal{L}_T = \mathcal{L}_{R^{1/2}C R^{1/2}}$, that is, $\mathcal{L}_T(f) = \mathcal{L}_{R^{1/2}}(\mathcal{L}_C(\mathcal{L}_{R^{1/2}}(f)))$. Since $\mathcal{L}_T$ is a bounded linear operator, there exist eigenvalues $\{s_1^T, s_2^T, \ldots\}$ in descending order and the corresponding eigenfunctions $\{\phi_1^T, \phi_2^T, \ldots\}$, such that $\mathcal{L}_T(\phi_k^T) = s_k^T \phi_k^T$, for $k = 1, 2, \ldots$.

Define $\omega_k = (s_k^T)^{-1/2}\mathcal{L}_{R^{1/2}}(\phi_k^T)$ and $\gamma_k = (1/s_k^T - 1)^{-1}$. The functions $\{\omega_k : k = 1, 2, \ldots\}$ are essential in the proof, since we will expand all the functions of interest onto these basis functions. The decay rate of γ_k plays a prominent role in the convergence rate.

We will impose the following conditions:

Condition 1: the values γ_k satisfy the decay rate,

$$\gamma_k \asymp k^{-2r}, \tag{13}$$

for some constant $0 < r < \infty$.

Condition 2: for any two functions $f, g \in \mathcal{L}_2(\mathcal{T})$, we further assume that the following fourth moment condition holds,

$$\mathbb{E}\left(\int_{\mathcal{T}\times\mathcal{T}} f(s)X(s,t)g(t)dsdt\right)^4 \leq M\left(\mathbb{E}\left(\int_{\mathcal{T}\times\mathcal{T}} f(s)X(s,t)g(t)dsdt\right)^2\right)^2, \tag{14}$$

for some constant $M > 0$.

Note that Condition 2 is satisfied with $M = 3$ when the process X is assumed to be normal and is therefore weaker than the Gaussian assumption.

Remark 1 (On Condition 1) *There are a few facts which can facilitate the understanding of this condition on the decay rate of γ_k. First, in the literature of nonparametric statistics, it is known that when Sobolev space is studied, the kernel K has eigenvalues decaying as $s_k^K \asymp k^{-2r_K}$ for some $r_K > 1/2$ (Micchelli and Wahba, 1981). Second, if the covariance C satisfies the Sacks-Ylvisaker conditions of order $r_C - 1$, then its eigenvalues decay as $s_k^C \asymp k^{-2r_C}$ (Sacks and Ylvisaker, 1966, 1968, 1970). Third, if the kernel K and the covariance C share the same set of ordered eigenfunctions $\phi_k^K = \phi_k^C$ for all k, that is, under the scenario with perfect alignment, Proposition 4 in Yuan and Cai (2010) shows that $\gamma_k = s_k^C s_k^K$. This implies that when perfect alignment happens and eigenvalues of C and K decay with the parameters r_C and r_K, then γ_k decays with parameter $r = r_C + r_K$. Fourth, even if the assumption of perfect alignment is violated, $\gamma_k \asymp s_k^C s_k^K$ holds in other situations such as Sobolev space $\mathcal{H}$ and C with Sacks-Ylvisaker condition (see Theorem 5 in Yuan and Cai (2010)), or when K and C are commutable. Lastly, the decay rate of γ_k*

depends on not only the decay rates of s_k^K and s_k^C, but also the alignment of the eigenfunctions of K and C. For instance, when $\phi_k^K = \phi_{k^2}^C$, and $s_k^C \asymp k^{-2r_C}, s_k^K \asymp k^{-2r_K}$, we have $\gamma_k \asymp k^{-(4r_C+2r_K)} \asymp k^{-2r}$, where $r = 2r_C + r_K$. Eventually, r will show up in the minimax upper and lower bounds.

3.2 Optimal Rate of Convergence

In this section, we study the asymptotic properties of the FBLR and provide justification of the methodology proposed in Section 2. We first establish a minimax upper bound for the smoothness regularization estimator in Theorem 2 and then derive a minimax lower bound for all possible estimators in Theorem 5. The upper bound matches the lower bound, and hence our proposed smoothness regularization estimator is rate optimal.

We assess the accuracy of the estimators by the excess prediction risk. Suppose (X^*, Y^*) has the same distribution as (X, Y) and is independent of the training data $\{(x_i(\cdot,\cdot), y_i)\}_{i=1}^n$. By taking expectation only with respect to (X^*, Y^*), denoted by $\mathbb{E}^*(\cdot)$, the excess prediction risk of the estimates $\widehat{\alpha}, \widehat{\beta}$ over the true coefficient functions α_0, β_0 is defined as follows,

$$\begin{aligned}\mathcal{E}(\widehat{\alpha}, \widehat{\beta}; \alpha_0, \beta_0) &= \mathbb{E}^*\left(Y^* - \int_{\mathcal{T}\times\mathcal{T}} \widehat{\alpha}(s)X^*(s,t)\widehat{\beta}(t)dsdt\right)^2 - \mathbb{E}^*\left(Y^* - \int_{\mathcal{T}\times\mathcal{T}} \alpha_0(s)X^*(s,t)\beta_0(t)dsdt\right)^2 \\ &= \mathbb{E}^*\left(\int_{\mathcal{T}\times\mathcal{T}} X^*(s,t)\left(\widehat{\alpha}(s)\widehat{\beta}(t) - \alpha_0(s)\beta_0(t)\right)dsdt\right)^2.\end{aligned}$$

Theorem 2 states the upper bound for the excess prediction risk of the smoothness regularized estimator with a properly chosen tuning parameter λ.

Theorem 2 *Under Conditions 1-2, the smoothness regularization estimators $(\widehat{\alpha}, \widehat{\beta})$ defined in (9) with Candidate 3 as the penalty and $\lambda = O(n^{-2r/(2r+1)})$ satisfies*

$$\lim_{A\to\infty}\lim_{n\to\infty}\sup_{\alpha_0\in\mathcal{H}(K),\,\beta_0\in\mathcal{H}(K)} \mathbb{P}\left(\mathcal{E}(\widehat{\alpha}, \widehat{\beta}; \alpha_0, \beta_0) \geq An^{-\frac{2r}{2r+1}}\right) = 0.$$

Remark 3 (On Theorem 2) *By assuming the same domains, kernels, and covariance structures along the two dimensions of the 2D functional covariates, the convergence rate is determined jointly by the joint properties of the covariance C and the kernel K and behaves like most of the non-parametric statistical problems. Furthermore, this result demonstrates that a faster decay rate of the eigenvalues of the covariance C will lead to a faster convergence rate of the estimator. Interestingly, this upper bound recovers the same convergence rate as the one-way FLR. We can relax these assumptions to allow different domains, kernels and/or covariance structures. The proof is essentially the same, but with more complicated notation. The convergence rate will be the maximum of the convergence rates from the two domains $\max\{n^{-\frac{2r_\alpha}{2r_\alpha+1}}, n^{-\frac{2r_\beta}{2r_\beta+1}}\}$, where the subscripts $_\alpha$ and $_\beta$ differentiate the discrepancies between the two domains corresponding to the α and β dimensions, respectively.*

Remark 4 (On the impact of alignment between kernel and covariance) *Since the convergence rate depends on r, defined in (13), in a form of $n^{-\frac{2r}{2r+1}}$, Theorem 2 suggests that, the larger the r, the faster the convergence. Remark 1 discussed the relationship between r and r_C, r_K, which measures the decay of the covariance and kernel, respectively. When the perfect alignment occurs, r takes its largest value of $r_C + r_K$, achieving the fastest convergence. When misalignment occurs, r is smaller, which leads to slower convergence.*

Theorem 5 provides the lower bound for the excess prediction risk over all possible estimators, which is achieved by our estimator.

Theorem 5 *Under the same assumptions as in Theorem 2, for any estimate* $(\tilde{\alpha}, \tilde{\beta})$ *based on the observations* $\{(x_i(\cdot,\cdot), y_i) : i = 1, 2, \ldots, n\}$*, we have the following lower bound,*

$$\lim_{a\to 0} \lim_{n\to\infty} \inf_{\tilde{\alpha},\tilde{\beta}} \sup_{\alpha_0\in\mathcal{H}(K),\, \beta_0\in\mathcal{H}(K)} \mathbb{P}\left(\mathcal{E}(\tilde{\alpha}, \tilde{\beta}; \alpha_0, \beta_0) \geq a n^{-\frac{2r}{2r+1}}\right) = 1.$$

The upper bound in Theorem 2 and the lower bound in Theorem 5 match each other. So we establish the rate optimality of our proposed smoothness regularization estimator.

4. Simulations

4.1 Simulation Settings

We now demonstrate numeric properties of our proposed methodology, in comparison with a few existing methods. In particular, we consider three model settings as discussed in the introduction: our proposed functional bilinear Model (2), the broader low-rank functional bilinear Model (4), and the broadest Model (3) for robustness verification. The last setting is unfavorable to our methodology as the true coefficient function is two-dimensional instead of the product of two 1D functions. However, it will be shown that our method still performs well with the deflation approach mentioned in Section 1, right after introducing Model (4).

We consider three streams of existing methods. The first stream is the Bayesian approach with Gaussian Markov random field (GMRF) priors (Happ et al., 2018). The second one is what we refer to as 2D-FPCR, which is based upon the regression of the scalar response on the estimated 2D functional PCs (2D-FPCA). Chen and Müller (2012); Park and Staicu (2015); Chen et al. (2017) all studied the problem of 2D-FPCA and we will adopt the last one because of its nice properties. Chen et al. (2017) described two versions of 2D-FPCA to estimate the PCs, namely, product FPCA and marginal FPCA, which lead to two estimators of $\beta_0(\cdot,\cdot)$ in Model (3) respectively, referred to as PFPCR and MFPCR accordingly. Implementation of FPCR requires the knowledge of the number of PCs. In what follows, we will compare various possibilities for the unknown ranks.[1]

The third stream targets at the estimation of $\beta_0(\cdot,\cdot)$ in Model (3) directly via an RKHS framework. Such a task can be accomplished via solving (7) while considering a four-dimensional reproducing kernel $K(\cdot,\cdot,\cdot,\cdot)$. To the best of our knowledge, there is no literature that specifically studies the theoretical properties for solving the problem of estimating Model (3) via adopting a four-dimensional kernel in (7). Although Sun et al. (2018); Sang and Li (2022) considered a four-dimensional kernel, they were for the problem with one-dimensional functional input, one-dimensional functional output, and two-dimensional coefficient function.

A natural choice for the four-dimensional kernel is associated with the tensor product RKHS. Suppose for a one-dimensional domain, $K(\cdot,\cdot) = K_0(\cdot,\cdot) + K_1(\cdot,\cdot)$, where the $K_0(\cdot,\cdot)$ and $K_1(\cdot,\cdot)$ correspond to the null space and its orthogonal complement, respectively, of some penalty. For the two-dimensional domain, consider the two marginal reproducing

1. For GMRF, the implementation is available on the webpage https://github.com/ClaraHapp/SOIR. For the calculation of 2D-FPCA, we use the `ProductFPCA` and `MarginalFPCA` functions in the `PACE` package at https://www.stat.ucdavis.edu/PACE/.

kernels $K^1(\cdot,\cdot) = K_0^1(\cdot,\cdot) + K_1^1(\cdot,\cdot)$ and $K^2(\cdot,\cdot) = K_0^2(\cdot,\cdot) + K_1^2(\cdot,\cdot)$. Note that whenever we discuss FLR+TPK in this section, the superscript denotes the domain and the subscript denotes either null or orthogonal complement, which is slightly different from the notations in Sections 2-3. The four-dimensional kernels corresponding to the null space and the orthogonal complement of the tensor product space satisfy that

$$K_0((s_1,t_1),(s_2,t_2)) = K_0^1(s_1,s_2)K_0^2(t_1,t_2), \quad \text{and,}$$

$$K_1((s_1,t_1),(s_2,t_2)) = K_0^1(s_1,s_2)K_1^2(t_1,t_2) + K_1^1(s_1,s_2)K_0^2(t_1,t_2) + K_1^1(s_1,s_2)K_1^2(t_1,t_2).$$

We refer to this approach as FLR+TPK, since it is functional linear regression with tensor product kernel. For more details on the tensor product RKHS, see Chapter 2 of Gu (2013). Note that FLR+TPK only has one tuning parameter that controls the overall smoothness of $\beta(\cdot,\cdot)$, unlike the two tuning parameters in our method that control the smoothness of $\alpha(\cdot)$ and $\beta(\cdot)$ separately.

We consider several other far less competitive competitors in the online supplementary materials (Appendix E). The first one is FLR of one-dimensional input function after vectorization. The second one is a naive implementation of Ridge after plain vectorization without considering the smoothness. The third one is bilinear regression with no penalty, which is a special case of FBLR when $\lambda_\alpha = \lambda_\beta = 0$ and will be called BLR. A quick summary of the message is that all these methods are much worse than FBLR, because they either do not keep matrix/tensor structure or do not take advantage of the smoothness.

Under all six settings, we adopt the same covariance structures from Cai and Yuan (2012), given by

$$C_\alpha(s,t) = C_\beta(s,t) = \sum_{i=1}^{200} 2i^{-2r_c}\cos(i\pi s)\cos(i\pi t), \tag{15}$$

where r_c controls the smoothness of the function. The parameter r_c appears implicitly in the upper bound of Theorem 2 and drives the convergence rate, according to Remark 1. Four choices of r_c are considered: 1, 1.5, 2, and 2.5.

Under all settings, the coefficient functions are set up differently for different purposes. The heat-map plots of $\alpha_0(s)\beta_0(t)$ in Model (2) for the first four settings, the heat-map plot of $\alpha_0^{[1]}(s)\beta_0^{[1]}(t) + \alpha_0^{[2]}(s)\beta_0^{[2]}(t)$ in Model (4) for Setting 5, and the heat-map plot of $\beta_0(s,t)$ in Model (3) for Setting 6 are given in Figure 1.

Specifically, for the first four settings, the coefficient functions are given by

$$\alpha_0(t) = \beta_0(t) = 4\sqrt{2}\sum_{i=1}^{n_{\text{eig}}}(-1)^i i^{-2}\cos\big((i+k_{\text{mis}})\pi t\big). \tag{16}$$

Here, n_{eig} controls the number of eigenfunctions in the coefficient function and will be set to 4 and 200 later. Furthermore, k_{mis} controls the degree of misalignment between the covariance eigen structure in (15) and the leading basis functions in the coefficient function (16). Because the leading eigenfunctions in (15) are ordered from the first to the last as $i = 1, 2, \ldots, 200$ in $\cos(i\pi t)$, while the first leading basis function in (16) is $\cos\big((1+k_{\text{mis}})\pi t\big)$. When $k_{\text{mis}} = 0$, there is no misalignment. As k_{mis} increases, the misalignment becomes more severe. Settings 1-4 consider combinations of $n_{\text{eig}} \in \{4, 200\}$ and $k_{\text{mis}} \in \{0, 4\}$, with detailed configurations in Table 1.

Settings 5-6 are based on Model (3), where the true coefficient function $\beta_0(s,t)$ is indeed two-dimensional, instead of the product of two 1D functions as our model as-

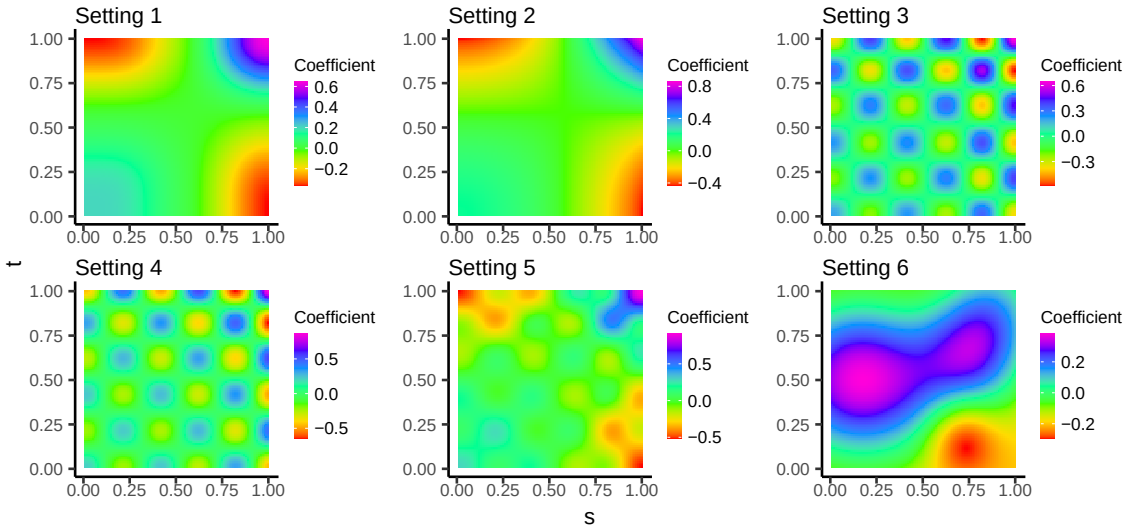

Figure 1: Settings 1-6: Heat-map plots of the true coefficient functions $\beta_0(s,t)$.

Setting	n_{eig}	k_{mis}	indices i of basis function $\cos(i\pi t)$ in the covariances C_α and C_β	indices i of basis function $\cos(i\pi t)$ in the coefficient $\alpha_0(\cdot), \beta_0(\cdot)$
1	4	0	$1, 2, \ldots, 200$	$1, 2, 3, 4$
2	200	0	$1, 2, \ldots, 200$	$1, 2, \ldots, 200$
3	4	4	$1, 2, \ldots, 200$	$5, 6, 7, 8$
4	200	4	$1, 2, \ldots, 200$	$5, 6, \ldots, 204$

Table 1: Settings 1-4: The configurations of the covariances and coefficient functions.

sumes. Setting 5 considers a two-dimensional coefficient function, as the sum of two terms, where each term is a product of two 1D functions, $\beta_0(s,t) = \alpha_0^{[1]}(s)\beta_0^{[1]}(t) + \alpha_0^{[2]}(s)\beta_0^{[2]}(t)$, where $\alpha_0^{[1]}(t) = \beta_0^{[1]}(t) = 4\sqrt{2}\sum_{i=1}^{4}(-1)^i i^{-2}\cos(i\pi t)$, and $\alpha_0^{[2]}(t) = \beta_0^{[2]}(t) = \sqrt{0.4} \times 4\sqrt{2}\sum_{i=1}^{4}(-1)^i i^{-2}\cos\big((i+4)\pi t\big)$.

Setting 6 uses a two-dimensional coefficient function borrowed from the GMRF literature (Happ et al., 2018). We magnified their coefficient function by four times to make its Frobenius norm maintain at the same level as those of the other five settings.

Under these six settings, the data are generated according to Models (2) or (3), where $\mu_0 = 0$, and noise level $\sigma = 0.5$. The predictor $X(s,t)$ follows a centered Gaussian process with the covariance structure described above. The sample size $n = 2^5$, 2^6, 2^7, and 2^8. The continuous functions are observed on a regular grid of length 100. For each value of r_c and each value of n, we repeat the experiments 100 times. The numerical results for different levels of smoothness are similar. Hence, to save space, we present the results for all four choices of r_c for Setting 1, and only for $r_c = 1$ for the other five settings. The kernel functions for the RKHS of FBLR and the marginal RKHS of FLR+TPK are the same,

$$K(s,t) = -B_4(|s-t|/2)/3 - B_4((s+t)/2)/3, \tag{17}$$

where $B_4(\cdot)$ is the 4th Bernoulli polynomial. This kernel indicates the Hilbert norm.

4.2 Simulation Results

Under Setting 1, we first examine the performance of FBLR. As discussed in Section 2.4, the implementation of our method needs special attention to the covariance structure estimation, tuning parameter selection, and proper initialization. We include both the true covariances C_α, C_β and the estimated ones $\hat{C}_\alpha, \hat{C}_\beta$, with the truth as the oracle benchmark. We also consider two choices of tuning parameter selections, CV and iGCV. These choices lead to FBLR+CV+true, FBLR+iGCV+true, FBLR+CV+est, and FBLR+iGCV+est. We further compare three choices of initialization methods: Ridge after vectorization, PFPCR and MFPCR, which lead to multiple versions of our methods: Ridge→FBLR, PFPCR→FBLR and MFPCR→FBLR, correspondingly.

Figure 2 shows the results of excess risk when considering tuning parameter selection via CV vs iGCV. Each panel demonstrates a linear relationship between log(risk) and log(n), and further reveals that the larger the r_c, the faster the convergence. These reconfirm the convergence rate developed theoretically in Theorem 2, where $r = r_c + 1$. The four panels are almost identical, implying that the computationally-inexpensive iGCV performs similar as the computationally-expensive CV, and FBLR with the estimated covariances performs as well as with the true covariances. Hereafter, FBLR means FBLR+iGCV+est for these reasons. Since multiple r_c's show similar messages, we will only use $r_c = 1$ from now on.

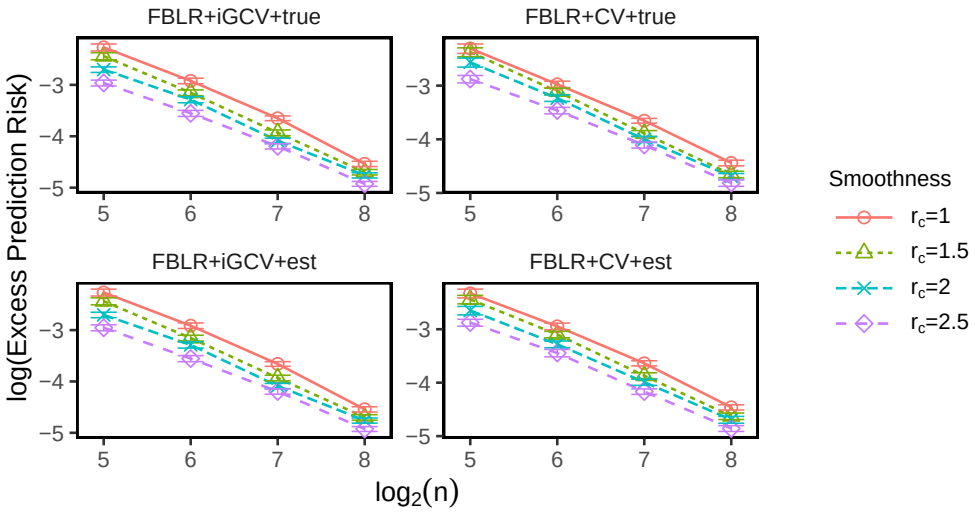

Figure 2: Plots of the excess prediction risk vs the sample size n with both axes in log scale under Setting 1. Four sample sizes and four values of r_c are considered. The error bars are generated according to mean ± one SE. The four panels are all for FBLR approaches, including FBLR+iGCV+true, FBLR+CV+true, FBLR+iGCV+est, and FBLR+CV+est.

Figure 3 shows the performance of FBLR with three different initialization methods under Setting 1 with $r_c = 1$. Initializing via Ridge, PFPCR or MFPCR produces indistinguishable results from the perspective of prediction risk. As for the computational time, it can be seen most of the time is spent on initialization, and FBLR itself takes little time to implement. Because of these observations and the fastest computation of Ridge, for the rest of this article, we will use Ridge as initialization, together with FBLR+iGCV+est.

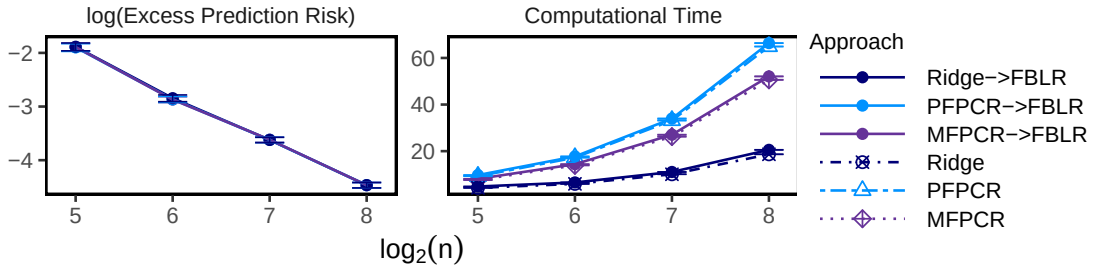

Figure 3: Plots of the logarithm of excess prediction risk and computation time vs the sample size n in log scale with $r_c = 1$ under Setting 1. The error bars are mean $\pm$ one SE.

For Settings 1-4, Figure 4 compares FBLR and three streams of existing methods, such as PFPCR, MFPCR, GMRF, and FLR+TPK, in terms of excess prediction risk and computation time. In the implementation of PFPCR and MFPCR, one needs to specify the maximum number $r^{\max}$ of PCs, and the package will automatically find the optimal number of PCs. According to Table 1, the theoretically optimal numbers of PCs to be provided to estimate the coefficient functions in the noiseless case are 4, 200, 8, and 204, for Settings 1-4, respectively. We tried two choices of $r^{\max} \in \{4, 8\}$. Additionally, in Appendix E of the Supplement, we examine the choice of $r^{\max} = \lfloor \sqrt{n-1} \rfloor$, which means that we provide the code with the largest possible number of PCs that the software can handle. The brief message is that $r^{\max} = \lfloor \sqrt{n-1} \rfloor$ is extremely time consuming and even less accurate than the other two. So, for this section, only $r^{\max} \in \{4, 8\}$ are compared with FBLR.

In Figure 4, it is unsurprising to see that GMRF performs the worst under all settings, albeit its fastest speed. Because it does not take advantage of the low-rank structure of the coefficient function. In terms of computational cost, FBLR is always much faster than 2D-FPCR with $r^{\max} = 8$ for all settings; much faster than 2D-FPCR with $r^{\max} = 4$ and FLR+ TPK under Settings 1-2; and faster than 2D-FPCR with $r^{\max} = 4$ and FLR+ TPK under Settings 3-4 for large sample sizes. Figure 4 also shows that the statistical performance of the prediction risk of FBLR dominates all the other methods under all four settings.

Let us compare FBLR and 2D-FPCR methods statistically. Under Setting 1, despite that $\text{PFPCR}_{r_{\max}=4}$ and $\text{MFPCR}_{r_{\max}=4}$ have the oracle knowledge of the true number of PCs and are perfectly aligned with the coefficient function, they are still worse than the penalized approach FBLR; $\text{PFPCR}_{r_{\max}=8}$ and $\text{MFPCR}_{r_{\max}=8}$ are worse than $\text{PFPCR}_{r_{\max}=4}$ and $\text{MFPCR}_{r_{\max}=4}$, because some unnecessary and noisy PCs are estimated. Under Setting 2, although more basis functions, 200 in total to be exact, are involved in the coefficient functions, $r^{\max} = 4$ still outperforms $r^{\max} = 8$ when $\log_2(n) = 7$ because of smaller variance given the small sample size; $r^{\max} = 8$ and $r^{\max} = 4$ are comparable when the sample size reaches $\log_2(n) = 8$. Under Settings 3-4, there is misalignment, and so 2D-FPCR with $r^{\max} = 4$ is completely off (the leading four PCs are orthogonal to the true coefficient functions). Therefore, $r^{\max} = 8$ performs better than $r^{\max} = 4$, but it is still not as accurate as FBLR. In summary, when the true model is indeed bilinear (2) under Settings 1-4, no matter whether misalignment exists or not, our penalized approach is more robust to the alignment structure than the PC-based approach, produces better prediction, and has less computational burden.

Between FBLR and FLR+TPK, under Settings 1-2, FLR+TPK ranks second and is significantly worse than FBLR; and under Settings 3-4, FLR+TPK is much worse than FBLR, even worse than 2D-FPCR with $r^{\max} = 8$. The phenomenon can be understood as follows. By the representer theorem, the estimate $\hat{\beta}(s,t)$ from FLR+TPK is a linear combination of the basis of the null space and $\int x_i(s_1,t_1)K_1((s_1,t_1),(s,t))ds_1dt_1$. Since tensor product RKHS is used, the estimated function partially depends upon linear combinations of $\int x_i(s_1,t_1)K_1^1(s_1,s)K_1^2(t_1,t)ds_1dt_1$, ignoring some less important terms. Assume the kernels $K_1^1(s,t)$ and $K_1^2(s,t)$ have spectral decompositions $\sum_{k=1}^{\infty} s_k^1\phi_k^1(s)\phi_k^1(t)$ and $\sum_{k=1}^{\infty} s_k^2\phi_k^2(s)\phi_k^2(t)$, respectively. Simplifying the representation theorem partially leads to linear combinations of

$$\int x_i(s_1,t_1)\left(\sum_{k=1}^{\infty} s_k^1\phi_k^1(s_1)\phi_k^1(s)\right)\left(\sum_{k=1}^{\infty} s_k^2\phi_k^2(t_1)\phi_k^2(t)\right)ds_1dt_1$$
$$=\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} s_j^1 s_k^2\left(\int x_i(s_1,t_1)\phi_j^1(s_1)\phi_k^2(t_1)ds_1dt_1\right)\phi_j^1(s)\phi_k^2(t).$$

Hence, the resulting estimator via FLR+TPK will be linear combinations of the products of two basis functions. However, the true coefficient function in Settings 1-4 is of the form of the product of two one-dimensional functions from (16). The FLR+TPK does not take advantage of this knowledge. This is magnified even more for Settings 3-4.

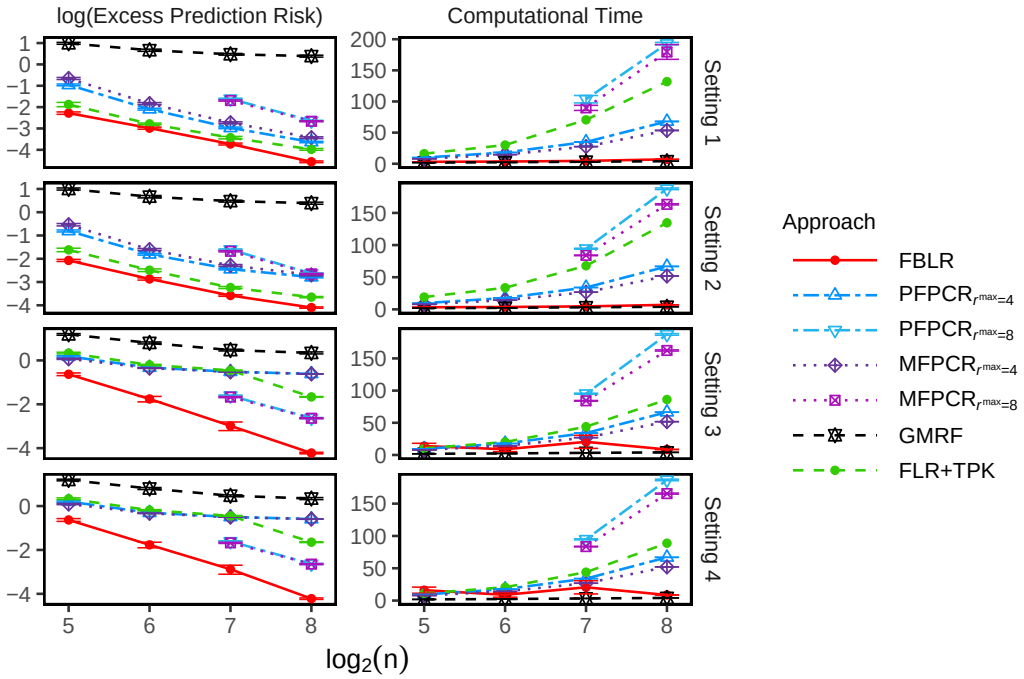

Figure 4: Plots of the logarithm of the excess prediction risk and computation time vs the sample size n in log scale for Settings 1-4 with $r_c = 1$. The error bars are generated according to mean $\pm$ one SE. PFPCR$_{r_{\max}=8}$ and MFPCR$_{r_{\max}=8}$ are only shown for $\log_2(n) = 7, 8$, because they require larger sample size.

Figure 5 shows the results for Settings 5-6, which follow Model (4) and Model (3), respectively. Both settings do not satisfy the model assumption (2), and therefore put our

FBLR in a disadvantageous situation. We used the iterative deflation idea to apply FBLR twice: once on the original data $\{y_i, x_i\}$ and once on the residual $\{e_i, x_i\}$. We denote this approach as FBLR$_{R=2}$, and obtain the estimate of the two-dimensional coefficient function of the form $\hat{\beta}(s,t) = \sum_{r=1}^{2} \hat{\alpha}_0^{[r]}(s)\hat{\beta}_0^{[r]}(t)$. Computationally, it is clear that FBLR, FBLR$_{R=2}$, and GMRF are the fastest among all.

Under Setting 5, FBLR$_{R=2}$ is supposed to be the best because it has the oracle knowledge of $R = 2$. Indeed, it performs better than PFPCR, MFPCR, GMRF and FLR+TPK for all sample sizes considered. It is also interesting to note that FBLR$_{R=2}$ is better than FBLR$_{R=1}$ for large sample sizes, but worse for small sample sizes, even though the true model consists of two terms. It implies that a simple model pays off when limited by sample size: even though Model (2) is more restrictive than Model (4), with limited data, estimating Model (2) when the underlying truth is Model (4) can still be beneficial.

Under Setting 6, the true two-dimensional coefficient function shown in Figure 1 is indeed not low-rank. Under this setting, FBLR$_{R=2}$ performs the best among all estimators for all sample sizes, followed by FLR+TPK, PFPCR$_{r^{\max}=4}$, and MFPCR$_{r^{\max}=4}$. Note that the advantage of FBLR$_{R=2}$ is significant because of the narrow SE shown in the figure. FBLR is worse than the other methods, except for small sample size. 2D-FPCRs with $r^{\max} = 8$ are worse than their counterparts with $r^{\max} = 4$, which suggests a simpler model is preferred when PC regression is considered under Setting 6. In short, although the low-rank Model (4) is more restrictive than the general Model (3), estimating Model (4) with FBLR$_{R=2}$ is still beneficial compared to estimating the two-dimensional coefficient function in Model (3) due to dimension reduction.

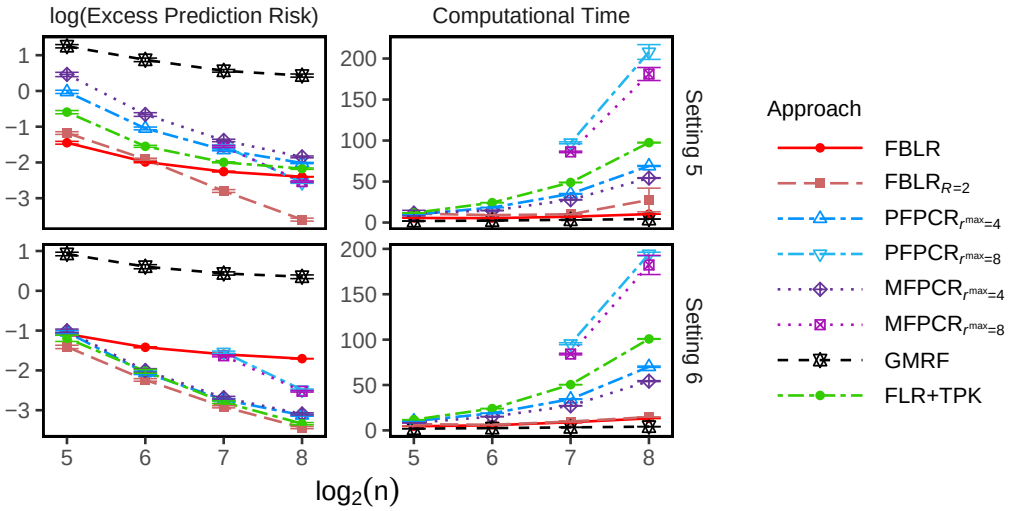

Figure 5: Plots of the logarithm of the excess prediction risk and computation time vs the sample size n in log scale for Settings 5-6 with $r_c = 1$. The error bars are generated according to mean $\pm$ one SE. PFPCR$_{r_{\max}=8}$ and MFPCR$_{r_{\max}=8}$ are only shown for $\log_2(n) = 7, 8$, because they require larger sample size.

In the Appendix, Section E provides extra simulations to study the performance of various vectorization approaches for FLR, other less competitive methods, more analysis on 2D-FPCR, the impact of the choice of the kernel on the performance, and for sparse data.

5. Real Data Analysis: Canadian Weather

We perform real data analysis on two data sets, the Canadian weather data in this section and the LIDAR data in Appendix G.

The Canadian weather data[2] has been widely used for FDA. Traditionally, it is typically used for 1D-FPCA, 1D-FPCR (Ramsay and Silverman, 2005b) or 1D-FLR (Cai and Yuan, 2012), where each vector is of length 365, containing the daily temperature averaged over 24 hours *and* averaged over a few years. We consider the matrices $x_i \in \mathbb{R}^{365\times 24}$, where $x_i(s,t)$ is the temperature in the t-th hour of the s-th day of the year averaged over 2002-2021. Hence, it contains extra hourly information compared to 1D analysis. Following Ramsay and Silverman (2005b); Cai and Yuan (2012), the response variable y_i is the logarithm of the average annual precipitation over 2002-2021, and 35 weather stations are included.

We compare the performances of FBLR and some existing methods, including PFPCR, MFPCR, GMRF, FLR+TPK, Ridge after vectorization, and two variants of 1D-FLR. The first variant is to adopt the FLR method in Cai and Yuan (2012) after matrix vectorization, denoted by FLR+vec. (See Appendix E.1 for further discussion on the potential twist of the vectorization approach.) The second variant is to apply FLR on the vectors of length 365, which are obtained by averaging temperatures over 24 hours, $\sum_{t=1}^{24} x_i(s,t)/24$, denoted by FLR+ave. Both FBLR and FLR-related methods choose the kernel $K_1(s,t) = K_2(s,t) = 1 - B_4(|s-t|)/24$, which is used in Cai and Yuan (2012) as well. For PFPCR and MFPCR, we use $r^{\max} = \lfloor\sqrt{n-1}\rfloor = 5$. BLR (FBLR with no penalty) was not compared because the sample size is not large enough to estimate the parameters.

We first compare all eight methods from the perspective of prediction accuracy and computational time in Figure 6. The leave-one-out method is used to calculate the out-of-sample squared error. FBLR works the best compared to the other 2D and 1D methods, and the differences are all significant because the p-value of the paired t-test between FBLR and each of the other methods is less than 0.05. GMRF performs the worst, followed by FLR+TPK. MFPCR is similar to and PFPCR is worse than three 1D methods including Ridge, FLR+vec, and FLR+ave. Furthermore, FLR+ave, the traditional 1D FDA method by Cai and Yuan (2012), performs worse than FBLR, suggesting that the temperature variations along the hour-of-the-day dimension contain extra information for predicting annual precipitation. For the computational time, it is unsurprising to see that FBLR takes longer than Ridge and FLR+ave, because FBLR uses Ridge as initialization, and FLR+ave has a 1D predictor of length 365 instead of a 2D predictor of size 365×24. But the FBLR is much faster compared to PFPCR, MFPCR, GMRF, FLR+TPK, and FLR+vec.

Figure 7 displays the heat-maps of the estimated 2D coefficient function $\hat{\beta}(s,t)$ in Model (3) for all approaches. For FLR+ave, the 2D function is generated from their estimated 1D coefficient function $\hat{\beta}(\cdot)$ in Model (1) by repeating the yearly pattern for each hour, i.e., $\hat{\beta}(s,t) = \hat{\beta}(s) \times \frac{1}{24}$ for $s = 1,\ldots,365$ and $t = 1,\ldots,24$.

As Figure 7 shows, FBLR produces a smoother coefficient function estimation compared with the other 2D methods. Furthermore, despite PFPCR and MFPCR being designed to generate smooth estimations, the actual estimates are not as smooth as expected, especially

2. The data can be downloaded from the official website of the government of Canada at `https://climate.weather.gc.ca/historical_data/search_historic_data_e.html`.

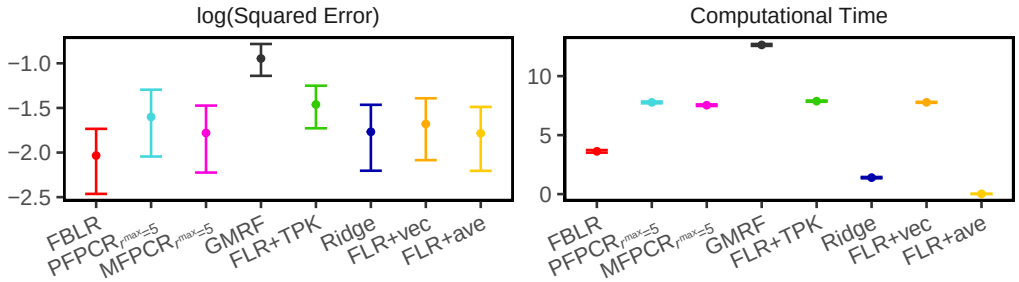

Figure 6: Plots of the out-of-sample performance on the Canadian weather data, which include the testing error and computational time. The error bars are mean ± one SE.

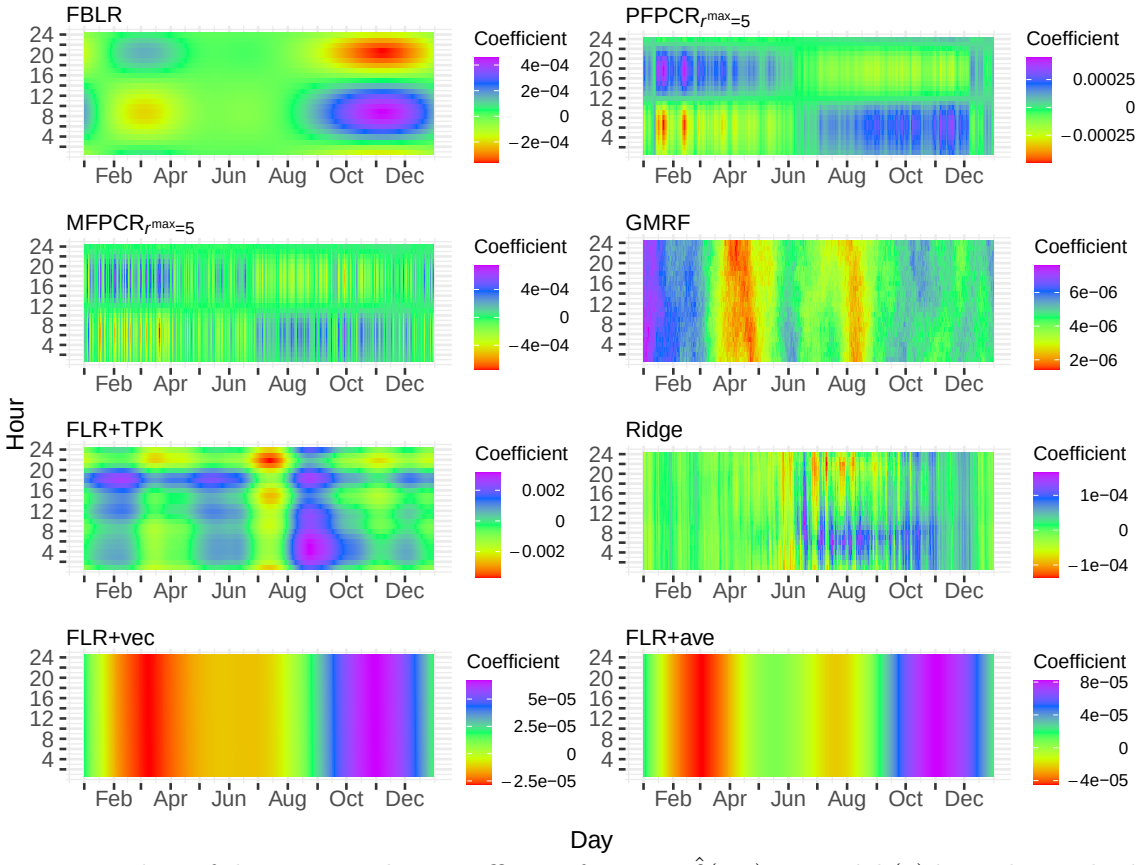

Figure 7: Plots of the estimated 2D coefficient function $\hat{\beta}(s,t)$ in Model (3) by eight methods for Canadian weather data. The x and y axes correspond to day and hour respectively.

for the day-of-the-year dimension. Among the 1D methods, Ridge estimation is non-smooth, and FLR+vec over-smoothes since the hour-of-day dimension has almost no variation.

Figure 8 provides the visualization of the estimated coefficient functions $\hat{\alpha}(\cdot)$ and $\hat{\beta}(\cdot)$ in Model (2). For those methods that do not estimate the 1D coefficient functions directly, the leading left and right singular vectors of the estimated 2D coefficient functions are plotted.

The left panels of Figure 8 show the day-of-the-year effect: PFPCR, MFPCR, and Ridge are not smooth; GMRF and FLR+TPK are not significant, since the confidence

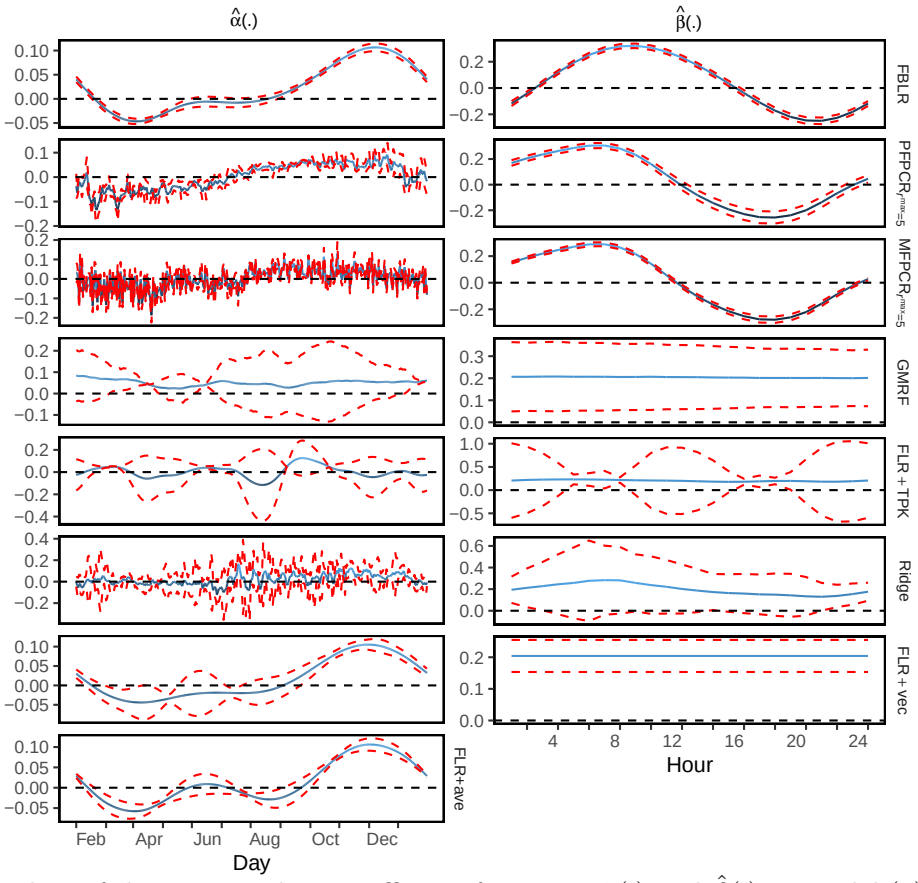

Figure 8: Plots of the estimated 1D coefficient functions $\hat{\alpha}(\cdot)$ and $\hat{\beta}(\cdot)$ in Model (2) by eight methods for Canadian weather data. The confidence intervals correspond to 95% confidence level.

interval covers 0; the width of the confidence interval of FLR+vec is very large, followed by FLR+ave, and FBLR is the narrowest; the shapes of FBLR and FLR+ave (Cai and Yuan, 2012) are rather similar with a peak near November and a trough near March. Such a contrast between Spring and Fall appears at moisture-laden coastal locations (Donohoe et al., 2020), which have warmer autumns and cooler springs than inland stations. As a consequence, they have more precipitation. However, there is a slight difference between FBLR and FLR+ave in that the confidence intervals throughout Summer contain zero for FBLR but not for FLR+ave. The results of FBLR imply that temperature variation in the summer has no effect on the precipitation. This makes sense, as the temperature in the summer at coastal and inland stations has no strong correlation with the maritime effect.

From the right panels of Figure 8 for the time-of-the-day effect, it can be seen that GMRF, FLR+TPK, Ridge, and FLR+vec estimations essentially suggest that there is no hourly effect, while FBLR, PFPCR, and MFPCR reveal and share a significant effect. The estimated $\hat{\beta}(\cdot)$'s by the latter three say that more precipitation occurs with warmer daytime and cooler nighttime. In other words, more precipitation accompanies larger diurnal temperature variation, which promotes the local breeze circulations caused by land-water

temperature differences (e.g., Yang and Smith, 2006), and is essential to convective precipitation in the areas near water bodies (e.g. sea, gulf, lake and river). The residuals of FBLR are examined in Appendix F, which suggests a fairly good fit of the data.

6. Discussion

This article studies the problem of regression of a scalar response on a two-dimensional functional predictor, which when observed on 2D grid becomes a matrix predictor. We propose a functional regression model where the two-dimensional coefficient function is assumed to adopt the form of a product of two 1D coefficient functions. We offer an iterative strategy for the circumstance when the two-dimensional coefficient function can be well approximated by the sum of a few products, which implies low rank for the matrix coefficient. We estimate the model via an innovative penalized approach and compare the penalized approach with the approach of regression on two-dimensional PCs and other methods. We show that the misalignment issue of the PC regression remains as in the 1D FLR case. Even if misalignment does not occur, the penalized approach still performs better than the PCR approach because of further shrinkage. Moreover, the penalized approach also exhibits a huge computational advantage. Real data application further demonstrates the "stableness" and smoothness of the penalized approach.

There are a few meaningful directions for future extensions. The first is to rigorously understand Model (4) with multiple terms, such as how to determine the optimal R. Empirically, one can choose this tuning parameter via CV or similar approaches. Theoretically, this is a challenging topic worth studying further. The second one is to extend from the scalar-on-matrix functional regression to the scalar-on-tensor functional regression, which takes the form of multilinear instead of bilinear. In this case, a careful design of the penalty function is necessary, and our work sheds light on this direction. The LIDAR task in Appendix G intrinsically requires such an extension to consider time, spatial range, and wavelength as a three-dimensional predictor. The fMRI data intrinsically requires such an extension to deal with four-dimensional functional input, corresponding to 3D brain and 1D time. The third one is to extend to a generalized linear model for classification and other purposes. For example, one may want to classify whether a person has Alzheimer's disease based on the fMRI data, which has spatial and temporal smooth input.

Lastly, it is interesting to study the problem when data are not fully observed, from the sparse to the ultra-dense regimes, as in Li and Hsing (2010); Zhang and Wang (2016); Guo et al. (2023). We implement our procedure based on principal component analysis through conditional expectation (PACE) from Yao et al. (2005), and the simulation results for not-fully-observed data are given in Appendix Section E.5. Our conjecture for the theoretical property is that the convergence rate will remain the same for the ultra-dense scenario, but will involve a non-parametric term combined together with the current rate for the sparse and semi-dense scenarios. The rates shall be ranked from the fastest to the slowest for the ultra-dense, semi-dense, and sparse scenarios.

Acknowledgments

We would like to thank the Action Editor and the anonymous referees for their detailed and insightful reviews, which helped to improve the paper substantially. Yang's research is

supported in part by NSF grant IIS-1741390, Hong Kong grant GRF 17301620 and Hong Kong grant CRF C7162-20GF. Shen's research is supported in part by Hong Kong grant CRF C7162-20GF, China Strategy Key Grant 2022SQGH10861, HKU BRC grant, and HKU FBE Shenzhen Research Institutes grant.

Supplementary Materials

The materials include proofs of main theorems and all lemmas, brief statement of the theorems and sketch of the proof for the case with two distinct domains, additional simulation results, additional results of the real data analysis of the Canadian weather, and the second real data example of the LIDAR data.

References

J. A. Aston, D. Pigoli, and S. Tavakoli. Tests for separability in nonparametric covariance operators of random surfaces. *The Annals of Statistics*, 45(4):1431–1461, 2017.

K. Balasubramanian, H. G. Müller, and B. K. Sriperumbudur. Unified rkhs methodology and analysis for functional linear and single-index models. *arXiv preprint arXiv:2206.03975*, 2022.

X. Bi, A. Qu, and X. Shen. Multilayer tensor factorization with applications to recommender systems. *The Annals of Statistics*, 46(6B):3308–3333, 2018.

X. Bi, X. Tang, Y. Yuan, Y. Zhang, and A. Qu. Tensors in statistics. *Annual review of statistics and its application*, 8(1):345–368, 2021.

T. T. Cai and P. Hall. Prediction in functional linear regression. *The Annals of Statistics*, 34(5):2159–2179, 2006.

T. T. Cai and M. Yuan. Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association*, 107(499):1201–1216, 2012.

H. Chen, G. Raskutti, and M. Yuan. Non-convex projected gradient descent for generalized low-rank tensor regression. *The Journal of Machine Learning Research*, 20(1):172–208, 2019.

K. Chen and H. G. Müller. Modeling repeated functional observations. *Journal of the American Statistical Association*, 107(500):1599–1609, 2012.

K. Chen, P. Delicado, and H. G. Müller. Modelling function-valued stochastic processes, with applications to fertility dynamics. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(1):177–196, 2017.

R. Chen, H. Xiao, and D. Yang. Autoregressive models for matrix-valued time series. *Journal of Econometrics*, 222(1):539–560, 2021.

R. Chen, D. Yang, and C. H. Zhang. Factor models for high-dimensional tensor time series. *Journal of the American Statistical Association*, 117(537):94–116, 2022.

X. Chen, D. Yang, Y. Xu, Y. Xia, D. Wang, and H. Shen. Testing and support recovery of correlation structures for matrix-valued observations with an application to stock market data. *Journal of Econometrics*, 232(2):544–564, 2023.

A. Donohoe, E. Dawson, L. McMurdie, D. S. Battisti, and A. Rhines. Seasonal asymmetries in the lag between insolation and surface temperature. *Journal of Climate*, 33(10):3921–3945, 2020.

M. Dyrholm, C. Christoforou, and L. C. Parra. Bilinear discriminant component analysis. *Journal of Machine Learning Research*, 8(5):1097–1111, 2007.

C. Gu. *Smoothing spline ANOVA models*, volume 297. Springer Science and Business Media, 2013.

S. Guillas and M. J. Lai. Bivariate splines for spatial functional regression models. *Journal of Nonparametric Statistics*, 22(4):477–497, 2010.

S. Guo, D. Li, X. Qiao, and Y. Wang. From sparse to dense functional data in high dimensions: Revisiting phase transitions from a non-asymptotic perspective. *arXiv preprint arXiv:2306.00476*, 2023.

C. M. Hafner, O. B. Linton, and H. Tang. Estimation of a multiplicative correlation structure in the large dimensional case. *Journal of Econometrics*, 217(2):431–470, 2020.

C. Happ, S. Greven, and V. J. Schmid. The impact of model assumptions in scalar-on-image regression. *Statistics in medicine*, 37(28):4298–4317, 2018.

P. D. Hoff. Multilinear tensor regression for longitudinal relational data. *The annals of applied statistics*, 9(3):1169, 2015.

J. Z. Huang, H. Shen, and A. Buja. The Analysis of Two-Way Functional Data Using Two-Way Regularized Singular Value Decompositions. *Journal of the American Statistical Association*, 104(488):1609–1620, 2009.

Y. Li and T. Hsing. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, 38(6): 3321–3351, 2010.

J. Liu, Z. Wu, L. Xiao, J. Sun, and H. Yan. Generalized tensor regression for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58(2):1244–1258, 2019.

E. F. Lock. Tensor-on-tensor regression. *Journal of Computational and Graphical Statistics*, 27(3):638–647, 2018.

C. A. Micchelli and G. Wahba. Design problems for optimal surface interpolation. In *Approximation Theory and Applications*, pages 329–348. Academic Press, New York, 1981.

S. Y. Park and A. M. Staicu. Longitudinal functional data analysis. *Stat*, 4(1):212–226, 2015.

J. O. Ramsay and C. J. Dalzell. Some Tools for Functional Data Analysis (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 53(3):539–572, 1991.

J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis: Methods and Case Studies*. Springer, New York, NY, 2002.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, New York, NY, 2nd edition, 2005a.

J. O. Ramsay and B. W. Silverman. Principal components analysis for functional data. *Functional data analysis*, pages 147–172, 2005b.

G. Raskutti, M. Yuan, and H. Chen. Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics*, 47(3):1554–1584, 2019.

P. T. Reiss and R. T. Ogden. Functional Generalized Linear Models with Images as Predictors. *Biometrics*, 66(1):61–69, 2010.

P. T. Reiss, L. Huo, Y. Zhao, C. Kelly, and R. T. Ogden. Wavelet-domain regression and predictive inference in psychiatric neuroimaging. *The Annals of Applied Statistics*, 9(2): 1076–1101, 2015.

P. T. Reiss, J. Goldsmith, H. L. Shang, and R. T. Ogden. Methods for scalar-on-function regression. *International Statistical Review*, 85(2):228–249, 2017.

J. Sacks and D. Ylvisaker. Designs for regression problems with correlated errors; many parameters. *The Annals of Mathematical Statistics*, 39(1):49–69, 1968.

J. Sacks and D. Ylvisaker. Designs for regression problems with correlated errors III. *The Annals of Mathematical Statistics*, 41(1):2057–2074, 1970.

J. Sacks and N. D. Ylvisaker. Designs for regression problems with correlated errors. *The Annals of Mathematical Statistics*, 37(1):66–89, 1966.

P. Sang and B. Li. Nonlinear function-on-function regression by rkhs. *arXiv preprint arXiv:2207.08211*, 2022.

L. M. Sangalli, J. O. Ramsay, and T. O. Ramsay. Spatial spline regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):681–703, 2013.

W. W. Sun and L. Li. Store: sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research*, 18(1):4908–4944, 2017.

X. Sun, P. Du, X. Wang, and P. Ma. Optimal penalized function-on-function regression under a reproducing kernel hilbert space framework. *Journal of the American Statistical Association*, 113(524):1601–1611, 2018.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

A. Volfovsky and P. D. Hoff. Testing for nodal dependence in relational data matrices. *Journal of the American Statistical Association*, 110(511):1037–1046, 2015.

G. Wahba. *Spline models for observational data*. SIAM, Philadelphia, PA, 1990.

J. L. Wang, J. M. Chiou, and H. G. Müller. Functional Data Analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295, 2016.

X. Wang, B. Nan, J. Zhu, and R. Koeppe. Regularized 3D functional regression for brain image data via Haar wavelets. *The Annals of Applied Statistics*, 8(2):1045–1064, 2014.

K. Werner, M. Jansson, and P. Stoica. On estimation of covariance matrices with Kronecker product structure. *IEEE Transactions on Signal Processing*, 56(2):478–491, 2008.

L. Xiao, Y. Li, and D. Ruppert. Fast bivariate p-splines: the sandwich smoother. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):577–599, 2013.

X. Xun, J. Cao, B. Mallick, A. Maity, and R. J. Carroll. Parameter estimation of partial differential equation models. *Journal of the American Statistical Association*, 108(503): 1009–1020, 2013.

S. Yang and E. A. Smith. Mechanisms for diurnal variability of global tropical rainfall observed from trmm. *Journal of climate*, 19(20):5190–5226, 2006.

F. Yao, H. G. Müller, and J. L. Wang. Functional data analysis for sparse longitudinal data. *Journal of the American statistical association*, 100(470):577–590, 2005.

M. Yuan and T. T. Cai. A reproducing kernel hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6):3412–3444, 2010.

X. Zhang and J. L. Wang. From sparse to dense functional data and beyond. *The Annals of Statistics*, 44(5):2281–2321, 2016.

H. Zhou, L. Li, and H. Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.

J. Zhou, W. W. Sun, J. Zhang, and L. Li. Partially observed dynamic tensor response regression. *Journal of the American Statistical Association*, 118(541):1–16, 2021.

S. Zhou. Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics*, 42(2):532–562, 2014.

Supplement to "Optimal Functional Bilinear Regression with Two-way Functional Covariates via Reproducing Kernel Hilbert Space"

Dan Yang, Jianlong Shao, Haipeng Shen, and Hongtu Zhu

In this supplement, we provide proofs of the main theorems in Section A, all lemmas and their detailed proofs in Sections B-C, and brief statement of the theorems and sketch of the proof for the case with two distinct domains in Section D. We also present additional simulation result in Section E, additional real data analysis on the Canadian weather in Section F, and the real data example of LIDAR data in Section G.

Appendix A. Proofs of Main Theorems

A.1 Proof of Theorem 2

In this section, we will show the proof of Theorem 2. We start by providing a result from Yuan and Cai (2010) that will be extensively used in the proof.

Lemma 6 (Theorem 3 of Yuan and Cai (2010)) *For any function $f \in \mathcal{H}(K)$, it can be written as $f = \sum_{k=1}^{\infty} f_k \omega_k$, where $f_k = s_k^T \langle f, \omega_k \rangle_R$. Moreover, the quadratic forms $\|\cdot\|_R^2$, $\|\cdot\|_K^2$, and $\|\cdot\|_0^2$ can be expressed as $\|f\|_R^2 = \sum_{k=1}^{\infty}(1+\gamma_k^{-1})f_k^2$, $\|f\|_0^2 = \sum_{k=1}^{\infty} f_k^2$, and $\|f\|_K^2 = \sum_{k=1}^{\infty}\gamma_k^{-1} f_k^2$.*

Lemma 6 demonstrates that the norms $\|\cdot\|_R^2$, $\|\cdot\|_K^2$, and $\|\cdot\|_0^2$ can be expressed on the basis $\omega_k, k = 1, 2, ...$, which was defined in Section 3.1. See Yuan and Cai (2010) for the elementary proof of Lemma 6.

Recall the definition of the excess prediction risk, from the identity, $\widehat{\alpha}(s)\widehat{\beta}(t) - \alpha_0(s)\beta_0(t) =$

$\big(\widehat{\alpha}(s) - \alpha_0(s)\big)\big(\widehat{\beta}(t) - \beta_0(t)\big) + \big(\widehat{\alpha}(s) - \alpha_0(s)\big)\beta_0(t) + \big(\widehat{\beta}(t) - \beta_0(t)\big)\alpha_0(s)$, and the definition of $\|\cdot\|_0$ norm as in (5) and the property (6), it follows that the excess prediction risk can be further bounded by three terms,

$$\mathcal{E}(\widehat{\alpha}, \widehat{\beta}; \alpha_0, \beta_0) \le 3\|\widehat{\alpha} - \alpha_0\|_0^2 \|\widehat{\beta} - \beta_0\|_0^2 + 3\|\widehat{\alpha} - \alpha_0\|_0^2 \|\beta_0\|_0^2 + 3\|\alpha_0\|_0^2 \|\widehat{\beta} - \beta_0\|_0^2. \tag{18}$$

Therefore, we only need to bound two terms $\|\widehat{\alpha} - \alpha_0\|_0^2$ and $\|\widehat{\beta} - \beta_0\|_0^2$. Due to symmetry in α and β, one bound on $\|\widehat{\alpha} - \alpha_0\|_0^2$ is sufficient for the problem. However, in what follows, we shall bound $\|\widehat{\alpha} - \alpha_0\|_a^2$ due to the necessity in the proof, where the norm $\|\cdot\|_a$ for $0 \le a \le 1$ is defined by $\|f\|_a^2 = \sum_{k=1}^{\infty}(1+\gamma_k^{-a})f_k^2$, when $f = \sum_{k=1}^{\infty} f_k \omega_k$. Clearly, $\|\cdot\|_a$ reduces to $\|\cdot\|_0$ by a factor of 2 when $a = 0$ due to Lemma 6.

Recall that the objective function for the optimization problem is $\ell_{n\lambda}(\alpha, \beta) = \ell_n(\alpha, \beta) + J(\alpha, \beta)$, and the smoothness regularized estimator is obtained via $(\widehat{\alpha}, \widehat{\beta}) = \arg\min \ell_{n\lambda}(\alpha, \beta)$. Write $\ell(\alpha, \beta) = \mathbb{E}\ell_n(\alpha, \beta)$, and $\ell_\lambda(\alpha, \beta) = \mathbb{E}\ell_{n\lambda}(\alpha, \beta)$, to be the expectations of $\ell_n(\alpha, \beta)$ and $\ell_{n\lambda}(\alpha, \beta)$ respectively. The convention is to use subscript n to denote the sample version and without subscript for the population counterpart. Denote the minimizer of $\ell_\lambda(\alpha, \beta)$ by $(\bar{\alpha}, \bar{\beta})$, that is, $(\bar{\alpha}, \bar{\beta}) = \arg\min \ell_\lambda(\alpha, \beta) = \arg\min \ell(\alpha, \beta) + J(\alpha, \beta)$. To bound $\|\widehat{\alpha} - \alpha_0\|_a$, one can bound $\|\bar{\alpha} - \alpha_0\|_a$ and $\|\widehat{\alpha} - \bar{\alpha}\|_a$, which can be thought of as the deterministic error (or bias) and stochastic error (or variance) respectively.

To bound the stochastic error term, another pair $(\tilde{\alpha}, \tilde{\beta})$ has to be introduced so that $\|\widehat{\alpha} - \bar{\alpha}\|_a \le \|\widehat{\alpha} - \tilde{\alpha}\|_a + \|\tilde{\alpha} - \bar{\alpha}\|_a$. Here $(\tilde{\alpha}, \tilde{\beta})$ can be thought of as the expansion and is defined by

$$\begin{pmatrix} \tilde{\alpha} \\ \tilde{\beta} \end{pmatrix} = \begin{pmatrix} \bar{\alpha} \\ \bar{\beta} \end{pmatrix} - H^{-1} \begin{pmatrix} D_\alpha \ell_{n\lambda}(\bar{\alpha}, \bar{\beta}) \\ D_\beta \ell_{n\lambda}(\bar{\alpha}, \bar{\beta}) \end{pmatrix}, \tag{19}$$

where

$$H = \begin{pmatrix} D^2_{\alpha\alpha} \ell_\lambda(\bar{\alpha}, \bar{\beta}) & D^2_{\alpha\beta} \ell_\lambda(\bar{\alpha}, \bar{\beta}) \\ D^2_{\beta\alpha} \ell_\lambda(\bar{\alpha}, \bar{\beta}) & D^2_{\beta\beta} \ell_\lambda(\bar{\alpha}, \bar{\beta}) \end{pmatrix}, \tag{20}$$

where the following operators are defined. The first set of operators are the first- and second-order derivatives of the sample and population loss functions ℓ_n and ℓ,

$$\begin{aligned}
D_\alpha \ell_n(\alpha,\beta) f &= -\frac{2}{n}\sum_{i=1}^n \left(y_i - \int_{\mathcal{T}\times\mathcal{T}} x_i(s,t)\alpha(s)\beta(t)\ dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}} x_i(s,t) f(s)\beta(t)\ dsdt\right), \\
D_\beta \ell_n(\alpha,\beta) f &= -\frac{2}{n}\sum_{i=1}^n \left(y_i - \int_{\mathcal{T}\times\mathcal{T}} x_i(s,t)\alpha(s)\beta(t)\ dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}} x_i(s,t)\alpha(s) f(t)\ dsdt\right), \\
D^2_{\alpha\alpha} \ell_n(\alpha,\beta) fg &= \frac{2}{n}\sum_{i=1}^n \left(\int_{\mathcal{T}\times\mathcal{T}} x_i(s,t) f(s)\beta(t)\ dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}} x_i(s,t) g(s)\beta(t)\ dsdt\right), \\
D^2_{\beta\beta} \ell_n(\alpha,\beta) fg &= \frac{2}{n}\sum_{i=1}^n \left(\int_{\mathcal{T}\times\mathcal{T}} x_i(s,t)\alpha(s) f(t)\ dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}} x_i(s,t)\alpha(s) g(t)\ dsdt\right), \\
D^2_{\alpha\beta} \ell_n(\alpha,\beta) fg &= -\frac{2}{n}\sum_{i=1}^n \left(y_i - \int_{\mathcal{T}\times\mathcal{T}} x_i(s,t)\alpha(s)\beta(t)\ dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}} x_i(s,t) f(s) g(t)\ dsdt\right) \\
&\quad + \frac{2}{n}\sum_{i=1}^n \left(\int_{\mathcal{T}\times\mathcal{T}} x_i(s,t) f(s)\beta(t)\ dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}} x_i(s,t)\alpha(s) g(t)\ dsdt\right), \\
D_\alpha \ell(\alpha,\beta) f &= -2\int_{\mathcal{T}^4} C(s_1,s_2)C(t_1,t_2)\big(\alpha_0(s_1)\beta_0(t_1) - \alpha(s_1)\beta(t_1)\big) f(s_2)\beta(t_2)\ ds_1 ds_2 dt_1 dt_2, \\
D_\beta \ell(\alpha,\beta) f &= -2\int_{\mathcal{T}^4} C(s_1,s_2)C(t_1,t_2)\big(\alpha_0(s_1)\beta_0(t_1) - \alpha(s_1)\beta(t_1)\big)\alpha(s_2) f(t_2)\ ds_1 ds_2 dt_1 dt_2, \\
D^2_{\alpha\alpha} \ell(\alpha,\beta) fg &= 2\|\beta\|_0^2 \int_{\mathcal{T}\times\mathcal{T}} C(s,t) f(s) g(t)\ dsdt, \\
D^2_{\beta\beta} \ell(\alpha,\beta) fg &= 2\|\alpha\|_0^2 \int_{\mathcal{T}\times\mathcal{T}} C(s,t) f(s) g(t)\ dsdt, \\
D^2_{\alpha\beta} \ell(\alpha,\beta) fg &= -2\left(\int_{\mathcal{T}\times\mathcal{T}} C(s,t)\alpha_0(s) f(t) dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}} C(s,t)\beta_0(s) g(t) dsdt\right) \\
&\quad + 4\left(\int_{\mathcal{T}\times\mathcal{T}} C(s,t)\alpha(s) f(t) dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}} C(s,t)\beta(s) g(t) dsdt\right),
\end{aligned}$$

and the second set of operators are first- and second-order derivatives of the sample and population objective functions $\ell_{n\lambda}$ and ℓ_λ,

$$\begin{aligned}
D_\alpha \ell_{n\lambda}(\alpha,\beta) &= D_\alpha \ell_n(\alpha,\beta) + D_\alpha J(\alpha,\beta), \\
D_\beta \ell_{n\lambda}(\alpha,\beta) &= D_\beta \ell_n(\alpha,\beta) + D_\beta J(\alpha,\beta), \\
D^2_{\alpha\alpha} \ell_{n\lambda}(\alpha,\beta) &= D^2_{\alpha\alpha} \ell_n(\alpha,\beta) + D^2_{\alpha\alpha} J(\alpha,\beta), \\
D^2_{\beta\beta} \ell_{n\lambda}(\alpha,\beta) &= D^2_{\beta\beta} \ell_n(\alpha,\beta) + D^2_{\beta\beta} J(\alpha,\beta), \\
D^2_{\alpha\beta} \ell_{n\lambda}(\alpha,\beta) &= D^2_{\alpha\beta} \ell_n(\alpha,\beta) + D^2_{\alpha\beta} J(\alpha,\beta), \\
D_\alpha \ell_\lambda(\alpha,\beta) &= D_\alpha \ell(\alpha,\beta) + D_\alpha J(\alpha,\beta), \\
D_\beta \ell_\lambda(\alpha,\beta) &= D_\beta \ell(\alpha,\beta) + D_\beta J(\alpha,\beta), \\
D^2_{\alpha\alpha} \ell_\lambda(\alpha,\beta) &= D^2_{\alpha\alpha} \ell(\alpha,\beta) + D^2_{\alpha\alpha} J(\alpha,\beta), \\
D^2_{\beta\beta} \ell_\lambda(\alpha,\beta) &= D^2_{\beta\beta} \ell(\alpha,\beta) + D^2_{\beta\beta} J(\alpha,\beta), \\
D^2_{\alpha\beta} \ell_\lambda(\alpha,\beta) &= D^2_{\alpha\beta} \ell(\alpha,\beta) + D^2_{\alpha\beta} J(\alpha,\beta).
\end{aligned}$$

By the triangle inequality, we have

$$\|\widehat{\alpha}-\alpha_0\|_a = \|(\widehat{\alpha}-\tilde{\alpha})+(\tilde{\alpha}-\bar{\alpha})+(\bar{\alpha}-\alpha_0)\|_a \le \|\widehat{\alpha}-\tilde{\alpha}\|_a + \|\tilde{\alpha}-\bar{\alpha}\|_a + \|\bar{\alpha}-\alpha_0\|_a. \quad (21)$$

Lemmas 7, 8 and 9 in Section B establish bounds for the three terms on the right-hand side of (21) respectively and together with (21) imply that, if $\lambda = O(n^{-2r/(2r+1)})$,

$$\lim_{A\to\infty}\lim_{n\to\infty}\sup_{\substack{\alpha_0\in\mathcal{H}(K),\\ \beta_0\in\mathcal{H}(K)}} \mathbb{P}(\|\widehat{\alpha}-\alpha_0\|_0^2 \ge An^{-\frac{2r}{2r+1}}) = 0, \quad (22)$$

$$\lim_{A\to\infty}\lim_{n\to\infty}\sup_{\substack{\alpha_0\in\mathcal{H}(K),\\ \beta_0\in\mathcal{H}(K)}} \mathbb{P}(\|\widehat{\beta}-\beta_0\|_0^2 \ge An^{-\frac{2r}{2r+1}}) = 0. \quad (23)$$

Combining (22), (23) and (18) completes the proof of Theorem 2. ■

Comment: Note that the proof to 2D FBLR differs much from the proof to 1D FLR since FLR only requires expansion of $\tilde{\beta}$ whereas FBLR relies on 2D expansion (19) which complicates the proofs of the lemmas extensively. To save the readers some detour, expanding $\tilde{\alpha}, \tilde{\beta}$ separately without considering their interaction cannot lead to the full proof.

A.2 Proof of Theorem 5

Note that although it has been assumed throughout that the noise ϵ has mean zero and finite variance, for the proof of lower bound, it suffices to assume the normal case $\epsilon \sim N(0,\sigma^2)$. This is because any lower bound for the normal case yields a lower bound for the general case without normality.

We will invoke Lemma 16 in Section C.1. To that end, we construct a parameter space Θ that contains elements $\theta = (\theta_{N+1},...,\theta_{2N})^T \in \{0,1\}^N$, where N is the smallest integer such that $N \ge c_1 n^{1/(2r+1)}$ for some constant $c_1 > 0$ whose value will be specified later. The slope function α depends on θ through

$$\alpha_\theta = N^{-1/2}\sum_{k=N+1}^{2N}\theta_k\gamma_k^{1/2}\omega_k. \quad (24)$$

Then it is easy to verify that

$$\|\alpha_\theta\|_K^2 = N^{-1}\sum_{k=N+1}^{2N}\theta_k^2\gamma_k\|\omega_k\|_K^2 = N^{-1}\sum_{k=N+1}^{2N}\theta_k^2 \le N^{-1}\sum_{k=N+1}^{2N}1 = 1,$$

which proves that $\alpha_\theta \in \mathcal{H}$.

Define the parameter space Θ as $\Theta = \{\theta^{(0)},\theta^{(1)},...,\theta^{(M)}\} \subset \{0,1\}^N$. For $N \ge 8$, Gilbert Shannon Varshamov bound guarantees that the following conditions hold simultaneously:

1. $\theta^{(0)} = (0,...,0)^T$,
2. $H(\theta,\theta') > N/8$ for any pair $\theta \ne \theta' \in \Theta$, where H is the Hamming distance,
3. The cardinality of the set is at least $M \ge 2^{N/8}$.

Denote by P_θ the joint distribution of $(X_i,Y_i), i = 1,...,n$ given $\alpha_0 = \alpha_\theta, \beta_0 = \omega_1$, then the ratio of the density becomes

$$\log\frac{P_\theta}{P_{\theta'}} = \frac{2\sum_{i=1}^n\left(Y_i - \int X_i\alpha_\theta\omega_1\right)\int X_i(\alpha_\theta\omega_1 - \alpha_{\theta'}\omega_1) + \sum_{i=1}^n\left(\int X_i(\alpha_\theta\omega_1 - \alpha_{\theta'}\omega_1)\right)^2}{2\sigma^2}. \quad (25)$$

Based on this expression, the Kullback-Leibler distance between P_θ and $P_{\theta'}$ can be computed $KL(P_\theta, P_{\theta'}) = \int \log \frac{P_\theta}{P_{\theta'}} P_\theta = \frac{n}{2\sigma^2}\|\alpha_\theta - \alpha_{\theta'}\|_0^2 \|\omega_1\|_0^2$. Plugging in α_θ (24) leads to

$$\begin{aligned} KL(P_\theta, P_{\theta'}) &= \frac{n}{2\sigma^2}\left\| N^{-1/2} \sum_{k=N+1}^{2N} (\theta_k - \theta'_k)\gamma_k^{1/2}\omega_k \right\|_0^2 = \frac{n}{2N\sigma^2} \sum_{k=N+1}^{2N} (\theta_k - \theta'_k)^2 \gamma_k \\ &\leq \frac{n\gamma_N}{2N\sigma^2} \sum_{k=N+1}^{2N} (\theta_k - \theta'_k)^2 \leq \frac{n\gamma_N}{2N\sigma^2} H(\theta, \theta') \leq \frac{n\gamma_N}{2\sigma^2}, \end{aligned}$$

since the Hamming distance is bounded by the dimension. Due to the rate assumption of γ, the cardinality of the set Θ, and the assumption on the size of N,

$$KL(P_\theta, P_{\theta'}) \leq \frac{c_2 n N^{-(2r)}}{2\sigma^2} \leq \frac{c_2 c_1^{-2r+1} N^{2r+1} N^{-(2r)}}{2\sigma^2} = \frac{c_2 c_1^{2(-r)+1} N}{2\sigma^2} \leq \delta \log 2^{N/8} \leq \delta \log M,$$

for any $0 < \delta < 1/8$ by taking c_1 large enough. This further proves that

$$\frac{1}{M}\sum_{j=1}^{M} KL(P_{\theta^{(j)}}, P_{\theta^{(0)}}) \leq \delta \log M, \tag{26}$$

which satisfies the second condition (ii) in Lemma 16.

Turning to the first condition (i), define a distance between (α, β) and (α', β') as $d\big((\alpha,\beta),(\alpha',\beta')\big) = E\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\big(\alpha(s)\beta(t) - \alpha'(s)\beta'(t)\big)dsdt\right)^2$. Due to the definition of $\|\cdot\|_0$ norm and the expression of α_θ, the distance can be lowered bounded by

$$\begin{aligned} d\big((\alpha_\theta,\beta_0),(\alpha_{\theta'},\beta_0)\big) &= \|\alpha_\theta - \alpha_{\theta'}\|_0^2 \|\beta_0\|_0^2 = \left\| N^{-1/2} \sum_{k=N+1}^{2N} (\theta_k - \theta'_k)\gamma_k^{1/2}\omega_k \right\|_0^2 \|\omega_1\|_0^2 \\ &= N^{-1} \sum_{k=N+1}^{2N} (\theta_k - \theta'_k)^2 \gamma_k \geq N^{-1}\gamma_{2N} \sum_{k=N+1}^{2N} (\theta_k - \theta'_k)^2 = N^{-1}\gamma_{2N} H(\theta, \theta'). \end{aligned}$$

Because of the second requirement of the construction of the set Θ, the rate assumption of γ, and the assumption on the size of N, we have

$$d\big((\alpha_\theta,\beta_0),(\alpha_{\theta'},\beta_0)\big) \geq \gamma_{2N}/8 \geq c_3 2^{-2r-3} N^{-2r} \geq 2c_4 \delta^{\frac{2r}{2r+1}} n^{-\frac{2r}{2r+1}}.$$

This inequality and (26) together imply that

$$\inf_{\widehat{\alpha},\widehat{\beta}} \sup_{\theta\in\Theta} P_\theta \left(d\big((\widehat{\alpha},\widehat{\beta}),(\alpha_\theta,\beta_0)\big) \geq c_4 \delta^{\frac{2r}{2r+1}} n^{-\frac{2r}{2r+1}} \right) \geq \frac{\sqrt{M}}{1+\sqrt{M}} \left(1 - 2\delta - \sqrt{\frac{2\delta}{\log M}} \right).$$

Letting $n \to \infty$, $\lim_{n\to\infty} \inf_{\widehat{\alpha},\widehat{\beta}} \sup_{\theta\in\Theta} P_\theta \left(d\big((\widehat{\alpha},\widehat{\beta}),(\alpha_\theta,\beta_0)\big) \geq c_4 \delta^{\frac{2r}{2r+1}} n^{-\frac{2r}{2r+1}} \right) \geq 1 - 2\delta$, which further implies that $\lim_{a\to 0} \lim_{n\to\infty} \inf_{\widehat{\alpha},\widehat{\beta}} \sup_{\theta\in\Theta} P_\theta \left(d\big((\widehat{\alpha},\widehat{\beta}),(\alpha_\theta,\beta_0)\big) \geq a n^{-2r/(2r+1)} \right) = 1$. Realizing $\mathcal{E}(\widehat{\alpha},\widehat{\beta};\alpha_0,\beta_0) = d\big((\widehat{\alpha},\widehat{\beta}),(\alpha_0,\beta_0)\big)$ completes the proof. ■

Appendix B. Main Lemmas and Their Proofs

B.1 Main Lemmas

Lemma 7 *If $\lambda = o(1)$, $0 \leq a \leq 1$, then*

$$\|\bar{\alpha} - \alpha_0\|_a^2 = O(\lambda^{1-a}), \text{ and } \|\bar{\beta} - \beta_0\|_a^2 = O(\lambda^{1-a}).$$

An immediate consequence of Lemma 7 is $O(\|\bar{\alpha}\|_a) = O(\|\alpha_0\|_a) = O(1)$ and $O(\|\bar{\beta}\|_a) = O(\|\beta_0\|_a) = O(1)$ because of the following observation

$$\begin{aligned} \|\bar{\alpha}\|_a = \|\bar{\alpha} - \alpha_0 + \alpha_0\|_a &\geq \|\alpha_0\|_a - \|\bar{\alpha} - \alpha_0\|_a \succeq \|\alpha_0\|_a - o(1) = O(1), \\ \|\bar{\alpha}\|_a = \|\bar{\alpha} - \alpha_0 + \alpha_0\|_a &\leq \|\alpha_0\|_a + \|\bar{\alpha} - \alpha_0\|_a \preceq \|\alpha_0\|_a + o(1) = O(1), \end{aligned}$$

and a parallel argument for $\bar{\beta}$ holds. From now on, there will be multiple appearances of $\|\bar{\alpha}\|_a, \|\bar{\beta}\|_a$, which will be treated as constants.

Lemma 8 *If $\lambda = o(1)$, $0 \le a \le 1$ and $r > 1/2$, then*
$\mathbb{E}\|\tilde{\alpha} - \bar{\alpha}\|_a^2 \preceq n^{-1}\lambda^{-(a+1/(2r))}$, *and* $\mathbb{E}\|\tilde{\beta} - \bar{\beta}\|_a^2 \preceq n^{-1}\lambda^{-(a+1/(2r))}$.

Lemma 9 *If there exists some constant c such that $1/(2r) < c \le 1$ and $n^{-1}\lambda^{-(c+1/(2r))} = o(1)$, then*
$\|\widehat{\alpha} - \tilde{\alpha}\|_a^2 = o_p(n^{-1}\lambda^{-(a+1/(2r))})$, *and* $\|\widehat{\beta} - \tilde{\beta}\|_a^2 = o_p(n^{-1}\lambda^{-(a+1/(2r))})$.

For the rest of Section B, we will provide Proofs of Lemmas 7-9.

B.2 Proof of Lemma 7

Expanding α, α_0, $\bar{\alpha}$, β, β_0, and $\bar{\beta}$ on the basis $\{\omega_k : k = 1, 2, \ldots\}$ and denoting

$$\begin{array}{lll} \alpha = \sum_{k=1}^{\infty} a_k\omega_k, & \alpha_0 = \sum_{k=1}^{\infty} a_{0k}\omega_k, & \bar{\alpha} = \sum_{k=1}^{\infty} \bar{a}_k\omega_k, \\ \beta = \sum_{k=1}^{\infty} b_k\omega_k, & \beta_0 = \sum_{k=1}^{\infty} b_{0k}\omega_k, & \bar{\beta} = \sum_{k=1}^{\infty} \bar{b}_k\omega_k. \end{array} \tag{27}$$

Substituting the expansions (27) into $\ell(\alpha, \beta)$ and together with identities

$$\langle \omega_j, \omega_k\rangle_R = \delta_{jk}(1/\gamma_k + 1), \quad \langle C\omega_j, \omega_k\rangle_{\mathcal{L}_2} = \delta_{jk}, \quad \text{and } \langle \omega_j, \omega_k\rangle_K = \delta_{jk}/\gamma_k, \tag{28}$$

it follows that

$$\ell(\alpha, \beta) = \sigma^2 + \left(\sum_{k=1}^{\infty} a_k^2\right)\left(\sum_{k=1}^{\infty} b_k^2\right) + \left(\sum_{k=1}^{\infty} a_{0k}^2\right)\left(\sum_{k=1}^{\infty} b_{0k}^2\right) - 2\left(\sum_{k=1}^{\infty} a_{0k}a_k\right)\left(\sum_{k=1}^{\infty} b_{0k}b_k\right).$$

Similarly, the penalty term $J(\alpha, \beta)$ can be re-expressed as

$$J(\alpha, \beta) = \lambda\left(\sum_{k=1}^{\infty} a_k^2\right)\left(\sum_{k=1}^{\infty} \gamma_k^{-1}b_k^2\right) + \lambda\left(\sum_{k=1}^{\infty} \gamma_k^{-1}a_k^2\right)\left(\sum_{k=1}^{\infty} b_k^2\right) + \lambda^2\left(\sum_{k=1}^{\infty} \gamma_k^{-1}a_k^2\right)\left(\sum_{k=1}^{\infty} \gamma_k^{-1}b_k^2\right).$$

Minimizing $\ell(\alpha, \beta) + J(\alpha, \beta)$ with respect to a_k and b_k leads to

$$\bar{a}_k = c\frac{a_{0k}}{1 + \lambda\gamma_k^{-1}}, \quad \bar{b}_k = c^{-1}\frac{b_{0k}}{1 + \lambda\gamma_k^{-1}}, \quad k = 1, 2, \ldots, \tag{29}$$

where c can be any nonzero real constant. For simplicity, we take $c = 1$ and hence $\bar{\alpha}$ and $\bar{\beta}$ can be written as follows, for all $k = 1, 2, \ldots$,

$$\bar{\alpha} = \sum_{k=1}^{\infty} \frac{a_{0k}}{1 + \lambda\gamma_k^{-1}}\omega_k, \quad \bar{\beta} = \sum_{k=1}^{\infty} \frac{b_{0k}}{1 + \lambda\gamma_k^{-1}}\omega_k. \tag{30}$$

Now we are ready to bound the $\|\bar{\alpha} - \alpha_0\|_a^2$ term in view of the definition of $\|\cdot\|_a$ norm,

$$\begin{aligned} \|\bar{\alpha} - \alpha_0\|_a^2 &= \|\sum_{k=1}^{\infty} \frac{\lambda\gamma_k^{-1}a_{0k}}{1 + \lambda\gamma_k^{-1}}\omega_k\|_a^2 = \sum_{k=1}^{\infty}(1 + \gamma_k^{-a})\left(\frac{\lambda\gamma_k^{-1}a_{0k}}{1 + \lambda\gamma_k^{-1}}\right)^2 \\ &\le \lambda^2 \sup_k \frac{\gamma_k^{-1}(1 + \gamma_k^{-a})}{(1 + \lambda\gamma_k^{-1})^2} \sum_{k=1}^{\infty} \gamma_k^{-1}a_{0k}^2 = \lambda^2\|\alpha_0\|_K^2 \sup_k \frac{\gamma_k^{-1}(1 + \gamma_k^{-a})}{(1 + \lambda\gamma_k^{-1})^2}. \end{aligned}$$

Replacing the maximum over non-negative integers by supremum over a continuous variable in $(0, \infty)$,

$$\sup_k \frac{\gamma_k^{-1}(1 + \gamma_k^{-a})}{(1 + \lambda\gamma_k^{-1})^2} \le \sup_{x>0} \frac{x^{-1}(1 + x^{-a})}{(1 + \lambda x^{-1})^2} = O(\lambda^{-(a+1)}).$$

Combining the last two displays completes the proof of Lemma 7. ■

B.3 Proof of Lemma 8

For brevity, we first introduce a few more notations. Define a new norm

$$\|\cdot\|_\lambda^2 = \|\cdot\|_0^2 + \lambda\|\cdot\|_K^2, \tag{31}$$

and write

$$\bar{\mathbf{a}} = (\bar{a}_1, \bar{a}_2, \ldots)^T, \quad \bar{\mathbf{b}} = (\bar{b}_1, \bar{b}_2, \ldots)^T, \tag{32}$$

for the vectors that contain the basis expansion coefficients of $\bar{\alpha}, \bar{\beta}$.

Since $\tilde{\alpha}, \tilde{\beta}$ defined in (19) depends on H^{-1}, in order to bound $\|\tilde{\alpha} - \bar{\alpha}\|_a^2$, $\|\tilde{\beta} - \bar{\beta}\|_a^2$, it is necessary to obtain the explicit form of H^{-1}. Note that H takes a block form, we decompose H defined in (20) and its inverse H^{-1} accordingly as follows,

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}, \quad H^{-1} = \begin{pmatrix} H^{11} & H^{12} \\ H^{21} & H^{22} \end{pmatrix}. \tag{33}$$

The two diagonal blocks of H^{-1} are further decomposed into two parts,

$$H^{11} = H^{11(1)} + H^{11(2)}, \quad H^{22} = H^{22(1)} + H^{22(2)}, \tag{34}$$

where the first parts $^{(1)}$ are the leading terms contributing to the error and the second parts $^{(2)}$ and off diagonal blocks G^{12}, G^{21} are negligible, which will be proved later.

Let G_{kl}, $k, l = 1, 2$, be matrices such that the ij-th entry of G_{kl} is given by, $(G_{kl})_{ij} = H_{kl} w_i w_j, \quad i, j = 1, 2, \ldots$. Write G^{kl} and $G^{kk(l)}$, $k, l = 1, 2$, in a similar way. Define G and G^{-1} as matrix counterparts of H and H^{-1} respectively,

$$G = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}, \quad G^{-1} = \begin{pmatrix} G^{11} & G^{12} \\ G^{21} & G^{22} \end{pmatrix}. \tag{35}$$

and G^{11} and G^{22} are further decomposed as,

$$G^{11} = G^{11(1)} + G^{11(2)}, \quad G^{22} = G^{22(1)} + G^{22(2)}. \tag{36}$$

All detailed expressions for each term in G^{-1} are provided in Lemma 10 in Section C.1.

Now we are ready to establish the upper bounds for $\|\tilde{\alpha} - \bar{\alpha}\|_a$ and $\|\tilde{\beta} - \bar{\beta}\|_a$. From the definitions of $\tilde{\alpha}$ and $\tilde{\beta}$ in (19), the difference $\tilde{\alpha} - \bar{\alpha}$ can be written as $\bar{\alpha} - \tilde{\alpha} = H^{11(1)} D_\alpha \ell_{n\lambda}(\bar{\alpha}, \bar{\beta}) + H^{11(2)} D_\alpha \ell_{n\lambda}(\bar{\alpha}, \bar{\beta}) + H^{12} D_\beta \ell_{n\lambda}(\bar{\alpha}, \bar{\beta})$, and hence it follows that

$$\|\tilde{\alpha} - \bar{\alpha}\|_a \le \|H^{11(1)} D_\alpha \ell_{n\lambda}(\bar{\alpha}, \bar{\beta})\|_a + \|H^{11(2)} D_\alpha \ell_{n\lambda}(\bar{\alpha}, \bar{\beta})\|_a + \|H^{12} D_\beta \ell_{n\lambda}(\bar{\alpha}, \bar{\beta})\|_a. \tag{37}$$

We now derive the upper bound for each term of the right-hand side in (37).

For the first term of (37), we have

$$\begin{aligned}
\mathbb{E}\|H^{11(1)} D_\alpha \ell_{n\lambda}(\bar{\alpha}, \bar{\beta})\|_a^2 &= 2^{-2}\|\bar{\beta}\|_\lambda^{-4} \mathbb{E}\|\mathrm{diag}\big((1 + \lambda\gamma_k^{-1})^{-1}\big) D_\alpha \ell_{n\lambda}(\bar{\alpha}, \bar{\beta})\|_a^2 && (38)\\
&= 2^{-2}\|\bar{\beta}\|_\lambda^{-4} \mathbb{E} \sum_{k=1}^{\infty} (1 + \gamma_k^{-a})(1 + \lambda\gamma_k^{-1})^{-2} (D_\alpha \ell_{n\lambda}(\bar{\alpha}, \bar{\beta})\omega_k)^2 && (39)\\
&\preceq n^{-1} \sum_{k=1}^{\infty} (1 + \gamma_k^{-a})(1 + \lambda\gamma_k^{-1})^{-2} && (40)\\
&\preceq n^{-1} \sum_{k=1}^{\infty} (1 + k^{2ar})(1 + \lambda k^{2r})^{-2} && (41)\\
&\asymp n^{-1} \lambda^{-(a+1/(2r))}, && (42)
\end{aligned}$$

where (38) relies on Lemma 10, (39) can be obtained by the definition of $\|\cdot\|_0$ norm, (40) comes from Lemma 11, (41) is based upon the rate assumptions on γ_k (13), and (42) holds if $4r > 2ar + 1$, which is valid so long as $r > 1/2$. The last line is obtained by replacing summation by integral approximation and the beta function.

The second term of (37) can be treated as follows

$$
\begin{aligned}
& \mathbb{E}\|H^{11(2)} D_\alpha \ell_{n\lambda}(\bar{\alpha},\bar{\beta})\|_a^2 \\
= \;& \mathbb{E}\|4^{-1}\|\bar{\alpha}\|_\lambda^{-2}\|\bar{\beta}\|_\lambda^{-2}\bar{\mathbf{a}}\bar{\mathbf{a}}^T D_\alpha \ell_{n\lambda}(\bar{\alpha},\bar{\beta})\|_a^2 && (43)\\
= \;& 4^{-2}\|\bar{\alpha}\|_\lambda^{-4}\|\bar{\beta}\|_\lambda^{-4}\|\bar{\alpha}\|_a^2 \mathbb{E}\big(\bar{\mathbf{a}}^T D_\alpha \ell_{n\lambda}(\bar{\alpha},\bar{\beta})\big)^2 = 4^{-2}\|\bar{\alpha}\|_\lambda^{-4}\|\bar{\beta}\|_\lambda^{-4}\|\bar{\alpha}\|_a^2 \mathbb{E}\left(\sum_{k=1}^{\infty}\bar{a}_k D_\alpha \ell_{n\lambda}(\bar{\alpha},\bar{\beta})\omega_k\right)^2 \\
\leq \;& 4^{-2}\|\bar{\alpha}\|_\lambda^{-4}\|\bar{\beta}\|_\lambda^{-4}\|\bar{\alpha}\|_a^2 \sum_{k=1}^{\infty}(1+\gamma_k^{-c})\bar{a}_k^2 \sum_{k=1}^{\infty}(1+\gamma_k^{-c})^{-1}\mathbb{E}(D_\alpha \ell_{n\lambda}(\bar{\alpha},\bar{\beta})\omega_k)^2 && (44)\\
\preceq \;& \|\bar{\alpha}\|_c^2 n^{-1}\sum_{k=1}^{\infty}(1+\gamma_k^{-c})^{-1} && (45)\\
\asymp \;& n^{-1} && (46)\\
= \;& o(\mathbb{E}\|H^{11(1)} D_\alpha \ell_{n\lambda}(\bar{\alpha},\bar{\beta})\|_a^2), && (47)
\end{aligned}
$$

where (43) plugs in the expression of $G^{11(2)}$ from Lemma 10, (44) is an application of Cauchy-Schwarz inequality, (45) makes use of Lemma 11, (46) always holds provided that $c > 1/2r$, and (47) demonstrates that this term is dominated by the first term of (37) when $\lambda = o(1)$.

Similarly, for the third term of (37),

$$
\begin{aligned}
\mathbb{E}\|H^{12} D_\beta \ell_{n\lambda}(\bar{\alpha},\bar{\beta})\|_a^2 &= \mathbb{E}\|4^{-1}\|\bar{\alpha}\|_\lambda^{-2}\|\bar{\beta}\|_\lambda^{-2}\bar{\mathbf{a}}\bar{\mathbf{b}}^T D_\beta \ell_{n\lambda}(\bar{\alpha},\bar{\beta})\|_a^2 \\
&= 4^{-2}\|\bar{\alpha}\|_\lambda^{-4}\|\bar{\beta}\|_\lambda^{-4}\|\bar{\alpha}\|_a^2 \mathbb{E}\big(\bar{\mathbf{b}}^T D_\beta \ell_{n\lambda}(\bar{\alpha},\bar{\beta})\big)^2 \\
&\asymp n^{-1} = o(\mathbb{E}\|H^{11(1)} D_\alpha \ell_{n\lambda}(\bar{\alpha},\bar{\beta})\|_a^2), \qquad (48)
\end{aligned}
$$

noticing the symmetry of the third last line of the display and (44).

Combining (37, 42, 47, 48), we have now proved Lemma 8 for α, the bound for β can be retrieved in parallel. ■

B.4 Proof of Lemma 9

In pursuance of $\|\widehat{\alpha}-\tilde{\alpha}\|_0^2$, $\|\widehat{\beta}-\tilde{\beta}\|_0^2$, we revisit $\tilde{\alpha}$, $\tilde{\beta}$ and $\widehat{\alpha},\widehat{\beta}$. By definition of $\tilde{\alpha},\tilde{\beta}$ defined in (19), $H\begin{pmatrix}\bar{\alpha}-\tilde{\alpha}\\ \bar{\beta}-\tilde{\beta}\end{pmatrix} = \begin{pmatrix} D_\alpha \ell_{n\lambda}(\bar{\alpha},\bar{\beta}) \\ D_\beta \ell_{n\lambda}(\bar{\alpha},\bar{\beta})\end{pmatrix}$. Plugging the definition of H in (20), we have

$$
\begin{pmatrix} D^2_{\alpha\alpha}\ell_\lambda(\bar{\alpha},\bar{\beta}) & D^2_{\alpha\beta}\ell_\lambda(\bar{\alpha},\bar{\beta}) \\ D^2_{\beta\alpha}\ell_\lambda(\bar{\alpha},\bar{\beta}) & D^2_{\beta\beta}\ell_\lambda(\bar{\alpha},\bar{\beta})\end{pmatrix}\begin{pmatrix}\bar{\alpha}-\tilde{\alpha}\\ \bar{\beta}-\tilde{\beta}\end{pmatrix} = \begin{pmatrix} D_\alpha \ell_{n\lambda}(\bar{\alpha},\bar{\beta}) \\ D_\beta \ell_{n\lambda}(\bar{\alpha},\bar{\beta})\end{pmatrix},
$$

which is equivalent to

$$
\begin{aligned}
D^2_{\alpha\alpha}\ell_\lambda(\bar{\alpha},\bar{\beta})(\bar{\alpha}-\tilde{\alpha}) + D^2_{\alpha\beta}\ell_\lambda(\bar{\alpha},\bar{\beta})(\bar{\beta}-\tilde{\beta}) &= D_\alpha \ell_{n\lambda}(\bar{\alpha},\bar{\beta}), \\
D^2_{\beta\alpha}\ell_\lambda(\bar{\alpha},\bar{\beta})(\bar{\alpha}-\tilde{\alpha}) + D^2_{\beta\beta}\ell_\lambda(\bar{\alpha},\bar{\beta})(\bar{\beta}-\tilde{\beta}) &= D_\beta \ell_{n\lambda}(\bar{\alpha},\bar{\beta}).
\end{aligned} \qquad (49)
$$

Taylor expansion of $D_\alpha \ell_{n\lambda}(\widehat{\alpha},\widehat{\beta})$, $D_\beta \ell_{n\lambda}(\widehat{\alpha},\widehat{\beta})$ around $D_\alpha \ell_{n\lambda}(\bar{\alpha},\bar{\beta})$, $D_\beta \ell_{n\lambda}(\bar{\alpha},\bar{\beta})$ implies that

$$
\begin{aligned}
0 = D_\alpha \ell_{n\lambda}(\widehat{\alpha},\widehat{\beta}) &= D_\alpha \ell_{n\lambda}(\bar{\alpha},\bar{\beta}) + D^2_{\beta\alpha}\ell_{n\lambda}(\bar{\alpha},\bar{\beta})(\widehat{\beta}-\bar{\beta}) + D^2_{\alpha\alpha}\ell_{n\lambda}(\bar{\alpha},\bar{\beta})(\widehat{\alpha}-\bar{\alpha}) + R_\alpha, \\
0 = D_\beta \ell_{n\lambda}(\widehat{\alpha},\widehat{\beta}) &= D_\beta \ell_{n\lambda}(\bar{\alpha},\bar{\beta}) + D^2_{\beta\alpha}\ell_{n\lambda}(\bar{\alpha},\bar{\beta})(\widehat{\alpha}-\bar{\alpha}) + D^2_{\beta\beta}\ell_{n\lambda}(\bar{\alpha},\bar{\beta})(\widehat{\beta}-\bar{\beta}) + R_\beta,
\end{aligned} \qquad (50)
$$

where the higher order residual terms are

$$
\begin{aligned}
R_\alpha &= \frac{1}{2}D^3_{\alpha\beta\beta}\ell_{n\lambda}(\bar{\alpha},\bar{\beta})(\widehat{\beta}-\bar{\beta})^2 + D^3_{\alpha\alpha\beta}\ell_{n\lambda}(\bar{\alpha},\bar{\beta})(\widehat{\alpha}-\bar{\alpha})(\widehat{\beta}-\bar{\beta}) + \frac{1}{2}D^4_{\alpha\alpha\beta\beta}\ell_{n\lambda}(\bar{\alpha},\bar{\beta})(\widehat{\alpha}-\bar{\alpha})(\widehat{\beta}-\bar{\beta})^2, \\
R_\beta &= \frac{1}{2}D^3_{\alpha\alpha\beta}\ell_{n\lambda}(\bar{\alpha},\bar{\beta})(\widehat{\alpha}-\bar{\alpha})^2 + D^3_{\alpha\beta\beta}\ell_{n\lambda}(\bar{\alpha},\bar{\beta})(\widehat{\alpha}-\bar{\alpha})(\widehat{\beta}-\bar{\beta}) + \frac{1}{2}D^4_{\alpha\alpha\beta\beta}\ell_{n\lambda}(\bar{\alpha},\bar{\beta})(\widehat{\alpha}-\bar{\alpha})^2(\widehat{\beta}-\bar{\beta}).
\end{aligned}
$$

Integrating (49,50), we arrive at

$$\begin{aligned}
& D^2_{\alpha\alpha}\ell_\lambda(\bar\alpha,\bar\beta)(\widehat\alpha-\tilde\alpha)+D^2_{\alpha\beta}\ell_\lambda(\bar\alpha,\bar\beta)(\widehat\beta-\tilde\beta)\\
= \;& D^2_{\alpha\alpha}\ell_\lambda(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)+D^2_{\alpha\beta}\ell_\lambda(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)+D^2_{\alpha\alpha}\ell_\lambda(\bar\alpha,\bar\beta)(\bar\alpha-\tilde\alpha)+D^2_{\alpha\beta}\ell_\lambda(\bar\alpha,\bar\beta)(\bar\beta-\tilde\beta)\\
= \;& D^2_{\alpha\alpha}\ell_\lambda(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)+D^2_{\alpha\beta}\ell_\lambda(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)-D^2_{\alpha\alpha}\ell_{n\lambda}(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)-D^2_{\alpha\beta}\ell_{n\lambda}(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)-R_\alpha\\
= \;& D^2_{\alpha\alpha}\ell(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)+D^2_{\alpha\beta}\ell(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)-D^2_{\alpha\alpha}\ell_n(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)-D^2_{\alpha\beta}\ell_n(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)-R_\alpha,\\
& D^2_{\beta\alpha}\ell_\lambda(\bar\alpha,\bar\beta)(\widehat\alpha-\tilde\alpha)+D^2_{\beta\beta}\ell_\lambda(\bar\alpha,\bar\beta)(\widehat\beta-\tilde\beta)\\
= \;& D^2_{\beta\alpha}\ell(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)+D^2_{\beta\beta}\ell(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)-D^2_{\beta\alpha}\ell_n(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)-D^2_{\beta\beta}\ell_n(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)-R_\beta,
\end{aligned}$$

which equates to the following given the definition of H in (20)

$$\begin{aligned}
H\begin{pmatrix}\widehat\alpha-\tilde\alpha\\ \widehat\beta-\tilde\beta\end{pmatrix} &= \begin{pmatrix} D^2_{\alpha\alpha}\ell(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)-D^2_{\alpha\alpha}\ell_n(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)\\ D^2_{\beta\alpha}\ell(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)-D^2_{\beta\alpha}\ell_n(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)\end{pmatrix}\\
&\quad+\begin{pmatrix} D^2_{\alpha\beta}\ell(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)-D^2_{\alpha\beta}\ell_n(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)\\ D^2_{\beta\beta}\ell(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)-D^2_{\beta\beta}\ell_n(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)\end{pmatrix}-\begin{pmatrix}R_\alpha\\ R_\beta\end{pmatrix}.
\end{aligned}$$

Multiplying both sides of the last display by H^{-1} leads to

$$\begin{aligned}
\widehat\alpha-\tilde\alpha &= H^{11(1)}\big(D^2_{\alpha\alpha}\ell(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)-D^2_{\alpha\alpha}\ell_n(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)\big)\\
&+ H^{11(2)}\big(D^2_{\alpha\alpha}\ell(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)-D^2_{\alpha\alpha}\ell_n(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)\big)\\
&+ H^{12}\big(D^2_{\beta\alpha}\ell(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)-D^2_{\beta\alpha}\ell_n(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)\big)\\
&+ H^{11(1)}\big(D^2_{\alpha\beta}\ell(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)-D^2_{\alpha\beta}\ell_n(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)\big)\\
&+ H^{11(2)}\big(D^2_{\alpha\beta}\ell(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)-D^2_{\alpha\beta}\ell_n(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)\big)\\
&+ H^{12}\big(D^2_{\beta\beta}\ell(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)-D^2_{\beta\beta}\ell_n(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)\big)\\
&- H^{11}\big(R_\alpha\big)-H^{12}\big(R_\beta\big).
\end{aligned}\tag{51}$$

To study the bound of $\|\widehat\alpha-\tilde\alpha\|_a$, one only needs to analyze the $\|\cdot\|_a$ norm of the eight terms in (51) separately, among which the first six terms can be bounded by Lemmas 14 and 15 and the last two terms are of smaller order. Therefore, $\|\widehat\alpha-\tilde\alpha\|_a^2=O(n^{-1}\lambda^{-(a+1/(2r))}\|\widehat\alpha-\bar\alpha\|_c^2+n^{-1}\lambda^{-(a+1/(2r))}\|\widehat\beta-\bar\beta\|_c^2)$. In particular, we obtain the following when letting $a=c$, $\|\widehat\alpha-\tilde\alpha\|_c^2=O(n^{-1}\lambda^{-(c+1/(2r))}\|\widehat\alpha-\bar\alpha\|_c^2+n^{-1}\lambda^{-(c+1/(2r))}\|\widehat\beta-\bar\beta\|_c^2)$. Under the condition $n^{-1}\lambda^{-(c+1/(2r))}=o(1)$, applying the triangle inequality yields $\|\tilde\alpha-\bar\alpha\|_c^2\geq\|\widehat\alpha-\bar\alpha\|_c^2-\|\widehat\alpha-\tilde\alpha\|_c^2=\big(1-o(1)\big)\|\widehat\alpha-\bar\alpha\|_c^2-o(1)\|\widehat\beta-\bar\beta\|_c^2$. Hence, $\|\widehat\alpha-\bar\alpha\|_c^2=O(\|\tilde\alpha-\bar\alpha\|_c^2+\|\tilde\beta-\bar\beta\|_c^2)$, which implies $\|\widehat\alpha-\tilde\alpha\|_c^2=O(n^{-1}\lambda^{-(c+1/(2r))}\|\tilde\alpha-\bar\alpha\|_c^2+n^{-1}\lambda^{-(c+1/(2r))}\|\tilde\beta-\bar\beta\|_c^2)$. Together with Lemma 8 and a parallel argument for β, completes the proof of Lemma 9. ■

Appendix C. Auxiliary Lemmas and Their Proofs

C.1 Auxiliary Lemmas

Lemma 10 *(Expression of G^{-1}) Suppose G^{-1} is decomposed as in* (35) *and* (36)*. Then it adopts the following form*

$$\begin{aligned}
G^{11(1)} &= 2^{-1}\|\bar\beta\|_\lambda^{-2}\mathrm{diag}\big((1+\lambda\gamma_k^{-1})^{-1}\big),\\
G^{11(2)} &= -4^{-1}\|\bar\alpha\|_\lambda^{-2}\|\bar\beta\|_\lambda^{-2}\bar{\mathbf{a}}\bar{\mathbf{a}}^T,\\
G^{22(1)} &= 2^{-1}\|\bar\alpha\|_\lambda^{-2}\mathrm{diag}\big((1+\lambda\gamma_k^{-1})^{-1}\big),\\
G^{22(2)} &= -4^{-1}\|\bar\alpha\|_\lambda^{-2}\|\bar\beta\|_\lambda^{-2}\bar{\mathbf{b}}\bar{\mathbf{b}}^T,\\
G^{12} &= -4^{-1}\|\bar\alpha\|_\lambda^{-2}\|\bar\beta\|_\lambda^{-2}\bar{\mathbf{a}}\bar{\mathbf{b}}^T,\\
G^{21} &= -4^{-1}\|\bar\alpha\|_\lambda^{-2}\|\bar\beta\|_\lambda^{-2}\bar{\mathbf{b}}\bar{\mathbf{a}}^T.
\end{aligned}$$

Lemma 11 *(Properties of first order operators) The first order operators have the following properties, for* $k = 1, 2, \ldots,$

$$\mathbb{E}\left(D_\alpha \ell_{n\lambda}(\bar{\alpha}, \bar{\beta})\omega_k\right)^2 = O(n^{-1}), \quad \mathbb{E}\left(D_\beta \ell_{n\lambda}(\bar{\alpha}, \bar{\beta})\omega_k\right)^2 = O(n^{-1}).$$

Lemma 12 *For any* $f_i \in \mathcal{L}_2(\mathcal{T})$, $i = 1, \ldots, 4$, *the following inequality holds,*

$$\mathbb{E}\left(\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t) f_1(s) f_2(t)\ dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t) f_3(s) f_4(t)\ dsdt\right)\right)^2 \leq M\|f_1\|_0^2\|f_2\|_0^2\|f_3\|_0^2\|f_4\|_0^2,$$

where the constant M *is defined in* (14).

Lemma 13 *(Property of second order derivative) The second order derivative operator can be bounded by*

$$\begin{aligned}
E(D^2_{\alpha\alpha}\ell(\bar{\alpha}, \bar{\beta})\omega_j\omega_k - D^2_{\alpha\alpha}\ell_n(\bar{\alpha}, \bar{\beta})\omega_j\omega_k)^2 &\asymp n^{-1},\\
E(D^2_{\beta\beta}\ell(\bar{\alpha}, \bar{\beta})\omega_j\omega_k - D^2_{\beta\beta}\ell_n(\bar{\alpha}, \bar{\beta})\omega_j\omega_k)^2 &\asymp n^{-1},\\
E(D^2_{\beta\alpha}\ell(\bar{\alpha}, \bar{\beta})\omega_j\omega_k - D^2_{\beta\alpha}\ell_n(\bar{\alpha}, \bar{\beta})\omega_j\omega_k)^2 &\asymp n^{-1},\\
E(D^2_{\alpha\beta}\ell(\bar{\alpha}, \bar{\beta})\omega_j\omega_k - D^2_{\alpha\beta}\ell_n(\bar{\alpha}, \bar{\beta})\omega_j\omega_k)^2 &\asymp n^{-1}.
\end{aligned}$$

Lemma 14 *The two leading terms in (51) can be bounded by*

$$\begin{aligned}
\|G^{11(1)}\big(D^2_{\alpha\alpha}\ell(\bar{\alpha}, \bar{\beta})(\widehat{\alpha} - \bar{\alpha}) - D^2_{\alpha\alpha}\ell_n(\bar{\alpha}, \bar{\beta})(\widehat{\alpha} - \bar{\alpha})\big)\|_a^2 &\asymp n^{-1}\lambda^{-(a+1/(2r))}\|\widehat{\alpha} - \bar{\alpha}\|_c^2,\\
\|G^{11(1)}\big(D^2_{\alpha\beta}\ell(\bar{\alpha}, \bar{\beta})(\widehat{\beta} - \bar{\beta}) - D^2_{\alpha\beta}\ell_n(\bar{\alpha}, \bar{\beta})(\widehat{\beta} - \bar{\beta})\big)\|_a^2 &\asymp n^{-1}\lambda^{-(a+1/(2r))}\|\widehat{\beta} - \bar{\beta}\|_c^2.
\end{aligned}$$

Lemma 15 *The four non-leading terms in (51) can be bounded by*

$$\begin{aligned}
\|G^{12}\big(D^2_{\beta\alpha}\ell(\bar{\alpha}, \bar{\beta})(\widehat{\alpha} - \bar{\alpha}) - D^2_{\beta\alpha}\ell_n(\bar{\alpha}, \bar{\beta})(\widehat{\alpha} - \bar{\alpha})\big)\|_a^2 &\asymp n^{-1}\|\widehat{\alpha} - \bar{\alpha}\|_c^2,\\
\|G^{12}\big(D^2_{\beta\beta}\ell(\bar{\alpha}, \bar{\beta})(\widehat{\beta} - \bar{\beta}) - D^2_{\beta\beta}\ell_n(\bar{\alpha}, \bar{\beta})(\widehat{\beta} - \bar{\beta})\big)\|_a^2 &\asymp n^{-1}\|\widehat{\beta} - \bar{\beta}\|_c^2,\\
\|G^{11(2)}\big(D^2_{\alpha\alpha}\ell(\bar{\alpha}, \bar{\beta})(\widehat{\alpha} - \bar{\alpha}) - D^2_{\alpha\alpha}\ell_n(\bar{\alpha}, \bar{\beta})(\widehat{\alpha} - \bar{\alpha})\big)\|_a^2 &\asymp n^{-1}\|\widehat{\alpha} - \bar{\alpha}\|_c^2,\\
\|G^{11(2)}\big(D^2_{\alpha\beta}\ell(\bar{\alpha}, \bar{\beta})(\widehat{\beta} - \bar{\beta}) - D^2_{\alpha\beta}\ell_n(\bar{\alpha}, \bar{\beta})(\widehat{\beta} - \bar{\beta})\big)\|_a^2 &\asymp n^{-1}\|\widehat{\beta} - \bar{\beta}\|_c^2.
\end{aligned}$$

Lemma 16 *(Theorem 2.5 of Tsybakov (2009)) Assume that* $M \geq 2$ *and suppose that the parameter space* Θ *contains* $\theta_0, \theta_1, ..., \theta_M$ *such that (i)* $d(\theta_j, \theta_k) \geq 2s > 0, \forall 0 \leq j < k \leq M$, *(ii)* $\forall j = 1, ..., M$

$$\frac{1}{M}\sum_{j=1}^{M} KL(P_{\theta_j}, P_{\theta_0}) \leq \delta \log M,$$

with $0 < \delta < 1/8$. *Then*

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} P_\theta(d(\hat{\theta}, \theta) \geq s) \geq \frac{\sqrt{M}}{1 + \sqrt{M}}\left(1 - 2\delta - \sqrt{\frac{2\delta}{\log M}}\right) > 0.$$

The readers are referred to Tsybakov (2009) for the proof of the lemma.

C.2 Proofs of Auxiliary Lemmas

Proof of Lemma 10

Recall the penalty function $J(\alpha, \beta) = \lambda\|\alpha\|_0^2\|\beta\|_K^2 + \lambda\|\beta\|_0^2\|\alpha\|_K^2 + \lambda^2\|\alpha\|_K^2\|\beta\|_K^2$. The operators related to the penalty function can be defined as follows,

$$\begin{aligned}
D^2_{\alpha\alpha}J(\alpha, \beta)fg &= 2\lambda\|\beta\|_0^2\langle f, g\rangle_K + 2\lambda\|\beta\|_K^2\langle Cf, g\rangle_{\mathcal{L}^2} + 2\lambda^2\|\beta\|_K^2\langle f, g\rangle_K,\\
D^2_{\beta\beta}J(\alpha, \beta)fg &= 2\lambda\|\alpha\|_0^2\langle f, g\rangle_K + 2\lambda\|\alpha\|_K^2\langle Cf, g\rangle_{\mathcal{L}^2} + 2\lambda^2\|\alpha\|_K^2\langle f, g\rangle_K,\\
D^2_{\alpha\beta}J(\alpha, \beta)fg &= 4\lambda\langle C\alpha, f\rangle_{\mathcal{L}^2}\langle\beta, g\rangle_K + 4\lambda\langle C\beta, g\rangle_{\mathcal{L}^2}\langle\alpha, f\rangle_K + 4\lambda^2\langle\alpha, f\rangle_K\langle\beta, g\rangle_K.
\end{aligned}$$

Therefore, $(G_{11})_{jk} = D^2_{\alpha\alpha}\ell_\lambda(\bar\alpha,\bar\beta)\omega_j\omega_k = 2\|\bar\beta\|^2_\lambda(1+\lambda\gamma_k^{-1})\delta_{jk}$, and similarly, $(G_{22})_{jk} = D^2_{\beta\beta}\ell_\lambda(\bar\alpha,\bar\beta)\omega_j\omega_k = 2\|\bar\alpha\|^2_\lambda(1+\lambda\gamma_k^{-1})\delta_{jk}$, where the definition of the $\|\cdot\|_\lambda$ norm is given in (31).

For the off diagonal blocks G_{12} and G_{21}, recall the expansions of α_0, $\bar\alpha$, β_0, and $\bar\beta$ in (27) and the coefficients of $\bar\alpha$ and $\bar\beta$ in (29), we get $(G_{12})_{jk} = D^2_{\alpha\beta}\ell_\lambda(\bar\alpha,\bar\beta)\omega_j\omega_k = -2a_{0j}b_{0k} + 4\bar a_j\bar b_k + 4\lambda\gamma_k^{-1}\bar a_j\bar b_k + 4\lambda\gamma_j^{-1}\bar a_j\bar b_k + 4\lambda^2\gamma_j^{-1}\gamma_k^{-1}\bar a_j\bar b_k = 2a_{0j}b_{0k}$, and $(G_{21})_{jk} = D^2_{\beta\alpha}\ell_\lambda(\bar\alpha,\bar\beta)\omega_j\omega_k = 2a_{0k}b_{0j}$. Put the above results in matrix form, it is clear that

$$\begin{array}{ll} G_{11} = 2\|\bar\beta\|^2_\lambda \mathrm{diag}\big((1+\lambda\gamma_k^{-1})\big), & G_{12} = 2\mathbf{a}_0\mathbf{b}_0^T, \\ G_{22} = 2\|\bar\alpha\|^2_\lambda \mathrm{diag}\big((1+\lambda\gamma_k^{-1})\big), & G_{21} = 2\mathbf{b}_0\mathbf{a}_0^T. \end{array}$$

The blocks in G^{-1} can be computed from the block matrix inversion formula as follows,

$$\begin{array}{ll} G^{11} = (G_{11} - G_{12}G_{22}^{-1}G_{21})^{-1}, & G^{12} = -G_{11}^{-1}G_{12}G^{22}, \\ G^{22} = (G_{22} - G_{21}G_{11}^{-1}G_{12})^{-1}, & G^{21} = -G_{22}^{-1}G_{21}G^{11}, \end{array}$$

which will be resolved one by one. We begin with the $G_{12}G_{22}^{-1}G_{21}$ term,

$$\begin{aligned} G_{12}G_{22}^{-1}G_{21} &= (2\mathbf{a}_0\mathbf{b}_0^T)(2^{-1}\|\bar\alpha\|_\lambda^{-2}\mathrm{diag}\big((1+\lambda\gamma_k^{-1})^{-1}\big)(2\mathbf{b}_0\mathbf{a}_0^T) \\ &= 2\|\bar\alpha\|_\lambda^{-2}\big(\mathbf{b}_0^T\mathrm{diag}\big((1+\lambda\gamma_k^{-1})^{-1}\big)\mathbf{b}_0\big)\mathbf{a}_0\mathbf{a}_0^T \\ &= 2\|\bar\alpha\|_\lambda^{-2}\|\bar\beta\|_\lambda^2\mathbf{a}_0\mathbf{a}_0^T, \end{aligned}$$

and hence,

$$G_{11} - G_{12}G_{22}^{-1}G_{21} = 2\|\bar\beta\|^2_\lambda\mathrm{diag}\big((1+\lambda\gamma_k^{-1})\big) - 2\|\bar\alpha\|_\lambda^{-2}\|\bar\beta\|^2_\lambda\mathbf{a}_0\mathbf{a}_0^T.$$

From the Woodbury matrix inversion identity, it follows that

$$G^{11} = (G_{11} - G_{12}G_{22}^{-1}G_{21})^{-1} = 2^{-1}\|\bar\beta\|_\lambda^{-2}\mathrm{diag}\big((1+\lambda\gamma_k^{-1})^{-1}\big) - 4^{-1}\|\bar\alpha\|_\lambda^{-2}\|\bar\beta\|_\lambda^{-2}\bar{\mathbf{a}}\bar{\mathbf{a}}^T.$$

The first term on the right-hand side is defined as $G^{11(1)}$ and the second one as $G^{11(2)}$.

The G^{22} term can be calculated in a similar fashion and we have

$$G^{22} = 2^{-1}\|\bar\alpha\|_\lambda^{-2}\mathrm{diag}\big((1+\lambda\gamma_k^{-1})^{-1}\big) - 4^{-1}\|\bar\alpha\|_\lambda^{-2}\|\bar\beta\|_\lambda^{-2}\bar{\mathbf{b}}\bar{\mathbf{b}}^T.$$

Similarly, the $G^{22(1)}$ and $G^{22(2)}$ terms are defined as the first and the second term on the right-hand side of the above equation. As for the G^{12} term, it follows that

$$\begin{aligned} G^{12} &= -G_{11}^{-1}G_{12}G^{22} \\ &= -2^{-1}\|\bar\alpha\|_\lambda^{-2}\|\bar\beta\|_\lambda^{-2}\mathrm{diag}\big((1+\lambda\gamma_k^{-1})^{-1}\big)\mathbf{a}_0\mathbf{b}_0^T\left(\mathrm{diag}\big((1+\lambda\gamma_k^{-1})^{-1}\big) - 2^{-1}\|\bar\beta\|_\lambda^{-2}\bar{\mathbf{b}}\bar{\mathbf{b}}^T\right) \\ &= -4^{-1}\|\bar\alpha\|_\lambda^{-2}\|\bar\beta\|_\lambda^{-2}\bar{\mathbf{a}}\bar{\mathbf{b}}^T, \end{aligned}$$

and similarly, $G^{21} = -4^{-1}\|\bar\alpha\|_\lambda^{-2}\|\bar\beta\|_\lambda^{-2}\bar{\mathbf{b}}\bar{\mathbf{a}}^T$. The proof of Lemma 10 is complete. ■

Proof of Lemma 11

Since $(\bar\alpha,\bar\beta)$ is the minimizer of $\ell_\lambda(\alpha,\beta)$, the stationary condition ensures that $D_\beta\ell_\lambda(\bar\alpha,\bar\beta) = 0$, and hence $D_\beta\ell_{n\lambda}(\bar\alpha,\bar\beta) = D_\beta\ell_{n\lambda}(\bar\alpha,\bar\beta) - D_\beta\ell_\lambda(\bar\alpha,\bar\beta) = D_\beta\ell_n(\bar\alpha,\bar\beta) - D_\beta\ell(\bar\alpha,\bar\beta)$. Therefore, for any positive integer k, we have

$$\begin{aligned} &\mathbb{E}(D_\beta\ell_{n\lambda}(\bar\alpha,\bar\beta)\omega_k)^2 = \mathbb{E}(D_\beta\ell_n(\bar\alpha,\bar\beta)\omega_k - D_\beta\ell(\bar\alpha,\bar\beta)\omega_k)^2 \\ =\ & \frac{4}{n}\mathrm{var}\left(\left(Y - \int_{\mathcal{T}\times\mathcal{T}} X(s,t)\bar\alpha(s)\bar\beta(t)\,dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\bar\alpha(s)\omega_k(t)\,dsdt\right)\right) \\ \le\ & \frac{4}{n}\mathbb{E}\left(\left(Y - \int_{\mathcal{T}\times\mathcal{T}} X(s,t)\bar\alpha(s)\bar\beta(t)\,dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\bar\alpha(s)\omega_k(t)\,dsdt\right)\right)^2 \\ =\ & \frac{4\sigma^2}{n}\mathbb{E}\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\bar\alpha(s)\omega_k(t)\,dsdt\right)^2 \\ +\ & \frac{4}{n}\mathbb{E}\left(\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\big(\bar\alpha(s)\bar\beta(t) - \alpha_0(s)\beta_0(t)\big)\,dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\bar\alpha(s)\omega_k(t)\,dsdt\right)\right)^2, \end{aligned}$$

For the first term, since $\|\omega_k\|_0^2 = 1$, we have

$$\mathbb{E}\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\bar{\alpha}(s)\omega_k(t)\ dsdt\right)^2 = \|\bar{\alpha}\|_0^2\|\omega_k\|_0^2 = \|\bar{\alpha}\|_0^2. \tag{52}$$

For the second term, by the Cauchy-Schwarz inequality, it follows that

$$\begin{aligned}
&\mathbb{E}\left(\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\big(\bar{\alpha}(s)\bar{\beta}(t) - \alpha_0(s)\beta_0(t)\big)\ dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\bar{\alpha}(s)\omega_k(t)\ dsdt\right)\right)^2 \\
\leq\ & \left(\mathbb{E}\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\big(\bar{\alpha}(s)\bar{\beta}(t) - \alpha_0(s)\beta_0(t)\big)\ dsdt\right)^4 \mathbb{E}\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\bar{\alpha}(s)\omega_k(t)\ dsdt\right)^4\right)^{1/2} \\
\leq\ & M\mathbb{E}\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\big(\bar{\alpha}(s)\bar{\beta}(t) - \alpha_0(s)\beta_0(t)\big)\ dsdt\right)^2 \mathbb{E}\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\bar{\alpha}(s)\omega_k(t)\ dsdt\right)^2 \\
=\ & M\|\bar{\alpha}\|_0^2\mathbb{E}\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\big(\bar{\alpha}(s)\bar{\beta}(t) - \alpha_0(s)\beta_0(t)\big)\ dsdt\right)^2,
\end{aligned} \tag{53}$$

where the second inequality uses the fourth moment condition (14). It is easy to see $\mathbb{E}\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\big(\bar{\alpha}(s)\bar{\beta}(t) - \alpha_0(s)\beta_0(t)\big)\ dsdt\right)^2 \leq 3\|\bar{\alpha}-\alpha_0\|_0^2\|\bar{\beta}-\beta_0\|_0^2 + 3\|\bar{\alpha}-\alpha_0\|_0^2\|\beta_0\|_0^2 + 3\|\alpha_0\|_0^2\|\bar{\beta}-\beta_0\|_0^2$. By Lemma 7, it is clear that

$$\mathbb{E}\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\big(\bar{\alpha}(s)\bar{\beta}(t) - \alpha_0(s)\beta_0(t)\big)\ dsdt\right)^2 = O(\lambda). \tag{54}$$

Now (52), (53) and (54) together yield $\mathbb{E}(D_\beta \ell_{n\lambda}(\bar{\alpha},\bar{\beta})\omega_k)^2 \leq \frac{4\sigma^2}{n}\|\bar{\alpha}\|_0^2 + \frac{4}{n}M\|\bar{\alpha}\|_0^2 O(\lambda) = O(n^{-1})$. An identical argument proves that $\mathbb{E}(D_\alpha \ell_{n\lambda}(\bar{\alpha},\bar{\beta})\omega_k)^2 = O(n^{-1})$. ■

Proof of Lemma 12

By Cauchy-Schwarz inequality and the fourth moment condition (14), it follows that,

$$\begin{aligned}
&\mathbb{E}\left(\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)f_1(s)f_2(t)\ dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)f_3(s)f_4(t)\ dsdt\right)\right)^2 \\
\leq\ & \left(\mathbb{E}\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)f_1(s)f_2(t)\ dsdt\right)^4 \mathbb{E}\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)f_3(s)f_4(t)\ dsdt\right)^4\right)^{1/2} \\
\leq\ & M\mathbb{E}\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)f_1(s)f_2(t)\ dsdt\right)^2 \mathbb{E}\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)f_3(s)f_4(t)\ dsdt\right)^2 \\
=\ & M\|f_1\|_0^2\|f_2\|_0^2\|f_3\|_0^2\|f_4\|_0^2.
\end{aligned}$$

■

Proof of Lemma 13

Since $D^2_{\alpha\alpha}\ell(\bar{\alpha},\bar{\beta}) = ED^2_{\alpha\alpha}\ell_n(\bar{\alpha},\bar{\beta})$,

$$\begin{aligned}
&\mathbb{E}(D^2_{\alpha\alpha}\ell(\bar{\alpha},\bar{\beta})\omega_j\omega_k - D^2_{\alpha\alpha}\ell_n(\bar{\alpha},\bar{\beta})\omega_j\omega_k)^2 \\
=\ & \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \left(\int_{\mathcal{T}\times\mathcal{T}} x_i(s,t)\omega_j(s)\bar{\beta}(t)\ dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}} x_i(s,t)\omega_k(s)\bar{\beta}(t)\ dsdt\right)\right. \\
& \left. - \|\bar{\beta}\|_0^2\int_{\mathcal{T}\times\mathcal{T}} C(s,t)\omega_j(s)\omega_k(t)\ dsdt\right)^2 \\
=\ & \frac{1}{n}\mathrm{var}\left(\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\omega_j(s)\bar{\beta}(t)\ dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\omega_k(s)\bar{\beta}(t)\ dsdt\right)\right) \\
\leq\ & \frac{1}{n}\mathbb{E}\left(\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\omega_j(s)\bar{\beta}(t)\ dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}} X(s,t)\omega_k(s)\bar{\beta}(t)\ dsdt\right)\right)^2 \leq \frac{M}{n}\|\bar{\beta}\|_0^4 \asymp O(n^{-1}),
\end{aligned}$$

where the second inequality is a simple application of Lemma 12. $E(D^2_{\beta\beta}\ell(\bar\alpha,\bar\beta)\omega_j\omega_k - D^2_{\beta\beta}\ell_n(\bar\alpha,\bar\beta)\omega_j\omega_k)^2 \asymp n^{-1}$ can be proved likewise. Notice that $D^2_{\beta\alpha}\ell(\bar\alpha,\bar\beta) = ED^2_{\beta\alpha}\ell_n(\bar\alpha,\bar\beta)$,

$$\begin{aligned}
&\mathbb{E}(D^2_{\beta\alpha}\ell(\bar\alpha,\bar\beta)\omega_j\omega_k - D^2_{\beta\alpha}\ell_n(\bar\alpha,\bar\beta)\omega_j\omega_k)^2\\
=\;& \frac{4}{n}\mathrm{var}\left(-\left(Y-\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\bar\alpha(s)\bar\beta(t)\,dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\omega_j(s)\omega_k(t)\,dsdt\right)\right.\\
&\left.+\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\omega_j(s)\bar\beta(t)\,dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\bar\alpha(s)\omega_k(t)\,dsdt\right)\right)\\
\le\;& \frac{4}{n}\mathbb{E}\left(-\left(Y-\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\bar\alpha(s)\bar\beta(t)\,dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\omega_j(s)\omega_k(t)\,dsdt\right)\right.\\
&\left.+\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\omega_j(s)\bar\beta(t)\,dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\bar\alpha(s)\omega_k(t)\,dsdt\right)\right)^2.
\end{aligned}$$

Plugging in the expression of Y and applying Cauchy-Schwarz inequality one more time,

$$\begin{aligned}
&E(D^2_{\beta\alpha}\ell(\bar\alpha,\bar\beta)\omega_j\omega_k - D^2_{\beta\alpha}\ell_n(\bar\alpha,\bar\beta)\omega_j\omega_k)^2\\
\le\;& \frac{4}{n}E\left(-\epsilon\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\omega_j(s)\omega_k(t)\,dsdt\right)\right.\\
&-\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\alpha_0(s)\beta_0(t)\,dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\omega_j(s)\omega_k(t)\,dsdt\right)\\
&+\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\bar\alpha(s)\bar\beta(t)\,dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\omega_j(s)\omega_k(t)\,dsdt\right)\\
&\left.+\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\omega_j(s)\bar\beta(t)\,dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\bar\alpha(s)\omega_k(t)\,dsdt\right)\right)^2\\
\le\;& \frac{16}{n}\left\{E\left(\epsilon\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\omega_j(s)\omega_k(t)\,dsdt\right)\right)^2\right.\\
&+E\left(\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\alpha_0(s)\beta_0(t)\,dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\omega_j(s)\omega_k(t)\,dsdt\right)\right)^2\\
&+E\left(\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\bar\alpha(s)\bar\beta(t)\,dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\omega_j(s)\omega_k(t)\,dsdt\right)\right)^2\\
&\left.+E\left(\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\omega_j(s)\bar\beta(t)\,dsdt\right)\left(\int_{\mathcal{T}\times\mathcal{T}}X(s,t)\bar\alpha(s)\omega_k(t)\,dsdt\right)\right)^2\right\}\\
\le\;& \frac{16}{n}(\sigma^2 + M\|\alpha_0\|_0^2\|\beta_0\|_0^2 + M\|\bar\alpha\|_0^2\|\bar\beta\|_0^2 + M\|\bar\alpha\|_0^2\|\bar\beta\|_0^2) \asymp n^{-1},
\end{aligned}$$

where the last inequality invokes Lemma 12 three times. $E(D^2_{\alpha\beta}\ell(\bar\alpha,\bar\beta)\omega_j\omega_k - D^2_{\alpha\beta}\ell_n(\bar\alpha,\bar\beta)\omega_j\omega_k)^2 \asymp n^{-1}$ can be proved analogously. ■

Proof of Lemma 14

Recall the expansion (27), we additionally write the expansion of $\widehat\alpha, \widehat\beta$ as

$$\widehat\alpha = \sum_{k=1}^{\infty}\hat a_k\omega_k, \quad \widehat\beta = \sum_{k=1}^{\infty}\hat b_k\omega_k. \tag{55}$$

To bound the first term in (51), recall the expression of the $G^{11(1)}$ in Lemma 10 and note the definition of the $\|\cdot\|_a$ norm

$$\begin{aligned}
&\|G^{11(1)}\big(D^2_{\alpha\alpha}\ell(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha) - D^2_{\alpha\alpha}\ell_n(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)\big)\|_a^2\\
=\;& 4^{-1}\|\bar\beta\|_\lambda^{-4}\sum_{k=1}^{\infty}(1+\gamma_k^{-a})(1+\lambda\gamma_k^{-1})^{-2}(D^2_{\alpha\alpha}\ell(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)\omega_k - D^2_{\alpha\alpha}\ell_n(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)\omega_k)^2.
\end{aligned}$$

Plugging in the expansion of functions $\widehat{\alpha}, \bar{\alpha}$ in (27) and (55), we get

$$\begin{aligned}
&\|G^{11(1)}\big(D^2_{\alpha\alpha}\ell(\bar{\alpha},\bar{\beta})(\widehat{\alpha}-\bar{\alpha}) - D^2_{\alpha\alpha}\ell_n(\bar{\alpha},\bar{\beta})(\widehat{\alpha}-\bar{\alpha})\big)\|_a^2 \\
= \;& 4^{-1}\|\bar{\beta}\|_\lambda^{-4}\sum_{k=1}^{\infty}(1+\gamma_k^{-a})(1+\lambda\gamma_k^{-1})^{-2}\left\{\sum_{j=1}^{\infty}(\hat{a}_j-\bar{a}_j)(D^2_{\alpha\alpha}\ell(\bar{\alpha},\bar{\beta})\omega_j\omega_k - D^2_{\alpha\alpha}\ell_n(\bar{\alpha},\bar{\beta})\omega_j\omega_k)\right\}^2.
\end{aligned}$$

Cauchy-Schwarz inequality produces

$$\begin{aligned}
&\|G^{11(1)}\big(D^2_{\alpha\alpha}\ell(\bar{\alpha},\bar{\beta})(\widehat{\alpha}-\bar{\alpha}) - D^2_{\alpha\alpha}\ell_n(\bar{\alpha},\bar{\beta})(\widehat{\alpha}-\bar{\alpha})\big)\|_a^2 \\
\leq \;& 4^{-1}\|\bar{\beta}\|_\lambda^{-4}\sum_{k=1}^{\infty}(1+\gamma_k^{-c})(1+\lambda\gamma_k^{-1})^{-2}\left\{\sum_{j=1}^{\infty}(1+\gamma_j^{-a})(\hat{a}_j-\bar{a}_j)^2\right\} \\
&\left\{\sum_{j=1}^{\infty}(1+\gamma_j^{-c})^{-1}\left(D^2_{\alpha\alpha}\ell(\bar{\alpha},\bar{\beta})\omega_j\omega_k - D^2_{\alpha\alpha}\ell_n(\bar{\alpha},\bar{\beta})\omega_j\omega_k\right)^2\right\}.
\end{aligned}$$

Lemma 13 generates

$$\begin{aligned}
&\|G^{11(1)}\big(D^2_{\alpha\alpha}\ell(\bar{\alpha},\bar{\beta})(\widehat{\alpha}-\bar{\alpha}) - D^2_{\alpha\alpha}\ell_n(\bar{\alpha},\bar{\beta})(\widehat{\alpha}-\bar{\alpha})\big)\|_a^2 \\
\preceq \;& \|\bar{\beta}\|_\lambda^{-4}\sum_{k=1}^{\infty}(1+\gamma_k^{-a})(1+\lambda\gamma_k^{-1})^{-2}\left(\sum_{j=1}^{\infty}(1+\gamma_j^{-c})(\hat{a}_j-\bar{a}_j)^2\right)\left(\sum_{j=1}^{\infty}(1+\gamma_j^{-c})^{-1}n^{-1}\right).
\end{aligned}$$

By the definition of $\|\cdot\|_c$ norm, $\sum_{j=1}^{\infty}(1+\gamma_j^{-c})(\hat{a}_j-\bar{a}_j)^2 = \|\widehat{\alpha}-\bar{\alpha}\|_c^2$, and $\sum_{j=1}^{\infty}(1+\gamma_j^{-c})^{-1} < \infty$, whenever $c > 1/2r$, and $\sum_{k=1}^{\infty}(1+\gamma_k^{-a})(1+\lambda\gamma_k^{-1})^{-2} = O(\lambda^{-(a+1/(2r))})$, we achieve

$$\|G^{11(1)}\big(D^2_{\alpha\alpha}\ell(\bar{\alpha},\bar{\beta})(\widehat{\alpha}-\bar{\alpha}) - D^2_{\alpha\alpha}\ell_n(\bar{\alpha},\bar{\beta})(\widehat{\alpha}-\bar{\alpha})\big)\|_a^2 \asymp n^{-1}\lambda^{-(a+1/(2r))}\|\widehat{\alpha}-\bar{\alpha}\|_c^2.$$

Following the same spirit,

$$\begin{aligned}
&\|G^{11(1)}\big(D^2_{\alpha\beta}\ell(\bar{\alpha},\bar{\beta})(\widehat{\beta}-\bar{\beta}) - D^2_{\alpha\beta}\ell_n(\bar{\alpha},\bar{\beta})(\widehat{\beta}-\bar{\beta})\big)\|_a^2 \\
= \;& 4^{-1}\|\bar{\beta}\|_\lambda^{-4}\sum_{k=1}^{\infty}(1+\gamma_k^{-a})(1+\lambda\gamma_k^{-1})^{-2}(D^2_{\alpha\beta}\ell(\bar{\alpha},\bar{\beta})(\widehat{\beta}-\bar{\beta})\omega_k - D^2_{\alpha\beta}\ell_n(\bar{\alpha},\bar{\beta})(\widehat{\beta}-\bar{\beta})\omega_k)^2 \\
= \;& 4^{-1}\|\bar{\beta}\|_\lambda^{-4}\sum_{k=1}^{\infty}(1+\gamma_k^{-a})(1+\lambda\gamma_k^{-1})^{-2} \\
&\left\{\sum_{j=1}^{\infty}(\hat{b}_j-\bar{b}_j)(D^2_{\alpha\beta}\ell(\bar{\alpha},\bar{\beta})\omega_j\omega_k - D^2_{\alpha\beta}\ell_n(\bar{\alpha},\bar{\beta})\omega_j\omega_k)\right\}^2 \\
&\leq \; 4^{-1}\|\bar{\beta}\|_\lambda^{-4}\sum_{k=1}^{\infty}(1+\gamma_k^{-c})(1+\lambda\gamma_k^{-1})^{-2}\left\{\sum_{j=1}^{\infty}(1+\gamma_j^{-a})(\hat{b}_j-\bar{b}_j)^2\right\} \\
&\left\{\sum_{j=1}^{\infty}(1+\gamma_j^{-c})^{-1}\left(D^2_{\alpha\beta}\ell(\bar{\alpha},\bar{\beta})\omega_j\omega_k - D^2_{\alpha\beta}\ell_n(\bar{\alpha},\bar{\beta})\omega_j\omega_k\right)^2\right\} \\
&\preceq \; \sum_{k=1}^{\infty}(1+\gamma_k^{-a})(1+\lambda\gamma_k^{-1})^{-2}\|\widehat{\beta}-\bar{\beta}\|_c^2\left\{\sum_{j=1}^{\infty}(1+\gamma_j^{-c})^{-1}n^{-1}\right\} \\
&\asymp \; n^{-1}\lambda^{-(a+1/(2r))}\|\widehat{\beta}-\bar{\beta}\|_c^2,
\end{aligned}$$

which finalizes the proof of Lemma 14. ■

Proof of Lemma 15

Plug in the expression of the G^{12} in Lemma 10,

$$\begin{aligned}&\|G^{12}\big(D^2_{\beta\alpha}\ell(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)-D^2_{\beta\alpha}\ell_n(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)\big)\|_a^2\\ =\;&\|-4^{-1}\|\bar\alpha\|_\lambda^{-2}\|\bar\beta\|_\lambda^{-2}\bar{\mathbf{a}}\bar{\mathbf{b}}^T\big(D^2_{\beta\alpha}\ell(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)-D^2_{\beta\alpha}\ell_n(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)\big)\|_a^2,\end{aligned}$$

which can be simplified as $4^{-2}\|\bar\alpha\|_\lambda^{-4}\|\bar\beta\|_\lambda^{-4}\|\bar\alpha\|_a^2(\bar{\mathbf{b}}^T\big(D^2_{\beta\alpha}\ell(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)-D^2_{\beta\alpha}\ell_n(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)\big))^2$. Replace $\widehat\alpha-\bar\alpha$ by its expansion (27) and (55),

$$\begin{aligned}&\|G^{12}\big(D^2_{\beta\alpha}\ell(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)-D^2_{\beta\alpha}\ell_n(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)\big)\|_a^2\\ =\;&4^{-2}\|\bar\alpha\|_\lambda^{-4}\|\bar\beta\|_\lambda^{-4}\|\bar\alpha\|_a^2\left\{\sum_{k=1}^\infty\sum_{j=1}^\infty\bar b_k(\hat a_j-\bar a_j)(D^2_{\beta\alpha}\ell(\bar\alpha,\bar\beta)\omega_j\omega_k-D^2_{\beta\alpha}\ell_n(\bar\alpha,\bar\beta)\omega_j\omega_k)\right\}^2.\end{aligned}$$

Due to the Cauchy-Schwarz inequality,

$$\begin{aligned}&\|G^{12}\big(D^2_{\beta\alpha}\ell(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)-D^2_{\beta\alpha}\ell_n(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)\big)\|_a^2\\ \le\;&4^{-2}\|\bar\alpha\|_\lambda^{-4}\|\bar\beta\|_\lambda^{-4}\|\bar\alpha\|_a^2\left\{\sum_{k=1}^\infty\sum_{j=1}^\infty(1+\gamma_k^{-c})(1+\gamma_j^{-c})\bar b_k^2(\hat a_j-\bar a_j)^2\right\}\\ &\left\{\sum_{k=1}^\infty\sum_{j=1}^\infty(1+\gamma_k^{-c})^{-1}(1+\gamma_j^{-c})^{-1}(D^2_{\beta\alpha}\ell(\bar\alpha,\bar\beta)\omega_j\omega_k-D^2_{\beta\alpha}\ell_n(\bar\alpha,\bar\beta)\omega_j\omega_k)^2\right\}.\end{aligned}$$

Bounding the second order derivative operator by Lemma 13 gives

$$\begin{aligned}&\|G^{12}\big(D^2_{\beta\alpha}\ell(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)-D^2_{\beta\alpha}\ell_n(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)\big)\|_a^2\\ \preceq\;&n^{-1}\|\bar\beta\|_c^2\|\widehat\alpha-\bar\alpha\|_c^2\left\{\sum_{k=1}^\infty\sum_{j=1}^\infty(1+\gamma_k^{-c})^{-1}(1+\gamma_j^{-c})^{-1}\right\}\asymp n^{-1}\|\widehat\alpha-\bar\alpha\|_c^2,\end{aligned}$$

because $1+\gamma_k^{-c}$ is summable provided $c>1/2r$.

The proof of the remaining three terms has a similar flavor: reexpress the inverse of the second order derivative operator, simplify the expression, expand the functions by basis, apply the Cauchy-Schwarz inequality, bound the second order derivative operator, and tidy the formula as follows

$$\begin{aligned}&\|G^{12}\big(D^2_{\beta\beta}\ell(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)-D^2_{\beta\beta}\ell_n(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)\big)\|_a^2\\ =\;&\|-4^{-1}\|\bar\alpha\|_\lambda^{-2}\|\bar\beta\|_\lambda^{-2}\bar{\mathbf{a}}\bar{\mathbf{b}}^T\big(D^2_{\beta\beta}\ell(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)-D^2_{\beta\beta}\ell_n(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)\big)\|_a^2\\ =\;&4^{-2}\|\bar\alpha\|_\lambda^{-4}\|\bar\beta\|_\lambda^{-4}\|\bar\alpha\|_a^2(\bar{\mathbf{b}}^T\big(D^2_{\beta\beta}\ell(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)-D^2_{\beta\beta}\ell_n(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)\big))^2\\ =\;&4^{-2}\|\bar\alpha\|_\lambda^{-4}\|\bar\beta\|_\lambda^{-4}\|\bar\alpha\|_a^2\left\{\sum_{k=1}^\infty\sum_{j=1}^\infty\bar b_k(\hat b_j-\bar b_j)(D^2_{\beta\beta}\ell(\bar\alpha,\bar\beta)\omega_j\omega_k-D^2_{\beta\beta}\ell_n(\bar\alpha,\bar\beta)\omega_j\omega_k)\right\}^2\\ \preceq\;&\|\bar\beta\|_c^2\|\widehat\beta-\bar\beta\|_c^2\left\{\sum_{k=1}^\infty\sum_{j=1}^\infty(1+\gamma_k^{-c})^{-1}(1+\gamma_j^{-c})^{-1}n^{-1}\right\}\asymp n^{-1}\|\widehat\beta-\bar\beta\|_c^2,\\ &\|G^{11(2)}\big(D^2_{\alpha\alpha}\ell(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)-D^2_{\alpha\alpha}\ell_n(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)\big)\|_a^2\\ =\;&\|-4^{-1}\|\bar\alpha\|_\lambda^{-2}\|\bar\beta\|_\lambda^{-2}\bar{\mathbf{a}}\bar{\mathbf{a}}^T\big(D^2_{\alpha\alpha}\ell(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)-D^2_{\alpha\alpha}\ell_n(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)\big)\|_a^2\\ =\;&4^{-2}\|\bar\alpha\|_\lambda^{-4}\|\bar\beta\|_\lambda^{-4}\|\bar\alpha\|_a^2\left(\bar{\mathbf{a}}^T\big(D^2_{\alpha\alpha}\ell(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)-D^2_{\alpha\alpha}\ell_n(\bar\alpha,\bar\beta)(\widehat\alpha-\bar\alpha)\big)\right)^2\\ =\;&4^{-2}\|\bar\alpha\|_\lambda^{-4}\|\bar\beta\|_\lambda^{-4}\|\bar\alpha\|_a^2\left\{\sum_{k=1}^\infty\sum_{j=1}^\infty\bar a_k(\hat a_j-\bar a_j)(D^2_{\alpha\alpha}\ell(\bar\alpha,\bar\beta)\omega_j\omega_k-D^2_{\alpha\alpha}\ell_n(\bar\alpha,\bar\beta)\omega_j\omega_k)\right\}^2\\ \preceq\;&\|\bar\alpha\|_c^2\|\widehat\alpha-\bar\alpha\|_c^2\left\{\sum_{k=1}^\infty\sum_{j=1}^\infty(1+\gamma_k^{-c})^{-1}(1+\gamma_j^{-c})^{-1}n^{-1}\right\}\asymp n^{-1}\|\widehat\alpha-\bar\alpha\|_c^2,\end{aligned}$$

$$
\begin{aligned}
& \|G^{11(2)}\big(D^2_{\alpha\beta}\ell(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta) - D^2_{\alpha\beta}\ell_n(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)\big)\|_a^2 \\
= \;& \| - 4^{-1}\|\bar\alpha\|_\lambda^{-2}\|\bar\beta\|_\lambda^{-2}\bar{\mathbf{a}}\bar{\mathbf{a}}^T\big(D^2_{\alpha\beta}\ell(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta) - D^2_{\alpha\beta}\ell_n(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)\big)\|_a^2 \\
= \;& 4^{-2}\|\bar\alpha\|_\lambda^{-4}\|\bar\beta\|_\lambda^{-4}\|\bar{\mathbf{a}}\|_a^2(\bar{\mathbf{a}}^T\big(D^2_{\alpha\beta}\ell(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta) - D^2_{\alpha\beta}\ell_n(\bar\alpha,\bar\beta)(\widehat\beta-\bar\beta)\big))^2 \\
= \;& 4^{-2}\|\bar\alpha\|_\lambda^{-4}\|\bar\beta\|_\lambda^{-4}\|\bar\alpha\|_a^2\left\{\sum_{k=1}^\infty\sum_{j=1}^\infty \bar a_k(\hat b_j-\bar b_j)(D^2_{\alpha\alpha}\ell(\bar\alpha,\bar\beta)\omega_j\omega_k - D^2_{\alpha\alpha}\ell_n(\bar\alpha,\bar\beta)\omega_j\omega_k)\right\}^2 \\
\preceq \;& \|\bar\alpha\|_c^2\|\widehat\beta-\bar\beta\|_c^2\left\{\sum_{k=1}^\infty\sum_{j=1}^\infty(1+\gamma_k^{-c})^{-1}(1+\gamma_j^{-c})^{-1}n^{-1}\right\} \asymp n^{-1}\|\widehat\beta-\bar\beta\|_c^2.
\end{aligned}
$$

■

Appendix D. Theoretical Results for the Distinct Version of Two Domains

In Section 3, for simplicity and notational convenience, we assumed everything related to the two domains are the same, including $\mathcal{T}_1 = \mathcal{T}_2 = \mathcal{T}$, $K_1 = K_2 = K$, $C_\alpha = C_\beta = C$ and $\lambda_\alpha = \lambda_\beta = \lambda$. In this section, we briefly sketch the preliminary, theorems, and proofs for the distinct version.

Recall in Section 3.1, we defined the following successively: kernel R associated with the norm $\|\cdot\|_R := (\|\cdot\|_0^2 + \|\cdot\|_K^2)^{1/2}$, where $\|\cdot\|_0$ and $\|\cdot\|_K$ depend on C and K respectively; linear operators $\mathcal{L}_R, \mathcal{L}_{R^{1/2}}$ and $\mathcal{L}_T = \mathcal{L}_{R^{1/2}CR^{1/2}}$; eigenvalues s_k^T's of $\mathcal{L}_T$; basis functions ω_k, and values $\gamma_k = (1/s_k^T - 1)^{-1}$. Suppose all of these quantities are defined again successively and differently for two domains with subscripts α and β corresponding to the two domains. Then we have $\gamma_{k,\alpha}$, $\gamma_{k,\beta}$, $\omega_{k,\alpha}$, and $\omega_{k,\beta}$. The functions, $\omega_{k,\alpha}$ and $\omega_{k,\beta}$, are essential in the proof, since we will expand all of the functions of interest onto these basis functions. The decay rates of $\gamma_{k,\alpha}$ and $\gamma_{k,\beta}$ play a prominent role in the convergence rate.

We will impose the following condition instead of Condition 1 in Section 3.1:

Condition 3: the values $\gamma_{k,\alpha}$ and $\gamma_{k,\beta}$ satisfy the following decay rates,

$$\gamma_{k,\alpha} \asymp k^{-2r_\alpha},\ \gamma_{k,\beta} \asymp k^{-2r_\beta}, \tag{56}$$

for some constants $0 < r_\alpha, r_\beta < \infty$.

Theorems 17 and 18 state the results of the matching upper and lower bounds for the distinct version and hence the optimality of our proposed estimator.

Theorem 17 *Under Conditions 2-3, the smoothness regularization estimators $(\widehat\alpha, \widehat\beta)$ defined in* (9) *with Candidate 3, $\lambda_\alpha = O(n^{-2r_\alpha/(2r_\alpha+1)})$ and $\lambda_\beta = O(n^{-2r_\beta/(2r_\beta+1)})$ satisfy*

$$\lim_{A\to\infty}\lim_{n\to\infty}\sup_{\alpha_0\in\mathcal{H}(K),\,\beta_0\in\mathcal{H}(K)}\mathbb{P}\left(\mathcal{E}(\widehat\alpha,\widehat\beta;\alpha_0,\beta_0) \geq A\max\left\{n^{-\frac{2r_\alpha}{2r_\alpha+1}}, n^{-\frac{2r_\beta}{2r_\beta+1}}\right\}\right) = 0.$$

Theorem 18 *Under the same assumptions as in Theorem 17, for any estimate $(\tilde\alpha, \tilde\beta)$ based on the observations $\{(x_i(\cdot,\cdot), y_i) : i = 1, 2, \ldots, n\}$, we have the following lower bound,*

$$\lim_{a\to 0}\lim_{n\to\infty}\inf_{\tilde\alpha,\tilde\beta}\sup_{\alpha_0\in\mathcal{H}(K),\,\beta_0\in\mathcal{H}(K)}\mathbb{P}\left(\mathcal{E}(\tilde\alpha,\tilde\beta;\alpha_0,\beta_0) \geq a\max\left\{n^{-\frac{2r_\alpha}{2r_\alpha+1}}, n^{-\frac{2r_\beta}{2r_\beta+1}}\right\}\right) = 1.$$

For the lower bound, proof of Theorem 18 is the same as proof of Theorem 5 when applying the same argument twice to α and β separately.

For the upper bound, proof of Theorem 17 follows the same logic as the proof of Theorem 2. Note that Lemma 6 still holds with extra subscripts α, β. Proof of Theorem 2 relies upon (18), proof of Theorem 17 relies upon the distinct version of (18), which is

$$\mathcal{E}(\widehat{\alpha}, \widehat{\beta}; \alpha_0, \beta_0) \leq 3\|\widehat{\alpha} - \alpha_0\|_{0\alpha}^2 \|\widehat{\beta} - \beta_0\|_{0\beta}^2 + 3\|\widehat{\alpha} - \alpha_0\|_{0\alpha}^2 \|\beta_0\|_{0\beta}^2 + 3\|\alpha_0\|_{0\alpha}^2 \|\widehat{\beta} - \beta_0\|_{0\beta}^2, \quad (57)$$

where the norms on the right-hand side are defined in (5).

Define the norms $\|\cdot\|_{a\alpha}$ and $\|\cdot\|_{a\beta}$ similarly as for $\|\cdot\|_a$. Previously for the version with the same domain, to bound (18), we bound the three terms on the right-hand side of (21). Now for the distinct version, to bound (57), we bound the three terms on the right-hand sides of (58) and (59)

$$\|\widehat{\alpha} - \alpha_0\|_{a\alpha} \leq \|\widehat{\alpha} - \tilde{\alpha}\|_{a\alpha} + \|\tilde{\alpha} - \bar{\alpha}\|_{a\alpha} + \|\bar{\alpha} - \alpha_0\|_{a\alpha}, \quad (58)$$

$$\|\widehat{\beta} - \beta_0\|_{a\beta} \leq \|\widehat{\beta} - \tilde{\beta}\|_{a\beta} + \|\tilde{\beta} - \bar{\beta}\|_{a\beta} + \|\bar{\beta} - \beta_0\|_{a\beta}. \quad (59)$$

Lemmas 7, 8 and 9 provide the bounds of the three terms on the right-hand side of (21) when the two domains are similar. Lemmas 19, 20 and 21 replace Lemmas 7, 8 and 9 for the distinct version.

Lemma 19 *If* $\lambda_\alpha = \lambda_\beta = o(1),\ 0 \leq a \leq 1$, *then*

$$\|\bar{\alpha} - \alpha_0\|_{a\alpha}^2 = O(\lambda_\alpha^{1-a}),\ \textit{and}\ \|\bar{\beta} - \beta_0\|_{a\beta}^2 = O(\lambda_\beta^{1-a}).$$

Lemma 20 *If* $\lambda_\alpha = \lambda_\beta = o(1),\ 0 \leq a \leq 1$ *and* $r_\alpha, r_\beta > 1/2$, *then*

$$\mathbb{E}\|\tilde{\alpha} - \bar{\alpha}\|_{a\alpha}^2 \preceq n^{-1}\lambda_\alpha^{-(a+1/(2r_\alpha))},\ \textit{and}\ \mathbb{E}\|\tilde{\beta} - \bar{\beta}\|_{a\beta}^2 \preceq n^{-1}\lambda_\beta^{-(a+1/(2r_\beta))}.$$

Lemma 21 *If there exists some constant* c *such that* $\max\{1/(2r_\alpha), 1/(2r_\beta)\} < c \leq 1$ *and* $n^{-1}\lambda_\alpha^{-(c+1/(2r_\alpha))} = o(1)$, $n^{-1}\lambda_\beta^{-(c+1/(2r_\beta))} = o(1)$, *then*

$$\|\widehat{\alpha} - \tilde{\alpha}\|_{a\alpha}^2 = o_p(n^{-1}\lambda_\alpha^{-(a+1/(2r_\alpha))}),\ \textit{and}\ \|\widehat{\beta} - \tilde{\beta}\|_{a\beta}^2 = o_p(n^{-1}\lambda_\beta^{-(a+1/(2r_\beta))}).$$

Since the roles of α's and β's can be switched, for the distinct version, we only need to prove Lemmas 19, 20 and 21 related to α's.

The proof of Lemma 19 remains almost the same as the proof of Lemma 7 except for adding subscripts $_\alpha$ and $_\beta$.

The proof of Lemma 20 for the distinct version relies upon the upper bounds of the three terms on the right hand side of (37), which is dominated by the first term. This first term is still bounded by the same rate as in (42) when an extra subscript is added, i.e., r becomes r_α and λ becomes λ_α. The reason is that (42) depends upon Lemma 10 and Lemma 11. Lemma 11 still holds for the distinct version and in Lemma 10, the two relevant equations become $G^{11(1)} = 2^{-1}\|\bar{\beta}\|_{\lambda_\beta}^{-2}\mathrm{diag}\big((1+\lambda_\alpha\gamma_{k,\alpha}^{-1})^{-1}\big)$ and $G^{22(1)} = 2^{-1}\|\bar{\alpha}\|_{\lambda_\alpha}^{-2}\mathrm{diag}\big((1+\lambda_\beta\gamma_{k,\beta}^{-1})^{-1}\big)$.

The proof of Lemma 21 for the distinct version still hinges upon (51) and its β version. The terms on the right-hand side of (51) need to be bounded by Lemmas 14 and 15. Lemma 15 still applies for the distinct version when $\|\cdot\|_c$ is replaced by $\|\cdot\|_{c\alpha}$ or $\|\cdot\|_{c\beta}$. Lemma 14 holds when adding subscripts appropriately to the RHS. Therefore, $\|\widehat{\alpha} - \tilde{\alpha}\|_{a\alpha}^2 = O(n^{-1}\lambda_\alpha^{-(a+1/(2r_\alpha))}\|\widehat{\alpha} - \bar{\alpha}\|_{c\alpha}^2 + n^{-1}\lambda_\beta^{-(a+1/(2r_\beta))}\|\widehat{\beta} - \bar{\beta}\|_{c\beta}^2)$. In particular, we obtain the following when letting $a = c$, $\|\widehat{\alpha} - \tilde{\alpha}\|_{c\alpha}^2 = O(n^{-1}\lambda_\alpha^{-(a+1/(2r_\alpha))}\|\widehat{\alpha} - \bar{\alpha}\|_{c\alpha}^2 + n^{-1}\lambda_\beta^{-(a+1/(2r_\beta))}\|\widehat{\beta} - \bar{\beta}\|_{c\beta}^2)$. Under the conditions $n^{-1}\lambda_\alpha^{-(c+1/(2r_\alpha))} = o(1)$ and $n^{-1}\lambda_\beta^{-(c+1/(2r_\beta))} = o(1)$, applying the triangle inequality yields $\|\tilde{\alpha} - \bar{\alpha}\|_{c\alpha}^2 \geq \|\widehat{\alpha} - \bar{\alpha}\|_{c\alpha}^2 - \|\widehat{\alpha} -$

$\tilde{\alpha}\|_{c\alpha}^2 = \big(1-o(1)\big)\|\widehat{\alpha}-\bar{\alpha}\|_{c\alpha}^2 - o(1)\|\widehat{\beta}-\bar{\beta}\|_{c\beta}^2$. Hence, $\|\widehat{\alpha}-\bar{\alpha}\|_{c\alpha}^2 = O(\|\tilde{\alpha}-\bar{\alpha}\|_{c\alpha}^2 + \|\tilde{\beta}-\bar{\beta}\|_{c\beta}^2)$, which implies $\|\widehat{\alpha}-\tilde{\alpha}\|_{c\alpha}^2 = O(n^{-1}\lambda_\alpha^{-(c+1/(2r_\alpha))}\|\tilde{\alpha}-\bar{\alpha}\|_{c\alpha}^2 + n^{-1}\lambda_\alpha^{-(c+1/(2r_\alpha))}\|\tilde{\beta}-\bar{\beta}\|_{c\beta}^2)$. Together with Lemma 20 and a parallel argument for β, completes the proof of Lemma 21.

Appendix E. Additional Simulation Results

In this section, we provide additional simulation results as a supplement to Section 4.2.

E.1 Additional Simulation Results on 1D FLR+vectorization

We assess the performance of FLR after various types of vectorization. Given the matrix-valued predictor, another natural but undesirable choice is to perform vectorization first and then apply existing methods which apply to vector-valued data. It is known in the literature of tensor data analysis that such vectorization is sub-optimal. With the additional feature of the functional data, to make a comprehensive comparison with existing methods, we still include the 1D FLR of Cai and Yuan (2012) after vectorization. In the literature, the default vectorization approach, denoted by vec here, is to stack all the columns of a matrix one by one. In the context of functional data, because of the requirement of smoothness, there are a few other ways of vectorization. For example, it might be worthwhile to consider flipping the even-numbered column vectors upside down and then stack all the columns together, which is denoted by vec*. Furthermore, maybe rows are smoother, and so stacking rows (and potentially flipping even-numbered rows) is more appropriate. These considerations lead to four ways of vectorization and result in FLR+vec$(X(s,t))$, FLR+vec$(X^T(s,t))$, FLR+vec$^*(X(s,t))$ and FLR+vec$^*(X^T(s,t))$.

The comparison of the four different vectorization approaches is shown in Figure E.1. Because in the simulation setup, the covariances C_α, C_β and coefficient functions $\alpha_0(\cdot), \beta_0(\cdot)$ are symmetric for the two domains, transposing X or not, i.e., stacking rows or columns, does not matter. However, concatenating head-to-tail or head-to-head does matter as revealed in the figure, where the performance of vec* dominates that of vec.

For these reasons, for the rest of this section, FLR refers to FLR+vec$^*(X(s,t))$; in Section 5 on the Canadian weather data, FLR refers to FLR+vec$(X^T(s,t))$, because it is natural to connect the last hour in the current day to the first hour in the next day; in Section Appendix G on the LIDAR data, where the two domains are different, we consider two choices: FLR+vec$^*(X(s,t))$ and FLR+vec$^*(X^T(s,t))$.

E.2 Additional Simulation Results on Other Competitors

Besides GMRF, PFPCR, MFPCR, FLR+TPK, we also compare the performance of FBLR with three more existing methods: FLR, Ridge regression after plain vectorization, and bilinear regression (BLR), which is a special case of FBLR when $\lambda_\alpha = \lambda_\beta = 0$, under Setting 1. Figure E.2 shows the clear advantage of FBLR over them due to their lack of smoothness or lack of matrix structure.

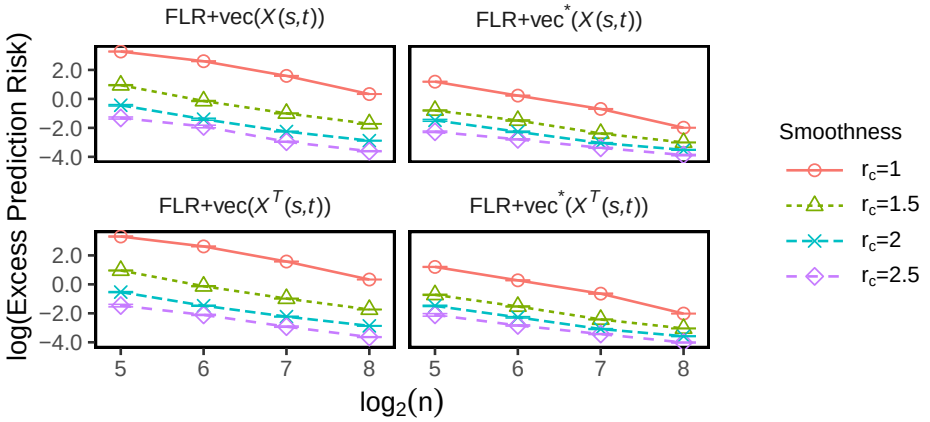

Figure E.1: Plots of the excess prediction risk vs the sample size with both axes in log scale under Setting 1. Four sample sizes and four values of r_c are considered. The error bars correspond to mean $\pm$ one SE. The four panels are for four types of vectorization methods.

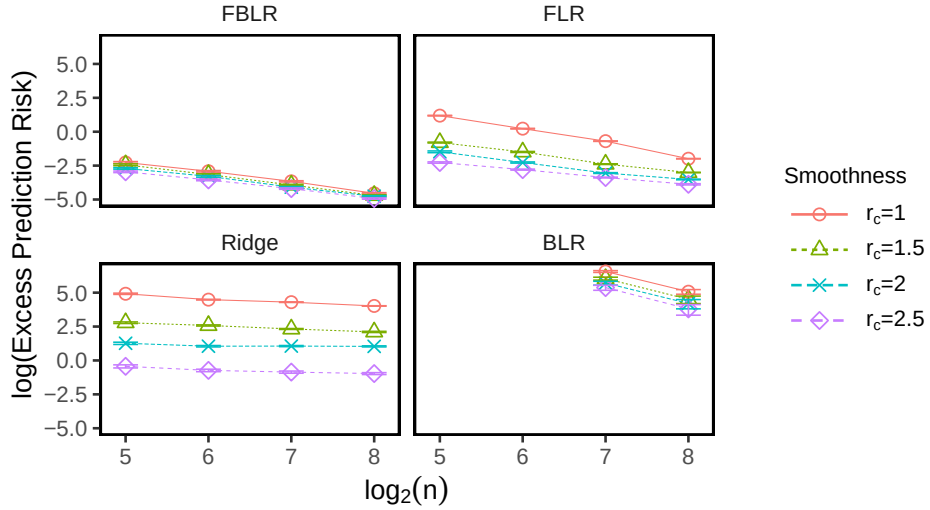

Figure E.2: Plots of the excess prediction risk vs the sample size in log scale for Settings 1. Four approaches are included. The error bars are generated according to mean $\pm$ one SE. BLR is only shown for $\log_2(n) = 7, 8$, because it requires a larger sample size.

E.3 Additional Simulation Results on 2D-FPCR

We examine the choice of $r^{\max}$ in 2D-FPCR. Figure E.3 demonstrates the prediction risk and computation time of PFPCR and MFPCR with three options of $r^{\max} \in \{4, 8, \lfloor\sqrt{n-1}\rfloor\}$. Here, $\lfloor\sqrt{n-1}\rfloor$ is the largest possible value for $r^{\max}$ in Chen et al. (2017). It is clear that $r^{\max} = \lfloor\sqrt{n-1}\rfloor$ is far more computationally expensive and even less accurate than the other two. Because given that the true coefficient function under Setting 1 only consists of the leading four basis functions, estimating more than four PCs will hurt the performance. Therefore, in Section 4, only $r^{\max} \in \{4, 8\}$ are compared with FBLR.

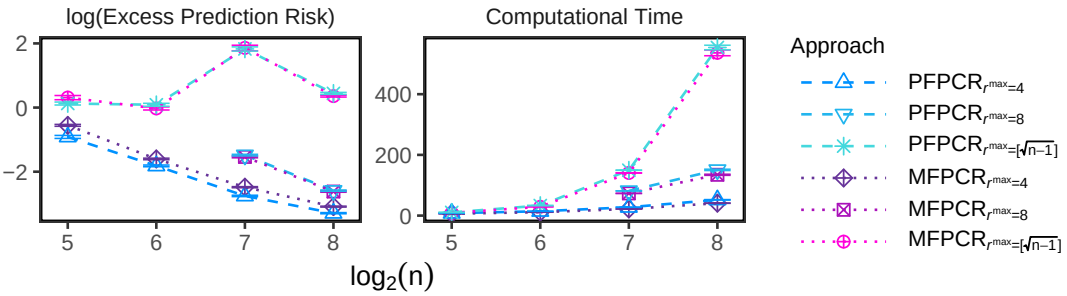

Figure E.3: Plots of the excess prediction risk and computational time vs the sample size in log scale with $r_c = 1$ under Setting 1. The error bars correspond to mean $\pm$ one SE.

E.4 Additional Simulation Results on FBLR with a Different Choice of Kernel

It is known that the estimation procedures that use the RKHS frameworks depend on the choice of the kernels, such as 1D FLR, FLR+TPK, and our FBLR. In this section, we use a simulation study to investigate how much influence the choice of the kernel has on the performance of FBLR. For the simulation study in Section 4 and the real data of LIDAR in Section G, we use kernel (17); for real data of Canadian weather in Section 5, we use $K(s,t) = 1 - B_4(|s-t|)/24$, both choices of kernels were used in Cai and Yuan (2012). In this section, we consider a universal Gaussian kernel $K(s,t) = \exp\left(-\frac{(s-t)^2}{2\sigma^2}\right)$, and refer to this approach as FBLR+GK. The parameter σ is selected through cross-validation. We provide simulation results for all Settings 1-6 (with $r_c = 1$) in Section 4.2. For Settings 5-6, we implement FBLR$_{R=2}$ with the Gaussian kernel, denoted as FBLR$_{R=2}$+GK.

Figure E.4 demonstrates the prediction risk of FBLR+GK and FBLR$_{R=2}$+GK, along with all the approaches shown in Section 4.2.

For Settings 1-4, where the true model follows a bilinear form (2), both $\alpha(\cdot)$ and $\beta(\cdot)$ are composed of multiple cosine basis functions. The aforementioned kernel (17) matches the linear span of these basis functions. See Cai and Yuan (2012) for more details on the RKHS associated with the kernel (17). It is expected that FBLR+GK performs worse than FBLR (with kernel (17)). However, FBLR+GK consistently outperforms PFPCR, MFPCR, GMRF, and FLR+TPK, except for Settings 1-2 with a very small sample size where FBLR+GK is slightly worse than FLR+TPK. Note that here, FLR+TPK still uses the tensor product kernel that depends on kernel (17).

For Settings 5-6, where the true models are based on Model (3), the true coefficient function $\beta_0(\cdot,\cdot)$ is a 2D function and hence we also consider FBLR$_{R=2}$+GK. Under Setting 5, the true coefficient function still depends on cosine basis function; therefore, FBLR+GK and FBLR$_{R=2}$+GK are slightly worse than FBLR and FBLR$_{R=2}$ respectively. But the comparison between FBLR+GK, FBLR$_{R=2}$+GK and all the other procedures remains the same as the comparison between FBLR, FBLR$_{R=2}$ and all the other procedures. In short, excluding FBLR and FBLR$_{R=2}$, FBLR+GK is the best among all for small sample sizes while FBLR$_{R=2}$+GK is the best among all for large sample sizes. Under Setting 6, where the two-dimensional coefficient function is not low-rank and we do not know the true basis function. It is seen that FBLR and FBLR+GK yield similar results, as do FBLR$_{R=2}$ and FBLR$_{R=2}$+GK. FBLR$_{R=2}$ and FBLR$_{R=2}$+GK dominate all the others.

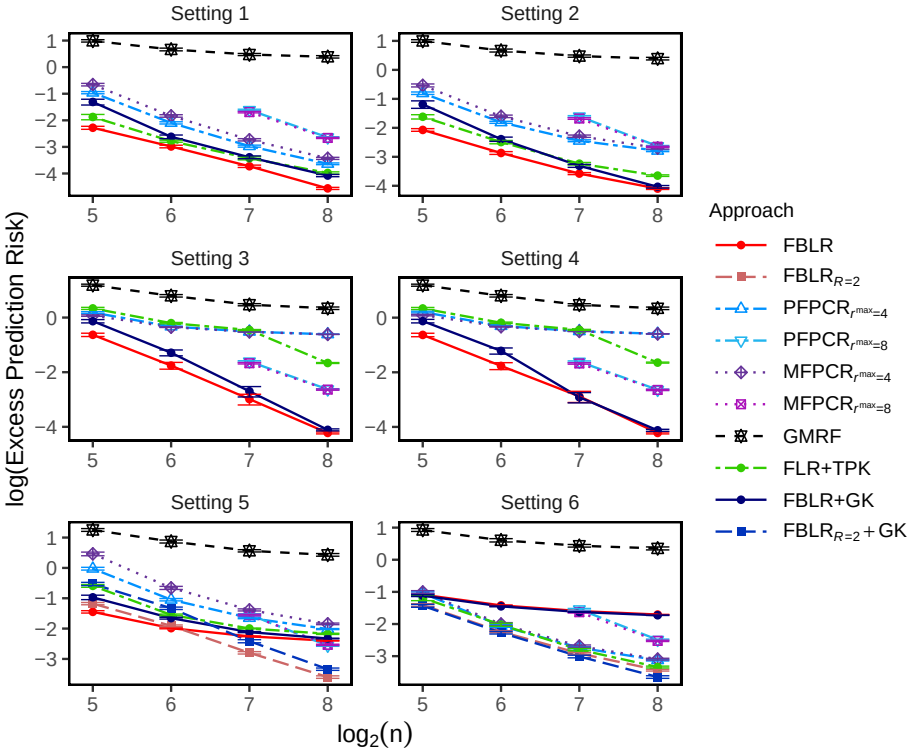

Figure E.4: Plots of the excess prediction risk vs the sample size in log scale with $r_c = 1$ under Settings 1-6. The error bars correspond to mean $\pm$ one SE.

In summary, no matter for FBLR with or without deflation, the implementation of the methodology certainly requires the specific choice of the kernel function, but the impact of the kernel function on the performance of the FBLR and its iterative deflation version is minimal and sometimes negligible. More importantly, the small impact of the kernel choice does not overshadow the strength of FBLR over other methods.

E.5 Additional Simulation Results on FBLR for Not-fully-observed Data

In this section, we study the numerical performance of our procedure when the data are not fully observed. We examine all six settings in Section 4. For each setting, we generate n, varying n, samples of the two-dimensional functional predictor on a 100×100 regular grid within $[0,1] \times [0,1]$. The predictor is fully observed in the first domain $\mathcal{T}_1$, but not fully observed in the second domain $\mathcal{T}_2$, being recorded at L random locations. The observations are denoted as $\{X_i(s, T_{ij}), 1 \le i \le n, 1 \le s \le 100, 1 \le j \le L\}$. We consider three sampling frequencies ($L = 10, 30, 50$) to achieve varying levels of sparsity.

To extend FBLR to not-fully-observed data, we first applied the principal component analysis through conditional expectation (PACE) method proposed by Yao et al. (2005) to impute data. Then we apply FBLR or FBLR$_{R=2}$ to the resulting dense data on the 100×100 grid.

Note that PACE is applicable to one-dimensional functional data. To adapt it for our two-dimensional data, we first convert n samples of sizes $100 \times L$ to $100n$ samples of one-

dimensional input of size L. Applying PACE will lead to $100n$ samples of one-dimensional input of size 100. We finally reshape the data back to n samples of sizes 100×100.

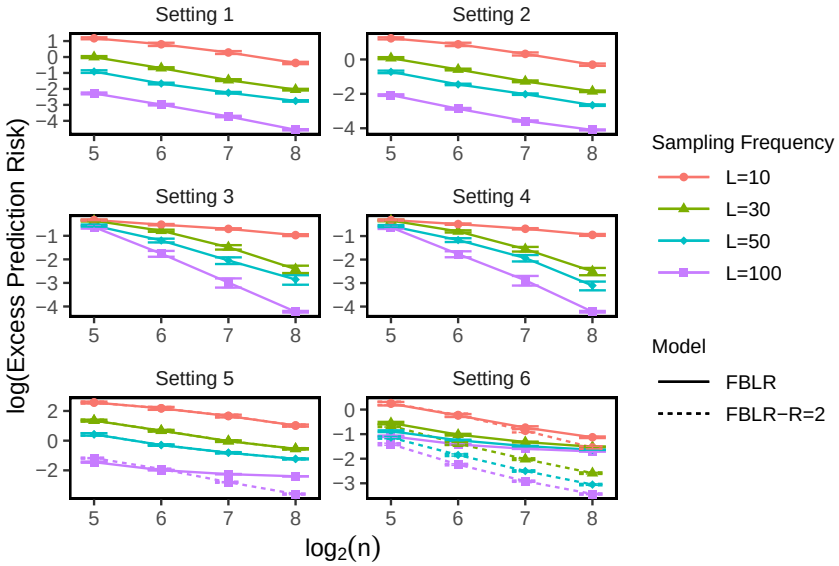

Figure E.5: Plots of the excess prediction risk and vs the sample size in log scale with $r_c = 1$ under Settings 1-6. The error bars correspond to mean $\pm$ one SE.

The results are presented in Figure E.5, together with dense data ($L = 100$). For Settings 1-4, FBLR performance is shown. For Settings 5-6, the performance of FBLR and $\text{FBLR}_{R=2}$ are both shown. The performance of either FBLR or $\text{FBLR}_{R=2}$ improves as sample size n increases or data get denser. Interestingly, in Setting 5, although the true coefficient function is rank 2, $\text{FBLR}_{R=2}$ outperforms FBLR for dense data ($L = 100$) but not for non-fully-observed data ($L = 10, 30, 50$). This further demonstrates that when data get sparser, a simpler model might be preferred. In Setting 6, the true coefficient function does not have low rank, $\text{FBLR}_{R=2}$ does outperform FBLR for all sparsity levels.

Appendix F. Additional Real Data Analysis on Canadian Weather

This section provides more information on the application to the Canadian weather data as a supplement to Section 5. Figure F.1 provides the residual diagnosis of FBLR, which suggests a fairly good fit of the data.

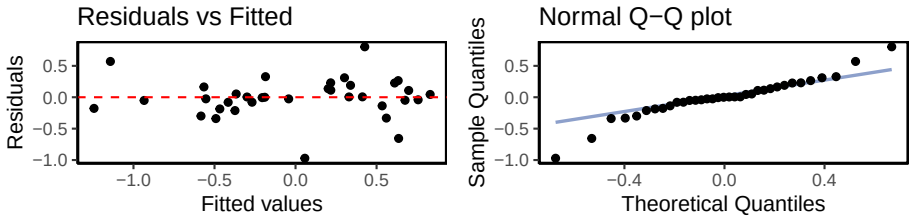

Figure F.1: The diagnosis plots of the residuals of FBLR for the Canadian weather data.

Appendix G. Real Data Analysis: LIDAR

We demonstrate the performance of FBLR and other methods on the LIDAR data. The goal is to discriminate biological threat aerosol clouds in the atmosphere from non-biological interferent aerosol clouds such as dust or smoke. We use the same data set as in Xun et al. (2013), where there are 28 aerosol clouds, half being biological and the other half non-biological. For each aerosol cloud at each time point, a set of 19 wavelength pulses is transmitted. The LIDAR receiver collects a fraction of the total optical power back-scattered over 60 equally-spaced range points (excluding background) at time $1, 2, \ldots, 20$ for wavelength $1, 2, \ldots, 19$. See Figure G.1 for an illustration of the data generation process.

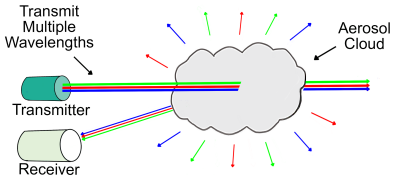

Figure G.1: A comic describing the LIDAR data generation.

Figure G.2 provides some visualization for one biological sample. It shows that the signal is smooth along three domains: time, range, and wavelength. In the literature of chemical biology, researchers have used approaches such as support vector machines after feature engineering or partial differential equations to solve this problem. Because of the smoothness of the surface, we will apply functional methods instead.

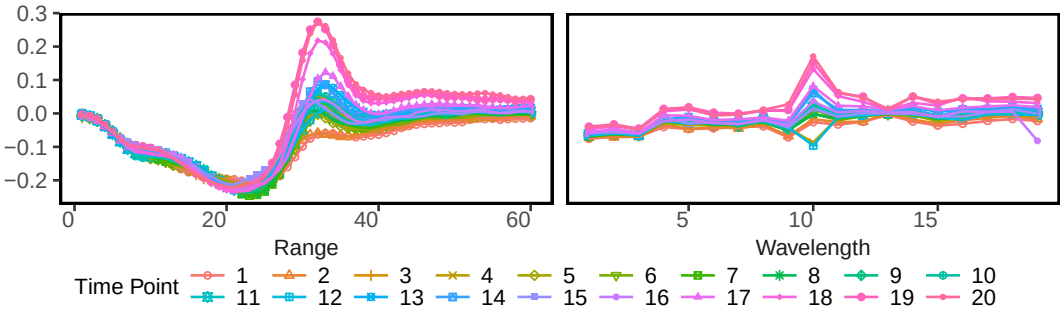

Figure G.2: Snapshots of one biological sample in the LIDAR data. 20 curves correspond to 20 time points. The left panel is the received signal over all 60 ranges at wavelength value 11. The right panel is the received signal over all 19 wavelengths at range value 45.

Ideally, one should use 3D input of size $60 \times 19 \times 20$ as the predictor to make predictions. Given that FBLR and most existing methods only apply to 2D data, we will perform the analysis 20 times corresponding to 20 time points separately. For each time point, we have 28 observations with input x_i of size 60×19 and response y_i taking value 0 or 1, standing for non-biological or biological aerosol, respectively. We adopt the regression approach to make classification: assign to class 1 if and only if the predicted response is larger than .5. We use the leave-one-out method to compute the out-of-sample testing misclassification

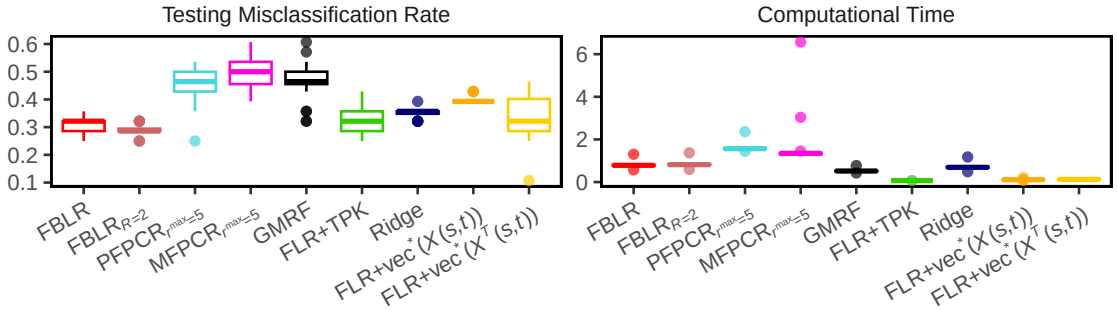

Figure G.3: Boxplots of the testing misclassification rates and computation time of 20 time points for multiple approaches in the LIDAR data.

rate. The boxplots of 20 testing misclassification rates for 20 time points and computational time for nine approaches are given in Figure G.3.

Both FBLR and FBLR$_{R=2}$ are included. We perform the test on the separability of the covariance by using Aston et al. (2017) and the separability is verified for all time points. As explained in Appendix E.1, FLR with two types of vectorization is compared with. For FBLR-related and FLR-related methods, we use the kernel in (17) again. For PFPCR and MFPCR, we set $r^{\max} = \lfloor \sqrt{n-1} \rfloor = 5$. BLR is not included because the sample size is not large enough. Figure G.3 shows that FBLR$_{R=2}$ is the best, followed by FBLR, and then FLR+TPK. The other 2D methods such as PFPCR, MFPCR and GMRF are close to random guesses and even worse than the 1D methods such as FLR+vec$^*(X(s,t))$, FLR+vec$^*(X^T(s,t))$ and Ridge.

Figures G.4 - G.7 further show the advantage of FBLR over other methods on LIDAR data in terms of interpretation, smoothness, and stableness. Figure G.4 shows the heat-maps of the estimated 2D coefficient function $\hat{\beta}(\cdot,\cdot)$ in Model (3) for these nine methods at time point 2, which is randomly selected (the other time points have similar message). It shows that both FBLR and FBLR$_{R=2}$ obtain smoother coefficient function estimations compared with other 2D methods. The small visual difference between FBLR and FBLR$_{R=2}$ indicates that the second term in Model (4) has a small magnitude. Furthermore, although PFPCR and MFPCR are supposed to provide smooth estimations, the resulting estimated coefficient function is not very smooth. For the 1D methods, the Ridge estimation inherently lacks smoothness, and FLR+vec$^*(X(s,t))$ (stacking columns) and FLR+vec$^*(X^T(s,t))$ (stacking rows) exhibit excessive smoothing because one dimension has nearly no variation.

Figure G.5 shows the estimated 2D coefficient functions $\hat{\beta}(\cdot,\cdot)$ in Model (3) by FBLR$_{R=2}$ for LIDAR data at 20 time points. Every 2D coefficient function at any time point is smooth and the 2D coefficient functions evolve smoothly over time.

We next compare the evolvement of the 2D coefficient functions over time of all nine methods. A direct comparison of the 2D surfaces is difficult, so we perform SVD of all 2D coefficient functions. Figures G.6 - G.7 show the leading left and right singular vectors, which correspond to $\hat{\alpha}^{[1]}(\cdot)$ and $\hat{\beta}^{[1]}(\cdot)$ in Model (2), respectively. Again, the estimated 1D coefficient functions by FBLR and FBLR$_{R=2}$ are smooth for each time point and evolve smoothly over time; in contrast, the estimated 1D coefficient functions by 2D methods such as PFPCR, MFPCR, and GMRF are in general not very smooth for each time point and do not evolve quite smoothly over time; the estimated 1D coefficient functions by 1D methods

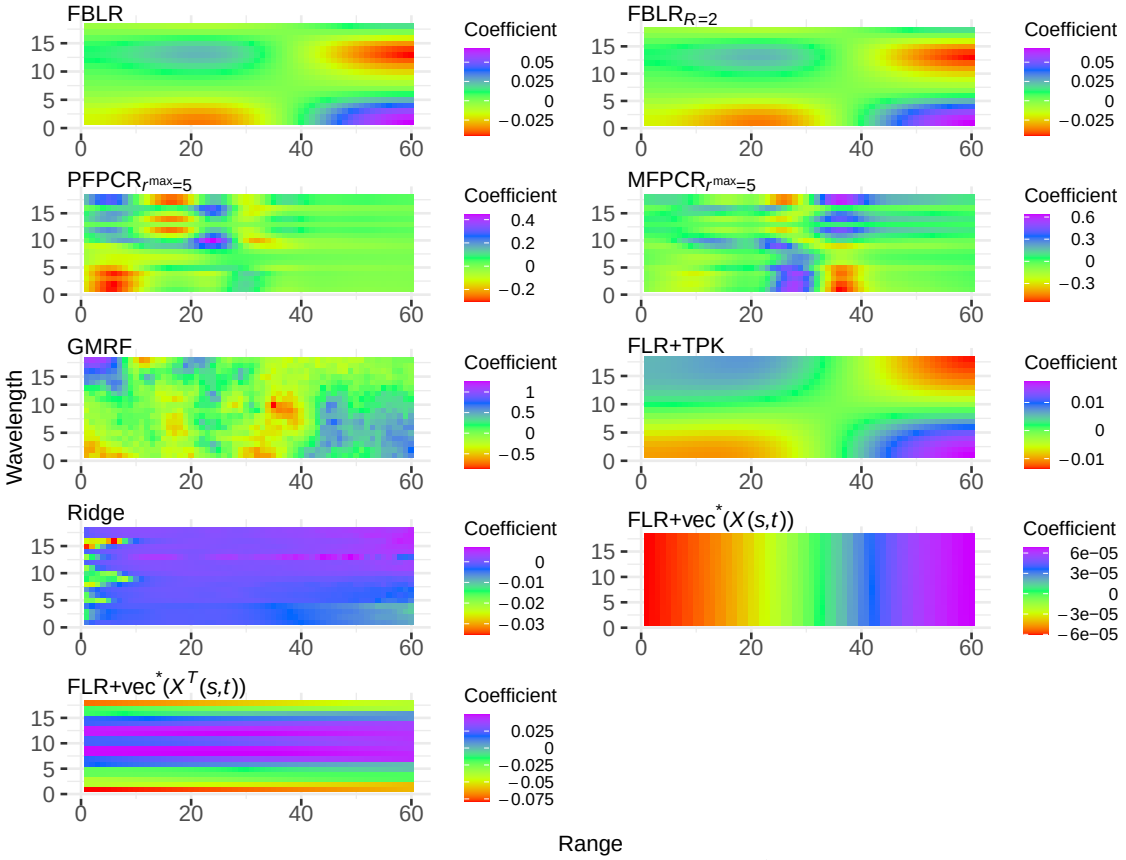

Figure G.4: Plots of the estimated 2D coefficient function $\hat{\beta}(\cdot, \cdot)$ in Model (3) for LIDAR data at time point 2. The two axes correspond to range and wavelength, respectively.

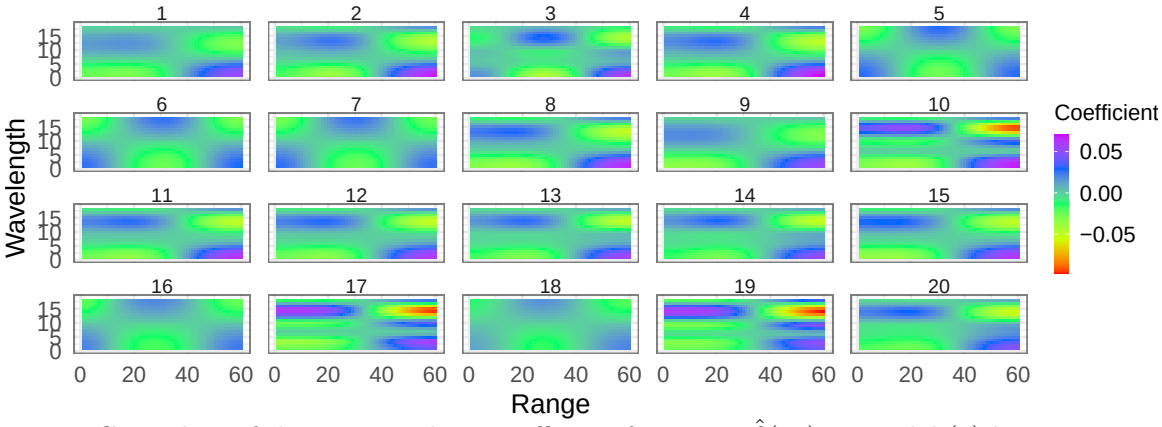

Figure G.5: Plots of the estimated 2D coefficient functions $\hat{\beta}(\cdot, \cdot)$ in Model (3) by FBLR$_{R=2}$ for LIDAR data at 20 time points. The two axes correspond to range and wavelength.

such as FLR+vec$^*(X(s,t))$ and FLR+vec$^*(X^T(s,t))$ tend to be over-smoothed for one of the two domains. This demonstrates the "stableness" of the (iterative) FBLR estimations. FLR+TPK is overly "stable" since the estimated functions do not evolve over time at all.

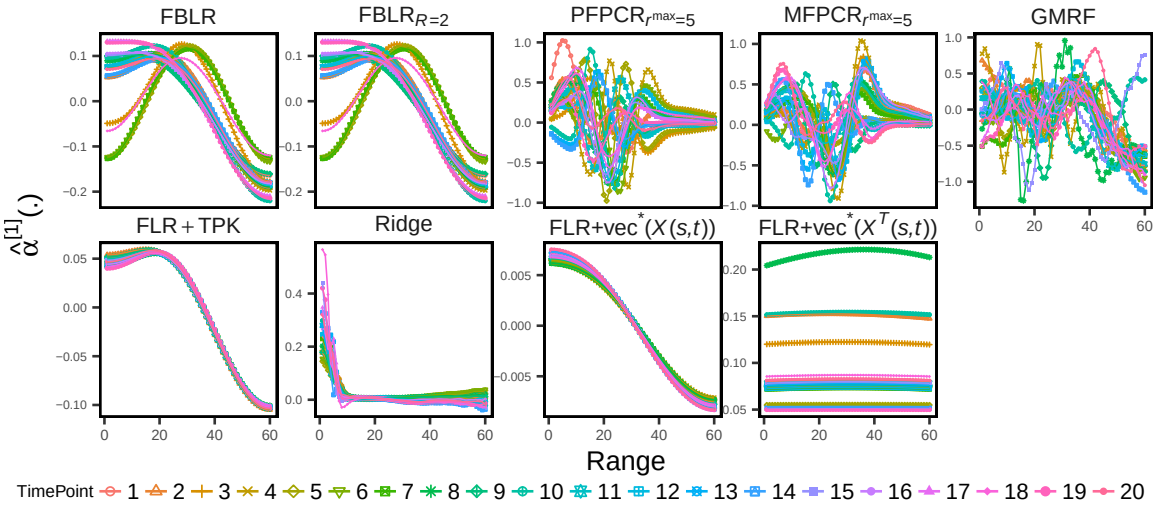

Figure G.6: Plots of the estimated 1D coefficient function $\hat{\alpha}^{[1]}(\cdot)$ that corresponds to the range domain in Model (2). 20 curves in each panel correspond to 20 time points.

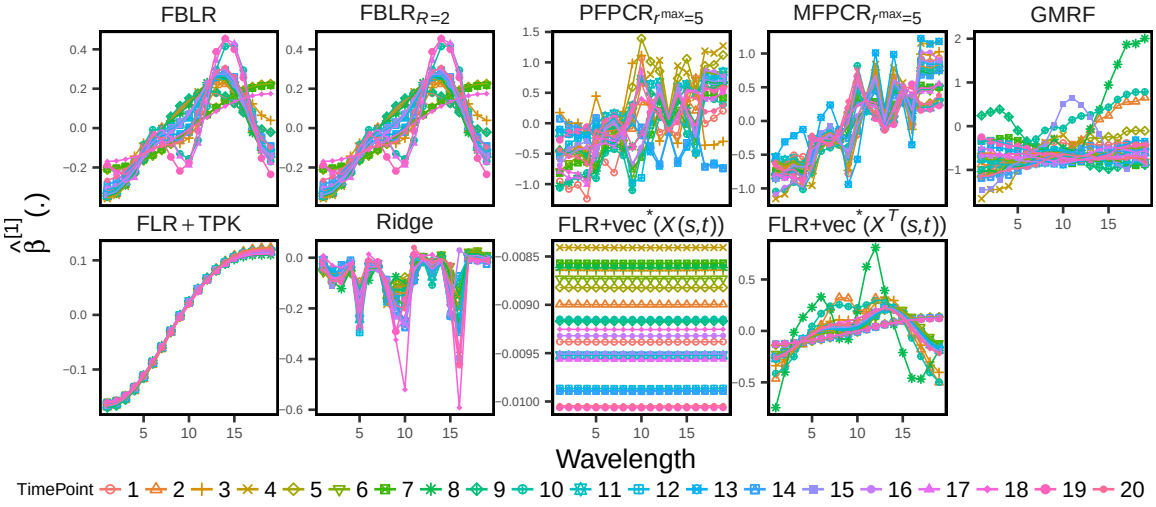

Figure G.7: Plots of the estimated 1D coefficient function $\hat{\beta}^{[1]}(\cdot)$ that corresponds to the wavelength domain in Model (2). 20 curves in each panel correspond to 20 time points.