Journal of Machine Learning Research 23 (2022) 1-28 Submitted 9/20; Revised 11/21; Published 5/22

Nonparametric Principal Subspace Regression

Yang Zhou YANGZ91@163.COM
School of Statistics
Beijing Normal University
Beijing 100875, China
Mark Koudstaal MKOUDSTAAL@GMAIL.COM
Dengdeng Yu DENGDENG.YU@UTORONTO.CA
Dehan Kong KONGDEHAN@UTSTAT.TORONTO.EDU
Department of Statistical Sciences
University of Toronto
Toronto, ON M5S 3G3, Canada

Fang Yao FYAO@MATH.PKU.EDU.CN
Department of Probability & Statistics
Center for Statistical Science
Peking University
Beijing 100871, China

Editor: Ambuj Tewari

Abstract

In scientific applications, multivariate observations often come in tandem with temporal or spatial covariates, with which the underlying signals vary smoothly. The standard approaches such as principal component analysis and factor analysis neglect the smoothness of the data, while multivariate linear or nonparametric regression fails to leverage the correlation information among multivariate response variables. We propose a novel approach named nonparametric principal subspace regression to overcome these issues. By decoupling the model discrepancy, a simple two-step estimation procedure is introduced, which takes advantage of the low-rank approximation while keeping smooth dynamics. The theoretical property of the proposed procedure is established under an increasing-dimension framework. We demonstrate the favorable performance of our method in comparison with its counterpart, the conventional nonparametric regression, from both theoretical and numerical perspectives.

Keywords: factor analysis, local polynomial smoothing, low-rank approximation, singular value decomposition, smoothness

1. Introduction

In scientific applications, one is often interested in predicting a multivariate response using one or a few predictor variables. The multivariate response linear regression is a conventional way to model this type of data. The usual procedure is the ordinary least squares, equivalent to performing an individual linear regression of each response variable on predictor variables, which fails to use the correlation information among the response variables. To incorporate the correlation information, Breiman and Friedman (1997) proposed a multivariate shrinkage method to

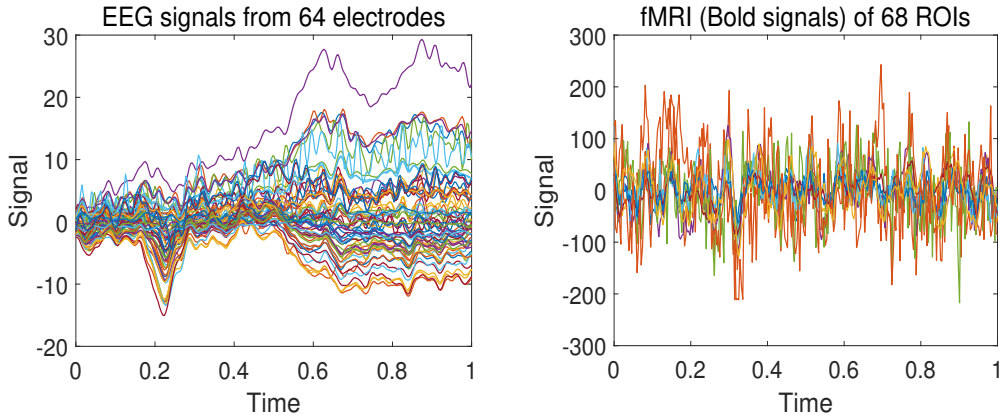

Figure 1: Left: EEG signals detected from 64 electrodes of the scalp at 256 Hz per second for a randomly selected subject; Right: fMRI of 68 Regions of Interest (ROIs, divided according to "Desikan-Killany" atlas) over 205 seconds including 284 frames for a randomly selected subject.

leverage information from the correlation structure, which helps to improve the predictive accuracy compared to the ordinary least squares. Although multivariate response linear regression is a useful tool, it may not work properly in some applications. For example, in the real data examples presented in Section 5, we are interested in modeling the dynamic changes of the electroencephalogram (EEG) signals and functional Magnetic Resonance Imaging (fMRI), shown in Figure 1, where both curves show nonlinear patterns. This indicates that the multivariate response linear regression model may not be adequate to characterize the relationship between the common predictor time and the multivariate signals. A natural rescue is to use nonparametric regression of the multivariate response variables on the common predictors, that is nonparametrically fit model (1) in Section 2.2. However, this solution is unsatisfactory as it is equivalent to performing curve-by-curve individual nonparametric regression for each component of the response, and thus does not capture correlations among the response variables.

Motivated by these applications, we propose a new nonparametric principal subspace regression model in which the essence is that the nonparametric function admits a singular value type decomposition with smooth dynamics. The new model allows flexible nonlinear structures of the regression functions, while takes into account the correlation among response variables at the same time.

Our proposal is related to the factor models, which have been frequently used to characterize the correlation structure in multivariate data (Gary and Rothschild, 1983; Fan et al., 2013). In factor models, the signal of interest is expressed as a linear combination of a few latent variables, and does not concern additional covariate information that may play a role in estimation or prediction. For instance, factor models are often employed in contexts such as multiple time series or correlated functional data (Engle and Watson, 1981; Huang et al., 2009), where useful information may be hidden in the form of smoothness with respect to some additional covariates,

for example, temporal or spatial variable. Neglecting such information in recovery and prediction potentially hinders the quality and performance of the resulting estimators. This has been noticed by Durante et al. (2014), which further proposed a locally adaptive factor process under the Bayesian framework for characterizing multivariate mean-covariance changes in continuous time, allowing locally varying smoothness in both the mean and covariance matrix of multivariate time series. However, theoretical guarantees are lacking for the approach, which may leave practitioners uncertain about the quality of resulting estimates.

In this work, we approach the problem from a different perspective that is intuitive and broadly applicable. The contributions are summarized as follows. First, we propose a new nonparametric principal subspace regression model with a diverging model dimension. This not only incorporates the correlation structure among multivariate responses, but also accounts for the nonlinear smooth trend of the data with respect to the covariates. Second, we introduce a simple two-step estimation framework, where the first step is to obtain the left singular vectors of the data matrix and the second step is to estimate the nonparametric functions by local polynomial regression. Third, we provide theoretical guarantees. Specifically, we show that the space spanned by the estimated singular vectors can consistently estimate their underlying space, and obtain a uniform error bound for a diverging number of function estimates, which together ensure the convergence of the nonparametric principal subspace estimate. Lastly, we show that our method outperforms its counterpart, the conventional nonparametric regression, from both theoretical and numerical perspectives. This is not surprising because our approach significantly reduces the model complexity and risk of overfitting compared to individual nonparametric regressions.

The rest of the paper is organized as follows. In Section 2, we propose the nonparametric principal subspace regression methodology with a fitting procedure, and the main theoretical results are presented in Section 3. Favorable finite-sample performance is illustrated through simulated and real data examples in Section 4 and 5, respectively. Proofs of main results and technical lemmas are contained in the Appendix.

2. Proposed Methodology

In this section, we first introduce some notation used throughout the whole paper. Then we propose the nonparametric principal subspace regression methodology. By the end of this section, we describe the two-step fitting procedure and parameters tuning.

2.1 Notation

We begin by listing some notation used throughout the paper. For two vectors $a, b \in \mathbb{R}^m$, denote the inner product by $\langle a, b\rangle = a^\top b = \sum_{i=1}^m a_i b_i$ and the corresponding norm $\|\cdot\|$. Define the rescaled inner product $\langle a, b\rangle_m = \frac{1}{m}\langle a, b\rangle$ and the induced norm $\|\cdot\|_m$. For two functions $f, g \in L^2$, the inner product and corresponding norm bear the subscript L^2, that is $\langle f, g\rangle_{L^2} = \int_T f(x)g(x)dx$ and $\|f\|_{L_2}^2 = \int_T f^2(x)dx$, where T is the domain of x. Let $\|\cdot\|_\infty$ denote the sup norm of vector or function. For a matrix $M \in \mathbb{R}^{p\times n}$, write the singular value decomposition as $M = U\Sigma V^\top$, where $\Sigma = \text{diag}\{\sigma_1(M), \sigma_2(M), \ldots\}$ with singular values $\sigma_1(M) \geq \sigma_2(M) \geq \cdots \geq 0$. In particular, we use $\sigma_{\min}(M) = \sigma_{\min(p,n)}(M), \sigma_{\max}(M) =$

$\sigma_1(M)$ as the smallest and largest nontrivial singular values of M. Denote $\|M\| = \sigma_1(M)$ the spectral norm and $\|M\|_F = \sqrt{\sum_j \sigma_j^2(M)} = \sqrt{\sum_{k,l} M_{kl}^2}$ the Frobenius norm, respectively. Let $\mathcal{P}_M = M(M^\top M)^\dagger M^\top$ be the projection matrix onto the column space of M, where $(\cdot)^\dagger$ represents the Moore-Penrose pseudo-inverse. For two $p \times q$ matrices A_1 and A_2 with $p \geq q$ and the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_q \geq 0$ of $A_1^\top A_2$, define their principal angles as $\Theta(A_1, A_2) = \text{diag}(\cos^{-1}(\sigma_1), \ldots, \cos^{-1}(\sigma_q))$. A measure of distance between A_1 and A_2 is given by $\|\sin\Theta(A_1, A_2)\|$ or $\|\sin\Theta(A_1, A_2)\|_F$. For any $a, b \in \mathbb{R}$, let $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. We use c and C to denote generic positive constants that may vary in the sequel. For two positive sequences a_n and b_n, $a_n \lesssim b_n$ means $a_n \leq Cb_n$ for large n, $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$, and $a_n \ll b_n$ if $\limsup_{n\to\infty} a_n/b_n = 0$.

2.2 Nonparametric Principal Subspace Regression

Let $\{(x_i, y_i)\}_{i=1}^n$ be independent and identically distributed observations, where $x_i \in [0,1]^d$ and $y_i \in \mathbb{R}^p$. We consider the following nonparametric model

$$y_i = F(x_i) + z_i, \tag{1}$$

where $z_i = (z_{i1}, \ldots, z_{ip})^\top \in \mathbb{R}^p$ is independent and identically distributed with mean 0 and covariance Σ, and $F : [0,1]^d \to \mathbb{R}^p$ so that $\mathbb{E}(y_i|x_i) = F(x_i)$. The data dimension p is allowed to grow with the sample size n, and our goal is to estimate the function F, which characterizes the relationship between x_i and y_i, under mild smoothness assumption on the components of F.

Motivated by factor analysis and the singular value decomposition as methods of accounting for correlation among variables in y_i, as we show in Proposition 1, F can be written as

$$F(x) = \sum_{k=1}^{q} u_k f_k(x), \qquad q \leq p, \tag{2}$$

where $u_k \in \mathbb{R}^p$ $(1 \leq k \leq q)$ are orthonormal vectors and $\{f_k, 1 \leq k \leq q\}$ are smooth functions which are orthogonal in $L^2[0,1]^d$. We refer to q as the (underlying) model dimension and allow q to grow with n, reflecting its nonparametric nature. If one takes $q = p$, every function $F : [0,1]^d \to \mathbb{R}^p$ with components in $L^2[0,1]^d$ has such a representation; hence the model has the simple interpretation of reducing dimension with smoothness on covariates. Thus, in addition to capturing correlations via factor type analysis, this model also nonparametrically incorporates smoothness information into the covariates. We refer to our model as "nonparametric principal subspace regression".

With the model in place, we aim to find an appropriate approximation to F from an r-dimensional function-valued vector subspace defined as

$$\mathcal{G}_r := \left\{ G(x) \middle| G(x) = \sum_{k=1}^{r} v_k g_k(x), v_k^\top v_l = \delta_{kl}, x \in [0,1]^d, 1 \leq k, l \leq r \right\},$$

where $v_k \in \mathbb{R}^p$ are orthonormal vectors, $g_k \in L^2[0,1]^d$ are smooth functions and δ_{kl} is the Kronecker delta function. Note that the elimination of orthogonality of g_k's facilitates our algorithm

(see *Step 2* below), but does not affect the orthogonality of minimizers in $\mathcal{G}_r$ approximating to F. The approximation dimension r serves as a tuning parameter that may vary with q and n, depending on the balance between the approximation and estimation errors. Given $G \in \mathcal{G}_r$, we use the sample discrepancy

$$R_n(G) = \frac{1}{n}\sum_{i=1}^{n} \|F(x_i) - G(x_i)\|^2 \tag{3}$$

as a reasonable approximation to $R(G) = \int_{[0,1]^d} \|F(x) - G(x)\|^2 dx$ to measure the error between G and F. Thus we can estimate g_k and v_k ($1 \le k \le r$) from the sample pairs $\{(x_i, y_i)\}_{i=1}^n$ by optimizing the following problem

$$\min_{G \in \mathcal{G}_r} R_n^{\mathcal{D}}(G), \quad \text{where} \quad R_n^{\mathcal{D}}(G) = \frac{1}{n}\sum_{i=1}^{n} \|y_i - G(x_i)\|^2.$$

Let $V = (v_1, \ldots, v_r) \in \mathbb{R}^{p\times r}$ and $g(x_i) = (g_1(x_i), \ldots, g_r(x_i))^\top \in \mathbb{R}^r$. Then for a given i, by orthogonal projection, we have

$$\begin{aligned}\|y_i - Vg(x_i)\|^2 &= \|(I_p - \mathcal{P}_V)y_i + \mathcal{P}_V\big(y_i - Vg(x_i)\big)\|^2 \\ &= \|(I_p - \mathcal{P}_V)y_i\|^2 + \|\mathcal{P}_V\big(y_i - Vg(x_i)\big)\|^2,\end{aligned}$$

where $I_p - \mathcal{P}_V$ is the orthogonal complement of the projection matrix $\mathcal{P}_V$. Setting $c(Y, V) = \frac{1}{n}\sum_{i=1}^n \|(I_p - \mathcal{P}_V)y_i\|^2$, which contains no information of g, and observing that

$$\|\mathcal{P}_V\left(y_i - Vg(x_i)\right)\|^2 = \|V^\top y_i - g(x_i)\|^2 = \sum_{k=1}^{r}\Big(v_k^\top y_i - g_k(x_i)\Big)^2,$$

we may decompose the objective $R_n^{\mathcal{D}}(G)$ as

$$R_n^{\mathcal{D}}(G) = c(Y, V) + \sum_{k=1}^{r} R_n^{\mathcal{D}}(v_k, g_k) \quad \text{where} \quad R_n^{\mathcal{D}}(v_k, g_k) = \frac{1}{n}\sum_{i=1}^{n}\Big(v_k^\top y_i - g_k(x_i)\Big)^2.$$

This decouples the optimization problem of minimizing $R_n^{\mathcal{D}}(G)$ over $\mathcal{G}_r$ into two separate problems: finding a sequence of orthonormal vectors v_k's, and then estimating g_k by considering individual optimization of the $R_n^{\mathcal{D}}(v_k, g_k)$ along the direction v_k.

According to model (2), a reasonable choice of v_k's is to find the empirical counterpart of the singular vectors u_k for $1 \le k \le r$. Let $Y = (y_1, \ldots, y_n) \in \mathbb{R}^{p\times n}$ be the response data matrix. A natural estimator of $U_{[r]} = (u_1, \ldots, u_r)$ is the top r left singular vectors of Y. For the estimation of g_k's, there exist many standard nonparametric smoothing methods. For simplicity, we adopt the local polynomial regression in the sequel for implementation and theoretical development. This suggests a two-step fitting procedure.

Step 1. For a given $r \le q$, let $\hat{U}_{[r]} = (\hat{u}_1, \ldots, \hat{u}_r)$ be the top r left singular vectors of data $Y = (y_1, \ldots, y_n) \in \mathbb{R}^{p\times n}$ from model (1);

Step 2. Plug in $\hat{U}_{[r]}$ into $R_n^{\mathcal{D}}(G)$ and find the corresponding minimizers of the $R_n^{\mathcal{D}}(\hat{u}_k, g_k)$ by applying local polynomial smoothing for $k = 1, \ldots, r$ separately, denoted by $\hat{f}_1, \ldots, \hat{f}_r$.

Then the estimate $\hat{F}$ is given by

$$\hat{F} = \sum_{k=1}^{r} \hat{u}_k \hat{f}_k. \tag{4}$$

There are two tuning parameters involved in our estimation procedure, the retained dimension r in *Step 1* and the bandwidth h_k ($1 \le k \le r$) in *Step 2*. To select these parameters, for each r, we choose h_k by the standard five-fold cross-validation for each function estimate individually. Let $\hat{F}^{(r)}$ be the corresponding estimator of F using the retained dimension r with selected bandwidths $\hat{h}_{k,r}$. In view of the nonparametric approximation nature of $\hat{F}^{(r)}$, a reasonable choice for selecting r is to minimize

$$\text{AIC}(r) = \log\left\{V(r, \hat{F}^{(r)})\right\} + \frac{2r}{n},$$

where $V(r, \hat{F}^{(r)}) = (2n)^{-1} \sum_{i=1}^{n} \|y_i - \hat{F}^{(r)}(x_i)\|^2$.

3. Theoretical Guarantees

We shall present the main theoretical result, followed by an explicit rate of convergence when the proposed method is coupled with local polynomial smoothing, while the proofs are deferred to Appendix. We first present a proposition that ensures that a reasonable $F : [0,1]^d \to \mathbb{R}^p$ has a singular value type representation and supports the form of function proposed in this paper.

Proposition 1 *Suppose that $F : [0,1]^d \to \mathbb{R}^p$, which can be written as $F = (F_1, \ldots, F_p)^\top$, satisfies $F_j \in L^2[0,1]^d$ for $1 \le j \le p$. Then F has a singular value type decomposition*

$$F(\cdot) = \sum_{k=1}^{q} \sigma_k u_k v_k(\cdot) = \sum_{k=1}^{q} u_k f_k(\cdot), \qquad q \le p,$$

where v_k's are orthonormal in $L^2[0,1]^d$, u_k's orthonormal in $\mathbb{R}^p$ and $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_q > 0$.

It is noted that Proposition 1 holds under a rather weak condition $F_j \in L^2[0,1]^d$ that is mostly satisfied in practice. To make model (2) feasible, we need an additional assumption for F,

$$\|f_k\|_{L_2}^2 = \sigma_k^2 \asymp k^{-\alpha}, \qquad \alpha > 1, \qquad 1 \le k \le q. \tag{5}$$

This type of polynomial decay condition on singular values is widely used in the field of high-dimensional statistics (Wainwright, 2019; Vershynin, 2019).

For simplicity, we assume the design points x_i's are independent and identically distributed with x_i following uniform distribution on $[0,1]^d$, that is $x_i \sim \mathcal{U}[0,1]^d$. This assumption can be relaxed with more technicality, see Remark 3. We also assume $z_i = (z_{i1}, \ldots, z_{ip})^\top \in \mathbb{R}^p$ are independent and identically distributed as $\mathcal{N}(0, \sigma^2 I_p)$, which puts us in the regime of p repeated nonparametric experiments. The assumption of uncorrelated Gaussian noise is commonly used

to facilitate model exploration and theoretical development (Cai, 2012; Donoho and Johnstone, 1994; Tsybakov, 2009; Johnstone, 2017) in the study of nonparametric experiments. This is in accordance to a common practice in nonparametric regression which incorporates the structural information into the underlying function F contaminated by independent noise.

Given the estimate $\hat{U}_{[r]}$ of $U_{[r]}$ in *Step 1*, we define the estimated rotated response as $\hat{y}^*_{ik} = \hat{u}_k^\top y_i$ and the oracle rotated data $\{(x_i, y^o_{ik})\}$ as

$$y^o_{ik} = u_k^\top y_i = f_k(x_i) + \epsilon_{ik}, \quad \text{for} \quad 1 \le k \le r,\ 1 \le i \le n,$$

where $\epsilon_{ik} = u_k^\top z_i$ are independent and identically distributed from $\mathcal{N}(0, \sigma^2)$ by the model assumptions. Denote the linear smoother by $\mathcal{L}$ and let $\hat{f}^o_k$ be the nonparametric estimates by applying $\mathcal{L}$ to the oracle rotated data $\{(x_i, y^o_{ik})\}$. Note that $\hat{f}^o_k$'s are not present in the estimation procedure, but serve as an intermediate quantity in theoretical analysis.

Therefore, in order to obtain the main result (7) in Theorem 2 below, we need to bound the rotation error between the estimated principal subspace $\hat{U}_{[r]}$ and its true counterpart $U_{[r]}$ as (6) states. This bound is of interest in its own rights given broad applicability in singular value type problems. It is also required to quantify the smoothing error for estimating f_k, where $1 \le k \le r$, where the difference between $\hat{f}^o_k$ and the actual estimates $\hat{f}_k$ is quantified via the rotation error (6). It is important to note that r is not fixed but may (slowly) diverge with n, which makes the classical theory on nonparametric regression for a fixed number of function estimates invalid. For the flow of exposition, we present here the needed condition and rate from the smoothing step and show their fulfillment afterward.

Theorem 2 *Assume that f_k's are uniformly bounded so that $\max_{1\le k\le q} \|f_k\|_\infty \le B$ and $\|f_k\|^2_{L_2}$'s satisfy a polynominal decay condition* (5)*. Let $\hat{U}_{[r]}$ be the estimate of $U_{[r]}$ in Step 1. If $r \ll (n/q^2 \log n)^{1/(2\alpha+2)}$, $q \ll \sqrt{p}$ and $p \ll n$, then*

$$\mathbb{E}\|\sin\Theta(\hat{U}_{[r]}, U_{[r]})\|^4 \lesssim \frac{p^2 r^{4\alpha+4} \log^2 n}{n^2}. \tag{6}$$

In addition, assume that the linear smoother $\mathcal{L}$ satisfies $\|\mathcal{L}\| \le C$ and $\mathbb{E}\|\hat{f}^o_k - f_k\|^2_n \lesssim k^\tau/n^\rho$ uniformly for $1 \le k \le q$, where τ and ρ are constants associated with the smoothness of f_k. Then the estimator $\hat{F} = \sum_{k=1}^r \hat{u}_k \hat{f}_k$ in (4) *satisfies*

$$\mathbb{E}[R_n(\hat{F})] \lesssim r^{-\alpha+1} + \frac{p r^{2\alpha+3} \log n}{n} + \frac{r^{\tau+1}}{n^\rho}, \tag{7}$$

where $R_n(\hat{F}) = n^{-1}\sum_{i=1}^n \|F_i(x_i) - \hat{F}(x_i)\|^2$ is the sample discrepancy defined in (3)*.*

Note that the assumption $r \ll (n/q^2 \log n)^{1/(2\alpha+2)}$ is to bound the gap between the estimated singular values, and the approximation error $\sum_{k=r+1}^q \sigma_k^2$ is bounded by $r^{-\alpha+1}$ regardless of the underlying model dimension q due to the decaying structure (5). The condition $q \ll \sqrt{p}$ is reasonable since q is usually much smaller than p in a low-rank approximation problem given the decay rate of singular values. Based on such observations, we see that the first term arises from the approximation error, while the second and third terms are due to estimation errors for

the principal subspace $U_{[r]}$ and the smoothing of f_k ($1 \le k \le r$), respectively. The proposed estimator is consistent, provided that

$$r \ll \left(\frac{n}{p \log n}\right)^{\frac{1}{2\alpha+3}} \wedge n^{\frac{\rho}{\tau+1}},$$

which reflects the essence of dimension reduction achieved by the proposed method. By comparing the order of each error term in (7), one may obtain the optimal convergence rate by choosing r appropriately. Specifically, if $pn^{\rho-1} \log n \lesssim r^{\tau-2\alpha-2}$ and $r \asymp n^{\rho/(\alpha+\tau)}$, the optimal rate is $n^{-\rho(\alpha-1)/(\alpha+\tau)}$; if $r^{\tau-2\alpha-2} \lesssim pn^{\rho-1} \log n$ and $r \asymp (n/p \log n)^{1/(3\alpha+2)}$, the optimal rate changes to $(p \log n/n)^{(\alpha-1)/(3\alpha+2)}$.

Remark 3 *If x_i are on a grid, saying $\{0, 1/m, \ldots, (m-1)/m, 1\}$. By defining $\mathbb{E}f^2(x_i) = m^{-1} \sum_{i=1}^m f^2(i/m)$ as a Riemann integral approximation to $\|f\|_{L_2}^2$, the result in this section remains valid by considering the additional integral approximation error. If x_i is sampled from some distribution π, we consider the transformation $y_i = F(\pi^{-1}(w_i)) + z_i = H(w_i) + z_i$, where w_i follows a uniform distribution on $[0,1]^d$. Let $\hat{\pi}$ be the empirical distribution based on x_i, then one can perform the proposed method to estimate H based on the sample pairs $\{(\hat{w}_i, y_i)\}$, where $\hat{w}_i = \hat{\pi}^{-1}(x_i)$. Consequently, the estimation of F is $\hat{F}(x) = \hat{H}(\hat{\pi}(x))$. For the fact that $\hat{\pi}$ enjoys a standard nonparametric rate only depending on n and d, the impact of transformation is negligible when $d \ll p$, thus the result also holds.*

We next turn to demonstrate that the requirements on a diverging number r of smoothing estimates of f_k ($1 \le k \le r$) are fulfilled. For conciseness, we focus on the common local polynomial regression (Fan and Gijbels, 1996; Tsybakov, 2009) that is implemented for numerical studies, while other smoothing methods can also be investigated with more technicality.

To bound the errors of a diverging number of nonparametric function estimators, the key is to extend and combine the smoothness classes to which such q functions belong. Recall that the smoothness class of primary concern in standard nonparametric regression is the Hölder class $\mathcal{H}(\beta, L)$, consisting of $l = \lfloor \beta \rfloor$ times differentiable functions f, where $\lfloor \beta \rfloor$ represents the largest integer strictly less than β, with the l-th derivative $f^{(l)}$ satisfying $|f^{(l)}(x) - f^{(l)}(y)| \le L|x-y|^{\beta-l}$ for x, y in the domain of interest. The idea arises from the fact that, for a sequence of orthonormal basis functions $\{v_k\}_{k=1}^{\infty}$, the smoothness of v_k deteriorates as index k increases. Thus we assume that the orthonormal functions v_k's in Proposition 1, hence f_k's, belong to different Hölder classes $\mathcal{H}(\beta, L_k)$ for $1 \le k \le q$. The Hölder constants L_k depict the amplitude of its derivatives which characterizes the function's frequency increment. Here are the examples of explicit forms of L_k for some commonly used basis functions with a given β.

Exp.1 Fourier Series: $\psi_0 = 1$, $\psi_{2k-1} = \sin(k\pi t)$ and $\psi_{2k} = \cos(k\pi t)$, then $L_k \asymp k^{\lfloor \beta \rfloor}$.

Exp.2 B-splines: $\{N_{jk}\}_{j=-k-1}^{J}$, where N_{jk} is defined in (6.20) in Hsing and Eubank (2015) and k is the order, then $L_k \asymp k!/(k - \lfloor \beta \rfloor)! \lesssim k^{\lfloor \beta \rfloor}$.

Exp.3 Wavelets: $\psi_{j,k} = 2^{j/2}\psi(2^j t - k)$ with a mother wavelet function ψ, then $L_k \asymp k^{\lfloor \beta \rfloor}$.

We adopt a similar argument as in Cai and Brown (1999) for the random design under consideration, and derive the minimax rate for the oracle functional estimates $\hat{f}_k^o$ $(1 \le k \le q)$ in the metric $\mathbb{E}_f\|\cdot\|_n^2$.

Theorem 4 *Suppose that $v_k \in \mathcal{H}(\beta, L_{lk})$ with $l > 1$ and $L_{1k} = \|v_k'\|_\infty$, and the kernel K satisfies the conditions outlined in Tsybakov (2009). Then the local polynomial smoother $\mathcal{L}$ is bounded and the resulting estimator $\hat{f}_k^o$ of $f_k = \sigma_k v_k$ satisfies*

$$\mathbb{E}_{f_k}\|\hat{f}_k^o - f_k\|_n^2 \lesssim \left(1 \vee (\sigma_k^2 L_{1k}^2)^{\frac{2\beta}{2\beta+1}}\right)\left(\sigma_k^2 L_{lk}^2\right)^{\frac{1}{2\beta+1}} n^{-\frac{2\beta}{2\beta+1}},$$

uniformly for $1 \le k \le q$, by choosing optimal bandwidth $h \asymp \left(\left(1 \vee \sigma_k^2 L_{1k}^2\right)/n\sigma_k^2 L_{lk}^2\right)^{1/(2\beta+1)}$. Moreover, if $L_{lk} \asymp k^l$ and $L_{1k} \asymp k$, then

$$\mathbb{E}_{f_k}\|\hat{f}_k^o - f_k\|_n^2 \lesssim \frac{k^\tau}{n^\rho} \quad \textit{with} \quad \tau = \frac{2(2-\alpha)\beta \vee 0 + 2l - \alpha}{2\beta+1}, \quad \rho = \frac{2\beta}{2\beta+1}. \tag{8}$$

We conclude this section by comparing with the curve-by-curve estimator of F without extracting the subspace. By Proposition 1 and Theorem 4, it is seen that $F_j \in \mathcal{H}(\beta, M_{lj})$, where $M_{lj} = \sum_{k=1}^q u_{kj}\sigma_k L_{lk}$ and $M_{1j} = \sum_{k=1}^q u_{kj}\sigma_k L_{1k}$. By Cauchy-Schwartz inequality, $M_{lj}^2 \le q^{2l-\alpha+1}\sum_{k=1}^q u_{kj}^2$ and $M_{1j}^2 \le q^{3-\alpha}\sum_{k=1}^q u_{kj}^2$. Following the same argument as in proof of Theorem 4, we have

$$\mathbb{E}_{F_j}\|\hat{F}_j - F_j\|_n^2 \lesssim \left(1 \vee M_{1j}^{\frac{4\beta}{2\beta+1}}\right) M_{lj}^{\frac{2}{2\beta+1}} n^{-\frac{2\beta}{2\beta+1}},$$

where $\hat{F}_j$ is the local polynomial estimator of F_j for $j = 1, \ldots, p$. Then the convergence rate of the curve-by-curve estimator $\hat{F}_{cbc} = \left(\hat{F}_1, \ldots, \hat{F}_p\right)$ is

$$\mathbb{E}[R_n(\hat{F}_{cbc})] \lesssim \left(q^{\frac{2l-\alpha+2}{2\beta+1}} p^{\frac{2\beta}{2\beta+1}} \vee q^{\frac{8\beta-2\alpha\beta+2l-\alpha+2}{2\beta+1}}\right) n^{-\frac{2\beta}{2\beta+1}}.$$

Note that $\beta \ge l, \alpha > 1$ and if $p \gtrsim q^{4-\alpha}$, it holds that

$$\mathbb{E}[R_n(\hat{F}_{cbc})] \lesssim q\left(\frac{p}{n}\right)^{\frac{2\beta}{2\beta+1}}.$$

In comparison with the proposed method, plugging (8) into (7),

$$\mathbb{E}[R_n(\hat{F})] \lesssim r^{-\alpha+1} + \frac{p\log n}{n} r^{2\alpha+3} + r^3 n^{-\frac{2\beta}{2\beta+1}}.$$

The essence of the proposed method is dimension reduction based on low-rank approximation, that is, the reduced dimension r is often much smaller than q and p, which implies that $\mathbb{E}[R_n(\hat{F})] \ll \mathbb{E}[R_n(\hat{F}_{cbc})]$.

4. Simulation Study

In this section, we perform simulation study to evaluate the performance of the proposed method. The $U = (u_1, \ldots, u_q)$ is generated by orthonormalizing a $p \times q$ matrix with all elements being independent and identically distributed standard normal. The x_i's are independently generated from uniform distribution on $[0, 1]$. The functions f_k's are independently generated from a zero mean Gaussian process with compactly supported covariance function

$$C_{a,b}(s,t) = b \max\{0, (1-h_a)^5\}(8h_a^2 + 5h_a + 1),$$

where $h_a(s,t) = |s-t|/a$, see Rasmussen and Williams (2006) for details. We set $a = 0.5$ and $b = 15$. We orthogonalize and scale f_k's by

$$\|f_k\|_{L_2}^2 = \sigma_k^2 = 5k^{-2}, \quad 1 \le k \le 6 \quad \text{and} \quad \sigma_k^2 = k^{-2}, \quad 6 < k \le q.$$

The error z_{ij} are independent and identically distributed standard normal $\mathcal{N}(0, \sigma^2)$ for $1 \le i \le n$ and $1 \le j \le p$. The response y_i is obtained by $y_i = \sum_{k=1}^q u_k f_k(x_i) + z_i$.

A natural comparison would be conducted against individual nonparametric regression of Y on x in a curve-by-curve manner. In particular, we compare with the method that fits the jth component of Y on x nonparametrically for each $1 \le j \le p$. We also use the local polynomial regression with a Gaussian kernel for curve-by-curve nonparametric recovery. For the tuning parameters, we use the selection method described in Section 2. For nonparametric principal subspace regression, we report the average values of selected $\hat{r}$ by AIC and the average estimation errors based on 100 Monte Carlo runs. For curve-by-curve nonparametric recovery, we only report the average estimation errors since the procedure fits each curve individually. We consider different combinations of (n, p, q) at two noise levels $\sigma = 0.5$ and $\sigma = 1$.

We first check the performance of two approaches under the typical low-rank scenario that p is much larger than q. The results in Table 1 suggest that our method outperforms the curve-by-curve nonparametric regression for all cases. Given the sample size n, the recovery results from nonparametric principal subspace regression tend to improve at a faster rate as p increases or q decreases. We then conduct the simulation when q increases to the extent close to p and the results are reported in Table 2. We see that our method still outperforms. In particular, the estimation error and the selected $\hat{r}$ become to level off even as q increases substantially.

The selection of r plays an important role in our method. As seen from Tables 1 and 2, the AIC is capable of extracting most of the important signals. Moreover, the selected $\hat{r}$ becomes smaller with a higher noise level that tends to blind small signals. To demonstrate the effectiveness of AIC, we plot the average estimation errors with increasing r and labeled the averages of selected $\hat{r}$ in Figure 2. It is observed that the AIC criterion indeed selects models comparable with those of minimal prediction error.

5. Real Data Applications

In this section, we apply our methodology to two data examples introduced in Section 1, which shows a favorable performance in comparison with the conventional nonparametric regression.

			$\sigma = 0.5$			$\sigma = 1$		
n	p	q	$\hat{r}$	NPSR	CBC	$\hat{r}$	NPSR	CBC
300	50	2	2.14	**0.171(0.003)**	0.546(0.017)	2.14	**0.634(0.013)**	1.728(0.024)
		4	4.09	**0.312(0.003)**	0.587(0.011)	4.09	**1.269(0.013)**	1.827(0.019)
		6	6.13	**0.471(0.004)**	0.611(0.011)	6.34	**2.100(0.018)**	1.877(0.019)
	100	3	3.01	**0.372(0.003)**	0.972(0.009)	3.01	**1.539(0.012)**	3.258(0.034)
		6	6.01	**0.755(0.004)**	1.053(0.010)	5.49	**3.446(0.024)**	3.452(0.033)
		9	6	**0.798(0.004)**	1.066(0.010)	5.48	**3.482(0.019)**	3.456(0.029)
	150	4	4	**0.668(0.004)**	1.422(0.014)	4	**2.936(0.018)**	4.780(0.048)
		8	6	**1.077(0.005)**	1.486(0.014)	4.72	**4.589(0.021)**	4.938(0.046)
		12	6.01	**1.115(0.005)**	1.488(0.012)	4.62	**4.624(0.022)**	4.932(0.045)
500	100	3	3.02	**0.239(0.002)**	0.718(0.007)	3.02	**0.978(0.007)**	2.472(0.025)
		6	6.02	**0.491(0.003)**	0.768(0.007)	6.05	**2.316(0.014)**	2.630(0.024)
		9	6.05	**0.537(0.003)**	0.768(0.007)	6.04	**2.374(0.015)**	2.625(0.025)
	200	4	4	**0.531(0.002)**	1.361(0.013)	4	**2.377(0.012)**	4.678(0.046)
		8	6	**0.871(0.003)**	1.420(0.013)	5.12	**3.875(0.014)**	4.851(0.044)
		12	6	**0.909(0.003)**	1.428(0.012)	5.03	**3.921(0.017)**	4.854(0.044)
	300	5	5	**0.955(0.004)**	1.984(0.018)	4.22	**4.313(0.016)**	6.790(0.063)
		10	6	**1.229(0.004)**	2.039(0.018)	4.18	**5.088(0.018)**	6.970(0.064)
		15	6	**1.262(0.004)**	2.052(0.019)	4.27	**5.106(0.017)**	6.992(0.066)

Table 1: The average estimation errors for our nonparametric principal subspace regression (NPSR) and the curve-by-curve nonparametric regression (CBC) with their associated standard errors in the parentheses, and the average values of selected $\hat{r}$ by AIC in 100 Monte Carlo runs are reported.

		$\sigma = 0.5$			$\sigma = 1$		
(n, p)	q	$\hat{r}$	NPSR	CBC	$\hat{r}$	NPSR	CBC
(300,150)	4	4	**0.668(0.004)**	1.422(0.014)	4	**2.936(0.018)**	4.780(0.048)
	40	6	**1.168(0.005)**	1.525(0.011)	4.77	**4.660(0.020)**	5.003(0.046)
	80	6	**1.174(0.005)**	1.537(0.012)	4.8	**4.670(0.021)**	5.010(0.049)
	120	6.01	**1.191(0.005)**	1.545(0.013)	4.77	**4.735(0.023)**	5.033(0.047)
(500,300)	5	5	**0.955(0.004)**	1.984(0.018)	4.22	**4.313(0.016)**	6.790(0.063)
	50	6	**1.307(0.004)**	2.078(0.019)	4.25	**5.141(0.016)**	7.005(0.066)
	100	6	**1.312(0.004)**	2.081(0.019)	4.25	**5.138(0.019)**	6.981(0.065)
	200	6	**1.315(0.004)**	2.086(0.018)	4.19	**5.145(0.017)**	6.993(0.063)

Table 2: The average estimation errors for our nonparametric principal subspace regression (NPSR) and the curve-by-curve nonparametric regression (CBC) with their associated standard errors in the parentheses, and the average values of selected $\hat{r}$ by AIC in 100 Monte Carlo runs are reported.

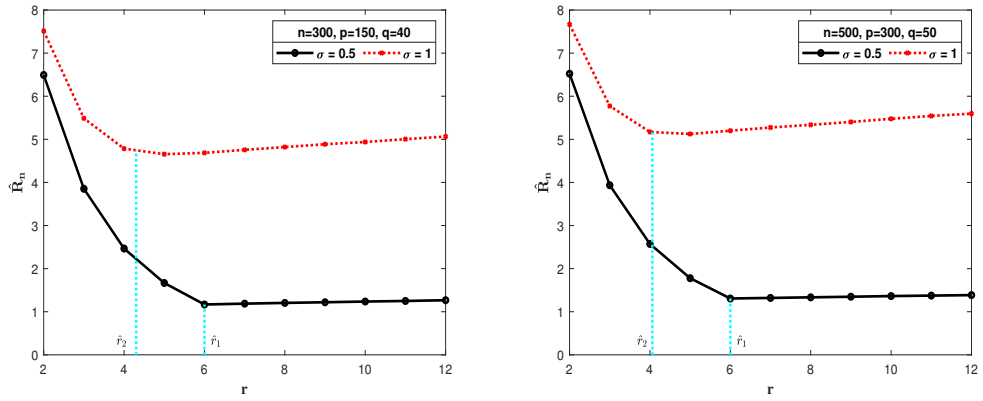

Figure 2: The average estimation errors $\hat{R}_n$ for our nonparametric principal subspace regression with different r under two noise levels. The average values of selected $\hat{r}$ by AIC in 100 Monte Carlo runs are depicted (vertical dotted lines), suggesting the effectiveness of AIC.

5.1 Application to an EEG Study

We apply the proposed method to an EEG data set, which is available at `https://archive.ics.uci.edu/ml/datasets/EEG+Database`. The data were collected by the Neurodynamics Laboratory and contain 122 subjects, where researchers measured the voltage values from 64 electrodes placed on each subject's scalps sampled at 256 Hz for 1 second. As EEG data are notoriously noisy while there are known to be strong relations between different electrodes, we model the data from each subject by the nonparametric framework (1). Thus, for each subject, we fit the nonparametric principal subspace regression to the data matrix $Y = (y_1, \ldots, y_n) \in \mathbb{R}^{p \times n}$ with $p = 64$ and $n = 256$. The average retained dimension selected by the proposed AIC among these 122 subjects is 7.270 with a standard error 0.112.

To compare the prediction performance, we also fit curve-by-curve nonparametric regression to the signals obtained from each of the 64 electrodes. For each subject, we randomly reserve 10% of data as the test set: $\mathcal{S}_{test} \subseteq \{1, \ldots, 256\}$ such that $|\mathcal{S}_{test}|/256 \approx 10\%$, while using the rest as the training set, and report the prediction errors $|\mathcal{S}_{test}|^{-1} \sum_{i \in \mathcal{S}_{test}} \|Y_i - \hat{F}(x_i)\|^2/64$ for both approaches. The average prediction error for nonparametric principal subspace regression over the 122 subjects is 1.153 with a standard error 0.072, while that obtained by the curve-by-curve nonparametric regression is 1.288 with a standard error 0.075.

5.2 Application to an fMRI Study

For another data application, we analyze the motor task-related fMRI data from the Human Connectome Project (HCP) Data `https://www.humanconnectome.org/` which includes behavioral and 3 Tesla magnetic resonance imaging data from 970 healthy adult participants collected from 2012 to spring 2015. The block-design motor task used in this study is adapted

from experiments by Buckner et al. (2011) and Yeo et al. (2011). The details on the HCP implementation can be referred to Barch et al. (2013). 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. For each subject, there are two runs of phase encoding scans (right-to-left and left-to-right), and we use the left-to-right phase encoding scan in this study. Each run of the motor task lasted for about 205 seconds including 284 frames, and we use the "Desikan-Killiany" atlas (Desikan et al., 2006) to divide the brain into 68 ROIs.

We use 869 subjects which had the motor task-related fMRI data. For each subject, we obtain the data matrix $Y \in \mathbb{R}^{p\times n}$ with $p = 68$ and $n = 284$. By fitting the nonparametric principal subspace regression, the average selected retained dimension among these 869 subjects is 8.353 with standard error 0.049. Same as the above example, we randomly select 10% of data as the test set and the rest of the data as the training set for each subject. The average prediction error for the proposed method over 869 subjects is 2.629 with a standard error 0.100, while that obtained by the curve-by-curve nonparametric regression is 2.727 with a standard error 0.103.

Acknowledgments

Yang Zhou is the first author, and Fang Yao is the corresponding author. Yang Zhou's research is partially supported by the Postdoctoral Science Foundation of China (Grant no. 2020M680226 and 2020TQ0014) and National Natural Science Foundation of China (Grant no. 11971048). Dengdeng Yu's research is partially supported by the Canadian Statistical Sciences Institute postdoctoral fellowship. Dehan Kong's research is partially supported by the Natural Science and Engineering Research Council of Canada. Fang Yao's research is partially supported by National Natural Science Foundation of China Grants 11931001 and 11871080, the National Key R&D Program of China Grant 2020YFE0204200, the LMAM, and the Key Laboratory of Mathematical Economics and Quantitative Finance (Peking University), Ministry of Education.

Appendix A. Notations

We first introduce the notation used in the appendix. Recall we have proposed a nonparametric model $y_i = \sum_{k=1}^{q} u_k f_k(x_i) + z_i \in \mathbb{R}^p$, where $z_i = (z_{i1}, \ldots, z_{ip})^\top$ follows independent and identically distributed $\mathcal{N}(0, \sigma^2 I_p)$. The design points x_i's are independent and identically distributed $\mathcal{U}[0,1]^d$. Let $Y = (y_1, \ldots, y_n) \in \mathbb{R}^{p\times n}$ be the response data matrix, $\tilde{F} = \big(F(x_1), \ldots, F(x_n)\big) \in \mathbb{R}^{p\times n}$ and $Z = (z_1, \ldots, z_n) \in \mathbb{R}^{p\times n}$, one can write $Y = \tilde{F} + Z$. Let $U_{[r]} = (u_1, \ldots, u_r) \in \mathbb{R}^{p\times r}$, and $\hat{U}_{[r]} = (\hat{u}_1, \ldots, \hat{u}_r) \in \mathbb{R}^{p\times r}$ the estimate of $U_{[r]}$. Define $\tilde{f}_k = \big(f_k(x_1), \ldots, f_k(x_n)\big)^\top \in \mathbb{R}^n$ and $\tilde{f} = (\tilde{f}_1, \ldots, \tilde{f}_q)^\top \in \mathbb{R}^{q\times n}$ so that $\tilde{F} = U\tilde{f}$. We further define $\hat{\tilde{f}}_k = \big(\hat{f}_k(x_1), \ldots, \hat{f}_k(x_n)\big)^\top \in \mathbb{R}^n$ and $\hat{\tilde{f}}_{[r]} = (\hat{\tilde{f}}_1, \ldots, \hat{\tilde{f}}_r)^\top \in \mathbb{R}^{r\times n}$ so that we may write $\hat{\tilde{F}} = \big(\hat{F}(x_1), \ldots, \hat{F}(x_n)\big) \in \mathbb{R}^{p\times n}$ as $\hat{\tilde{F}} = \hat{U}_{[r]}\hat{\tilde{f}}_{[r]}$.

Appendix B. Proofs of the Proposition and Main Theorems

Proof. [**Proposition 1**] Given $F : [0,1]^d \to \mathbb{R}^p$, we can define an operator $\mathcal{F} : L^2[0,1]^d \to \mathbb{R}^p$ mapping $h \in L^2[0,1]^d$ to

$$\mathcal{F}h = \langle F, h\rangle_{L^2} := \left(\langle F_1, h\rangle_{L^2}, \ldots, \langle F_p, h\rangle_{L^2}\right)^\top \in \mathbb{R}^p.$$

In this case we can represent the operator $\mathcal{F}$ as

$$\mathcal{F} = \sum_{j=1}^{p} e_j \otimes F_j \quad \text{and} \quad \mathcal{F}^*\mathcal{F}h = \sum_{j=1}^{p} \langle F_j, h\rangle_{L^2} F_j.$$

where $\otimes$ is the Kronecker product. Thus $\mathcal{F}^*\mathcal{F}$ is of finite rank and hence compact. As it is also symmetric, it has an eigendecomposition

$$\mathcal{F}^*\mathcal{F} = \sum_{k=1}^{\infty} \sigma_k^2 v_k \otimes v_k$$

with at most p of the $\sigma_k \neq 0$ and v_k's forming an orthonormal basis of $L^2[0,1]^d$. We order the nonzero σ_k decreasingly by $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_q > 0$ with $q \leq p$. Note that we may write

$$h = \sum_{k=1}^{\infty} \langle h, v_k\rangle_{L^2} v_k.$$

Setting $w_k = \mathcal{F}v_k$ we have that $\langle w_k, w_l\rangle = \langle v_k, \mathcal{F}^*\mathcal{F}v_l\rangle_{L^2} = \sigma_l^2 \langle v_k, v_l\rangle_{L^2} = \sigma_l^2 \delta_{kl}$. Hence, the w_k's are orthogonal vectors for $1 \leq k \leq q$. Letting $u_k = \sigma_k^{-1} w_k$, we may write

$$\mathcal{F}h = \sum_{k=1}^{\infty} \langle h, v_k\rangle_{L^2} \mathcal{F}v_k = \sum_{k=1}^{q} \sigma_k \langle h, v_k\rangle_{L^2} u_k = \left(\sum_{k=1}^{q} \sigma_k u_k \otimes v_k\right) h.$$

Since $F_j = \sum_{k=1}^{\infty} \langle F_j, v_k\rangle_{L^2} v_k$, we have

$$F = \sum_{k=1}^{\infty} \langle F, v_k\rangle_{L^2} v_k = \sum_{k=1}^{\infty} \left(\mathcal{F}v_k\right) v_k = \sum_{k=1}^{q} \sigma_k u_k v_k,$$

which completes the proof. ■

Proof. [**Theorem 2**] As the $\tilde{f}_k$'s are not necessarily orthogonal in $\mathbb{R}^n$, we need to consider that the singular value decomposition of $\tilde{F}$ is of the form $\tilde{F} = P\Lambda Q^\top$, with P possibly spanning a different subspace from U. Let Y admit the singular value decomposition $Y = \hat{U}\hat{\Lambda}\hat{V}^\top$. Recall that $Y = \tilde{F} + Z = U\tilde{f} + Z = P\Lambda Q^\top + Z$ and note that

$$\|\sin\Theta(\hat{U}_{[r]}, U_{[r]})\|^4 \leq C\left[\|\sin\Theta(\hat{U}_{[r]}, P_{[r]})\|^4 + \|\sin\Theta(P_{[r]}, U_{[r]})\|^4\right],$$

we shall bound $\mathbb{E}\|\sin\Theta(\hat{U}_{[r]}, P_{[r]})\|^4$ and $\mathbb{E}\|\sin\Theta(P_{[r]}, U_{[r]})\|^4$ separately below.

Define the event $\mathcal{E}$ as

$$\mathcal{E} = \left\{ \max_{1\le k,l\le q} |\langle f_k, f_l\rangle_n - \langle f_k, f_l\rangle_{L_2}| \le 4B^2\sqrt{\frac{\log n}{n}} \right\}.$$

By Lemma 7, it holds that $\mathbb{P}\big(\mathcal{E}^c\big) \le q(q+1)/n^2$. Denote $\mathbb{E}(\cdot|\mathcal{D})$ the expectation conditional on the design $\{x_1,\ldots,x_n\}$. As $\|\sin\Theta(\hat{U}_{[r]}, P_{[r]})\| \le 1$, we decompose $\mathbb{E}\|\sin\Theta(\hat{U}_{[r]}, P_{[r]})\|^4$ as

$$\begin{aligned}\mathbb{E}\|\sin\Theta(\hat{U}_{[r]}, P_{[r]})\|^4 &= \mathbb{E}\big[\mathbb{E}\big(\|\sin\Theta(\hat{U}_{[r]}, P_{[r]})\|^4\big|\mathcal{D}\big)\big] \\ &\le \mathbb{P}(\mathcal{E}^c) + \mathbb{E}\big[\mathbb{E}\big(\|\sin\Theta(\hat{U}_{[r]}, P_{[r]})\|^4\big|\mathcal{D}\big)\big|1_{\mathcal{E}}\big].\end{aligned}$$

On the event $\mathcal{E}$, Lemma 8 ensures (12) holds. Since $q \ll \sqrt{p}$ and $p \ll n$ there exists a constant $C_{\max}$ such that

$$\sigma^2_{\max}(\tilde{F}) \le n\sigma^2_{\max} + 4B^2\sqrt{q^2 n\log n} \le C_{\max} n.$$

Under the assumption that $r \ll (n/q^2\log n)^{1/(2\alpha+2)}$, it holds that

$$4B^2\sqrt{\frac{q^2\log n}{n}} \le \frac{1}{3}(\sigma_r^2 - \sigma_{r+1}^2).$$

Combing this with Lemma 10 we get

$$\mathbb{E}\left(\|\sin\Theta(\hat{U}_{[r]}, P_{[r]})\|^4\big|\mathcal{D}\right) \le \frac{Cp^2(\sigma_r^2(\tilde{F})+n)^2}{\big(\sigma_r^2(\tilde{F}) - \sigma_{r+1}^2(\tilde{F})\big)^4} \le \frac{Cp^2(2n\sigma_r^2+n)^2}{n^4(\sigma_r^2-\sigma_{r+1}^2)^4} \le \frac{Cp^2 r^{4\alpha+4}}{n^2}.$$

As a result, we obtain that

$$\mathbb{E}\|\sin\Theta(\hat{U}_{[r]}, P_{[r]})\|^4 \le \frac{Cq^2}{n^2} + \frac{Cp^2r^{4\alpha+4}}{n^2} \lesssim \frac{p^2r^{4\alpha+4}}{n^2}.$$

Using Proposition 1 in Cai and Zhang (2018), it holds that

$$\|\sin\Theta(P_{[r]}, U_{[r]})\| \le \frac{\sigma_r(\tilde{F}^\top U_{[r]})\|\mathcal{P}_{(\tilde{F}^\top U_{[r]})}\tilde{F}^\top U_{[r]}^\perp\|}{\sigma_r^2(\tilde{F}^\top U_{[r]}) - \sigma_{r+1}^2(\tilde{F}^\top)}.$$

Note that $\sigma_r^2(\tilde{F}^\top U_{[r]}) = \sigma_r^2(\tilde{f}_{[r]})$, $\sigma_{r+1}^2(\tilde{F}^\top) = \sigma_{r+1}^2(\tilde{f})$ and

$$\|\mathcal{P}_{(\tilde{F}^\top U_{[r]})}\tilde{F}^\top U_{[r]}^\perp\| = \|\tilde{f}_{[r]}^\top(\tilde{f}_{[r]}\tilde{f}_{[r]}^\top)^{-1}\tilde{f}_{[r]}(\tilde{f}_{[r]}^\perp)^\top\| \le \sigma_{\min}^{-1}(\tilde{f}_{[r]}^\top)\|\tilde{f}_{[r]}(\tilde{f}_{[r]}^\perp)^\top\|$$

where we use (8.18) in Cai and Zhang (2018) in the inequality. Hence,

$$\|\sin\Theta(P_{[r]}, U_{[r]})\|^4 \le \frac{C\sigma_r^4(\tilde{f}_{[r]})\sigma_{\min}^{-4}(\tilde{f}_{[r]}^\top)\|\tilde{f}_{[r]}(\tilde{f}_{[r]}^\perp)^\top\|^4}{\big(\sigma_r^2(\tilde{f}_{[r]}) - \sigma_{r+1}^2(\tilde{f})\big)^4} = \frac{C\|\tilde{f}_{[r]}(\tilde{f}_{[r]}^\perp)^\top\|^4}{\big(\sigma_r^2(\tilde{f}_{[r]}) - \sigma_{r+1}^2(\tilde{f})\big)^4}.$$

On the event $\mathcal{E}$, following the proof of Lemma 8, it is easily seen that

$$\|\tilde{f}_{[r]}(\tilde{f}_{[r]}^\perp)^\top\|^4 \le \|\tilde{f}_{[r]}(\tilde{f}_{[r]}^\perp)^\top\|_F^4 \le Cq^2r^2n^2\log^2 n \lesssim p^2n^2\log^2 n$$

which deduces that

$$\mathbb{E}\|\sin\Theta(P_{[r]}, U_{[r]})\|^4 \leq \frac{Cq^2}{n^2} + \frac{Cp^2 r^{4\alpha+4}\log^2 n}{n^2} \lesssim \frac{p^2 r^{4\alpha+4}\log^2 n}{n^2}.$$

Piecing together what has been shown finishes the proof of (6).

To prove (7), we divide the error $\mathbb{E}[R_n(\hat{F})]$ into two parts

$$\mathbb{E}\left[R_n(\hat{F})\right] \leq 2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|F(x_i) - F_{[r]}(x_i)\|^2\right] + 2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|F_{[r]}(x_i) - \hat{F}(x_i)\|^2\right]$$

where $F_{[r]} = \sum_{k=1}^r u_k f_k$ is the first r truncation of F. The first part is the approximation error and the second part is the estimation error.

For the approximation error, By the assumption (5) it is easily seen that

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|F(x_i) - F_{[r]}(x_i)\|^2\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left\|\sum_{k=r+1}^q u_k f_k(x_i)\right\|^2\right] \\ &= \frac{1}{n}\sum_{i=1}^n \sum_{k=r+1}^q \mathbb{E} f_k^2(x_i) = \sum_{k=r+1}^q \sigma_k^2 \lesssim r^{-\alpha+1},\end{aligned} \tag{9}$$

where $\mathbb{E}f_k^2(x_i) = \|f_k\|_{L_2}^2 = \sigma_k^2$ since $x_i \sim \mathcal{U}[0,1]^d$.

Now we consider the estimation error. Given the definitions in the paper and at the outset of the appendix, the estimation error can be rewritten as $\frac{1}{n}\mathbb{E}\|\tilde{F}_{[r]} - \tilde{\hat{F}}\|_F^2$ where $\tilde{F}_{[r]} = U_{[r]}\tilde{f}_{[r]}$ and $\tilde{\hat{F}} = \hat{U}_{[r]}\tilde{\hat{f}}_{[r]}$. Recall that $Y_{\cdot k}^o = Y^\top u_k$ and $\hat{Y}_{\cdot k}^* = Y^\top \hat{u}_k$ and for a linear smoother $\mathcal{L}$, the estimates are $\hat{f}_k^o = \mathcal{L}Y_{\cdot k}^o$ and $\hat{f}_k = \mathcal{L}\hat{Y}_{\cdot k}^*$, respectively. Consequently $\tilde{\hat{f}}_{[r]} = \hat{U}_{[r]}^\top Y\mathcal{L}^\top$ and $\tilde{\hat{F}} = \hat{U}_{[r]}\hat{U}_{[r]}^\top Y\mathcal{L}^\top$. With this notation, we have

$$\mathbb{E}\|\tilde{F}_{[r]} - \tilde{\hat{F}}\|_F^2 \leq 2\mathbb{E}\|U_{[r]}\tilde{f}_{[r]} - U_{[r]}\tilde{\hat{f}}_{[r]}^o\|_F^2 + 2\mathbb{E}\|U_{[r]}\tilde{\hat{f}}_{[r]}^o - \hat{U}_{[r]}\tilde{\hat{f}}_{[r]}\|_F^2.$$

By the assumption that $\mathbb{E}\|\hat{f}_k^o - f_k\|_n^2 \leq Ck^\tau/n^\rho$, we have

$$\begin{aligned}\frac{1}{n}\mathbb{E}\|U_{[r]}\tilde{f}_{[r]} - U_{[r]}\tilde{\hat{f}}_{[r]}^o\|_F^2 = \frac{1}{n}\mathbb{E}\|\tilde{f}_{[r]} - \tilde{\hat{f}}_{[r]}^o\|_F^2 &= \mathbb{E}\sum_{k=1}^r \|f_k - \hat{f}_k^o\|_n^2 \\ &\leq C\sum_{k=1}^r \frac{k^\tau}{n^\rho} \lesssim \frac{r^{\tau+1}}{n^\rho}.\end{aligned} \tag{10}$$

Also notice that

$$\begin{aligned}\frac{1}{n}\mathbb{E}\|U_{[r]}\tilde{\hat{f}}_{[r]}^o - \hat{U}_{[r]}\tilde{\hat{f}}_{[r]}\|_F^2 &= \frac{1}{n}\mathbb{E}\|(U_{[r]}U_{[r]}^\top - \hat{U}_{[r]}\hat{U}_{[r]}^\top)Y\mathcal{L}^\top\|_F^2 \\ &\leq \frac{1}{n}\mathbb{E}\left[\|U_{[r]}U_{[r]}^\top - \hat{U}_{[r]}\hat{U}_{[r]}^\top\|_F^2\|Y\|^2\|\mathcal{L}\|^2\right] \\ &\leq \frac{C}{n}\sqrt{\mathbb{E}\|U_{[r]}U_{[r]}^\top - \hat{U}_{[r]}\hat{U}_{[r]}^\top\|_F^4}\sqrt{\mathbb{E}\|Y\|^4},\end{aligned}$$

where in the last inequality we use $\|\mathcal{L}\| \leq C$. Using (6), Lemma 6 and together with the fact that $\|U_{[r]}U_{[r]}^\top - \hat{U}_{[r]}\hat{U}_{[r]}^\top\|_F \leq 2\sqrt{r}\|\sin\Theta(\hat{U}_{[r]}, U_{[r]})\|$ give that

$$\frac{1}{n}\mathbb{E}\|U_{[r]}\tilde{\hat{f}}^o_{[r]} - \hat{U}_{[r]}\tilde{\hat{f}}_{[r]}\|_F^2 \leq \frac{Cpr^{2\alpha+3}\log n}{n}\max\left(1, \frac{q}{\sqrt{n}}, \frac{p}{n}\right) \lesssim \frac{pr^{2\alpha+3}\log n}{n}, \tag{11}$$

since $q \ll \sqrt{p}$ and $p \ll n$. Combing (9), (10) and (11), we obtain (7) and conclude the proof. ■

Proof. [**Theorem 4**] It is well known that in the fixed design case, where $x_i = i/n$, the local polynomial estimator enjoys the minimax optimal rate $\mathbb{E}_f\|\hat{f} - f\|_n^2 \lesssim n^{-2\beta/(1+2\beta)}$; see Proposition 1.13 and Theorem 1.6 in Tsybakov (2009). However, in the random design case, there seem no results on convergence in the metric $\mathbb{E}_f\|\cdot\|_n^2$. One remedy is to adopt a similar approach in Cai and Brown (1999), and slightly modify the local polynomial regression strategy.

Represent the estimated rotated data $\{(x_i, \hat{y}^*_{ik})\}$ as $\{(x_{(i)}, \hat{y}^*_{(i)k})\}$, where $x_{(i)}$ is the ith order statistic of x_i's and $\hat{y}^*_{(i)k}$ is the corresponding response. In the recovery procedure, we perform local polynomial regression on the equispaced data $\{(\delta_i, \hat{y}^*_{(i)k})\}$ for $i = 1, \ldots, n$, where $\delta_i = \mathbb{E}x_{(i)} = i/(n+1)$. Then for a given x, $\hat{f}_k(x)$ can be represented as $\hat{f}_k(x) = \sum_{i=1}^n W_{n,i}(x)\hat{y}^*_{(i)k} = \mathcal{L}\hat{Y}^*$ where the $W_{n,i}(x)$ are defined in (1.67) in Tsybakov (2009) and completely deterministic, satisfying all of the properties derived therein. Similarly, for the oracle equispaced data $\{(\delta_i, \hat{y}^o_{(i)k})\}$, we denote the oracle estimate function by $\hat{f}^o_k = \mathcal{L}\hat{Y}^o$. Next we turn to the main proof of the theorem.

Let $b(x) = \mathbb{E}_{f_k,\mathcal{D}}\hat{f}^o_k(x) - f_k(x)$ denote the bias, conditioned on design, of the estimator $\hat{f}^o_k(x)$ at x. Then we find that

$$\begin{aligned} b(x) &= \sum_{i=1}^n f_k(x_{(i)})W_{n,i}(x) - f_k(x) = \sum_{i=1}^n \left\{f_k(x_{(i)}) - f_k(x)\right\}W_{n,i}(x) \\ &= \sum_{i=1}^n \left\{f_k(x_{(i)}) - f_k(\delta_i)\right\}W_{n,i}(x) + \sum_{i=1}^n \left\{f_k(\delta_i) - f_k(x)\right\}W_{n,i}(x) \end{aligned}$$

and hence

$$|b(x)| \leq \underbrace{\left|\sum_{i=1}^n \left\{f_k(x_{(i)}) - f_k(\delta_i)\right\}W_{n,i}(x)\right|}_{\mathrm{I}(x)} + \underbrace{\left|\sum_{i=1}^n \left\{f_k(\delta_i) - f_k(x)\right\}W_{n,i}(x)\right|}_{\mathrm{II}(x)}.$$

Starting from the fact that for x_i we have

$$\begin{aligned} \mathbb{E}_{f_k}\{\hat{f}^o_k(x_i) - f_k(x_i)\}^2 &= \mathbb{E}_{f_k}\mathbb{E}_{f_k,\mathcal{D}}\{\hat{f}^o_k(x_i) - f_k(x_i)\}^2 \\ &= \mathbb{E}_{f_k}\mathbb{E}_{f_k,\mathcal{D}}\{\hat{f}^o_k(x_i) - \mathbb{E}_{f_k,\mathcal{D}}\hat{f}^o_k(x_i) + \mathbb{E}_{f_k,\mathcal{D}}\hat{f}^o_k(x_i) - f_k(x_i)\}^2 \\ &= \mathbb{E}_{f_k}\left[\mathrm{var}_{f_k,\mathcal{D}}\{\hat{f}^o_k(x_i)\} + b^2(x_i)\right]. \end{aligned}$$

Conditioned on design $\mathcal{D}$, as the same as the proof in Tsybakov (2009) we can show that

$$\mathrm{II}^2(x_i) \leq \left(\frac{C\sigma_k L_{lk}}{l!}\right)^2 h^{2\beta} \quad \text{and} \quad \mathrm{var}_{f_k,\mathcal{D}}\{\hat{f}^o_k(x_i)\} \leq \frac{C\sigma^2}{nh},$$

which together give that

$$\mathbb{E}_{f_k}\{\hat{f}_k^o(x_i) - f_k(x_i)\}^2 \leq \left(\frac{C\sigma_k L_{lk}}{l!}\right)^2 h^{2\beta} + \frac{C\sigma^2}{nh} + 2\mathbb{E}_{f_k}\mathrm{I}^2(x_i).$$

Since $v_k \in \mathcal{H}(\beta, L_{lk})$ and $l > 1$, by Theorem 1.34 in Adams and Fournier (2003), we have $L_{1k} < \infty$. For each x_i, it holds that

$$\mathrm{I}(x_i) \leq \sigma_k L_{1k} \sum_{i=1}^{n} |\delta_i - x_{(i)}||W_{n,i}(x_i)|.$$

Applying Cauchy-Schwarz to the right hand side and using the properties of $W_{n,i}(x)$ from Tsybakov (2009) gives

$$\mathrm{I}^2(x_i) \leq \sigma_k^2 L_{1k}^2 \sum_{i=1}^{n} |\delta_i - x_{(i)}|^2 \sum_{i=1}^{n} |W_{n,i}(x_i)|^2 \leq \frac{C\sigma_k^2 L_{1k}^2}{nh} \sum_{i=1}^{n} |\delta_i - x_{(i)}|^2.$$

Thus

$$\mathbb{E}_{f_k}\mathrm{I}^2(x_i) \leq \frac{C\sigma_k^2 L_{1k}^2}{nh} \sum_{i=1}^{n} \mathrm{var}\{x_{(i)}\} \leq \frac{C\sigma_k^2 L_{1k}^2}{nh} \sum_{i=1}^{n} \frac{i}{n^2} \leq \frac{C\sigma_k^2 L_{1k}^2}{nh}.$$

Together with these we show that

$$\begin{aligned}
\mathbb{E}_{f_k}\|\hat{f}_k^o - f_k\|_n^2 &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{f_k}\{\hat{f}_k^o(x_i) - f_k(x_i)\}^2 \\
&\leq \frac{C\sigma_k^2 L_{lk}^2}{(l!)^2} h^{2\beta} + \frac{C\sigma^2}{nh} + \frac{C\sigma_k^2 L_{1k}^2}{nh} \\
&\lesssim \max\{1, (\sigma_k^2 L_{1k}^2)^{\frac{2\beta}{2\beta+1}}\}(\sigma_k^2 L_{lk}^2)^{\frac{1}{2\beta+1}} n^{-\frac{2\beta}{2\beta+1}}
\end{aligned}$$

by choosing optimal bandwidth $h \asymp \left(\max\{1, \sigma_k^2 L_{1k}^2\}/(n\sigma_k^2 L_{lk}^2)\right)^{1/(2\beta+1)}$.

Now we verify the boundness of $\mathcal{L}$. This follows from a result for bounds of eigenvalues of matrices. Let $A = (a_{kl})_{k,l=1}^n$ be an $n \times n$ matrix and set

$$R_k = \sum_l |a_{kl}| \quad \text{and} \quad C_l = \sum_k |a_{kl}|.$$

Then one can show that the eigenvalues of A, $\mu(A)$, are bounded by

$$\mu(A) \leq \min\left(\max_k R_k, \max_l C_l\right) \leq \max_k R_k.$$

Note that the $\mathcal{L}$ satisfies $L_{ij} = W_{n,j}(x_i)$ and from Tsybakov (2009) we know that

$$R_i = \sum_j |L_{ij}| = \sum_j |W_{n,j}(x_i)| \leq C.$$

Thus we have that $\mu(\mathcal{L}) \leq C$ and hence $\|\mathcal{L}\| \leq C$. ■

Appendix C. Auxiliary Lemmas for Main Theorems

In the Appendix 5.2, we introduce the auxiliary lemmas for main theorems. Lemma 6 bounds fourth moments of $\|Y\|$ in which the proof needs Lemma 5. Lemma 7 quantifies the discrepancy between $\langle\cdot,\cdot\rangle_n$ and $\langle\cdot,\cdot\rangle_{L^2}$, which is crucial to the proof of Lemma 8. Lemma 8 shows that the scaled singular values of $\tilde{F}$ are close to the true counterparts of F. Lemma 9 is crucial to the proofs of Lemma 10, which in turn is crucial to the proofs of the main theorems of the paper. Lemma 10 extends the results of Theorem 3 in Cai and Zhang (2018) by considering fourth moment perturbation bounds for the top r singular vectors.

Lemma 5 *If $X \geq 0$ is a positive random variable and for $a, b > 0$ we have $\mathbb{P}(X > a + bt) \leq 2\exp(-t^2)$ for all $t \geq 0$, then it follows that $\mathbb{E}(X^4) \leq C\max(a^4, b^4)$.*

Proof. Separating on the value of a we find that

$$\mathbb{E}(X^4) = \mathbb{E}\{X^4 1_{(X\leq a)}\} + \mathbb{E}\{X^4 1_{(X>a)}\} \leq a^4 + \mathbb{E}\{X^4 1_{(X>a)}\}.$$

Now notice that

$$\begin{aligned}
\mathbb{E}\{X^4 1_{(X>a)}\} &= \int_\Omega X^4(\omega) 1_{\{X(\omega)>a\}} dP(\omega) \\
&= \int_\Omega \left(a^4 + 4\int_a^{X(\omega)} s^3 ds\right) dP(\omega) \\
&= \int_\Omega \left(a^4 + 4\int_a^{\infty} s^3 1_{\{s<X(\omega)\}} ds\right) dP(\omega) \\
&= a^4 + 4\int_a^\infty s^3 \mathbb{P}(X > s) ds \\
&= a^4 + 4b\int_0^\infty (a+bt)^3 \mathbb{P}(X > a + bt) dt \\
&\leq a^4 + 8b\max(a^3, b^3)\int_0^\infty (1+t)^3 \exp(-t^2) dt \\
&\leq C\max(a^4, b^4),
\end{aligned}$$

which concludes the proof of the lemma. ■

Lemma 6 *With $Y = \tilde{F} + Z \in \mathbb{R}^{p\times n}$ denoting the data matrix and $\|Y\| = \max_i \sigma_i(Y)$ the operator norm, or maximum singular value of Y, we have that*

$$\mathbb{E}\|Y\|^4 \leq C\max(nq^2, p^2, n^2)$$

holds when $\max_k \|f_k\|_\infty \leq B$.

Proof. Since $\tilde{F} = U\tilde{f}$ as defined above, we have

$$\|\tilde{F}\|^2 \leq \|\tilde{F}\|_F^2 = \mathrm{tr}(\tilde{f}^\top U^\top U \tilde{f}) = \mathrm{tr}(\tilde{f}^\top \tilde{f}) = \sum_{i=1}^{n}\sum_{k=1}^{q} f_k^2(x_i).$$

Note that $\mathbb{E}f_k^2(x_i) = \|f_k\|^2 = \sigma_k^2$ and $\sum_{k=1}^{q} f_k^2(x_i) \leq qB^2$ by the assumption, one has

$$\begin{aligned}\mathbb{E}\|\tilde{F}\|^4 \leq \mathbb{E}\left[\sum_{i=1}^{n}\sum_{k=1}^{q} f_k^2(x_i)\right]^2 &= n(n-1)\left(\sum_{k=1}^{q}\mathbb{E}\left[f_k^2(x_i)\right]\right)^2 + n\mathbb{E}\left[\sum_{k=1}^{q} f_k^2(x_i)\right]^2 \\ &\leq Cn(n-1) + B^4 nq^2.\end{aligned}$$

Now, if $Z \in \mathbb{R}^{p\times n}$ is composed of independent and identically distributed $\mathcal{N}(0,\sigma^2)$ entries, then according to Theorem 4.4.5 in Vershynin (2019) there is a constant C so that for all $t > 0$,

$$\mathbb{P}\left\{\|Z\| > C\sigma(p^{1/2} + n^{1/2} + t)\right\} \leq 2e^{-t^2}.$$

By Lemma 5, this implies that $\mathbb{E}\|Z\|^4 \leq C\max(p^2, n^2)$ which completes the proof by

$$\mathbb{E}\|Y\|^4 \leq C(\mathbb{E}\|\tilde{F}\|^4 + \mathbb{E}\|Z\|^4) \leq C\max(nq^2, p^2, n^2).$$

■

Lemma 7 *Suppose that f_k's are bounded and orthogonal in $L^2[0,1]^d$, satisfying $\max_k \|f_k\|_\infty \leq B$. Then*

$$\mathbb{P}\left\{\max_{k,l}|\langle f_k, f_l\rangle_n - \langle f_k, f_l\rangle_{L_2}| > 2\delta B^2\right\} \leq q(q+1)\exp\left(-n\delta^2/2\right)$$

and so with the probability at least $1 - q(q+1)/n^2$,

$$\max_{k,l}|\langle f_k, f_l\rangle_n - \langle f_k, f_l\rangle_{L_2}| \leq 4B^2\sqrt{\frac{\log n}{n}}.$$

Proof. Noting that for any $1 \leq k, l \leq q$ we have

$$\langle f_k, f_l\rangle_n - \langle f_k, f_l\rangle_{L_2} = \frac{1}{n}\sum_{i=1}^{n}\{f_k(x_i)f_l(x_i) - \langle f_k, f_l\rangle_{L_2}\}$$

guarantees that $\langle f_k, f_l\rangle_n - \langle f_k, f_l\rangle_{L_2}$ is expressible as the sum of n independent and identically distributed mean 0 random variables, each bounded by $2B^2/n$. Hoeffding then gives that

$$\mathbb{P}\left\{|\langle f_k, f_l\rangle_n - \langle f_k, f_l\rangle_{L_2}| > 2\delta B^2\right\} \leq 2\exp\left(-\frac{n\delta^2}{2}\right).$$

Symmetry of inner product guarantees that there are $q(q+1)/2$ distinct sums $|\langle f_k, f_l\rangle_n - \langle f_k, f_l\rangle_{L_2}|$ as we vary k, l over $1, \ldots, q$ and so the first inequality of the theorem follows from a union bound. ■

Lemma 8 *Let $\tilde{F}$ be the sampled version of F admitting a singular value type decomposition*

$$F = \sum_{k=1}^{q} \sigma_k u_k v_k = \sum_{k=1}^{q} u_k f_k.$$

Under the conditions of Lemma 7, the k-th singular value of $\tilde{F}$ satisfies

$$\max_k \left| \frac{1}{\sqrt{n}} \sigma_k(\tilde{F}) - \sigma_k \right| \leq 2B \left(\frac{q^2 \log n}{n} \right)^{1/4} \tag{12}$$

with probability at least $1 - q(q+1)/n^2$.

Proof. Recall that the matrix $\tilde{f} \in \mathbb{R}^{q \times n}$ collects the sampled values of the f_k in its rows. Note that the matrix $\tilde{f}\tilde{f}^\top = (a_{kl})_{q \times q}$ with $a_{kl} = \sum_{i=1}^{n} f_k(x_i) f_l(x_i) = n \langle f_k, f_l \rangle_n$. Furthermore, we may write

$$\tilde{f}\tilde{f}^\top = \sum_{k=1}^{q} n \langle f_k, f_k \rangle_{L^2} e_k e_k^\top + \Delta = n \sum_{k=1}^{q} \sigma_k^2 e_k e_k^\top + \Delta,$$

where the matrix Δ is composed of elements $\Delta_{kl} = n(\langle f_k, f_l \rangle_n - \langle f_k, f_l \rangle_{L^2})$. Thus we have

$$\tilde{F}\tilde{F}^\top = U \tilde{f}\tilde{f}^\top U^\top = n \sum_{k=1}^{q} \sigma_k^2 u_k u_k^\top + U \Delta U^\top$$

and $\tilde{F}\tilde{F}^\top$, Δ, and thus $U\Delta U^\top$, are both real and symmetric. Therefore a well known perturbation result for matrices (Weyl, 1912) implies that

$$\max_k |\sigma_k^2(\tilde{F}) - n\sigma_k^2| \leq \|U \Delta U^\top\|.$$

Since $\|U\Delta U^\top\|^2 = \|\Delta\|^2 \leq \|\Delta\|_F^2 = \sum_{k,l}^{q} \Delta_{k,l}^2$, by Lemma 7 it implies that, with probability at least $1 - q(q+1)/n^2$,

$$\|U \Delta U^\top\|^2 \leq 16 B^4 q^2 n \log n,$$

which completes the proof by collecting the above results. ■

Lemma 9 *Suppose that the design $\mathcal{D} = \{x_1, \ldots, x_n\}$ is fixed. Let $Y = \tilde{F} + Z \in \mathbb{R}^{p \times n}$ denote the data matrix and $\tilde{F}$ admits a singular value decomposition $\tilde{F} = P \Lambda Q^\top$. Then it holds that*

$$\mathbb{P}\left\{ \sigma_r^2(Y^\top P_{[r]}) \leq (\sigma_r^2(\tilde{F}) + n)(1 - t) \right\} \leq C \exp\left(Cr - c(\sigma_r^2(\tilde{F}) + n) t \wedge t^2 \right),$$
$$\mathbb{P}\left\{ \sigma_{r+1}^2(Y^\top) \geq (\sigma_{r+1}^2(\tilde{F}) + n)(1 + t) \right\} \leq C \exp\left(Cp - c(\sigma_p^2(\tilde{F}) + n) t \wedge t^2 \right).$$

Moreover, there exists C_0 only depending on C and c, such that whenever $\sigma_r^2(\tilde{F}) \geq C_0 p$, for any $t > 0$ we have

$$\mathbb{P}\left\{ \|\mathcal{P}_{(Y^\top P_{[r]})} Y^\top P_{[r]}^\perp\| \geq t \right\} \leq C \exp\left(Cp - ct^2 \wedge \sqrt{\sigma_r^2(\tilde{F}) + nt} \right) + C \exp\left(-c(\sigma_r^2(\tilde{F}) + n) \right).$$

Proof. The proof mainly uses random matrix theory and follows the lines of proof of Lemma 4 in Cai and Zhang (2018).
(A1) Low bounds for $\sigma_r^2(Y^\top P_{[r]})$
Note that $\mathbb{E}(YY^\top) = \tilde{F}\tilde{F}^\top + nI_p = P\Lambda^2 P^\top + nI_p$, we have

$$\mathbb{E}(P_{[r]}^\top YY^\top P_{[r]}) = P_{[r]}^\top P\Lambda^2 P^\top P_{[r]} + nP_{[r]}^\top I_p P_{[r]} = \Lambda_{[r]}^2 + nI_r,$$

where $\Lambda_{[r]}$ is a diagonal matrix consisting the first r diagonal elements of Λ. Let $M_{[r]} = (\Lambda_{[r]}^2 + nI_r)^{-1/2}$ and then $\mathbb{E}(M_{[r]}^\top P_{[r]}^\top YY^\top P_{[r]} M_{[r]}) = I_r$. Since $\sigma_r^2(Y^\top P_{[r]} M_{[r]}) \le \sigma_r^2(Y^\top P_{[r]})\sigma_{\max}^2(M_{[r]})$, $\sigma_r^2(Y^\top P_{[r]} M_{[r]}) = \sigma_r(M_{[r]}^\top P_{[r]}^\top YY^\top P_{[r]} M_{[r]})$ and

$$1 - \sigma_r^2(Y^\top P_{[r]} M_{[r]}) = \sigma_r(I_r - M_{[r]}^\top P_{[r]}^\top YY^\top P_{[r]} M_{[r]}) \le \|I_r - M_{[r]}^\top P_{[r]}^\top YY^\top P_{[r]} M_{[r]}\|,$$

we have

$$\sigma_r^2(Y^\top P_{[r]}) \ge (\sigma_r^2(\tilde{F}) + n)\big[1 - \|I_r - M_{[r]}^\top P_{[r]}^\top YY^\top P_{[r]} M_{[r]}\|\big]. \tag{13}$$

In the following we need an upper bound for $\|I_r - M_{[r]}^\top P_{[r]}^\top YY^\top P_{[r]} M_{[r]}\|$. For any unit vector $u \in \mathbb{R}^r$,

$$\begin{aligned} u^\top\big(M_{[r]}^\top P_{[r]}^\top YY^\top P_{[r]} M_{[r]} - I_r\big)u &= u^\top M_{[r]}^\top P_{[r]}^\top \tilde{F}\tilde{F}^\top P_{[r]} M_{[r]} u - \mathbb{E}\big(u^\top M_{[r]}^\top P_{[r]}^\top \tilde{F}\tilde{F}^\top P_{[r]} M_{[r]} u\big) \\ &\quad + 2\big[u^\top M_{[r]}^\top P_{[r]}^\top \tilde{F} Z^\top P_{[r]} M_{[r]} u - \mathbb{E}\big(u^\top M_{[r]}^\top P_{[r]}^\top \tilde{F} Z^\top P_{[r]} M_{[r]} u\big)\big] \\ &\quad + u^\top M_{[r]}^\top P_{[r]}^\top ZZ^\top P_{[r]} M_{[r]} u - \mathbb{E}\big(u^\top M_{[r]}^\top P_{[r]}^\top ZZ^\top P_{[r]} M_{[r]} u\big) \\ &= 0 + 2u^\top M_{[r]}^\top P_{[r]}^\top \tilde{F} Z^\top P_{[r]} M_{[r]} u \\ &\quad + u^\top M_{[r]}^\top P_{[r]}^\top \big[ZZ^\top - \mathbb{E}(ZZ^\top)\big] P_{[r]} M_{[r]} u. \end{aligned}$$

Due to $u^\top M_{[r]}^\top P_{[r]}^\top \tilde{F} Z^\top P_{[r]} M_{[r]} u = \mathrm{tr}\big[Z^\top P_{[r]} M_{[r]} u\big(\tilde{F}^\top P_{[r]} M_{[r]} u\big)^\top\big]$, using general Hoeffding inequality (see Theorem 2.6.3 in Vershynin, 2019), it holds that

$$\begin{aligned} \mathbb{P}\Big\{\big|\mathrm{tr}\big[Z^\top P_{[r]} M_{[r]} u\big(\tilde{F}^\top P_{[r]} M_{[r]} u\big)^\top\big]\big| \ge t\Big\} &\le 2\exp\bigg(-\frac{ct^2}{\|P_{[r]} M_{[r]} u\big(\tilde{F}^\top P_{[r]} M_{[r]} u\big)^\top\|_F^2}\bigg) \\ &\le 2\exp\left(-\frac{ct^2}{\|M_{[r]}\|^2\|\Lambda_{[r]} M_{[r]}\|^2}\right) \\ &\le 2\exp\big(-ct^2(\sigma_r^2(\tilde{F}) + n)\big). \end{aligned}$$

On the other hand, using Hanson-Wright inequality (see Theorem 6.2.1 in Vershynin, 2019), it holds that

$$\begin{aligned} \mathbb{P}\Big\{\big|u^\top M_{[r]}^\top P_{[r]}^\top \big[ZZ^\top - \mathbb{E}(ZZ^\top)\big] P_{[r]} M_{[r]} u\big| \ge t\Big\} &\le 2\exp\left(-c\frac{t^2}{\|P_{[r]} M_{[r]} u\|^4 n} \wedge \frac{t}{\|P_{[r]} M_{[r]} u\|^2}\right) \\ &\le 2\exp\left(-c\frac{t^2(\sigma_r^2(\tilde{F}) + n)^2}{n} \wedge t(\sigma_r^2(\tilde{F}) + n)\right). \end{aligned}$$

Together with these we obtain that

$$\mathbb{P}\left\{\left|u^\top\left(M_{[r]}^\top P_{[r]}^\top YY^\top P_{[r]}M_{[r]}-I_r\right)u\right|\geq t\right\}\leq C\exp\left(-c(\sigma_r^2(\tilde{F})+n)t\wedge t^2\right).$$

The ϵ-net argument in Lemma 5 in Cai and Zhang (2018) leads to

$$\mathbb{P}\left\{\|M_{[r]}^\top P_{[r]}^\top YY^\top P_{[r]}M_{[r]}-I_r\|\geq t\right\}\leq C\exp\left(Cr-c(\sigma_r^2(\tilde{F})+n)t\wedge t^2\right). \tag{14}$$

which together with (13) deduces that

$$\mathbb{P}\left\{\sigma_r^2(Y^\top P_{[r]})\geq(\sigma_r^2(\tilde{F})+n)(1-t)\right\}\geq 1-C\exp\left(Cr-c(\sigma_r^2(\tilde{F})+n)t\wedge t^2\right).$$

(A2) Upper bounds for $\sigma_{r+1}^2(Y^\top)$
Using the best rank-r approximation of $Y^\top$ (see Eckart-Young-Mirsky Theorem on page 73 in Vershynin, 2019),

$$\sigma_{r+1}(Y^\top)=\min_{\mathrm{rank}(A)\leq r}\|Y^\top-A\|\leq\|Y^\top-Y^\top P_{[r]}P_{[r]}^\top\|=\|Y^\top P_{[r]}^\perp(P_{[r]}^\perp)^\top\|=\|Y^\top P_{[r]}^\perp\|,$$

which switch our focus from $\sigma_{r+1}(Y^\top)$ to $\sigma_{\max}(Y^\top P_{[r]}^\perp)$.

Since $\mathbb{E}\big[(P_{[r]}^\perp)^\top YY^\top P_{[r]}^\perp\big]=(\Lambda_{[r]}^\perp)^2+nI_{p-r}$ where $\Lambda_{[r]}^\perp$ denotes the diagonal matrix with eliminating the first r diagonal elements of Λ, let $M_{[r]}^\perp=\big((\Lambda_{[r]}^\perp)^2+nI_{p-r}\big)^{-1/2}$ and then

$$\begin{aligned}\sigma_{\max}^2(Y^\top P_{[r]}^\perp)&\leq\left\|(P_{[r]}^\perp M_{[r]}^\perp)^\top YY^\top P_{[r]}^\perp M_{[r]}^\perp\right\|\left(\sigma_{r+1}^2(\tilde{F})+n\right)\\&\leq\left(\left\|(P_{[r]}^\perp M_{[r]}^\perp)^\top YY^\top P_{[r]}^\perp M_{[r]}^\perp-I_{p-r}\right\|+1\right)\left(\sigma_{r+1}^2(\tilde{F})+n\right).\end{aligned}$$

Following the same arguments for the proof of (14), we have

$$\mathbb{P}\left\{\left\|(P_{[r]}^\perp M_{[r]}^\perp)^\top YY^\top P_{[r]}^\perp M_{[r]}^\perp-I_{p-r}\right\|\geq t\right\}\leq C\exp\left(C(p-r)-c(\sigma_p^2(\tilde{F})+n)t\wedge t^2\right),$$

which deduces that

$$\mathbb{P}\left\{\sigma_{r+1}^2(Y^\top)\leq(\sigma_{r+1}^2(\tilde{F})+n)(1+t)\right\}\geq 1-C\exp\left(Cp-c(\sigma_p^2(\tilde{F})+n)t\wedge t^2\right).$$

(A3) Upper bounds for $\|\mathcal{P}_{(Y^\top P_{[r]})}Y^\top P_{[r]}^\perp\|$
Since

$$\begin{aligned}\|\mathcal{P}_{(Y^\top P_{[r]})}Y^\top P_{[r]}^\perp\|&=\|\mathcal{P}_{(Y^\top P_{[r]}M_{[r]})}Y^\top P_{[r]}^\perp\|\\&=\|(Y^\top P_{[r]}M_{[r]})\left((Y^\top P_{[r]}M_{[r]})^\top(Y^\top P_{[r]}M_{[r]})\right)^{-1}(Y^\top P_{[r]}M_{[r]})^\top Y^\top P_{[r]}^\perp\|\\&\leq\sigma_{\min}^{-1}(Y^\top P_{[r]}M_{[r]})\|M_{[r]}^\top P_{[r]}^\top YY^\top P_{[r]}^\perp\|,\end{aligned}$$

we shall analyze $\sigma_{\min}(Y^\top P_{[r]}M_{[r]})$ and $\|M_{[r]}^\top P_{[r]}^\top YY^\top P_{[r]}^\perp\|$ separately below. Similar to (13),

$$\sigma_{\min}^2(Y^\top P_{[r]}M_{[r]})\geq 1-\|I_r-M_{[r]}^\top P_{[r]}^\top YY^\top P_{[r]}M_{[r]}\|$$

and by (14) it implies that

$$\mathbb{P}\left\{\sigma_{\min}^2(Y^\top P_{[r]}M_{[r]}) \geq 1-t\right\} \geq 1 - C\exp\left(Cr - c(\sigma_r^2(\tilde{F})+n)t \wedge t^2\right).$$

Setting $t = 1/2$ and choosing C_0 large enough such that $\sigma_r^2(\tilde{F}) \geq C_0 p \geq C_0 r$, we have $Cr - c(\sigma_r^2(\tilde{F})+n)t \wedge t^2 \leq -c(\sigma_r^2(\tilde{F})+n)/8$, leading to that

$$\mathbb{P}\left\{\sigma_{\min}^2(Y^\top P_{[r]}M_{[r]}) \geq 1/2\right\} \geq 1 - C\exp\left(-c(\sigma_r^2(\tilde{F})+n)\right). \tag{15}$$

For $\|M_{[r]}^\top P_{[r]}^\top YY^\top P_{[r]}^\perp\|$, since $P_{[r]}^\top \tilde{F}\tilde{F}^\top P_{[r]}^\perp = 0$, we have the decomposition

$$u^\top M_{[r]}^\top P_{[r]}^\top YY^\top P_{[r]}^\perp v = u^\top\left(M_{[r]}^\top P_{[r]}^\top \tilde{F}Z^\top P_{[r]}^\perp + M_{[r]}^\top P_{[r]}^\top Z\tilde{F}^\top P_{[r]}^\perp + M_{[r]}^\top P_{[r]}^\top ZZ^\top P_{[r]}^\perp\right)v.$$

Following the same proof of (14) again, we can show that

$$\mathbb{P}\left\{\|M_{[r]}^\top P_{[r]}^\top YY^\top P_{[r]}^\perp\| \geq t\right\} \leq C\exp\left(C(p-r) - ct^2 \wedge \sqrt{\sigma_r^2(\tilde{F})+nt}\right). \tag{16}$$

Combing (15) and (16), we obtain

$$\mathbb{P}\left\{\|\mathcal{P}_{(Y^\top P_{[r]})}Y^\top P_{[r]}^\perp\| \geq t\right\} \leq C\exp\left(Cp - ct^2 \wedge \sqrt{\sigma_r^2(\tilde{F})+nt}\right) + C\exp\left(-c(\sigma_r^2(\tilde{F})+n)\right).$$

Then the proof is complete. ■

Lemma 10 *Denote the design $\mathcal{D} = \{x_1, \ldots, x_n\}$. Assume that there exists a constant $C_{\max}$ large enough such that $\sigma_{\max}^2(\tilde{F}) \leq C_{\max} n$, then it holds that*

$$\mathbb{E}\left(\|\sin\Theta(\hat{U}_{[r]}, P_{[r]})\|^4 \big| \mathcal{D}\right) \lesssim \frac{p^2(\sigma_r^2(\tilde{F})+n)^2}{\left(\sigma_r^2(\tilde{F}) - \sigma_{r+1}^2(\tilde{F})\right)^4} \wedge 1.$$

Proof. Since the left singular vectors of Y are just the right singular vectors of $Y^\top$, we can apply the same arguments for the right singular vectors of $Y^\top$ to get bounds for estimation of the left singular vectors of Y. Thus, by Proposition 1 in Cai and Zhang (2018),

$$\|\sin\Theta(\hat{U}_{[r]}, P_{[r]})\| \leq \frac{\sigma_r(Y^\top P_{[r]})\|\mathcal{P}_{(Y^\top P_{[r]})}Y^\top P_{[r]}^\perp\|}{\sigma_r^2(Y^\top P_{[r]}) - \sigma_{r+1}^2(Y^\top)}.$$

Since $||\sin\Theta(\hat{U}_{[r]}, P_{[r]})|| \leq 1$, to complete the proof we only need focus on the case that $\left(\sigma_r^2(\tilde{F}) - \sigma_{r+1}^2(\tilde{F})\right)^2 \geq C_0 p\left(n + \sigma_r^2(\tilde{F})\right)$ for large C_0 only depending on $C_{\max}, C, c$. Note that in this case it holds that $\sigma_r^2(\tilde{F}) \geq C_0 p$, then by Lemma 9 we have

$$\mathbb{P}\left\{\sigma_r^2(Y^\top P_{[r]}) \leq \frac{2\sigma_r^2(\tilde{F})}{3} + \frac{\sigma_{r+1}^2(\tilde{F})}{3} + n\right\} \leq C\exp\left(Cr - c\frac{\left(\sigma_r^2(\tilde{F}) - \sigma_{r+1}^2(\tilde{F})\right)^2}{\sigma_r^2(\tilde{F})+n}\right),$$

$$\mathbb{P}\left\{\sigma_{r+1}^2(Y^\top) \geq \frac{2\sigma_{r+1}^2(\tilde{F})}{3} + \frac{\sigma_r^2(\tilde{F})}{3} + n\right\} \leq C\exp\left(Cp - c\frac{\left(\sigma_p^2(\tilde{F})+n\right)\left(\sigma_r^2(\tilde{F}) - \sigma_{r+1}^2(\tilde{F})\right)^2}{\left(\sigma_r^2(\tilde{F})+n\right)^2}\right),$$

$$\mathbb{P}\left\{\|\mathcal{P}_{(Y^\top P_{[r]})}Y^\top P_{[r]}^\perp\| \geq t\right\} \leq C\exp\left(Cp - ct^2 \wedge \sqrt{\sigma_r^2(\tilde{F})+nt}\right) + C\exp\left(-c(\sigma_r^2(\tilde{F})+n)\right).$$

Furthermore,

$$\mathbb{P}\left\{\|\mathcal{P}_{(Y^\top P_{[r]})}Y^\top P_{[r]}^\perp\| \geq \sqrt{\sigma_r^2(\tilde{F})+n}\right\} \leq C\exp\left(Cp - c(\sigma_r^2(\tilde{F})+n)\right).$$

Denote the event $\mathcal{Q}$ as

$$\mathcal{Q} = \Bigg\{\sigma_r^2(Y^\top P_{[r]}) \geq \frac{2\sigma_r^2(\tilde{F})}{3} + \frac{\sigma_{r+1}^2(\tilde{F})}{3} + n,\ \sigma_{r+1}^2(Y^\top) \leq \frac{2\sigma_{r+1}^2(\tilde{F})}{3} + \frac{\sigma_r^2(\tilde{F})}{3} + n, \\ \|\mathcal{P}_{(Y^\top P_{[r]})}Y^\top P_{[r]}^\perp\| \leq \sqrt{\sigma_r^2(\tilde{F})+n}\Bigg\}.$$

Now, under the event $\mathcal{Q}$, it holds that

$$\|\sin\Theta(\hat{U}_{[r]}, P_{[r]})\|^4 \leq \frac{\sigma_r^4(Y^\top P_{[r]})\|\mathcal{P}_{(Y^\top P_{[r]})}Y^\top P_{[r]}^\perp\|^4}{\left(\sigma_r^2(Y^\top P_{[r]}) - \sigma_{r+1}^2(Y^\top)\right)^4} \leq \frac{C(\sigma_r^2(\tilde{F})+n)^2\|\mathcal{P}_{(Y^\top P_{[r]})}Y^\top P_{[r]}^\perp\|^4}{\left(\sigma_r^2(\tilde{F}) - \sigma_{r+1}^2(\tilde{F})\right)^4},$$

where we use the fact that $u^2/(u^2-v^2)^2$ is a decreasing function of u and increasing function of v when $u > v > 0$. Thus for fixed $\mathcal{D}$, it gives that

$$\begin{aligned}\mathbb{E}\|\sin\Theta(\hat{U}_{[r]}, P_{[r]})\|^4 &= \mathbb{E}\big[\|\sin\Theta(\hat{U}_{[r]}, P_{[r]})\|^4 1_{\mathcal{Q}}\big] + \mathbb{E}\big[\|\sin\Theta(\hat{U}_{[r]}, P_{[r]})\|^4 1_{\mathcal{Q}^c}\big] \\ &\leq \frac{C(\sigma_r^2(\tilde{F})+n)^2}{\left(\sigma_r^2(\tilde{F}) - \sigma_{r+1}^2(\tilde{F})\right)^4}\mathbb{E}\big[\|\mathcal{P}_{(Y^\top P_{[r]})}Y^\top P_{[r]}^\perp\|^4 1_{\mathcal{Q}}\big] + \mathbb{P}\big(\mathcal{Q}^c\big).\end{aligned}$$

For large C_0, it holds that

$$Cr \leq Cp \leq \frac{cC_0 p}{2C_{\max}} \leq \frac{c}{2C_{\max}}\frac{\left(\sigma_r^2(\tilde{F}) - \sigma_{r+1}^2(\tilde{F})\right)^2}{\sigma_r^2(\tilde{F})+n}.$$

By the condition $\sigma_r^2(\tilde{F}) \leq C_{\max}n$, note that

$$\frac{\sigma_p^2(\tilde{F})+n}{\sigma_r^2(\tilde{F})+n} \geq \frac{n}{\sigma_r^2(\tilde{F})+n} \geq \frac{1}{C_{\max}+1} > 0.$$

Together with these we can show that

$$\begin{aligned}&Cr - c\frac{\left(\sigma_r^2(\tilde{F}) - \sigma_{r+1}^2(\tilde{F})\right)^2}{\sigma_r^2(\tilde{F})+n} \leq -c_0\frac{\left(\sigma_r^2(\tilde{F}) - \sigma_{r+1}^2(\tilde{F})\right)^2}{\sigma_r^2(\tilde{F})+n}, \\ &Cp - c\frac{(\sigma_p^2(\tilde{F})+n)\left(\sigma_r^2(\tilde{F}) - \sigma_{r+1}^2(\tilde{F})\right)^2}{(\sigma_r^2(\tilde{F})+n)^2} \leq -c_0\frac{\left(\sigma_r^2(\tilde{F}) - \sigma_{r+1}^2(\tilde{F})\right)^2}{\sigma_r^2(\tilde{F})+n}, \\ &Cp - c(\sigma_r^2(\tilde{F})+n) \leq Cp - c\frac{\left(\sigma_r^2(\tilde{F}) - \sigma_{r+1}^2(\tilde{F})\right)^2}{\sigma_r^2(\tilde{F})+n} \leq -c_0\frac{\left(\sigma_r^2(\tilde{F}) - \sigma_{r+1}^2(\tilde{F})\right)^2}{\sigma_r^2(\tilde{F})+n},\end{aligned}$$

where $c_0 > 0$ only depends on $C_0, C_{\max}, C, c$. Using the basic property of exponential function, one can see that

$$\mathbb{P}\left(\mathcal{Q}^c\right) \leq C\exp\bigg(- c_0 \frac{\big(\sigma_r^2(\tilde{F}) - \sigma_{r+1}^2(\tilde{F})\big)^2}{\sigma_r^2(\tilde{F}) + n}\bigg) \leq \frac{Cp^2(\sigma_r^2(\tilde{F}) + n)^2}{\big(\sigma_r^2(\tilde{F}) - \sigma_{r+1}^2(\tilde{F})\big)^4}.$$

Thus for the desired extension, it remains to show that $\mathbb{E}\big[\|\mathcal{P}_{(Y^\top P_{[r]})} Y^\top P_{[r]}^\perp\|^4 1_{\mathcal{Q}}\big] \leq Cp^2$. As in the proof, we let $T = \|\mathcal{P}_{(Y^\top P_{[r]})} Y^\top P_{[r]}^\perp\|$ and apply Lemma 9 again,

$$\begin{aligned}
\mathbb{E}T^4 1_{\mathcal{Q}} &\leq \mathbb{E}T^4 1_{\{T^2 \leq \sigma_r^2(\tilde{F}) + n\}} = \int_0^\infty \mathbb{P}\big(T^4 1_{\{T^2 \leq \sigma_r^2(\tilde{F}) + n\}} \geq t\big) dt \\
&\leq \delta^2 p^2 + \int_{\delta^2 p^2}^{(\sigma_r^2(\tilde{F}) + n)^2} \mathbb{P}\big(T \geq t^{1/4}\big) dt \\
&\leq \delta^2 p^2 + \int_{\delta^2 p^2}^{(\sigma_r^2(\tilde{F}) + n)^2} C\big(e^{Cp - c\sqrt{t}} + e^{-c(\sigma_r^2(\tilde{F}) + n)}\big) dt \\
&\leq \delta^2 p^2 + C(\sigma_r^2(\tilde{F}) + n)^2 e^{-c(\sigma_r^2(\tilde{F}) + n)} + \frac{2C(1 + c\delta p)}{c^2} e^{(C - c\delta)p} \\
&\leq \delta^2 p^2 + C + \frac{2C(1 + c\delta p)}{c^2} e^{(C - c\delta)p}.
\end{aligned}$$

It is seen that as long as we choose δ large enough, but only depending on C and c, it is guaranteed that $\mathbb{E}T^4 1_{\mathcal{Q}} \leq \delta^2 p^2$. ■

References

Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*. Academic Press, 2th edition, 2003.

Deanna M. Barch, Gregory C. Burgess, Michael P. Harms, Steven E. Petersen, and et al. Function in the human connectome: Task-fmri and individual differences in behavior. *NeuroImage*, 80:169–189, 2013.

Leo Breiman and Jerome H. Friedman. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59 (1):3–54, 1997.

Randy L. Buckner, Fenna M. Krienen, Angela Castellanos, Julio C. Diaz, and B. T. Thomas Yeo. The organization of the human cerebellum estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(5):2322–2345, 2011.

T. Tony Cai. Minimax and adaptive inference in nonparametric function estimation. *Statistical Science*, 27(1):31–50, 2012.

T. Tony Cai and Lawrence D. Brown. Wavelet estimation for samples with random uniform design. *Statistics & Probability Letters*, 42(3):313–321, 1999.

T. Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *Annals of Statistics*, 46(1):60–89, 2018.

Rahul S. Desikan, Florent Ségonne, Bruce Fischl, Brian T. Quinn, and et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *NeuroImage*, 31(3):968–980, 2006.

David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

Daniele Durante, Bruno Scarpa, and David B. Dunson. Locally adaptive factor processes for multivariate time series. *Journal of Machine Learning Research*, 15(1):1493–1522, 2014.

Robert Engle and Mark Watson. A one-factor multivariate time series model of metropolitan wage rates. *Journal of the American Statistical Association*, 76(376):774–781, 1981.

Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London, 1st edition, 1996.

Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680, 2013.

Chamberlain Gary and Michael J. Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51:1305–1324, 1983.

Tailen Hsing and Randall L. Eubank. *Theoretical Foundations of Functional Data Analysis, with An Introduction to Linear Operators*. Wiley, West Sussex, 2015.

Jianhua Z. Huang, Haipeng Shen, and Andreas 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.

Iain M. Johnstone. *Gaussian Estimation: Sequence and Wavelet Models*. Unpublished manuscript, 2017.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press Cambridge, MA, 2006.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, 1st edition, 2009.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 1st edition, 2019.

Martin J. Wainwright. *High-dimensional Statistics: A Non-asymptotic Viewpoint*. Cambridge University Press, 1st edition, 2019.

Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

B. T. Thomas Yeo, Fenna M. Krienen, Jorge Sepulcre, Mert R. Sabuncu, and et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, 2011.