Multivariate functional response low-rank regression with an application to brain imaging data

Xiucai DING¹, Dengdeng YU², Zhengwu ZHANG³, and Dehan KONG²*

¹Department of Statistics, University of California, Davis, Davis, CA 95616, U.S.A.
²Department of Statistical Sciences, University of Toronto, Toronto, CA M5G 1X6, Canada
³Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, U.S.A.

Key words and phrases: functional data; functional magnetic resonance imaging; low rank; sieve regression.

MSC 2010: Primary 62G08; secondary 62G20.

Abstract: We propose a multivariate functional response low-rank regression model with possible high-dimensional functional responses and scalar covariates. By expanding the slope functions on a set of sieve bases, we reconstruct the basis coefficients as a matrix. To estimate these coefficients, we propose an efficient procedure using nuclear norm regularization. We also derive error bounds for our estimates and evaluate our method using simulations. We further apply our method to the Human Connectome Project neuroimaging data to predict cortical surface motor task-evoked functional magnetic resonance imaging signals using various clinical covariates to illustrate the usefulness of our results.

1. INTRODUCTION

The advancement of neuroimaging technology has produced massive imaging data observed over both time and space, including functional magnetic resonance imaging (fMRI), electroencephalography (EEG), diffusion tensor imaging (DTI), positron emission tomography (PET), and single-photon emission-computed tomography (SPECT), among others. Scientists are often interested in characterizing the association between imaging data and clinical predictors.

Additional Supporting Information may be found in the online version of this article at the publisher’s website.

*Author to whom correspondence may be addressed.
E-mail: dehan.kong@utoronto.ca
The functional regression models are widely used to achieve this goal, for instance, the functional linear regression (Ramsay & Dalzell, 1991; Ramsay & Silverman, 2005; Yao, Müller, & Wang, 2005) and functional response regression model (Faraway, 1997; Ramsay & Silverman, 2005). We refer the readers to Wang, Chiou, & Müller (2016) for a recent review.

In this article, we are interested in predicting the blood-oxygen level-dependent (BOLD) signals obtained from different regions of interest (ROIs) of the brain using clinical covariates. Specifically, we propose the following multivariate functional response regression model

\[ Y(t) = \sum_{j=1}^{s} X_j \beta_j(t) + \epsilon(t), \quad (1) \]

where \( Y(t) = (Y_1(t), \ldots, Y_p(t))^T \in \mathbb{R}^p \) represents the BOLD signals from \( p \) ROIs where \( \beta_j(t) = (\beta_{j1}(t), \ldots, \beta_{jp}(t))^T \in \mathbb{R}^p \) for \( 1 \leq j \leq s \) represents the coefficient functions, characterizing the effect of the \( j \)th predictor \( X_j \) on the responses, and where \( \epsilon(t) \in \mathbb{R}^p \) is the random error that is independent of \( X_j \) for \( 1 \leq j \leq s \). In the current article, we assume that both \( Y(t) \) and \( X_j \) are centered and focus on the case when model (1) does not have an intercept term. Indeed, our methodology can be extended to the case when it has nonzero known mean functions. We refer the readers to our discussion in Remark 3.

A similar multivariate varying coefficient model (MVCM) (Zhu et al., 2011; Zhu, Li, & Kong, 2012) has been studied to delineate the association between multiple diffusion properties along major white matter fiber bundles with a set of covariates of interest. They assume that the error term \( \epsilon(t) \) can be decomposed into two independent terms, where the first term depicts the error correlations between two time points and the second term depicts the individual curve variation. Under this assumption, they proposed a weighted least-squares procedure based on a local polynomial kernel-smoothing technique (Fan & Gijbels, 1996) to estimate the coefficient functions \( \beta_j(t) \). They also employed the functional principal component analysis to delineate the structure of the variability in fiber bundle diffusion properties.

There are several key differences between our proposal and the MVCM. First, the task-evoked fMRI data often have a nonstationary nature (Jones et al., 2012). Motivated by this perspective, unlike the MVCM, our error process can cover a wide range of nonstationary processes. Second, in neuroimaging studies, the dimensions of the responses and covariates can be quite large. In the present article, we allow the dimensions to be divergent from the sample size, while the MVCM considers the case when the dimensions are fixed. Third, the MVCM uses a local kernel smoothing method to estimate the coefficient function, which can be computationally slow as it needs to estimate the coefficient functions in a pointwise manner. To overcome this computational difficulty, our method employs the state-of-art sieve regression, which utilizes the global information among all the time points. By imposing a low-rank structure of the coefficient matrix, our proposal can obtain a global fit of the coefficient curves, which significantly improves the computational efficiency.

The rest of the article is organized as follows. In Section 2, we introduce our multivariate functional response low-rank regression model and propose a low-rank estimation procedure with an efficient algorithm. Section 3 investigates the theoretical properties of our method. Simulations are conducted in Section 4 to evaluate the finite-sample performance of the proposed approach. Section 5 illustrates an application of our method using data from the Human Connectome Project. We end with some discussion in Section 6. Technical proofs and additional simulation and real data results are given in the Appendix and the Supplementary Material.

DOI: 10.1002/cjs.11604 The Canadian Journal of Statistics / La revue canadienne de statistique
2. MODEL SETUP AND ESTIMATION PROCEDURE

Denote \( \{y_i(t), (x_{ij}, 1 \leq j \leq s), \epsilon_i(t) : i = 1, \ldots, n \} \) independent and identically distributed (i.i.d.) realizations from the population \( \{Y(t), (X_j, 1 \leq j \leq s), \epsilon(t)\} \) generated from the model (1). Without loss of generality, we assume \( t \in [0, 1] \). In practice, we cannot observe the entire trajectories of \( \{y_i(t)\} \). Instead, we can collect intermittent measurements \( \{y_i(t_k)\} \) for \( 0 \leq t_1 \leq t_2 \leq \ldots \leq t_T \leq 1 \) for each \( i \), where \( T \in \mathbb{N} \) is the number of time points. In this article, we assume that each subject is observed at the same time points \( t_1, \ldots, t_T \), and this assumption is valid for our fMRI data. In light of model (1), we can write

\[
y_i(t_k) = \sum_{j=1}^{s} x_{ij} \beta_j(t_k) + \epsilon_i(t_k), \quad k = 1, 2, \ldots, T, \; i = 1, 2, \ldots, n. \tag{2}
\]

We are interested in estimating the coefficient functions \( \{\beta_j(t), 1 \leq j \leq s\} \). In the current article, we assume that the functions \( \beta_j(t) \) are smooth in \( t \), which is a realistic assumption for fMRI data. We also allow \( s, p, \) and \( T \) to diverge with the sample size \( n \).

To estimate \( \beta_j(t) \), we approximate \( \beta_j(t) \) using sieve expansion (Chen, 2007). Examples of sieve bases include trigonometric series, orthogonal polynomials, and the orthogonal wavelet bases. In particular, according to Section 2.3 of Chen (2007), we have

\[
\beta_j(t) \approx \sum_{h=1}^{c} M_{jl,h} b_h(t) + \sum_{h=c+1}^{\infty} M_{jl,h} b_h(t) \approx \sum_{h=1}^{c} M_{jl,h} b_h(t), \tag{3}
\]

where \( \{b_h(t)\}_{h=1}^{\infty} \) is a set of prechosen sieve basis functions, \( \{M_{jl,h} : 1 \leq j \leq s, 1 \leq l \leq p, 1 \leq h \leq c\} \) are coefficients to be estimated, and \( c \) is the truncation number of sieve basis functions. For simplicity, we use the same \( c \) for all \( 1 \leq j \leq s \) and \( 1 \leq l \leq p \).

Plugging (3) into (2), we obtain the approximation

\[
y_{il}(t_k) \approx \sum_{j=1}^{s} x_{ilj} \sum_{h=1}^{c} M_{jl,h} b_h(t_k), \tag{4}
\]

for \( k = 1, 2, \ldots, T, \; l = 1, 2, \ldots, p, \; i = 1, 2, \ldots, n \). Based on this approximation, the estimation of \( \{\beta_j(t), 1 \leq j \leq s\} \) boils down to estimating \( M_{jl,h} \).

Let \( Y_i \in \mathbb{R}^{p \times T} \) with \( lk \)th entry \( y_{il}(t_k) \), and \( E_i \in \mathbb{R}^{p \times T} \) with \( lk \)th entry \( \epsilon_{il}(t_k) \). Let \( \otimes \) be the Kronecker product. We define

\[
X_i = x_i \otimes B \in \mathbb{R}^{sc \times T}, \tag{5}
\]

where \( x_i = (x_{i1}, \ldots, x_{is})^T \in \mathbb{R}^s \) and \( B \in \mathbb{R}^{c \times T} = (b(t_1), \ldots, b(t_T)) \) with \( b(t) = (b_1(t), \ldots, b_p(t))^T \in \mathbb{R}^c \). Furthermore, denote \( M_j \in \mathbb{R}^{p \times c} \), whose entry satisfies \( (M_j)_{lh} = M_{jl,h} \), \( 1 \leq l \leq p, \; 1 \leq h \leq c \), and \( M = (M_1, M_2, \ldots, M_s) \in \mathbb{R}^{p \times sc} \). One can rewrite (4) as a matrix form

\[
Y_i \approx MX_i + E_i, \; i = 1, 2, \ldots, n. \tag{6}
\]

The model (6) is a multivariate response linear regression model, and the parameter of interest is the coefficient matrix \( M \).

The conventional approach to estimate \( M \) is ordinary least squares (OLS). However, OLS may perform suboptimally as it does not utilize the information that the entries of \( Y_i \) are related, especially when both \( p \) and \( T \) diverge with the sample size \( n \). Recently, Yuan et al. (2007), and Chen, Dong, & Chan (2013) proposed reduced rank regression models by assuming the low
rankness of $M$. They introduced nuclear norm penalized regression methods to estimate $M$, which can achieve parsimonious models with enhanced interpretability. The low-rank assumption has been commonly used in neuroimaging applications; see Zhu et al. (2014); Zhou & Li (2014); Kong et al. (2020); Yu et al. (2020); Hu et al. (2020b); Hu et al. (2020a) for examples. In the current article, we also assume that $M$ is of low rank. As we will see in the discussion in Section 3, the low-rank assumption of $M$ indicates that $\beta_i$, $1 \leq i \leq s$ can be viewed as a finite linear combination of dynamic factors.

In particular, we solve

$$\min_M \left[ \frac{1}{2nT} \sum_{i=1}^n \text{Tr} \left\{ (Y_i - MX_i)(Y_i - MX_i)^T \right\} + \lambda \|M\|_s \right],$$

(7)

where $\|M\|_s$ is the nuclear norm, defined as the summation of all the singular values of $M$, and where $\lambda$ is a tuning parameter.

If we let $\mathcal{Y} = \{Y_1^T, \ldots, Y_n^T\} \in \mathbb{R}^{nT \times p}$, $\mathcal{X} = \{X_1^T, \ldots, X_n^T\} \in \mathbb{R}^{nT \times sc}$, the optimization problem (7) is equivalent to

$$\min_M \left( \frac{1}{2nT} \|\mathcal{Y} - \mathcal{X} M^T\|_F^2 + \lambda \|M\|_s \right).$$

(8)

Denote the solution of (8) as $\hat{M}$. It is easy to see that the estimate of the coefficient function can be written as $\hat{\beta}_j(t) = \sum_{b=1}^c \hat{M}_{jb}(t) t^b$.

The optimization problem (8) can be solved by the proximal gradient algorithm. For simplicity, we use $D = M^T$. Let $\mathcal{L}(D) = \frac{1}{2nT} \|\mathcal{Y} - \mathcal{X} D\|_F^2$ and $P(D) = \lambda \|D\|_s$. The objective function $Q(D)$ can be decomposed as $Q(\cdot) = \mathcal{L}(\cdot) + P(\cdot)$. Define $\nabla \mathcal{L}(S^{(i)}) = \nabla \|\mathcal{Y} - \mathcal{X} S^{(i)}\|_F^2 = 2 \mathcal{X}^T (\mathcal{X} S^{(i)} - \mathcal{Y})$. We utilize Nesterov’s gradient descent method (Beck & Teboulle, 2009; Nesterov, 2013) to solve (8). In particular, we propose the following algorithm:

Algorithm 1.

1. **Initialize:** $D^{(0)} = D^{(1)}$, $a^{(0)} = 0$ and $a^{(1)} = 1$, $\delta = 1 / \lambda_{\max}(\mathcal{X}^T \mathcal{X})$.
2. **Repeat:**
   i. $S^{(t)} = D^{(t)} + \frac{a^{(t-1)} - 1}{a^{(t)}} (D^{(t)} - D^{(t-1)})$;
   ii. $A_{\text{temp}} = S^{(t)} - \delta \nabla \mathcal{L}(S^{(t)})$;
   iii. **Singualar value decomposition:** $A_{\text{temp}} = U \text{diag}(a)V^T$;
   iv. $d = (a - \delta \cdot 1)_+$;
   v. $D^{(t+1)} = U \text{diag}(d)V^T$;
   vi. $a^{(t+1)} = \left[ 1 + \sqrt{1 + (2a^{(t)})^2} \right] / 2$;
3. **Until objective function $Q(D^{(t)})$ converges.**

From steps ii to v, the gradient descent is based on the first-order approximation to the loss function $\mathcal{L}$ at the current search point $S^{(t)}$. Specifically,

$$g(D|S^{(t)}, \delta) = \mathcal{L}(S^{(t)}) + \langle \nabla \mathcal{L}(S^{(t)}), D - S^{(t)} \rangle + \frac{1}{2\delta} \|D - S^{(t)}\|_F^2 + P(D),$$

$$= \frac{1}{2\delta} \|D - (S^{(t)} - \delta \nabla \mathcal{L}(S^{(t)}))\|_F^2 + P(D) + c^{(t)},$$

DOI: 10.1002/cj.11604
where $c^{(t)}$ collects the term irrelevant to the optimization, and the constant $\delta$ is chosen such that the relation between the surrogate and target functions always holds: $g(D|S^{(t)}, \delta) \geq Q(D)$. We set $\delta$ as a Lipschitz constant for $\nabla L(\cdot)$ with $\delta = 1/\lambda_{\text{max}}(\lambda^T \lambda)$. A solution to the surrogate optimization problem is given by Proposition 1 of Zhou & Li (2014). Singular value decomposition is performed on the intermediate matrix $A_{\text{temp}} = S^{(t)} - \delta \nabla L(S^{(t)})$. The next iteration $D^{(t+1)}$ shares the same singular vectors as $A_{\text{temp}}$, and its singular values $d^{(t+1)}$ are determined by minimizing $\frac{1}{2n} || d - a ||_F^2 + f(d)$, where $a = \sigma(A_{\text{temp}})$. For the nuclear norm regularization $f(d) = \lambda \sum |b_j|$, the solution is given by soft thresholding the singular values $b_j^{(t+1)} = (a_j - \lambda \delta)_+$ as suggested by Corollary 1 of Zhou & Li (2014).

There are two tuning parameters involved in our estimation procedure, the truncation number $c$ and the regularization parameter $\lambda$. In this article, we use fivefold cross validation to select the optimal values based on two-dimensional grid search.

3. THEORETICAL RESULTS

We begin with some notations. Define $z = (z_1, \ldots, z_d)^T$ as a subgaussian random vector with some parameter $\sigma > 0$ if for all $\alpha \in \mathbb{R}^s$,

$$\mathbb{E} \left[ \exp(\alpha^T z) \right] \leq \exp(||\alpha||_2^2 \sigma^2/2).$$

We next introduce the locally stationary time series. Consider the time series (Zhou & Wu, 2009, 2010)

$$z_i = G \left( \frac{i}{n}, F_i \right),$$

(9)

Here, $F_i = (\eta_M, \eta_{M+1}, \ldots, \eta_{l-1}, \eta_l)$ and $\eta_i$, $i \in \mathbb{Z}$ are i.i.d. random variables, where $M$ is any arbitrarily small negative integer that is independent of $i$, and $G : [0, 1] \times \mathbb{R}^\infty \to \mathbb{R}$ is a measurable function such that $\xi_i(t) := G(t, F_i)$ is a properly defined random variable for all $t \in [0, 1]$. We introduce the following dependence measure to quantify the temporal dependence of (9).

Definition 1. Let $\{\eta'_i\}$ be an i.i.d. copy of $\{\eta_i\}$. We assume that, for some $q > 2$, $||x||_q < \infty$, where $|| \cdot ||_q = [\mathbb{E} | \cdot |^q]^{1/q}$ is the $L_q$ norm of a random variable. For $j \geq 0$, we define the physical dependence measure by

$$\delta(j, q) := \sup_{t \in [0, 1]} \max_{i} \| G(t, F_i) - G(t, F_{i,j}) \|_q,$$

(10)

where $F_{i,j} := (F_{i,j-1}, \eta'_{i-j}, \eta_{i-j+1}, \ldots, \eta_l)$.

The measure $\delta(j, q)$ quantifies the changes in the system’s output when the input of the system $j$ steps before $\eta_{i-j}$ is changed to an i.i.d. copy. If the change is small, then we have short-range dependence.

With the above notation, we introduce the assumptions for the theoretical development.

Assumption 1. We assume that $x_i, i = 1, 2, \ldots, n$, are i.i.d. centered sub-Gaussian random vectors independent of $\epsilon_{i}(t_k), i = 1, 2, \ldots, n$, for all $k = 1, 2, \ldots, T$. Moreover, we assume that, for each $i = 1, 2, \ldots, n$, and $l = 1, 2, \ldots, p$, $\{\epsilon_{j}(t_k)\}_{k=1}^T$ is a centered locally stationary time series of the form (9). Finally, for some large $q, \gamma > 0$, there exists some universal constant $C > 0$, such that

$$\delta(j, q) \leq C j^{-\gamma}, \quad j \geq 1.$$

(11)
We note that the assumption (9) represents a wide class of stationary, locally stationary linear, and nonlinear processes (Zhou & Wu, 2009, 2010). As we mentioned earlier, previous works Zhu et al. (2011); Zhu, Li, & Kong (2012) focus on fitting the coefficients of functional regression locally. Hence, they do not need to consider the temporal relation for the underlying stochastic process. In contrast, our estimation relies on (4), which utilizes the global information for all the time points. A natural assumption is the short-range temporal dependence, that is, (11), which requires the temporal correlation between the process $\epsilon_i(\cdot)$ to have a polynomial decay. Moreover, as a technical by-product, we only require the existence of a second moment of $\epsilon_i(\cdot)$. This improves the assumption of a finite fourth moment in Zhu, Li, & Kong (2012).

Finally, we mention that (11) can be satisfied by many stochastic processes, for instance, the Ornstein-Uhlenbeck process and the linear process $\epsilon_i(t) = \sum_{k=1}^{\infty} a_{k,i} b(t_k)$, where $\{b_i\}$ are independent standard Gaussian random variables and $\sup_t |a_{k,i}(t)|^2 \leq C k^{-\gamma}$ for some constant $C > 0$.

We also utilize the following assumption on the smoothness of $\beta_{jl}(\cdot)$s.

**Assumption 2.** For $j = 1, 2, \ldots, s, l = 1, 2, \ldots, p, \beta_{jl}(\cdot)$s are smooth functions of time such that $\beta_{jl}(\cdot) \in C^d([0, 1])$, where $C^d([0, 1])$ is the function space on $[0, 1]$ of continuous functions that have continuous first $d$ derivatives.

Using Assumption 2, $\beta_{jl}(t)$ can be well approximated by sieve expansion (Chen, 2007). Specifically, in light of (3), we find that

$$\beta_{jl}(t) = \sum_{h=1}^{c} M_{j,l,h} b_{h}(t) + O(c^{-d}), \quad j = 1, 2, \ldots, s,$$

where the error $O(c^{-d})$ is entrywise. Plugging (12) into (2), with high probability, we have

$$y_i(t_k) = \sum_{j=1}^{s} x_{ij} M_{j} b(t_k) + e_i(t_k) + O(sc^{-d}).$$

When the error term $O(sc^{-d})$ is negligible, we can approximate $\beta_{jl}(t_k)$ using

$$\tilde{\beta}_{jl}(t_k) = M_{j} b(t_k).$$

For a rigorous justification, we refer the readers to Theorem 1 and Corollary 1 and their proofs.

Recall that $Y_i \in \mathbb{R}^{p \times T}$ is a matrix with $lk$th entry $y_{il}(t_k)$ and $E_i \in \mathbb{R}^{p \times T}$ with $lk$th entry $e_i(t_k)$. We can write the model (2) as follows:

$$Y_i = MX_i + E_i + o(1), \quad i = 1, 2, \ldots, n,$$

where $M$ is defined under equation (5). Therefore, our estimation problem boils down to estimating the coefficient matrix $M$.

**Remark 1.** As $\tilde{\beta}_{jl}(t)$ can approximate $\beta_{jl}$ well, we now connect the structure of $\tilde{\beta}_{jl}(t)$ with the matrix $M$ to show that $\beta_{jl}$ will have a dynamic factor model structure when $M$ is approximately
of low rank. Note that $M$ is a rectangular matrix stacking $M_j \in \mathbb{R}^{p \times c}, j = 1, 2, \ldots, s$. For each $j = 1, 2, \ldots, s$, we write the singular value decomposition of $M_j$ as

$$M_j = \sum_{l=1}^{\min\{p,c\}} \sigma_l u_l v_l^\top,$$

where $\{\sigma_l\}, \{u_l\}$, and $\{v_l\}$ are the singular values, left singular vectors, and right singular vectors of $M_j$, respectively. Consequently, we find that

$$\tilde{\beta}_j(t_k) = \sum_{l=1}^{\min\{p,c\}} \sigma_l(v_l^\top b(t_k))u_l.$$

If we further denote $\alpha_l(t_k) := \sigma_l(v_l^\top b(t_k))$, then $\tilde{\beta}_j(t_k)$ can be further written as

$$\tilde{\beta}_j(t_k) = \sum_{l=1}^{\min\{p,c\}} \alpha_l(t_k)u_l.$$

This implies that $\tilde{\beta}_j(t_k)$ is a time-varying linear combination of the basis $\{u_l\}$, which we can consider a dynamic factor model. In the current article, we follow the common low-rank assumption in the literature of approximate factor models Bai & Ng (2002) and assume that only a few of $\{u_l\}$ are useful for our estimation and prediction. As a result, $M_j$ is of a low-rank structure.

Suppose that the rank of $M_j$ is $r_j, j = 1, 2, \ldots, s$. As

$$\text{rank}(M) \leq \sum_{j=1}^{s} r_j,$$

and $s$ is slowly divergent, we can assume that $M$ is of an approximately low-rank structure. This is formally stated in Assumption 3.

Denote

$$\xi := \sup_{1 \leq h \leq c} \sup_{t \in [0,1]} |b_h(t)|. \quad (16)$$

As mentioned in Section 4.2 of Ding & Zhou (2020), $\xi$ can be well controlled for the commonly used sieve basis functions. For instance, $\xi = O(1)$ for the trigonometric series and orthogonal polynomials, and $\xi = O(\sqrt{c})$ for the orthogonal wavelet basis.

**Assumption 3.** We assume that $M$ is of an approximately low-rank structure, that is, there exists a constant $\kappa > 0$ such that

$$\sum_{i=1}^{\min\{p,c\}} \sigma_i(M) \leq \kappa,$$

where $\sigma_i(M), i = 1, 2, \ldots, \min\{p,c\}$ are the singular values of $M$. Moreover, we assume that for any arbitrarily small constant $\tau > 0$,

$$\frac{\sqrt{r}\xi p n^\tau}{csT} = o(1), \quad \text{where } r = \text{rank}(M). \quad (17)$$
Remark 2. The assumption in (17) is mild. Denote
\[ c = O(n^{\alpha_1}), T = O(n^{\alpha_2}), p = O(n^{\alpha_3}), s = O(n^{\alpha_4}). \] (18)
If we choose the trigonometric series or orthogonal polynomials, (17) reads as
\[ \mu^{1/2} n^{\alpha_3 - \alpha_2/2 + \tau - \alpha_1 - \alpha_4} = o(1). \]
In other words, when the true rank \( r \) is finite, and \( p = O(sc) \), we only need to have \( O(n^{2\tau}) \) time points observed from the stochastic process \( \epsilon_i(\cdot) \).

To guarantee the consistency of the estimation, we require the following assumption on the parameters.

Assumption 4. We assume that
\[ \xi p s^2 n^{2\tau} c^{-d} = o(1). \] (19)
Assumption 4 is mild. When we use the trigonometric series or orthogonal polynomials and assume \( \beta_j(\cdot) \) is infinitely differentiable, (19) will always hold, and we can allow \( p s^2 \) to diverge fast. In our article, we require our sieve bases to satisfy (17) and (19) and belong to \( C^d([0, 1]) \) defined in Assumption 2. Note that the parameter \( \xi \) in (17) and (19) is directly related to the sieve bases via (16). Indeed, all the sieve bases listed in Section 2.3 of Chen (2007) satisfy these assumptions, for instance, Fourier basis, orthogonal polynomials, Daubenchies orthogonal wavelets, and splines.

Finally, we introduce the following assumption to guarantee that the covariance matrix of \( x_i \) is regular. We will see later that the following condition is a sufficient condition for the restricted strong convexity condition (c.f. Definition 3).

Assumption 5. Denote \( \Sigma_s \) as the covariance matrix of \( x_i \). We assume that \( \Sigma_s \) is bounded, and there exists some constant \( \delta > 0 \) such that
\[ \lambda_{\min}(\Sigma_s) \geq \delta, \]
where \( \lambda_{\min}(\Sigma_s) \) is the smallest eigenvalue of \( \Sigma_s \).

Armed with the above assumptions, we now present our main result. Denote \( \lambda_n \) as the regularization parameter of the optimization problem (8) and \( M^* \) as the true value of \( M \). Recall that \( \hat{M} \) is the solution of (8), and we have the following result. It can be seen that, even though our approach may be suboptimal, it can achieve consistency under mild conditions.

Theorem 1. Suppose Assumptions 1–5 hold. For any given arbitrarily small constant \( \tau > 0 \) defined in Assumption 3, when both \( n \) and \( T \) are large enough, there exists some \( C_q > 0 \) depending on \( q \) in Assumption 1, with probability at least \( 1 - C_q n^{-\mu \tau} \); we have for some constants \( C, C_1 > 0 \), when \( \lambda_n \geq C_1 p \xi n^{\tau} T^{-1/4} \),
\[ \|M^* - \hat{M}\|_F \leq C \left( \frac{\sqrt{p \xi n^{\tau}}/2}{\sqrt{\text{tr}(\Sigma_s)} T^{1/4}} + \sqrt{\frac{rp \xi n^{\tau}}{\text{tr}(\Sigma_s) \sqrt{T}}} \right). \] (20)
One thing to note here is that \( T \) also diverges with the sample size \( n \). As the role of \( \tau \) is to control the probability and it can be arbitrary, we can obtain a consistent estimator for
a reasonably large $T$. Specifically, in the setting of (18), our estimator is consistent under Assumption 4.

**Remark.** In our theoretical development, we borrow the idea of the regularized M-estimator developed in Negahban et al. (2012). However, one main challenge of our proof is that we need to account for the approximation errors caused by truncation of the basis expansion in (3).

Next, we provide some insights of the above results when we use either the trigonometric series or the orthogonal polynomials. From Assumption 5, we have $\text{tr}(\Sigma_s) = O(s)$. Hence, by Assumption 3, the second term of the right-hand side of (20) is of the order $o(1)$. On one hand, if the second term of the right-hand side of (20) dominates the first one,

$$\sqrt{rp\xi n^r \text{tr}(\Sigma_s)T^{1/4}},$$

which implies $\sqrt{r\sqrt{pn^{r/2}}} > \sqrt{csT^{1/4}}$. If we further let the matrix be of a fixed rank $r$ and the functions $\hat{\beta}_j(\cdot)$ be infinitely differentiable such that $c$ can be chosen at an order of $\log T$, then we obtain an upper bound for number of time points as follows:

$$T \ll n^{2r} \left( \frac{P}{s} \right)^2.$$

On the other hand, when the first term of the right-hand side dominates, we need to have

$$\frac{\sqrt{pn^{r/2}}}{\sqrt{csT^{1/4}}} = o(1),$$

which basically requires that

$$T \gg n^{2r} \left( \frac{P}{s} \right)^2.$$

In this sense, the choice of $T$ will not significantly affect the consistency of our estimators. As our assumptions are mild as explained in Section 2, once $T$ and $n$ are reasonably large, we obtain a consistent estimator.

Let $\hat{M} = (\hat{M}_1, \ldots, \hat{M}_s)$. Denote the estimator

$$\hat{\beta}_j(t) = \hat{M}_j b(t), \quad 1 \leq j \leq s, \quad t \in [0, 1].$$

We now state the convergence result for the coefficient functions.

**Corollary 1.** Suppose that the assumptions of Theorem 1 hold. Then, for $1 \leq j \leq s$ and some universal constant $C > 0$, with probability of at least $1 - C_q n^{-\eta r}$, we have

$$\sup_t \|\beta_j(t) - \hat{\beta}_j(t)\| \leq C \sqrt{c} \left( \frac{\sqrt{p\xi n^{r/2}}}{\sqrt{\text{tr}(\Sigma_s)T^{1/4}}} + \frac{\sqrt{rp\xi n^r}}{\sqrt{\text{tr}(\Sigma_s)\sqrt{T}}} \right).$$

Compared to Theorem 1, we obtain an extra $\sqrt{c}$ factor in Corollary 1 as our estimate involves the sieve basis functions. Similarly, we can obtain a consistent estimator when both $T$ and $n$ are reasonably large.
Remark 3. It is remarkable that, in the high-dimensional setting, it is not trivial to center the high-dimensional responses. However, in many applications, we can assume that there exists a time-varying mean function for each \( Y_i(t) \), \( 1 \leq l \leq p \). Specifically, we can assume that, for some functions, \( m_l(\cdot) \in C^d([0, 1]) \) such that

\[
E(Y_i(t)) = m_l(t), \quad 1 \leq l \leq p.
\]

In this setting, we can rewrite our model (1) as

\[
Y(t) - m(t) = \sum_{j=1}^{s} X_j \beta_j(t) + \epsilon(t),
\]

where \( m(t) = (m_1(t), \ldots, m_p(t)) \). Moreover, if we set \( \beta_0(t) = m(t) \) and \( X_0 = 1 \), we can further write

\[
Y(t) = \sum_{j=0}^{s} X_j \beta_j + \epsilon(t).
\]

As \( m_l(t) \in C^d([0, 1]), l = 1, 2, \ldots, p \), we can expand them on a set of basis functions and \( m_l(t) \approx \sum_{h=1}^{c} k_{lh} b_h(t) \). Therefore, we can apply our current methodology to estimate the coefficients.

4. SIMULATIONS

In the section, we perform simulation studies to evaluate our method. We consider a set of Fourier bases

\[
b_j(t) = \begin{cases} 
1, & \text{if } j = 1; \\
\sqrt{2} \sin(\pi j t), & \text{if } j \text{ is even}; \\
\sqrt{2} \cos(\pi (j - 1) t), & \text{Otherwise}.
\end{cases}
\]

The \( x_i \) is generated from a multivariate normal distribution with mean zero and covariance \( \Sigma \) with \( j_1 j_2 \text{th entry } \Sigma_{j_1 j_2} = 0.5 |j_1 - j_2| \) for \( 1 \leq j_1 \leq j_2 \leq s \). The \( X_i = x_i \otimes B \), where \( B \in \mathbb{R}^{c \times T} = (b(t_1), \ldots, b(t_T)) \) with \( b(t) = (b_1(t), \ldots, b_c(t))^T \in \mathbb{R}^c \). Here, we set \( t_k = \frac{k-1}{T} \) for \( k = 1, \ldots, T \).

The response is generated from

\[
Y_i = MX_i + \nu \cdot E_i, \quad i = 1, 2, \ldots, n,
\]

where \( \nu \) is a constant, and each row of \( E_i \in \mathbb{R}^{n \times T} \) is a time series with an autoregressive structure. In particular, let \( E_{ij} \in \mathbb{R}^T \) be the \( j \)th entry of \( E_i \) for \( j = 1, \ldots, p \) and \( k = 1, \ldots, T \). We set \( E_{ij}(k) = 0.3 E_{ij}((k-1)T) + \epsilon_{ij}(k) \), with \( E_{ij}(0) = 0 \) for all \( j = 1, \ldots, p \), where \( \epsilon_{ij}(k) \) is a series of i.i.d Gaussian random variables with mean 0 and variance 1. We consider \( (n, p, T, c, s) = (100, 32, 256, 4, 8) \). In this case, the matrix \( M \in \mathbb{R}^{32 \times 32} \), and we consider three different shapes for \( M \): a square shape, a T shape, and a cross shape, shown in Figure 1(a), (d), and (g), respectively. The true coefficient functions are generated from \( \beta_{jl}(t) = \sum_{h=1}^{c} M_{jl,h} b_h(t) \).

We define the signal-to-noise ratio (SNR) as

\[
\text{SNR} = \frac{\sum_{i=1}^{n} \text{Tr}(X_i^* M^* X_i)}{\sum_{i=1}^{n} \nu^2 \text{Tr}(E_i^* E_i)}.
\]

DOI: 10.1002/cjs.11604
Figure 1: Simulation results for the case $n = 100$, $p = 32$, $T = 256$, $c = 4$, and SNR = 1 from a randomly selected Monte Carlo run. The first column plots the shapes of the true $M$, the second column plots estimated $\hat{M}$ based on the Fourier basis, and the third column plots estimated $\hat{M}_{OLS}$ using OLS. The true basis is Fourier. The fitted basis is Chebyshev2.

We consider three cases SNR = 1, 5, and 10, where we change $\nu$ to obtain different SNRs.

To fit the model, we consider two sets of sieve bases. One is the same set of Fourier bases, and we evaluate the performance of our method if one can choose the basis correctly. In practice, we may not know the true basis; therefore, we also fit a different basis when applying our procedure to reflect the scenarios where the true underlying basis does not align with the fitted basis. In the simulation, we consider the Chebyshev basis of the second kind. In particular, the basis is defined as

$$b_j(t) = \begin{cases} 
2 \cdot (1 - [2(t - 1/2)]^2)^{1/4} / \sqrt{\pi}, & \text{if } j = 1; \\
2tb_1(t), & \text{if } j = 2; \\
2tb_{j-1}(t) - b_{j-2}(t), & \text{Otherwise.}
\end{cases}$$
Table 1: Simulation results for \(n = 100, p = 32, T = 256, c = 4,\) and SNR = 5 for 100 Monte Carlo runs. The mean values of MISEs \((10^{-2})\) for eight functional slope estimates are listed for the proposed method, as well as the OLS method. Their associated standard errors are reported in the parentheses. For each method, the matrices \(M \in \mathbb{R}^{32 \times 32}\) are chosen to be 32 by 32 pictures of Square, \(T,\) and Cross correspondingly. The true basis is Fourier. The fitted basis is Chebyshev2.

| MISE (sieve) | Square | \(T\) | Cross |
|--------------|--------|-------|-------|
| \(\beta_1(t)\) | 1.944 (0.049) | 1.201 (0.003) | 2.984 (0.018) |
| \(\beta_2(t)\) | 1.946 (0.045) | 0.416 (0.002) | 2.987 (0.018) |
| \(\beta_3(t)\) | 1.958 (0.048) | 0.416 (0.002) | 2.984 (0.019) |
| \(\beta_4(t)\) | 1.952 (0.048) | 0.414 (0.002) | 1.425 (0.008) |
| \(\beta_5(t)\) | 1.954 (0.048) | 0.535 (0.003) | 1.426 (0.006) |
| \(\beta_6(t)\) | 1.937 (0.045) | 1.170 (0.004) | 1.783 (0.019) |
| \(\beta_7(t)\) | 1.944 (0.049) | 1.169 (0.004) | 1.799 (0.019) |
| \(\beta_8(t)\) | 4.198 (0.067) | 1.169 (0.004) | 1.798 (0.019) |

| MISE (OLS) | Square | \(T\) | Cross |
|------------|--------|-------|-------|
| \(\beta_1(t)\) | 24.410 (1.402) | 2.075 (0.006) | 8.028 (0.020) |
| \(\beta_2(t)\) | 24.598 (1.379) | 0.931 (0.005) | 8.120 (0.022) |
| \(\beta_3(t)\) | 24.704 (1.403) | 0.935 (0.005) | 8.108 (0.023) |
| \(\beta_4(t)\) | 24.622 (1.391) | 0.938 (0.006) | 4.235 (0.015) |
| \(\beta_5(t)\) | 24.680 (1.410) | 1.450 (0.007) | 4.257 (0.017) |
| \(\beta_6(t)\) | 24.554 (1.382) | 2.260 (0.008) | 6.949 (0.021) |
| \(\beta_7(t)\) | 24.589 (1.401) | 2.267 (0.008) | 6.957 (0.024) |
| \(\beta_8(t)\) | 37.464 (2.070) | 2.230 (0.007) | 6.881 (0.021) |

For each case, we report the mean integrated squared errors (MISEs) of the estimates of \(\beta_j(\cdot) \in \mathbb{R}^p, j = 1, \ldots, s\) defined as

\[
\text{MISE}_j := \frac{1}{p} \sum_{l=1}^{p} \int_0^1 (\beta_{jl}(t) - \hat{\beta}_{jl}(t))^2 dt.
\]

We also compare with the OLSs, where we set \(\lambda = 0\) in (7) and solve the optimization problem. All the results are based on 100 Monte Carlo runs.

We include the cases (SNR = 5), where the true basis is the Fourier basis and the fitted basis is also the Fourier basis in Table 2, and the true basis is the Fourier basis with the fitted basis Chebyshev of the second kind in Table 1. For SNR = 1 and SNR = 10, the results are included in Tables S.1–S.4 in the Supplementary Material. In particular, the MISEs for eight functional slope estimates for our proposed methods are smaller than those for OLS methods. As expected, when the true basis is Fourier, fitting using Fourier basis results in better estimation accuracy (smaller MISEs) compared to using a Chebyshev basis. We have also plotted the estimated \(\hat{M}\) from one randomly selected Monte Carlo run in Figure 1 for SNR = 1 with a Fourier basis fit. From the results, we can see that our estimates can achieve much better estimation accuracy for those coefficient functions compared with OLS.
TABLE 2: Simulation results for $n = 100$, $p = 32$, $T = 256$, $c = 4$, and SNR $= 5$ for 100 Monte Carlo runs using a Fourier basis. The mean values of MISEs ($10^{-2}$) for eight functional slope estimates are listed for the proposed method, as well as the OLS method. Their associated standard errors are reported in the parentheses. For each method, the matrices $M \in \mathbb{R}^{32 \times 32}$ are chosen to be 32 by 32 pictures of Square, T, and Cross correspondingly. The true basis type is Fourier. The fitted basis is Fourier.

| MISE(sieve) | Square | $T$ | Cross |
|-------------|--------|-----|-------|
| $\beta_1(t)$ | 0.037 (0.004) | 0.020 (0.001) | 0.028 (0.001) |
| $\beta_2(t)$ | 0.038 (0.004) | 0.019 (0.001) | 0.029 (0.001) |
| $\beta_3(t)$ | 0.039 (0.005) | 0.016 (0.001) | 0.032 (0.001) |
| $\beta_4(t)$ | 0.040 (0.005) | 0.018 (0.001) | 0.032 (0.001) |
| $\beta_5(t)$ | 0.038 (0.005) | 0.020 (0.001) | 0.029 (0.001) |
| $\beta_6(t)$ | 0.040 (0.005) | 0.020 (0.001) | 0.032 (0.001) |
| $\beta_7(t)$ | 0.038 (0.005) | 0.019 (0.001) | 0.030 (0.001) |
| $\beta_8(t)$ | 0.051 (0.007) | 0.018 (0.000) | 0.029 (0.001) |

| MISE(OLS) | Square | $T$ | Cross |
|-----------|--------|-----|-------|
| $\beta_1(t)$ | 0.659 (0.130) | 0.089 (0.002) | 0.139 (0.003) |
| $\beta_2(t)$ | 0.714 (0.130) | 0.109 (0.002) | 0.175 (0.004) |
| $\beta_3(t)$ | 0.702 (0.132) | 0.109 (0.002) | 0.175 (0.004) |
| $\beta_4(t)$ | 0.713 (0.134) | 0.114 (0.002) | 0.179 (0.004) |
| $\beta_5(t)$ | 0.708 (0.132) | 0.109 (0.002) | 0.174 (0.003) |
| $\beta_6(t)$ | 0.717 (0.135) | 0.107 (0.002) | 0.176 (0.004) |
| $\beta_7(t)$ | 0.709 (0.133) | 0.104 (0.002) | 0.177 (0.004) |
| $\beta_8(t)$ | 0.890 (0.195) | 0.084 (0.002) | 0.137 (0.003) |

In addition, we also perform a simulation study where the true basis is the Chebyshev basis of the second kind, defined previously in this study. The results of fitting our method using both Chebyshev basis of the second kind and Fourier basis are included in Tables S.5–S.10 in the Supplementary Material. The findings are similar.

When the fitted basis and the true basis align with each other, we also report the average number of bases selected using fivefold cross validation in Table 3. As observed in the results, if we know the true basis, cross validation can select the right number of bases for most scenarios. When SNR increases, the average number of bases selected is closer to the truth ($c = 4$).

To investigate how the truncation number $c$ affects the performance of the proposed method, we add a simulation study, where four Fourier bases are used to generate the data, but we fit our model by setting $c = 6$, and the $\lambda$ is still chosen by fivefold cross validation. Compared with the result where ($c, \lambda$) are chosen by fivefold cross validation, we find that the rank of the estimated $\hat{M}$ becomes smaller. The results are as shown in Tables S.11–S.13 in the Supplementary Material. Taking SNR $= 5$ for example (Table S.12), the average ranks of estimated $\hat{M}$ when $c = 6$ (1.00, 2.00, and 2.00) are smaller than the average ranks of $\hat{M}$ when $c$ is determined by cross validation (9.79, 13.41, and 13.56). However, the MISEs of the estimated coefficient functions using $c = 6$ are actually greater than the MISEs of the estimated functional slopes from using cross validation, which shows the importance of using cross validation to select the truncation number $c$. 

The Canadian Journal of Statistics / La revue canadienne de statistique DOI: 10.1002/cjs.11604
Table 3: Simulation results for $n = 100$, $p = 32$, $T = 256$, and $c = 4$ for 100 Monte Carlo runs. The average number of bases selected, as well as the oracle number of bases, is listed for each of basis type; Fourier or Chebyshev2; and SNR 1, 5, or 10. Their associated standard errors are reported in the parentheses.

| SNR | Basis        | Square  | $T$          | Cross        | ORACLE |
|-----|--------------|---------|--------------|--------------|--------|
| 1   | Fourier      | 5.190 (0.049) | 4.000 (0.000) | 4.000 (0.000) | 4      |
| 5   | Fourier      | 4.220 (0.052) | 4.000 (0.000) | 4.000 (0.000) | 4      |
| 10  | Fourier      | 4.100 (0.030) | 4.000 (0.000) | 4.000 (0.000) | 4      |
| 1   | Chebyshev2   | 5.210 (0.050) | 4.000 (0.000) | 4.000 (0.000) | 4      |
| 5   | Chebyshev2   | 4.270 (0.057) | 4.000 (0.000) | 4.000 (0.000) | 4      |
| 10  | Chebyshev2   | 4.120 (0.036) | 4.000 (0.000) | 4.000 (0.000) | 4      |

In addition, we perform a simulation study, where an equivalent basis (Chebyshev2) is used in the fitting. The results are included in Table S.14 in the Supplementary Material. We find that the average ranks of estimated $\hat{M}$ are smaller than the average ranks of $\hat{M}$ when the true basis (Fourier) is used. Taking SNR = 5 (Table S.14) for example, the average ranks of estimated $\hat{M}$ when the equivalent basis is used (1.00, 2.00, and 2.00) are smaller than the average ranks of $\hat{M}$ when $c$ is determined by cross validation (9.79, 13.41, and 13.56). We have found that the MISEs obtained by fitting using the Chebyshev2 basis are still reasonably small, which shows the robustness of the proposed method when an equivalent basis is used.

To mimic the case where the coefficient functions lie in an infinite dimensional space, we add an additional simulation study with a modified setting $(n, p, T, c, s) = (100, 32, 256, 50, 4)$, where the coefficient function $\beta_{jl}(t)$s are generated from 50 basis functions such that $\beta_{jl}(t) = \sum_{h=1}^{8} M_{jl,h} \omega_{h} b_{h}(t) + \sum_{h=9}^{50} w_{h} b_{h}(t)$, where $\omega_{1} = 1$, $\omega_{2} = 0.8$, $\omega_{3} = 0.6$, $\omega_{4} = 0.5$, and $\omega_{h} = 8(h - 2)^{-4}$ for $h \geq 5$. We consider three cases SNR = 1, 5, and 10. As shown in Table S.15 of the Supplementary Material, the number of bases selected by cross validation is much smaller than 50 due to the decay of $\omega_{h}s$. Taking SNR = 5 for example, when the Fourier basis is used, the average number of bases selected for the T shape is 4.790, with a standard error 0.041. We include the cases (SNR = 5) where the true basis is the Fourier basis and fit our method using the Fourier basis in Table S.26 and the Chebyshev basis of the second kind in Table S.23 in the Supplementary Material. For SNR = 1 and SNR = 10, the results are included in Tables S.22, S.24, S.25, and S.27 in the Supplementary Material. In particular, the MISEs for four functional slope estimates for our proposed methods are smaller than those for OLS methods. As expected, when the true basis is Fourier, fitting using a Fourier basis results in better estimation accuracy (smaller MISEs) compared with using a Chebyshev basis.

In addition, we also perform simulation studies where the true basis is the Chebyshev basis of the second kind, previously defined in this study, and we fit our method using both a Chebyshev basis of the second kind and a Fourier basis. The results are included in Tables S.16–21 in the Supplementary Material. The findings are similar.

5. REAL DATA APPLICATIONS

We apply our method to the cortical surface motor task-related fMRI data from the Human Connectome Project (HCP) Dataset (https://www.humanconnectome.org/). We include the data usage acknowledgement in Section S.1 of the Supplementary Material. We use the 900-subject release that includes behavioural and 3 Tesla magnetic resonance imaging data from 970 healthy adult participants collected from 2012 to spring 2015. We focus on the 845 subjects who had the
FIGURE 2: Real data results: The scree plot for the first 10 singular values for the estimated $\hat{M}$ (red solid) and $\hat{M}_{OLS}$ using OLS (black dashed). The fitted basis is Fourier.

FIGURE 3: Real data results: Panels (a)–(d) plot, for the left superior frontal region, the estimated $\hat{\beta}_j(t)$ for $s = 4$ motor instrument covariates: “Endurance-AgeAdj,” “GaitSpeed-Comp,” “Dexterity-AgeAdj,” and “Strength-AgeAdj.” The fitted basis is Fourier.
Figure 4: Real data results: Panel (a)–(d) plot, for the right superior frontal region, the estimated \( \hat{\beta}_j(t) \) corresponding to four motor instrument covariates: “Endurance-AgeAdj,” “GaitSpeed-Comp,” “Dexterity-AgeAdj,” and “Strength-AgeAdj,” respectively. The fitted basis is Fourier.

cortical surface motor task-related fMRI data. This task was adapted from the one developed by Buckner and colleagues (Buckner et al., 2011; Yeo et al., 2011).

In the motor task, participants are presented with visual cues that ask them to either tap their left or right fingers or squeeze their left or right toes or move their tongue to map motor areas. Each block of a movement type lasted 12 s (10 movements) and was preceded by a 3-s cue. In each of the two runs, there are 13 blocks, with two tongue movements, four hand movements (two right and two left), and four foot movements (two right and two left). In addition, there are three 15-second fixation blocks per run. This task includes the following events, each of which is computed against the fixation baseline. For each subject, the number of frames per run of the motor task is 284, with a run duration of 3.57 min (Wu-Minn, 2017). For each subject, two motor task-related fMRI scans are available: One run was acquired with right-to-left phase encoding and a second run with left-to-right phase encoding. In this article, we use the left-to-right phase encoding scan for each subject.

We use the “Desikan-Killiany” atlas (Desikan et al., 2006) to divide the brain into 68 ROIs. For each subject \( i \), we average the BOLD time series of all pixels in each ROI, which results in a functional curve \( y_{il}(t) \) for \( 1 \leq i \leq n \) and \( 1 \leq l \leq p \). For each curve \( y_{il}(t) \), we do not observe their full trajectory but instead observe the realization of the curve on 284 equal space–time points: \( t_k = 2.16 * (k - 1)/283 \text{ min} \) (\( 1 \leq k \leq 284 \)). We consider \( s = 4 \) motor instrument covariates measured using tests adapted from the American Thoracic Society’s 6-min walk test (Enright, 2003), the 9-hole pegboard test (Wang et al., 2015), and the American Society of Hand Therapy’s grip strength test (MacDermid et al., 1994). In the test adapted from the American Thoracic Society’s 6-min walk test, the submaximal cardiovascular endurance is measured by recording the distance that the participant is able to walk on a 50-foot course in 2 min and the time that the participant is able to walk a 4-m distance at his or her usual pace. In the 9-hole pegboard test,
FIGURE 5: Real Data Results: Panels (a)–(d) plot the $\hat{\beta}_j(t)$s of 68 ROIs for $s=4$ motor instrument covariates: “Endurance-AgeAdj,” “GaitSpeed-Comp,” “Dexterity-AgeAdj,” and “Strength-AgeAdj.” The fitted basis is Fourier.

The manual dexterity is measured by the time required for the participant to accurately place and remove nine plastic pegs into a plastic pegboard. In the test adapted from the American Society of Hand Therapy’s grip strength test, participants are seated in a chair with their feet touching the ground. With the elbow bent to 90 degrees and the arm against the trunk and wrist at neutral, participants squeeze the Jamar Plus Digital dynamometer as hard as they can for a count of three. The dynamometer records a digital reading of force in pounds. The four covariates we consider are “Endurance-AgeAdj,” “GaitSpeed-Comp,” “Dexterity-AgeAdj,” and “Strength-AgeAdj,” where “GaitSpeed-Comp” is the distance walked in 2 min, and “Endurance-AgeAdj,” “Dexterity-AgeAdj,” and “Strength-AgeAdj” are submaximal cardiovascular endurance, manual dexterity, and grip strength, respectively, adjusted by the participant’s age.

To implement our method, we first standardize the functional responses $y_{ij}(t)$s and centre the covariates $x_{ij}$s. We apply our method by fitting the model using a Fourier basis and select the optimal regularization parameter and truncation number by fivefold cross validation. After selecting nine Fourier basis functions, the rank of estimated $\hat{M}$, that is, $\hat{\hat{M}}$, is 4.

We have also obtained the OLS estimate $\hat{M}_{OLS}$ by setting $\lambda = 0$ in Equation 7. We plotted the first 10 singular values of the $\hat{M}$ (red solid) and $\hat{M}_{OLS}$ (black dashed) in Figure 2. Inspecting the figure reveals that the first four singular values of $\hat{M}_{OLS}$ dominate the remaining ones, which verifies the low-rank assumption in this article.

Previous literature (Martino et al., 2011) suggested that the left and right superior frontal regions are strongly associated with motor function. Therefore, we plot the estimated coefficient functions $\{\hat{\beta}_j(t), 1 \leq j \leq 4\}$ corresponding to the left superior frontal regions in Figure 3 and the estimated coefficient functions $\{\hat{\beta}_j(t), 1 \leq j \leq 4\}$ corresponding to the right superior frontal region in Figure 4. From the figures, we can see that the estimated coefficient functions $\hat{\beta}_j(t)$ for
right and left superior frontal regions have similar patterns for each $1 \leq j \leq 4$. This is explained by the symmetry of the brain.

To summarize the result of the performance of all 68 regions, we plot the standardized $\hat{\beta}_j(t)$ for all 68 ROIs in Figure 5(a). Here, the standardized $\hat{\beta}_j(t)$ is defined as $\hat{\beta}_{j,\text{stand}}(t) = \{\hat{\beta}_j(t) - \int_0^1 \hat{\beta}_j(s)ds\}/\{\int_0^1 \{\hat{\beta}_j(u) - \int_0^1 \hat{\beta}_j(s)ds\}^2du\}^{1/2}$ for $1 \leq j \leq 4$. Similar plots for standardized versions of $\hat{\beta}_2(t)$, $\hat{\beta}_3(t)$, and $\hat{\beta}_4(t)$ are included in Figure 5(b)–(d), respectively.

We have also tested the nonstationary assumption of the error processes in real data application. In particular, we apply the Kwiatkowski–Phillips-Schmidt–Shin (KPSS) tests (Kwiatkowski et al., 1992) on the fitted residual time series for 845 individuals, which yield $845 \times 68 = 57,640$ error processes. We find that 96.7% of them are not (trend) stationary, with a significance level 0.05. This indicates that most of the error processes in the application are not stationary.

6. DISCUSSION

In this article, we propose a multivariate functional response low-rank regression model with possible high-dimensional functional responses and scalar covariates. To estimate the nonparametric coefficient functions, our method employs the state-of-art sieve regression. By imposing a low-rank structure of the coefficient matrix, our proposal can obtain a global fit of the coefficient estimates. We have shown that our method performs well in both simulation and the HCP fMRI data application.

There are several important directions for future work. First, we assume that the covariates affect the responses linearly with only main effects. Further investigation is warranted to extend the proposed approach to the case with interaction effects and/or nonlinear effects. Second, it is interesting to further develop an inference procedure for our approach, which can characterize the uncertainty of estimates. One may consider using either bootstrap or debiased approaches to construct simultaneous confidence bands for the coefficient curves.

APPENDIX A: TECHNICAL PROOFS

In the appendix, we provide some preliminaries for our technical derivations and auxiliary lemmas, as well as the proofs for the main theorem and the corollary.

A.1 Some Preliminaries

We write (12) as

$$\hat{\beta}_j(t) = \sum_{h=1}^{c} M_{j,l} b_h(t) + \sum_{h=c+1}^{\infty} M_{j,l} b_h(t), \quad (A1)$$

where $\sum_{h=c+1}^{\infty} M_{j,l} b_h(t)$ corresponds to the error $O(c^{-d})$ in (12). Similar to the definition of $M$, we denote $M^{\dagger}$ as the collection of the entries corresponding to the second term of the right-hand side of (A1) and $\overline{M}$ as the matrix containing all the entries $M_{j,l,h,j=1,2,\ldots,s, h=1,2,\ldots}$, in (A1). Strictly speaking, $\overline{M}$ is not a matrix, but for notational convenience, we denote it as a matrix of dimension $p \times \infty$. Therefore, we find that $\overline{M} = (M, M^{\dagger})$.

In light of (2) with (A1), for the sequence of observation pairs $(X_i, Y_i), i = 1, 2, \ldots, n$, we denote the loss function for $\overline{M}$ as

$$\mathcal{L}(\overline{M}; (X_i, Y_i), i = 1, 2, \ldots, n) = \frac{1}{nT} \sum_{i=1}^{n} \| Y_i - \overline{M} X_i \|^2_F, \quad (A2)$$
where \( \overline{MX}_i = MX_i + M^\dagger X_i^\dagger \). Recall that \( X_i \) is defined in (5), that \( X_i^\dagger \) is defined in a similar fashion by using the basis \( \{ b_j(i) \}_{j>c} \), and that \( X_i \) is defined accordingly. Ideally, \( M^\dagger X_i^\dagger \) corresponds to the \( o(1) \) part in (15). Therefore, the true value \( \overline{M}^* \) is defined as

\[
\overline{M}^* = \arg \min_M \mathbb{E} \mathcal{L}(\overline{M}; (X_i, Y_i), i = 1, 2, \ldots, n).
\]

As discussed in Section 2, we want to estimate \( M \), and hence, we can treat \( M^\dagger \) as nuisance parameters. In this sense, the true value of \( M \) is defined as

\[
M^* = \arg \min_M \mathbb{E} \mathcal{L}(M; (X_i, Y_i), i = 1, 2, \ldots, n)(M^\dagger)^*),
\]

where

\[
\mathcal{L}(M; (X_i, Y_i), i = 1, 2, \ldots, n)(M^\dagger)^*) = \frac{1}{nT} \sum_{i=1}^{n} \| Y_i - MX_i - (M^\dagger)^* X_i^\dagger \|_F^2.
\]

To ease the notation, we introduce a diagonal block matrix \( M \in \mathbb{R}^{np \times nsc} \) with \( n \) blocks, and each diagonal block is \( M \) and \( \mathcal{X} \in \mathbb{R}^{ns \times c} \); \( \mathcal{Y} \in \mathbb{R}^{np \times T} \) contain the sequences of \( X_i \) and \( Y_i, i = 1, 2, \ldots, n \) respectively. Similarly, we can define \( M^\dagger \) and \( \overline{M} \). As a consequence, we can rewrite (A3) as

\[
M^* = \arg \min_M \mathbb{E} \mathcal{L}(M; \mathcal{X}, \mathcal{Y})(M^\dagger)^*),
\]

where

\[
\mathcal{L}(M; \mathcal{X}, \mathcal{Y})(M^\dagger)^*) = \frac{1}{nT} \| \mathcal{Y} - M\mathcal{X}_c - M^\dagger \mathcal{X}_c^\dagger \|_F^2,
\]

with \( \mathcal{X}_c \in \mathbb{R}^{nc \times T} \) containing the matrices \( X_i, i = 1, 2, \ldots, n \) and \( \mathcal{X}_c \) is defined similarly.

Next, we provide an estimate of \( M^* \) denoted by \( \hat{M} \). For any given regularizer \( R \) and regularization penalty \( \lambda_n \), let

\[
\hat{M} = \arg \min_M \left[ \mathcal{L}_1(M; \mathcal{X}, \mathcal{Y}) + \lambda_n R(M) \right],
\]

where \( \mathcal{L}_1 \) is an approximate loss function for \( \mathcal{L} \) and is defined as

\[
\mathcal{L}_1(M; \mathcal{X}, \mathcal{Y}) = \frac{1}{nT} \| \mathcal{Y} - M\mathcal{X}_c \|_F^2,
\]

\( \mathcal{R} \) is the nuclear norm for the rectangular matrix. Our goal is to derive a bound for \( \| M - \hat{M} \|_F \). We state such results in the following subsection.

We start by decomposing the loss function \( \mathcal{L}(M; \mathcal{X}, \mathcal{Y})(M^\dagger)^*) \). Note that

\[
\mathcal{L}(M; \mathcal{X}, \mathcal{Y})(M^\dagger)^*) = \mathcal{L}_1(M; \mathcal{X}, \mathcal{Y}) - 2 \frac{1}{nT} \text{tr} \left( (\mathcal{Y} - M\mathcal{X}_c)(M^\dagger)^* \mathcal{X}_c^\dagger \right)^T \]

\[+ \frac{1}{nT} \| (M^\dagger)^* \mathcal{X}_c^\dagger \|_F^2.
\]
First, as \((M^\dagger)^*\) is a nuisance parameter, the third term on the right-hand side of (A7) can be regarded as a constant term with respect to \(M\). Hence, it suffices to minimize \(\mathcal{E}(M) = \frac{2}{nT} \text{tr} \left( (\mathcal{Y} - MA_c)(M^\dagger)^* A_c^\dagger \right)^T \). Denote
\[
\mathcal{E}(M) = \frac{2}{nT} \text{tr} \left( (\mathcal{Y} - MA_c)(M^\dagger)^* A_c^\dagger \right)^T .
\] (A8)

In order to state our results and clarify how our work differs from Negahban et al. (2012), we follow the notation of Negahban et al. (2012) and let \(\theta \equiv M\). Based on the above discussion, in view of the definition of \(\hat{M}\) in (A5), it suffices to consider the following optimization problem
\[
\hat{\theta} = \arg \min_{\theta} \left[ \mathcal{L}(\theta; \mathcal{X}, \mathcal{Y})(M^\dagger)^* + \lambda \mathcal{R}(\theta) + \mathcal{E}(\theta) \right] .
\] (A9)

We introduce some notations and assumptions, which are also used in Negahban et al. (2012). Let \(\mathcal{M}\) be the model subspace to capture the constraints — for instance, the subspace of low-rank matrices under Assumption 3 in our problem. Let \(\overline{\mathcal{M}}\) be the completion of \(\mathcal{M}\) and \(\overline{\mathcal{M}}^\perp\) be the orthogonal complement of \(\overline{\mathcal{M}}\). It is remarkable that \(\overline{\mathcal{M}}^\perp\) is referred to as the perturbation subspace, representing deviations away from the model subspace \(\mathcal{M}\).

We need the following definition, taken from Definition 1 in Negahban et al. (2012).

**Definition A.2.** (Decomposability of \(\mathcal{R}\)). Given a pair of subspaces \(\mathcal{M} \subseteq \overline{\mathcal{M}}\), a norm-based regularizer \(\mathcal{R}\) is decomposable with respect to \((\mathcal{M}, \overline{\mathcal{M}}^\perp)\) if
\[
\mathcal{R}(\theta + \gamma) = \mathcal{R}(\theta) + \mathcal{R}(\gamma),
\]
for all \(\theta \in \mathcal{M}\) and \(\gamma \in \overline{\mathcal{M}}^\perp\).

It has been shown in Example 3 of Negahban et al. (2012) that the nuclear norm is decomposable with respect to appropriately chosen subspaces (see equations (13a) and (13b) of Negahban et al. (2012)).

We then introduce the restricted strong convexity (RSC) condition, which is taken from Definition 2 of Negahban et al. (2012). Denote the error of Taylor series of \(\mathcal{L}\) at \(\theta^*\) as
\[
\delta \mathcal{L}(\Delta, \theta^*) := \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) - \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle .
\]

**Definition A.3.** (Restricted strong convexity). The loss function satisfies an RSC condition with curvature \(\kappa_{\mathcal{L}} > 0\) and tolerance function \(\tau_{\mathcal{L}}\) if
\[
\delta \mathcal{L}(\Delta, \theta^*) \geq \kappa_{\mathcal{L}} \|\Delta\|^2 - \tau^2_{\mathcal{L}}(\theta^*),
\]
for all \(\Delta \in \mathbb{C}\) defined in (B2) or (B3).

Finally, we introduce the subspace compatibility constant to control \(\mathcal{R}(\cdot)\) (Definition 3 in Negahban et al. (2012)).

**Definition A.4.** (Subspace compatibility constant). For any subspace \(\mathcal{M}\) of \(\mathbb{R}^p\), the subspace compatibility constant with respect to the pair \((\mathcal{R}, \|\cdot\|)\) is given by
\[
\Psi(\mathcal{M}) := \sup_{u \in \mathcal{M}\setminus\{0\}} \frac{\mathcal{R}(u)}{\|u\|} .
\]
We are ready to state our main results. The following result deals with the general M-estimator of the form (A9). We define the projection operator as follows

$$\Pi_M(u) := \arg \min_{v \in \mathcal{M}} \|u - v\|,$$

with the projection $$\Pi_{M^\perp}$$ defined in an analogous way. For simplicity, we use the following shorthand notations $$u_M = \Pi_M(u)$$ and $$u_{M^\perp} = \Pi_{M^\perp}(u)$$.

**Theorem A.2.** Suppose that the loss function $$\mathcal{L}$$ is convex and differentiable and satisfies the RSC condition in Definition A.3 with curvature $$\kappa$$ and tolerance $$\tau$$. We also assume that the regularizer $$\mathcal{R}$$ is a norm and is decomposable with respect to the subspace pair $$(\mathcal{M}, \mathcal{M}^\perp)$$, where $$\mathcal{M} \subset \mathcal{M}^\perp$$. Denote

$$\zeta(\kappa, \tau, \nu) \equiv \zeta(\lambda_n, \Psi, \kappa, \tau^*, \nu) := \frac{\lambda_n}{\kappa^2} \Psi^2(\mathcal{M}) + \frac{\lambda_n}{\kappa} \{2\tau^2(\theta^*) + 4\mathcal{R}(\theta_{M^\perp}^*) + 2\nu\}. \quad (A10)$$

Suppose that there exists some linear function $$\mathcal{R}$$ in $$\mathcal{D}$$ and independent of $$\theta^*$$ such that

$$\mathcal{R}(\Delta) = \mathcal{E}(\theta^* + \Delta) - \mathcal{E}(\theta^*).$$

Furthermore, let $$\epsilon > 0$$ such that

$$\sup_{\Delta} |\mathcal{D}(\Delta)| \leq \epsilon. \quad (A11)$$

If the strictly positive regularization constant satisfies $$\lambda_n \geq 2\mathcal{R}^*(\nabla \mathcal{L}(\theta^*))$$, when conditional on the observation $$(\mathcal{X}, \mathcal{Y})$$ and when $$n$$ is large enough, we find that

$$||\hat{\theta}_{\lambda_n} - \theta^*||^2 \leq C\zeta(\kappa, \tau, 2\epsilon), \quad (A12)$$

where the error norm is the same as $$\mathcal{L}(\cdot)$$, and $$C > 0$$ is some universal constant.

We remark that the error bound in (Negahban et al., 2012, Theorem 1) is $$\zeta(\kappa, \tau, 0)$$ as they do not have the error term $$\mathcal{E}$$. Moreover, in our setup for sieve regression, $$\mathcal{E}(\cdot)$$ satisfies (i) of Theorem A.2 as

$$\text{tr} \left( (\mathcal{Y} - (\theta^* + \Delta)\mathcal{X})((\theta^*)^*\mathcal{X}^\perp)^T \right) - \text{tr} \left( (\mathcal{Y} - \theta^*\mathcal{X})((\theta^*)^*\mathcal{X}^\perp)^T \right)$$

$$= \text{tr} \left( \Delta \mathcal{X}_c[(\theta^*)^*\mathcal{X}^\perp]^T \right). \quad (A13)$$

**APPENDIX B: AUXILIARY LEMMAS AND PROOFS OF LEMMAS AND THEOREMS**

To make it convenient for the reader, we use Table B1 to list the matrices and their associated dimensions.

We use the following abbreviations for our proof

$$\mathcal{L}(\cdot) = \mathcal{L}(\theta; \mathcal{X}, \mathcal{Y}|(\theta^*)^*).$$

We will make use of the function $$\mathcal{F}$$ given by

$$\mathcal{F}(\Delta) := \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) + \lambda_n(\mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*)) + \mathcal{E}(\theta^* + \Delta) - \mathcal{E}(\theta^*). \quad (B1)$$
We denote the optimal error by
\[ \hat{\Delta} = \hat{\theta} - \theta^*. \]

We notice that \( F(0) = 0 \) and \( F(\hat{\Delta}) \leq 0 \).

**B.1 Some Auxiliary Lemmas**

In this section, we provide some auxiliary lemmas that will be used in the proofs of Theorems A.2 and 1. We first provide some preliminary results.

**Lemma B.1.** (Deviation inequalities). For any decomposable regularizer and \( \theta^* \) and \( \Delta \), we have
\[
R(\theta^* + \Delta) - R(\theta^*) \geq R(D_{\mathcal{M}}) - R(D_{\mathcal{M}^\perp}) - 2R(\theta^*_{\mathcal{M}^\perp}).
\]
Moreover, as long as \( \lambda_n \geq 2R^*(\nabla L(\theta^*)) \) and \( L \) is convex, we have
\[
L(\theta^* + \Delta) - L(\theta^*) \geq \frac{\lambda_n}{2} [R(D_{\mathcal{M}}) + R(D_{\mathcal{M}^\perp})].
\]

**Proof.** See Lemma 3 of Negahban et al. (2012).

**Lemma B.2.** Suppose \( \mathcal{L} \) is a convex and differentiable function, and consider any optimal solution \( \hat{\theta} \) to the optimization problem (A9) with a strictly positive regularization parameter satisfying
\[
\lambda_n \geq 2R^*(\nabla L(\theta^*)).
\]
Then, for any pair \( (\mathcal{M}, \mathcal{M}^\perp) \) over which \( R \) is decomposable, the error \( \hat{\Delta} = \hat{\theta} - \theta^* \) belongs to the set
\[
\mathcal{C}(\mathcal{M}, \mathcal{M}^\perp; \theta^* : = \left\{ \Delta | R(D_{\mathcal{M}^\perp}) \leq \frac{\lambda_n}{2} \left( L(\theta^*) - L(\theta^* + \Delta) \right) + 3R(D_{\mathcal{M}}) + 4R(\theta^*_{\mathcal{M}^\perp}) \right\}. (B2)
\]

**Table B1:** Matrices and their dimensions. Here, we use the shorthand that \( sc = s \times c \). For instance, \( X_i \) contains \( T \) matrices of dimension \( s \times c \) as a stack.

| Matrix     | Dimension  |
|------------|------------|
| \( Y_i, E_i \) | \( p \times T \) |
| \( X_i \) | \( sc \times T \) |
| \( M \) | \( p \times sc \) |
| \( X_i^\dagger \) | \( s \infty \times T \) |
| \( M^\dagger \) | \( p \times s \infty \) |
| \( X_c^\circ \) | \( nsc \times T \) |
| \( M^\circ \) | \( np \times nsc \) |
| \( X_c^\perp \) | \( ns \infty \times T \) |
| \( M^\perp \) | \( np \times ns \infty \) |
| \( \mathcal{Y}, E \) | \( np \times T \) |
Moreover, if $\mathcal{E}(\cdot)$ is some convex and differentiable norm on the metric space of the parameter, we find that if

$$
\lambda_n \geq 2 \mathcal{R}^*(\nabla \mathcal{L}(\theta^*) + \nabla \mathcal{E}(\theta^*)),
$$

then for any pair $(\mathcal{M}, \overline{\mathcal{M}})$ over which $\mathcal{R}$ is decomposable, the error $\hat{\theta} = \hat{\theta}_n - \theta^*$ belongs to the set

$$
\mathcal{C}(\mathcal{M}, \overline{\mathcal{M}}; \theta^*) := \left\{ \Delta | \mathcal{R}(\Delta_{\mathcal{M}}) \leq 3 \mathcal{R}(\Delta_{\overline{\mathcal{M}}}) + 4 \mathcal{R}(\theta^*_{\mathcal{M}^\perp}) \right\}. \tag{B3}
$$

**Proof.** By the expansion (B1), the fact $\mathcal{F}(\hat{\Delta}) \leq 0$, and Lemma B.1, we readily find that

$$
0 \geq \mathcal{F}(\hat{\Delta}) \geq \lambda_n \left\{ \mathcal{R}(\Delta_{\mathcal{M}^\perp}) - \mathcal{R}(\Delta_{\overline{\mathcal{M}}}) - 2 \mathcal{R}(\theta^*_{\mathcal{M}^\perp}) \right\} - \frac{\lambda_n}{2} \left[ \mathcal{R}(\Delta_{\mathcal{M}}) + \mathcal{R}(\Delta_{\overline{\mathcal{M}}}) \right] + \mathcal{E}(\theta^*) + \hat{\Delta} - \mathcal{E}(\theta^*)
$$

$$
= \frac{\lambda_n}{2} \left\{ \mathcal{R}(\Delta_{\mathcal{M}^\perp}) - 3 \mathcal{R}(\Delta_{\overline{\mathcal{M}}}) - 4 \mathcal{R}(\theta^*_{\mathcal{M}^\perp}) \right\} + \mathcal{E}(\theta^*) + \hat{\Delta} - \mathcal{E}(\theta^*).
$$

This concludes the proof of (B2). For the proof of (B3), we can apply Lemma B.1 to the convex and differentiable function $\mathcal{L}_1 := \mathcal{L} + \mathcal{E}$. As $\nabla$ is a linear operator, when $\lambda_n \geq 2 \mathcal{R}^*(\nabla \mathcal{L}(\theta^*) + \nabla \mathcal{E}(\theta^*))$, we have

$$
0 \geq \mathcal{F}(\hat{\Delta}) \geq \lambda_n \left\{ \mathcal{R}(\Delta_{\mathcal{M}^\perp}) - \mathcal{R}(\Delta_{\overline{\mathcal{M}}}) - 2 \mathcal{R}(\theta^*_{\mathcal{M}^\perp}) \right\} - \frac{\lambda_n}{2} \left[ \mathcal{R}(\Delta_{\mathcal{M}}) + \mathcal{R}(\Delta_{\overline{\mathcal{M}}}) \right]
$$

$$
= \frac{\lambda_n}{2} \left\{ \mathcal{R}(\Delta_{\mathcal{M}^\perp}) - 3 \mathcal{R}(\Delta_{\overline{\mathcal{M}}}) - 4 \mathcal{R}(\theta^*_{\mathcal{M}^\perp}) \right\},
$$

from which the proof of (B3) follows.

**Remark.** The counterpart of the above lemma (Negahban et al., 2012, Lemma 1) does not have the approximation error term $\mathcal{E}(\cdot)$. In our setup, $\mathcal{E}(\cdot)$ is not a properly defined norm. Hence, we need to apply (B2) whenever it is needed. For an interpretation of Lemma B.2, we refer to Figure 1 of Negahban et al. (2012).

Recall the sets $\mathcal{C}$ defined in (B2) or (B3). Let $\delta$ be a given error radius and denote $\mathbb{K}(\delta) := \mathcal{C} \cap \{ ||\Delta|| = \delta \}$. We have the following lemma, the counterpart of which is (Negahban et al., 2012, Lemma 4).

**Lemma B.3.** Suppose $\mathcal{R}(\cdot)$ is decomposable and convex, and $\mathcal{L}$ is differentiable and convex. Then

(i). When $\mathcal{C}$ is defined in (B2), suppose that there exists some linear function $D$ in $\Delta$ that is independent of $\theta^*$ such that

$$
D(\Delta) = \mathcal{E}(\theta^* + \Delta) - \mathcal{E}(\theta^*).
$$

If $\mathcal{F}(\Delta) > 0$ for all vectors $\Delta \in \mathbb{K}(\delta)$, then $||\hat{\Delta}|| \leq \delta$.

(ii). When $\mathcal{C}$ is defined in (B3), that is, when $\mathcal{E}(\cdot)$ is some convex differentiable norm, if $\mathcal{F}(\Delta) > 0$ for all vectors $\Delta \in \mathbb{K}(\delta)$, then $||\hat{\Delta}|| \leq \delta$. 

The Canadian Journal of Statistics / La revue canadienne de statistique DOI: 10.1002/cjs.11604
Proof. We start with the proof of (i). We prove the contrapositive statement: In particular, if for some optimal solution \( \hat{\theta} \) such that \( \| \hat{\theta} \| > \delta \), there must be some vector \( \Delta \in \mathbb{R}(\delta) \) such that \( \mathcal{F}(\hat{\Delta}) \leq 0 \). To achieve this goal, it suffices to prove the following claim:

**Claim 1.** If \( \hat{\Delta} \in \mathbb{C} \), then the entire line \( \{ t\hat{\Delta} | t \in (0, 1) \} \) connecting \( \hat{\Delta} \) with the all-zero vector is contained within \( \mathbb{C} \).

We first show how we can construct such a \( \Delta \) using the above claim. If \( \| \hat{\Delta} \| > \delta \), then the line joining \( \hat{\Delta} \) and \( 0 \) must intersect the set \( \mathbb{K}(\delta) \) at some intermediate point \( t^* \Delta \), for some \( t^* \in (0, 1) \) (i.e., after some proper scaling). By Claim 1, we find that \( t^* \hat{\Delta} \in \mathbb{C} \). As \( D(\Delta) \) is linear, and both \( \mathcal{L} \) and \( \mathcal{R} \) are convex, by Jensen’s inequality, we have

\[
\mathcal{F}(t^* \hat{\Delta}) = \mathcal{F}(t^* \Delta + (1 - t^*)0) \leq t^* \mathcal{F}(\hat{\Delta}) + (1 - t^*) \mathcal{F}(0) = t^* \mathcal{F}(\hat{\Delta}),
\]

where in the last equality, we use the fact that \( \mathcal{F}(0) = 0 \). As \( \hat{\Delta} \) is optimal, we have that \( \mathcal{F}(t^* \hat{\Delta}) \leq 0 \). Hence, we can find \( \hat{\Delta} = t^* \hat{\Delta} \) and conclude the proof of (i).

Finally, we prove Claim 1. First, when \( \theta^* \in \mathcal{M} \), we find that \( \mathcal{R}(\theta^*_{\mathcal{M}^\perp}) = 0 \), and the proof follows immediately. Second, when \( \theta^* \notin \mathcal{M} \), it is easy to see that, for any \( t \in (0, 1) \), we have

\[
\Pi_{\mathcal{M}^\perp}(t\Delta) = \arg \min_{\gamma \in \mathcal{M}} \| t\Delta - \gamma \| = t \arg \min_{\gamma \in \mathcal{M}} \left\| \Delta - \frac{\gamma}{t} \right\| = t \Pi_{\mathcal{M}^\perp}(\Delta),
\]

where we use the fact that \( \gamma/t \) also belongs to the subspace \( \overline{\mathcal{M}} \). Similarly, we can show that

\[
\Pi_{\mathcal{M}^\perp}(t\Delta) = t \Pi_{\mathcal{M}^\perp}(\Delta).
\]

Hence, we find that, for all \( \Delta \in \mathbb{C} \),

\[
\mathcal{R}(\Pi_{\mathcal{M}^\perp}(t\Delta)) = \mathcal{R}(t\Pi_{\mathcal{M}^\perp}(\Delta)) = t \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\Delta)) \leq t \left\{ 3 \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\Delta)) + 4 \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*_{\Delta})) + D(\Delta) \right\},
\]

where we use the fact that \( \mathcal{R}(\cdot) \) is a norm and the definition of \( \mathbb{C} \) in (B2). We observe that

\[
3t \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\Delta)) = 3 \mathcal{R}(\Pi_{\mathcal{M}^\perp}(t\Delta)) \quad \text{and} \quad 4t \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*_{\Delta})) \leq 4 \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*_{\Delta})), \quad t \in (0, 1).
\]

Moreover, as \( D(\cdot) \) is linear in \( \Delta \), we have \( D(t\Delta) = t D(\Delta) \). Putting all these together, we find that

\[
\mathcal{R}(\Pi_{\mathcal{M}^\perp}(t\Delta)) \leq 3 \mathcal{R}(\Pi_{\mathcal{M}^\perp}(t\Delta)) + 4t \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*_{\Delta})) + D(t\Delta)
\]

\[
\leq 3 \mathcal{R}(\Pi_{\mathcal{M}^\perp}(t\Delta)) + 4 \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*_{\Delta})) + D(t\Delta).
\]

This concludes the proof of Claim 1.

For (ii), we can apply (Negahban et al., 2012, Lemma 4) for the loss function \( \mathcal{L}_1 := \mathcal{L} + \mathcal{E} \) to finish the proof.

Next, we provide some matrix inequalities that will be used in the proof of Theorem 1 to bind \( \mathcal{E}(\cdot) \).

**Lemma B.4.** Suppose \( A \in \mathbb{R}^{m \times n} \). Denote

\[
r_j := \sum_{1 \leq j \neq i \leq n} |a_{ij}|, \quad c_i := \sum_{1 \leq j \neq i \leq m} |a_{ji}|,
\]

DOI: 10.1002/cjs.11604
and

\[ s_i := \max \{ r_i, c_i \}, \quad a_i := |a_{ij}|, \]

for \( i = 1, 2, \ldots, \min\{m, n\} \). Moreover, for \( m \neq n \), we define

\[
s := \begin{cases} 
\max_{n+1 \leq i \leq m} \left\{ \sum_{j=1}^{n} |a_{ij}| \right\}, & m > n \\
\max_{m+1 \leq i \leq n} \left\{ \sum_{j=1}^{n} |a_{ij}| \right\}, & m < n.
\end{cases}
\]

For \( m \geq n \), we find that each singular value of \( A \) lies in one of the real intervals, defined as

\[ B_i = [\max \{ a_i - s_i, 0 \}, a_i + s_i], \quad i = 1, 2, \ldots, n; \quad B_{n+1} = [0, s]. \]

If \( m = n \) or if \( m > n \) and \( a_i \geq s_i + s, \quad i = 1, \ldots, n \), then \( B_{n+1} \) is not needed in the above statement. Similar results hold when \( m \leq n \).

**Proof.** See Theorem 2 of Qi (1984).

**Lemma B.5.** Suppose \( A \) and \( B \) are positive definite square matrices. Then, we have

\[ \lambda_{\min}(A) \text{tr}(B) \leq \text{tr}(AB) \leq \lambda_{\max}(A) \text{tr}(B), \]

where \( \lambda_{\max}(A) \) is the largest eigenvalue of \( A \), and \( \lambda_{\min}(A) \) is the smallest eigenvalues of \( A \).

**Proof.** See equation (1) of Fang, Loparo, & Feng (1994).

**Lemma B.6.** Consider a sequence of i.i.d. sub-Gaussian vectors \( z_1, \ldots, z_n \) in \( \mathbb{R}^s \) with covariance matrix \( \Sigma_s \). Let \( \epsilon \in (0, 1), t \geq 1 \); then, for some constant \( C > 0 \), with probability at least \( 1 - 2 \exp(-t^2n) \), we have if \( n \geq C(t/\epsilon)^2s \)

\[ ||\Sigma^n_s - \Sigma_s|| \leq \epsilon, \]

where \( \Sigma^n_s \) is the sample covariance matrix of \((z_i)\).

**Proof.** See Corollary 5.50 of Vershynin (2011).

**Lemma B.7.** Let \([a, b]\) be a bounded closed interval. We take an \( n \)-division \( \Delta \) of \([a, b]\) as

\[ \Delta : \quad a = s_0 \leq s_1 \leq \ldots \leq s_{n-1} \leq s_n = b, \]

and hence, \( s_i = a + i(b - a)/n \). If \( f \) is twice differentiable and \( f'' \) is bounded and almost everywhere continuous on \([a, b]\), then

\[ \lim_{n \to \infty} n^2 \left\{ \int_a^b f(x)dx - \sum_{i=1}^{n} (s_i - s_{i-1}) f \left( \frac{s_{i-1} + s_i}{2} \right) \right\} = \frac{(b - a)^2}{24} (f'(b) - f'(a)). \quad (B4) \]

**Proof.** See Theorem 1.1 of Tasaki (2009).
Lemma B.8. Let $A = (a_{ij})$ be a real $n \times n$ matrix. For $1 \leq i \leq n$, let $R_i = \sum_{j \neq i} |a_{ij}|$ be the sum of the absolute values of the nondiagonal entries in the $i$-th row. Let $D(a_{ii}, R_i) \subseteq \mathbb{R}$ be a closed disc centered at $a_{ii}$ with radius $R_i$. Such a disc is called a Gershgorin disc. Every eigenvalue of $A = (a_{ij})$ lies within at least one of the Gershgorin discs $D(a_{ii}, R_i)$, where $R_i = \sum_{j \neq i} |a_{ij}|$.

Proof. See Lemma D.1 of Ding & Zhou, (2020).

Finally, we provide a concentration inequality for the locally stationary time series \{\epsilon_{ik}(t)\}, $1 \leq i \leq n, 1 \leq k \leq p$.

Lemma B.9. Let $x_i = G_i(T_i)$, where $G_i(\cdot)$ is a measurable function. In addition, let $T_i = (\eta_M, \eta_{M+1}, \ldots, \eta_{i-1}, \eta_i)$ and $\eta_i, i \in \mathbb{Z}$ be i.i.d. random variables, where $M$ is any arbitrarily small negative integer that is independent of $i$. Suppose that $\mathbb{E}x_i = 0$ and $\max_{1 \leq i \leq n} \|G_i(T_i) - G_i(T_{i-\dot})\|_q < \infty$ for some $q > 1$. For some $k > 0$, let $\delta(k) := \max_{1 \leq i \leq n} \|G_i(T_i) - G_i(T_{i-k})\|_q$. We further let $\delta(k) = 0$ if $k < 0$. Write $\gamma_k = \sum_{i=0}^k \delta(i)$. Let $S_i = \sum_{j=1}^i x_j$.

(i). For $q' = \min(2, q)$, 
$$\left\|S_i\right\|_{q'}^{q'} \leq C_q \sum_{i=0}^\infty (\gamma_{i+n} - \gamma_i)^{q'}.$$ 

(ii). If $\Delta := \sum_{j=0}^\infty \delta(j) < \infty$, we then have 
$$\left\|\max_{1 \leq i \leq n} |S_i|\right\|_q \leq C_q n^{1/q} \Delta.$$ 

In (i) and (ii), $C_q$ are generic finite constants that only depend on $q$ and can vary from place to place.

Proof. See Lemma D.6 of Ding & Zhou (2020).

B.2 Proof of Theorem A.2

With the preparation in Section B.1, especially Lemma B.3, we now proceed to finish the proof of Theorem A.2.

Proof. The proof is essentially similar to that of (Negahban et al., 2012, Theorem 1); for the self-completeness, we also provide the complete proof. In light of Lemma B.3, it suffices to establish a lower bound on $F(\Delta)$ over $\mathbb{K}(\delta)$ for an appropriately chosen radius $\delta > 0$. Indeed, for an arbitrary $\Delta \in \mathbb{K}(\delta)$, using the definition of $F$ in (B1), we have

$$F(\Delta) \geq \langle \nabla L(\theta^*), \Delta \rangle + \kappa_L \|\Delta\|^2 - \tau_L^2(\theta^*) + \lambda_n \{R(\theta^* + \Delta) - R(\theta^*)\} + D(\Delta),$$

$$\geq \langle \nabla L(\theta^*), \Delta \rangle + \kappa_L \|\Delta\|^2 - \tau_L^2(\theta^*)$$

$$+ \lambda_n \{R\left(\Delta_{\lambda_{\Delta}}\right) - R(\Delta_{\lambda^*}) - 2R(\theta^{\star}_{\lambda_{\Delta}})\} + D(\Delta),$$

where the first inequality follows from RSC and the second inequality follows from Lemma B.1. Moreover, by the Cauchy–Schwarz inequality and the definition of dual norm, we readily find that

$$\left|\langle \nabla L(\theta^*), \Delta \rangle\right| \leq R^\ast(\nabla L(\theta^*))R(\Delta).$$
As \( \lambda_n \geq 2 R^*(\nabla L(\theta^*)) \) by assumption, we find that
\[
|\langle \nabla L(\theta^*), \Delta \rangle| \leq \frac{\lambda_n}{2} R(\Delta),
\]
and hence, we have
\[
F(\Delta) \geq \kappa_\varepsilon \|\Delta\|^2 - \tau_\varepsilon^2(\theta^*) + \lambda_n \{ R(\Delta_{\mathcal{M}^*}) - R(\Delta_{\mathcal{M}^*}) - 2 R(\theta^*_{\mathcal{M}^*}) \} + D(\Delta) - \frac{\lambda_n}{2} R(\Delta).
\]
Together with \( R(\Delta) \leq R(\Delta_{\mathcal{M}^*}) + R(\Delta_{\mathcal{M}^*}) \), we find that
\[
F(\Delta) \geq \kappa_\varepsilon \|\Delta\|^2 - \tau_\varepsilon^2(\theta^*) - \frac{\lambda_n}{2} \{ 3 R(\Delta_{\mathcal{M}^*}) + 4 R(\theta^*_{\mathcal{M}^*}) \} + D(\Delta).
\]
As \( 0 \in \mathcal{M}^* \), it is easy to see by the equation below, eq. (55) of Negahban et al. (2012), that \( \|\Delta_{\mathcal{M}^*}\| \leq \|\Delta\| \). Moreover, by Definition A.4, we find that\( \varepsilon(\mathcal{M}^*) \leq \psi(\mathcal{M}^*) \|\Delta\| \). As \( \sup_{\Delta} |D(\Delta)| \leq \varepsilon \), this leads to
\[
F(\Delta) \geq \kappa_\varepsilon \|\Delta\|^2 - \tau_\varepsilon^2(\theta^*) - \frac{\lambda_n}{2} \{ 3 \psi(\mathcal{M}^*) \|\Delta\| + 4 R(\theta^*_{\mathcal{M}^*}) \} - \varepsilon.
\]
Note that the right-hand side of the above inequality is a strictly defined quadratic form in \( \|\Delta\| \) and hence will be positive for large \( \|\Delta\| \). The proof then follows from some elementary computations on the quadratic equation.

\[\blacksquare\]

**B.3 Proof of Theorem 1**

We will employ Theorem A.2 to prove Theorem 1. The key ingredient is to provide a bound for \( \mathcal{E}(\cdot) \) using (12). We will require the following facts about matrix differentiation. For details, these facts can be found in Petersen & Pedersen (2007). For any two \( m \times n \) rectangular matrices \( A, B \) and any matrix function \( f \), we have
\[
\nabla_A \text{tr}(AB^T) = B, \quad \nabla_A f(A) = (\nabla_A f(A))^T.
\]
We prepare some computations on the derivatives using (B5). Recall (A6) as
\[
\mathcal{L}_1 = \frac{1}{nT} \text{tr} \left( \mathcal{Y} \mathcal{Y}^T - \mathcal{Y}(M\mathcal{X}_c)^T - M\mathcal{X}_c \mathcal{Y}^T + M\mathcal{X}_c \mathcal{X}_c^T M^T \right),
\]
By (B5), we readily find that
\[
\nabla_M \mathcal{L}_1 = \frac{1}{nT} \left( -\mathcal{Y} \mathcal{X}_c^T - \mathcal{X}_c \mathcal{Y}^T + M\mathcal{X}_c \mathcal{X}_c^T + \mathcal{X}_c \mathcal{X}_c^T M^T \right),
\]
and so, the Hessian matrix of \( \mathcal{L}_1 \) at \( M \) is
\[
H_M \mathcal{L}_1 = \frac{2}{nT} \mathcal{X}_c \mathcal{X}_c^T.
\]
Recall (A8), with which we find that
\[
\nabla_M \mathcal{E} = -\frac{2}{nT} (M^*)^T \mathcal{X}_c \mathcal{X}_c^T,
\]
and so, the Hessian matrix of \( \mathcal{E} \) at \( M \) is
\[
H_M \mathcal{E} = 0.
\]
Proof. In view of (A13), we need to apply (i) of Theorem A.2. We prepare two important facts for our proof. First, as $\mathcal{R}$ is the nuclear norm, we find that in Negahban et al. (2012) Section 2.3,

$$\mathcal{R}^*(M) = \|M\|,$$

where $\|M\|$ stands for the largest singular value of $M$. Second, from the proof of (Negahban et al., 2009, Corollary 5), we know that, under Assumption 3, $\Psi(\hat{M}) = 2\sqrt{r}$ (Recall Definition A.4). Armed with the above results, we now proceed to check the conditions of Theorem A.2 and the computation of the inputs there.

In what follows, we denote $\Delta = M - \hat{M}$ and $\Delta$ to be a diagonal matrix with $n$ blocks whose diagonals are $\Delta$.

Checking of the decomposability and differentiability: It is clear that $\mathcal{L}$ is differentiable with respect to $M$. $\mathcal{R}$ is the nuclear norm, and it is decomposable with respect to suitable subspaces defined in equations (13a) and (13b) of Negahban et al. (2012).

Checking of the RSC condition: By (B7) and (B9), we find that the first-order Taylor expansion from Definition A.3 is exact, such that

$$\delta \mathcal{L} = \frac{2}{nT} \|X_i \Delta\|^2_F.$$

It suffices to provide a lower bound for $\frac{2}{nT} \|X_i \Delta\|^2_F$. Note that

$$\delta \mathcal{L} = \frac{2}{nT} \sum_{i=1}^{n} \|X_i \Delta\|^2_F \geq \frac{2}{nT} \left\| \sum_{i=1}^{n} X_i \Delta \right\|^2_F \geq 2 \|\Delta\|^2_F \lambda_{\min}\left(\frac{1}{nT} \sum_{i=1}^{n} X_i X_i^T\right),$$

where $X = (x_1, \ldots, x_n)$. Note that, in the first inequality, we use the property of matrix norm, and in the second inequality, we use Lemma B.5.

By Assumption 4 and Lemma B.6, for some constant $C > 0$, with probability at least $1 - 2 \exp(-n)$, we have

$$\left\| \frac{1}{n} XX^T - \Sigma_i \right\| \leq C \sqrt{\frac{s}{n}},$$

(B11)

Together with Assumptions 4 and 5, when $n$ is sufficiently large, with probability at least $1 - 2 \exp(-n)$, we have

$$\lambda_{\min}\left(\frac{1}{n} XX^T\right) \geq \frac{\sqrt{2}}{2} \lambda_{\min}(\Sigma_i).$$

Moreover, by Lemmas B.7 and B.8, we find that, for some constant $C > 0$,

$$\left\| \frac{1}{T} BB^T - I \right\| \leq \frac{C}{T^2},$$

(B12)

where we use the smoothness of the basis functions and the fact that $\int_0^T b_i(t)b_j(t)dt = \delta_{ij}$ and that the $ij$th entry of $\frac{1}{T} BB^T$ is $\frac{1}{T} \sum_{k=1}^{T} b_i(t_k)b_j(t_k)$.

Hence, when $T$ is large enough, we have

$$\lambda_{\min}\left(\frac{1}{T} BB^T\right) \geq \frac{\sqrt{2}}{2}.$$
This shows that, with probability of at least $1 - \exp(-n)$, we have

$$\delta \mathcal{L} \geq c \lambda_{\min}(\Sigma_e)\|\Delta\|^2_F,$$

where we use the fact that the eigenvalues of $A \otimes B$ are the products of the eigenvalues of $A$ and $B$.

Hence, we can consider $\kappa = C \lambda_{\min}(\Sigma_e) > 0$.

**Computation of $\lambda_n$**: Let $E$ be the matrix containing $E_i$, $i = 1, 2, \ldots, n$. By (B6) and (B8), we find that

$$\nabla_M \mathcal{L} = -\frac{1}{nT} (E \dot{X}_C^T + X_C E^T) - \frac{3}{nT} (M^\dagger)^* X_C^\dagger \dot{X}_C^T - \frac{1}{nT} X_C [(M^\dagger)^* X_C^\dagger]^T,$$

where we use the fact that $Y - M X_C = (M^\dagger)^* X_C^\dagger + E$. Using (B10), it suffices to provide an upper bound for the largest singular value of the right-hand side of (B13) using Lemmas B.4 and B.7. First, by Lemmas B.4 and B.9, under Assumption 1, for some constant $C_q > 0$, with probability at least $1 - C_q n^{-q}$, and for some constant $C > 0$,

$$\frac{1}{nT} \|E \dot{X}_C^T\| \leq \frac{p \xi n^\tau}{\sqrt{T}}.$$ 

Similarly, we have

$$\frac{1}{nT} \|X_C E^T\| \leq \frac{p \xi n^\tau}{\sqrt{T}}.$$ 

Second, by (12), (13), and Assumption 5, where $x_i$, $i = 1, 2, \ldots, n$, are sub-Gaussian, for some small constant $\tau > 0$, we have that with probability at least $1 - 2 \exp(-n^{2\tau})$,

$$\frac{1}{nT} \|(M^\dagger)^* X_C^\dagger \dot{X}_C^T\| \leq \frac{p \xi n^\tau C^{-d}}{\sqrt{T}},$$

where we use Lemma B.4. Similarly, we have

$$\frac{1}{nT} \|X_C [(M^\dagger)^* X_C^\dagger]^T\| \leq \frac{p \xi n^\tau C^{-d}}{\sqrt{T}}.$$ 

This implies that, with probability at least $1 - C_q n^{-q}$, we can choose

$$\lambda_n \geq C \frac{p \xi n^\tau}{\sqrt{T}}.$$ 

**Computation of $\epsilon$ in (A11)**: We simply show that $\epsilon$ can be chosen as a bounded constant value with high probability, which is sufficient for our proof. This is done by using the Cauchy–Schwartz inequality. Note that

$$D(\Delta) = \frac{1}{nT} \text{tr} (\Delta X_C((\theta^\dagger)^* X_C^\dagger))^T) \leq \frac{1}{nT} \left(\|\Delta X_C\|^2_F + \|((\theta^\dagger)^* X_C^\dagger)^\dagger\|^2_F\right),$$

where $D(\Delta)$ is the trace of the squared Frobenius norm of $\Delta$.
where we use the Cauchy–Schwartz inequality. On one hand, by (B11) and Assumption 5 that $\Sigma_s$ is bounded, when $n$ is large enough, we have that with a probability of at least $1 - 2 \exp(-n)$ find

$$
\frac{1}{nT} \|\Delta^c \|^2_F \leq C_1 \lambda_1(\Sigma_s) \|\Delta\|^2_F \leq C_2,
$$

(B16)

where in the first inequality, we use Lemma B.5, and in the second inequality, we use Assumption 3 and where $C_1, C_2 > 0$ are some constants. On the other hand, by (12), (13), and Assumption 5 that $x_i, i = 1, 2, \ldots, n$, are sub-Gaussian, for some small constant $\tau > 0$, we find that, with a probability of at least $1 - 2 \exp(-n^2\tau)$,

$$
\frac{1}{nT} \|(\theta^* \phi^c)^\dagger\|^2_F \leq \xi ps^2 n^{2\tau} e^{-d} = o(1),
$$

(B17)

where in the first inequality, we use the fact that, for sub-Gaussian random variable $x$, we have $P(|x| \geq n^{\tau}) \leq 1 - 2 \exp(-n^2\tau)$ and the definition of $\| \cdot \|_F$, and for the second equality, we use Assumption 4. By (B16) and (B17), we have shown that, with a probability of at least $1 - 2 \exp(-n^2\tau)$,

$$
D(\Delta) \leq 2C_2.
$$

After checking the conditions of Theorem A.2 and the computation of $\lambda_n$ and $\epsilon$, together with the fact that $R(\theta^*_M) = 0$, we complete the proof.

B.4 Proof of Corollary 1

In this section, we prove Corollary 1.

Proof. Note that $\hat{\beta}_j(t) - \beta_j(t) = (\hat{M}_j - M_j)b(t)$. Since

$$
\|\hat{\beta}_j(t) - \beta_j(t)\|^2 = \text{tr}((\hat{\beta}_j(t) - \beta_j(t))(\hat{\beta}_j(t) - \beta_j(t))^\top)
$$

$$
= \text{tr}((\hat{M}_j - M_j)(\hat{M}_j - M_j)^\top(b(t)b^\top(t))),
$$

by Lemma B.5 and Theorem 1, with probability at least $1 - C_q n^{-q\tau}$, we have

$$
\|\hat{\beta}_j(t) - \beta_j(t)\|^2 \leq C \left( \frac{\sqrt{\rho^2 \xi n^{\tau/2}}}{\text{tr}(\Sigma_s)^{1/4}} + \frac{\sqrt{\rho^2 \xi n^{\tau}}}{\text{tr}(\Sigma_s)^{1/4}} \right)^2 \|b(t)b^\top(t)\|.
$$

Moreover, for any $t \in [0, 1]$, we have

$$
\|b(t)b^\top(t)\| = \sum_{h=1}^c b_h^2(t) \leq c.
$$

We can therefore conclude our proof.

ACKNOWLEDGEMENTS

The authors thank the editor, associate editor, and the referees for their constructive comments, which have substantially improved the article. Yu’s research was partially supported by the Canadian Statistical Sciences Institute postdoctoral fellowship. Zhang’s research was partially
supported by the National Institutes of Health of the United States (MH118927 and AG066970). Kong’s research was partially supported by Natural Sciences and Engineering Research Council of Canada.

BIBLIOGRAPHY

Bai, J. & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70, 191–221.

Beck, A. & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2, 183–202.

Buckner, R. L., Krienen, F. M., Castellanos, A., Diaz, J. C., & Yeo, B. T. (2011). The organization of the human cerebellum estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106: 2322–2345.

Chen, K., Dong, H., & Chan, K. S. (2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100, 901–920.

Chen, X. (2007). Chapter 76: Large sample sieve estimation of semi-nonparametric models. Amsterdem and Oxford: Elsevier, North–Holland, vol. 6 of *Handbook of Econometrics*, pp. 5549 – 5632.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31, 968–980.

Ding, X. & Zhou, Z. (2020). Estimation and inference for precision matrices of nonstationary time series. *Annals of Statistics*, 48, 2455–2477.

Enright, P. L. (2003). The six-minute walk test. *Respiratory Care*, 48, 783–785.

Fan, J. & Gijbels, I. (1996), Local polynomial modelling and its applications, vol. 66 of *Monographs on Statistics and Applied Probability*, Chapman & Hall, London.

Fang, Y., Loparo, K. A., & Feng, X. (1994). Inequalities for the trace of matrix product. *IEEE Transactions on Automatic Control*, 39, 2489–2490.

Faraway, J. J. (1997). Regression analysis for a functional response. *Technometrics*, 39, 254–261.

Hu, W., Pan, T., Kong, D., & Shen, W. (2020a). Nonparametric matrix response regression with application to brain imaging data analysis. *Biometrics*, in press.

Hu, W., Shen, W., Zhou, H., & Kong, D. (2020b). Matrix linear discriminant analysis. *Technometrics*, 62, 196–205.

Jones, D. T., Vemuri, P., Murphy, M. C., Gunter, J. L., Senjem, M. L., Machulda, M. M., Przybelski, S. A., Gregg, B. E., Kantarci, K., Knopman, D. S., et al. (2012). Non-stationarity in the “resting brain’s” modular architecture. *PLoS One*, 7, e39731.

Kong, D., An, B., Zhang, J., & Zhu, H. (2020). L2RM: Low-rank linear regression models for high-dimensional matrix responses. *Journal of the American Statistical Association*, 115, 403–424.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., Shin, Y., et al. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54, 159–178.

MacDermid, J. C., Kramer, J. F., Woodbury, M. G., McFarlane, R. M., & Roth, J. H. (1994). Interrater reliability of pinch and grip strength measurements in patients with cumulative trauma disorders. *Journal of Hand Therapy*, 7, 10–14.

Martino, J., Gabarrós, A., Deus, J., Juncadella, M., Acebes, J., Torres, A., & Puig, J. (2011). Intrасural mapping of complex motor function in the superior frontal gyrus. *Neuroscience*, 179, 131–142.

Negahban, S., Yu, B., Wainwright, M. J., & Ravikumar, P. K. (2009). A unified framework for high-dimensional analysis of $M$-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, Vol. 22, pp. 1348–1356.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B., et al. (2012). A unified framework for high-dimensional analysis of $M$-estimators with decomposable regularizers. *Statistical Science*, 27, 538–557.

Nesterov, Y. (2013). *Introductory lectures on convex optimization: A basic course*, Vol. 87. Springer Science & Business Media, Berlin.

Petersen, K. B. & Pedersen, M. S. (2007). *The Matrix Cookbook*, Technical report. Technical University of Denmark, p. 2007.
Qi, L. (1984). Some simple estimates for singular values of a matrix. *Linear Algebra and its Applications*, 56, 105–119.

Ramsay, J. O. & Dalzell, C. J. (1991). Some tools for functional data analysis (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 53, 539–572.

Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York.

Tasaki, H. (2009). Convergence rates of approximate sums of Riemann integrals. *Journal of Approximation Theory*, 161, 477–490.

Vershynin, R. (2011), *Introduction to the non-asymptotic analysis of random matrices*, arXiv preprint arXiv 1011.3027.

Wang, J.-L., Chiou, J.-M., & Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3, 257–295.

Wang, Y. C., Bohannon, R. W., Kapellusch, J., Garg, A., & Gershon, R. C. (2015). Dexterity as measured with the 9-Hole Peg Test (9-HPT) across the age span. *Journal of Hand Therapy*, 28, 53–60.

Wu-Minn, H. (2017), *1200 subjects data release reference manual*, https://www.humanconnectome.org.

Yao, F., Müller, H. G., & Wang, J. L. (2005). Functional linear regression analysis for longitudinal data. *Annals of Statistics*, 33, 2873–2903.

Zhou, H. & Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 463–483.

Zhou, Z. & Wu, W. (2009). Local linear quantile estimation for non-stationary time series. *Annals of Statistics*, 37, 2696–2729.

Zhu, H., Li, R., & Kong, L. (2012). Multivariate varying coefficient model for functional responses. *Annals of Statistics*, 40, 2634–2666.

---

Received 6 February 2020
Accepted 7 October 2020