Abstract

Functional Linear Regression in High Dimensions

Kaijie Xue
Doctor of Philosophy
Graduate Department of Statistical Science
University of Toronto
2017

Functional linear regression has occupied a central position in the area of functional data analysis, and attracted substantial research attention in the past decade. With increasingly complex data of this type collected in modern experiments, we conduct further investigations in response to the great need of statistical tools that are capable of handling functional objects in high-dimensional spaces.

In the first project, we deal with the situation that functional and non-functional data are encountered simultaneously when observations are sampled from random processes and a potentially large number of scalar covariates. It is difficult to apply existing methods for model selection and estimation. We propose a new class of partially functional linear models to characterize the regression between a scalar response and those covariates, including both functional and scalar types. The new approach provides a unified and flexible framework to simultaneously take into account multiple functional and ultra-high dimensional scalar predictors, identify important features and improve interpretability of the estimators. The underlying processes of the functional predictors are considered to be infinite-dimensional, and one of our contributions is to characterize the impact of regularization on the resulting estimators. We establish consistency and oracle properties under mild conditions, illustrate the performance of the proposed method with simulation studies, and apply it to air pollution data.

In the second project, we further explore the linear regression by focusing on the large-scale scenario that the scalar response is related to potentially an ultra-large number of functional pre-
dictors, leading to a more challenging model framework. The emphasis of our investigation is to establish valid testing procedures for general hypothesis on an arbitrary subset of regression coefficient functions. Specifically, we exploit the techniques developed for post-regularization inference, and propose a score test for the large-scale functional linear regression based on the so-called de-correlated score function that separates the primary and nuisance parameters in functional spaces. The proposed score test is shown uniformly convergent to the prescribed significance, and its finite sample performance is illustrated via simulation studies.
Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor, Professor Fang Yao, for his guidance and patience through my PHD program. Without his precious help and support, it is not possible for me to finish the thesis.

Secondly, I would also like to show my gratitude to my thesis committee members, Professor Keith Knight and Professor Zhou Zhou for their insightful and constructive comments.

Thirdly, I am also very grateful to all the faculty members, especially Professor Mike Evans, Professor Nancy Reid, Professor Keith Knight, Professor Zhou Zhou, Professor Andrey Feuerverger and Professor Lawrence J. Brunner for teaching me insightful statistical courses.

In addition, I would also like to thank all the staff members, especially Andrea Carter, Christine Bulguryemez and Dermot Whelan for their help and support during my research.

Finally, I would like to thank my parents for their help and support.
# Contents

1 Partially Functional Linear Regression in High Dimensions

1.1 Introduction ........................................... 1

1.2 Regularized Partially Functional Linear Regression .................. 3
    1.2.1 Classical functional linear model via principal components ... 3
    1.2.2 Partially functional linear regression with regularization ... 3
    1.2.3 Algorithms and parameter tuning .......................... 7

1.3 Asymptotic Properties .................................. 10

1.4 Simulation Studies ..................................... 13

1.5 Application ........................................... 20

1.6 Proof of Lemmas and Main Theorems .......................... 21
    1.6.1 Regularity Conditions .................................. 21
    1.6.2 Auxiliary Lemmas ...................................... 23
    1.6.3 Proof of Lemmas ...................................... 27
    1.6.4 Proofs of Main Theorems ............................... 37

2 Testing General Hypothesis in Large-scale FLR ..................... 47

2.1 Introduction ........................................... 47

2.2 Proposed Methodology ................................... 51
    2.2.1 Regularized estimation by group penalized least squares ... 51
    2.2.2 General hypothesis and score decorrelation .................. 53
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2.3 Bootstrapped score test for general hypothesis in large-scale FLR</td>
<td>55</td>
</tr>
<tr>
<td>2.3 Asymptotic Properties</td>
<td>57</td>
</tr>
<tr>
<td>2.4 Simulation Studies</td>
<td>61</td>
</tr>
<tr>
<td>2.5 Proof of Lemmas and Main Theorems</td>
<td>63</td>
</tr>
<tr>
<td>2.5.1 Algorithm and parameter tuning</td>
<td>63</td>
</tr>
<tr>
<td>2.5.2 Technical Conditions on Regularizers</td>
<td>64</td>
</tr>
<tr>
<td>2.5.3 Notations</td>
<td>64</td>
</tr>
<tr>
<td>2.5.4 Auxiliary Lemmas</td>
<td>66</td>
</tr>
<tr>
<td>2.5.5 Proof of Lemmas</td>
<td>69</td>
</tr>
<tr>
<td>2.5.6 Proofs of Main Theorems</td>
<td>94</td>
</tr>
</tbody>
</table>

Bibliography

99
1.1 Simulation results with sample size $n = 200$ based on 200 Monte Carlo replicates for Designs I and II. Shown are the Monte Carlo averages (standard errors in parentheses) for the number of false zero functional predictors ($FZ_f$), the number of the false nonzero functional predictors ($FN_f$), the functional mean squared error ($MSE_f$), the number of false zero scalar covariates ($FZ_s$), the number of false nonzero scalar covariates ($FN_s$), the scalar mean squared error ($MSE_s$), and the prediction error (PE). We first use $\text{ABIC}$ to choose the tuning parameter $\lambda_n$ and a common truncation $s_n$, then tune $s_{nj}$ jointly with $\text{AIC}$ by refitting the selected model using ordinary least squares. In Design II, Step 1 results are based on the original sample in each Monte Carlo run, while Step 2 contains the improved results by fitting the penalized procedure to the selected model in Step 1 with an additional sample of $n = 200$. 

vii
1.2 Simulation results with sample size $n = 200$ based on 200 Monte Carlo replicates for Designs III and IV. Shown are the Monte Carlo averages (standard errors in parentheses) for the number of false zero functional predictors ($FZ_f$), the number of the false nonzero functional predictors ($FN_f$), the functional mean squared error ($MSE_f$), the number of false zero scalar covariates ($FZ_s$), the number of false nonzero scalar covariates ($FN_s$), the scalar mean squared error ($MSE_s$), and the prediction error (PE). We first use ABIC to choose the tuning parameter $\lambda_n$ and a common truncation $s_n$, then tune $s_{nj}$ jointly with AIC by refitting the selected model using ordinary least squares. In Design IV, Step 1 results are based on the original sample in each Monte Carlo run, while Step 2 contains the improved results by fitting the penalized procedure to the selected model in Step 1 with an additional sample of $n = 200$.

1.3 Simulation results using the same settings as Design I & II in Section 4 of the paper and Design III & IV as above, based on 200 Monte Carlo replicates, except the sample size $n = 400$.

1.4 Simulation results using the same settings as Design I & II in Section 4 of the paper and Design III & IV as above, based on 200 Monte Carlo replicates, except the variance of the regression error $\sigma^2 = 2$.

2.1 Results for the several different combinations of null hypothesis and settings of $\beta_j(t)$ that were measured over 300 simulations.
List of Figures

1.1 The estimated regression parameter functions and 95% confidence bands based on 1000 bootstrap samples for PM2·5 and temperature with \( \hat{s}_{n1} = 2 \) and \( \hat{s}_{n2} = 2 \), respectively. The solid line denotes the estimated regression parameter functions and the dashed lines denote the 95% bootstrap confidence bands. The left and right panels are for the PM2·5 and temperature, respectively. . . . 22
Chapter 1

Partially Functional Linear Regression in High Dimensions

1.1 Introduction

Functional linear regression is widely used to model the prediction of a functional predictor through a linear operator, often realized by an integral form of a regression parameter function; see Ramsay and Dalzell [1991], Cardot et al. [2003], Cuevas et al. [2002], Yao et al. [2005a], and Ramsay and Silverman [2005]. To capture the regression relation of the response with a functional predictor, regularization is necessary. One common approach is functional principal component analysis, which has been studied by Rice and Silverman [1991], and recently by Yao et al. [2005b], Hall et al. [2006], Cai and Hall [2006], Zhang and Chen [2007], and Hall and Horowitz [2007], among others. Functional linear models have been extended to generalized functional linear models [Escabias et al., 2004, Cardot and Sarda, 2005, Müller and Stadtmüller, 2005], varying-coefficient models [Fan and Zhang, 2000, Fan et al., 2003], wavelet-based functional models [Morris et al., 2003], functional additive models [Müller and Yao, 2008] and quadratic models [Yao and Müller, 2010].

Classical functional linear regression is designed to describe the relation between a real-
valued response and one functional explanatory variable. However, in many real problems, it is common to also collect a large number of non-functional predictors. How to incorporate scalar predictors in functional linear regression and perform model selection/regularization is an important issue. For a standard linear regression with scalar covariates only, various penalization procedures have been proposed and studied, including the lasso [Tibshirani, 1996], the smoothly clipped absolute deviation [Fan and Li, 2001] and the adaptive lasso [Zou, 2006].

In this work, we develop a class of partially functional linear regression models, to handle multiple functional and non-functional predictors and automatically identify important risk factors by suitable regularization. Shin [2009] and Lu et al. [2014] have considered similar partially functional linear and quantile models, respectively, but did not deal with variable selection or with multiple functional predictors and high-dimensional scalar covariates. We propose a unified framework that regularizes each functional predictor as a whole, combined with a penalty on high-dimensional scalar covariates. Due to the differences between the functional and scalar predictors, we use two regularizing operations. Shrinkage penalties are imposed on the effects of both functional predictors and scalar covariates to achieve model selection and enhance interpretability, while a data-adaptive truncation that plays the role of a tuning parameter is applied to functional predictors. We treat the functional predictors as infinite-dimensional processes, which distinguishes our work from work that fixes the number of principal components [Li et al., 2010]. A main contribution is to quantify the theoretical impact of functional principal component estimation with diverging truncation, especially when the number of scalar covariates is permitted to diverge at an exponential order of the sample size.
1.2 Regularized Partially Functional Linear Regression

1.2.1 Classical functional linear model via principal components

Let $X(\cdot)$ be a square-integrable random function defined on a closed interval $T$ of the real line with continuous mean and covariance functions, denoted by $E\{X(t)\} = \mu(t)$ and $\text{cov}\{X(s), X(t)\} = K(s, t)$, respectively. The classical functional linear model is

$$Y = \mu_Y + \int_T \{X(t) - \mu(t)\} \beta(t) dt + \epsilon,$$

where the regression parameter function $\beta(\cdot)$ is assumed to be square-integrable, and $\epsilon$ is a random error independent of $X(t)$. Mercer’s theorem implies that there exists a complete and orthonormal basis $\{\phi_k\}$ in $L^2(T)$ and a non-increasing sequence of non-negative eigenvalues $\{w_k\}$ such that $K(s, t) = \sum_{k=1}^{\infty} w_k \phi_k(s) \phi_k(t)$ with $\sum_{k=1}^{\infty} w_k < \infty$. We further assume that $w_1 > w_2 > \cdots \geq 0$. Let $\{(y_i, x_i), i = 1, \ldots, n\}$ be independent and identically distributed random samples from $(Y, X)$. The Karhunen–Loève expansion $x_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik} \phi_k(t)$ forms the foundation of functional principal component analysis, where the coefficients $\xi_{ik} = \int_T \{x_i(t) - \mu(t)\} \phi_k(t) dt$ are uncorrelated random variables with mean zero and variances $E(\xi_{ik}^2) = w_k$, also called the functional principal component scores. Expanded on the orthonormal eigenbasis $\{\phi_k\}$, the regression function becomes $\beta(t) = \sum_{k=1}^{\infty} b_k \phi_k(t)$, and the functional linear model (1.1) can be written as $y_i = \mu_Y + \sum_{k=1}^{\infty} b_k \xi_{ik} + \epsilon_i$. The basis with respect to which the regression parameter $b$ is expanded is determined by the covariance function $K$. This is not unnatural since $\{\phi_k\}$ is the unique canonical basis leading to a generalized Fourier series which gives the most rapidly convergent representation of $X$ in $L^2$ sense.

1.2.2 Partially functional linear regression with regularization

We now consider functional linear regression with multiple functional and scalar predictors. Suppose the data are $\{Y, X(\cdot), Z\}$, where $Y$ is a scalar continuous response, $X(\cdot) = \{X_j(\cdot) :
CHAPTER 1. PARTIALLY FUNCTIONAL LINEAR REGRESSION IN HIGH DIMENSIONS

$j = 1, \ldots, d$ are $d$ functional predictors, and $Z = (Z_1, \ldots, Z_{p_n})^T$ is a $p_n$-dimensional vector of scalar covariates. This is motivated by commonly encountered situations where both functional and non-functional predictors may have significant impacts on the response. As with many real examples, we assume the number of functional predictors $d$ to be fixed, while the number of scalar covariates $p_n$ may grow with the sample size. Specifically we allow $p_n$ to be ultra-high dimensional, such that $\log p_n = O(n^\alpha)$ for some $\alpha > 0$. Without loss of generality, we assume that the response $Y$, the functional predictors $\{X_j : j = 1, \ldots, d\}$ and the scalar covariates $\{Z_l : l = 1, \ldots, p_n\}$ have been centered to have mean zero. We then model the linear relationship between $Y$ and $(X, Z)$ by

$$Y = \sum_{j=1}^{d} \int_T X_j(t) \beta_j(t) dt + Z^T \gamma + \epsilon,$$

where $\{\beta_j(\cdot) : j = 1, \ldots, d\}$ are square-integrable regression parameter functions, $\gamma = (\gamma_1, \ldots, \gamma_{p_n})^T$ contains the regression coefficients of non-functional covariates, and $\epsilon$ is the random error independent of $\{X_j(\cdot) : j = 1, \ldots, d\}$ and $Z$ with $E(\epsilon) = 0$ and $\text{var}(\epsilon) = \sigma^2$. For convenience, assume that the first $q_n$ scalar covariates are significant, while the rest are not. In other words, the true values of regression coefficients $\gamma_0^T$ equals $(\gamma_0^{(1)}^T, \gamma_0^{(2)}_0^T)$, where $\gamma_0^{(1)}$ is a $q_n \times 1$ vector corresponding to significant effects and $\gamma_0^{(2)}$ is a $(p_n - q_n) \times 1$ zero vector. We also assume that only the first $g$ functional predictors are significant, equivalently, the true values of regression functions $\beta_j(\cdot)$ is an infinite-dimensional process and requires regularization. Therefore the proposed model has a partially functional structure that combines the multiple functional and high-dimensional scalar components into one linear framework.

Let $\{(y_i, x_i, z_i) : i = 1, \ldots, n\}$ denote independent and identically distributed realizations from the population $(Y, X, Z)$. Let $x_{ij}$ denote the $j$th component of $x_i$ for $j = 1, \ldots, d$, and let $z_{ij}$ be the $l$th component of $z_i$ for $l = 1, \ldots, p_n$. We further write $Y_M = (y_1, \ldots, y_n)^T$, and $Z_M = (z_1, \ldots, z_n)^T$. To estimate the functions $\{\beta_j(\cdot) : j = 1, \ldots, d\}$ and the regres-
sion coefficients \( \{ \gamma_j : j = 1, \ldots, p_n \} \), we consider the least squares loss, which couples
\[ \beta_j(t) = \sum_k b_{jk} \phi_{jk}(t) \] with \( x_{ij}(t) = \sum_k \xi_{ijk} \phi_{jk}(t) \) for each \( j = 1, \ldots, d \) given the complete
orthonormal basis series \( \{ \phi_{jk} \}_{k=1,2,\ldots} \),

\[
L(b, \gamma | \mathcal{D}_n) = \sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{d} \int_{T} x_{ij}(t) \beta_j(t) dt - z_i^T \gamma \right\}^2 
\]
(1.3)

where \( \mathcal{D}_n = \{(y_i, x_i, z_i) : i = 1, \ldots, n\} \), and \( b = (b_1^T, \ldots, b_d^T)^T \) with \( b_j = (b_{j1}, b_{j2}, \ldots)^T \)
for each \( j \). It is evident that the loss function (1.3) should not be directly minimized due
to the infinite expansions of the functional predictors and high-dimensional scalar covariates,
requiring suitable regularization for both \( X \) and \( Z \).

One primary goal for (1.2) is to extract useful information from \( Z \) and \( X \), whereas the
classical functional linear model focuses only on a single functional predictor. It is thus essen-
tial to select and estimate the nonzero coefficients in \( \gamma \) and nonzero functions in \( b_1, \ldots, b_d \)
to enhance model prediction and interpretability. To achieve simultaneous variable selection and
estimation, we introduce a shrinkage penalty function \( J_\lambda(\cdot) \) associated with a tuning parameter
\( \lambda \). Many penalty choices are available for variable selection. We use the smoothly clipped
absolute deviation penalty of Fan and Li [2001], whose derivative is
\[
J'_\lambda(|\gamma|) = \lambda I(|\gamma| \leq \lambda) + I(|\gamma| > \lambda)(a\lambda - |\gamma|)_+/(a - 1)\lambda
\]
with \( a = 3.7 \) suggested by Fan and Li [2001] for implementation.

Due to the infinite dimensionality of the functional predictors, smoothing and regulariza-
tion are necessary in estimation. It is sensible to control the complexity of \( \beta_j(t) \) as a whole
function, rather than treating its basis terms as separate predictors. We adopt the simple yet ef-
effective truncation approach in the spirit of controlling smoothness as in classical nonparametric
regression. Denote the truncated form by \( X_{s_j}(t) = \mu_j(t) + \sum_{k=1}^{s_{jk}} \xi_{jk} \phi_{jk}(t) \) for \( j = 1, \ldots, d \),
where \( s_j \) is the truncation parameter. Correspondingly, for \( \beta_j(t) = \sum_{k=1}^{\infty} b_{jk} \phi_{jk}(t) \), write
CHAPTER 1. PARTIALLY FUNCTIONAL LINEAR REGRESSION IN HIGH DIMENSIONS

\[ b_j = (b_j^{(1)T}, b_j^{(2)T})^T, \text{ where } b_j^{(1)} = (b_{j1}, ..., b_{j_s})^T \text{ and } b_j^{(2)} = (b_{j,s_j+1}, ...)^T. \]

Unlike with non-functional effects, there is no underlying separation between \( b_j^{(1)} \) and \( b_j^{(2)} \). For the sake of adaptivity, we allow \( s_j \) to vary with the sample size, \( s_j \equiv s_{nj} \), such that it behaves like a smoothing parameter that balances the trade-off between bias and variance. The coefficients in \( b_j^{(2)} \) associated with higher-order basis functions are nonzero but decay rapidly. It is of interest to study the impact of \( s_{nj} \) on the convergence rates of the resulting estimators.

In practice we do not observe the entire trajectories \( x_{ij} \), and only have intermittent noisy measurements \( W_{ijl} = x_{ij}(t_{ijl}) + \epsilon_{ijl} \), where \( \{\epsilon_{ijl}, i = 1, ..., n\} \) are independent and identically distributed measurement errors independent of \( x_{ij} \), satisfying \( E(\epsilon_{ijl}) = 0 \), \( \text{var}(\epsilon_{ijl}) = \sigma^2_{xj} \) for \( i = 1, ..., n \) and \( l = 1, ..., m_{ij} \). When the repeated observations are sufficiently dense for each subject, a common practice is to run a smoother through \( \{(t_{ijl}, W_{ijl}), l = 1, ..., m_{ij}\} \), and then the estimates \( \{\hat{x}_{ij}, i = 1, ..., n, j = 1, ..., d\} \) are used to construct the covariance, eigenvalues/basis, and functional principal component scores; details are given in section 1.6. A theoretical justification of the asymptotic equivalence between the estimators obtained from \( \hat{x}_{ij} \) and those from the true \( x_{ij} \) is also given in section 1.6. The unobservable functional principal component scores \( \{\xi_{ijk} : k = 1, ..., s_{nj}; j = 1, ..., d; i = 1, ..., n\} \) are estimated by functional principal component analysis based on the observed data \( \{(t_{ijl}, W_{ijl}) : l = 1, ..., m_{ij}; j = 1, ..., d; i = 1, ..., n\} \). Therefore we minimize

\[
\min_{b_j^{(1), \gamma}} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{d} \sum_{k=1}^{s_{nj}} \xi_{ijk} b_{jk} - z_i^T \gamma)^2 + 2n \sum_{j=1}^{d} J_{\lambda_jn}(\|b_j^{(1)}\|) + 2n \sum_{l=1}^{p_n} J_{\lambda_n}(|\gamma_l|), \tag{1.4}
\]

given suitable choices of \( s_{nj}, \lambda_n \) and \( \lambda_{jn} \), where \( \|b_j^{(1)}\| \) is the Euclidean norm invoking a group penalty that shrinks the regression functions of unimportant functional predictors to zero. To regularize all predictors on a comparable scale, one often standardizes the predictors before imposing a penalty associated with a common tuning parameter [Fan and Li, 2001]. Thus we standardize \( (z_{1l}, ..., z_{nl})^T \) to have unit variance. The variability of the \( j \)th functional predictor can be approximated by \( \sum_{k=1}^{s_{nj}} \hat{w}_{jk} \), where \( \hat{w}_{jk} \) is the \( k \)th estimated eigenvalue of the \( j \)th predic-
CHAPTER 1. PARTIALLY FUNCTIONAL LINEAR REGRESSION IN HIGH DIMENSIONS

Since standardization is equivalent to adding weights to the penalty function, we suggest using
\[ \lambda_{jn} = \lambda_n (\sum_{k=1}^{s_{nj}} \hat{w}_{jk})^{1/2}, \]
which simplifies both the computation and theoretical analysis. The estimated regression parameter functions are
\[ \hat{\beta}_j(t) = \sum_{s_{nj}=1}^{s_{nj}} \hat{b}_{jk} \hat{\phi}_{jk}(t). \]

1.2.3 Algorithms and parameter tuning

The optimization of (1.4) can be seen as a group smoothly clipped absolute deviation problem with different weights on penalties, and the individual \( \gamma_l \) can be treated as group of size one. We propose two algorithms to solve the minimization problem (1.4), which adapts to the dimension \( p_n \). Generally, when \( p_n \) is moderately large, say \( p_n < n \), we modify the local linear approximation algorithm [Zou and Li, 2008], which inherits the computational efficiency and sparsity of lasso-type solutions. For ultra-high \( p_n \), especially \( p_n \gg n \), the local linear approximation algorithm may not be applicable, and then we modify the concave convex procedure used in Kim et al. [2008]. The following gives the details. Recall that \( Y_M = (y_1, \ldots, y_n)^T \) and \( Z_M = (z_1, \ldots, z_n)^T \), where \( z_i = (z_{i1}, \ldots, z_{ip_n})^T \). In addition, \( M_j \) is a \( n \times s_n \) matrix with \((i,k)\)th element \( \xi_{ijk} \), \( M = (M_1, \ldots, M_d) \), and \( N = (M, Z_M) = (N_1, \ldots, N_s)^T \) is a \( n \times (d s_n + p_n) \) matrix. Further, \( \eta = (b^{(1)}_r, \gamma^T)^T \). The solution to (4) is equivalent to

\[
\arg\min_{\eta} \left\{ (2n)^{-1} ||Y_M - N\eta||^2 + \sum_{r=1}^{R} J'_{\lambda_r}(||\eta_r||) ||\eta_r|| \right\},
\]

where \( R = d + p_n \). The tuning parameter \( \lambda_r = \lambda_{rn} \) with group size \( K_r = s_n \) if \( r = 1, \ldots, d \), and \( \lambda_r = \lambda_n \) with group size \( K_r = s_n \) if \( r = d+1, \ldots, d+p_n \).

When \( p_n \) is moderately large, say \( p_n < n \), one can modify the local linear approximation algorithm of Zou and Li [2008] which inherits the computational efficiency and sparsity of lasso-type solutions. Denote the initial estimate from the ordinary least square solution by \( \hat{\eta}^{(0)} \), and we solve \( \hat{\eta}^{(1)} = \arg\min_{\eta} \left\{ (2n)^{-1} ||Y_M - N\eta||^2 + \sum_{r=1}^{R} J'_{\lambda_r}(||\eta_r^{(0)}||) ||\eta_r|| \right\} \). Since some of the \( J'_{\lambda_r}(||\eta_r^{(0)}||) \) are zero, we use similar algorithm proposed by Zou and Li [2008]. Denote \( V = \{ r : J'_{\lambda_r}(||\eta_r^{(0)}||) = 0 \} \), \( W = \{ r : J'_{\lambda_r}(||\eta_r^{(0)}||) > 0 \} \), \( N = (N_V, N_W) \) and
\( \eta^{(1)} = (\eta_V^{(1)\top}, \eta_W^{(1)\top})\). Our algorithm is as follows:

1. Reparameterize the response vector by \( Y_M^* = N\eta^{(0)} \), and reparameterize the observed data matrix by \( N_r^* = N_r K_r^{1/2} / J_{\lambda_r} (\| \eta_r^{(0)} \|) \) for \( r \in W \); and \( N_r^* = N_r \) for \( r \in V \).

2. Let \( P_V \) denote the projection matrix of the space \( \{ N_r^*, r \in V \} \), where \( P_V = N_V (N_V^T N_V)^{-1} N_V^T \).

Then, calculate \( Y_M^* = Y_M^* - P_V Y_M^* \) and \( N_W^* = N_W^* - P_V N_W^* \).

3. Find \( \hat{\eta}_W^* = \arg\min_{\eta} \{ (2n)^{-1} \| Y_M^* - N_W^* \eta \|^2 + \sum_{r \in W} K_r^{1/2} \| \eta_r \| \} \).

4. Compute \( \hat{\eta}_V^* = (N_V^T N_V)^{-1} N_V^T (Y_M^* - N_W^* \hat{\eta}_W^*) \).

5. Let \( \eta^{(0)} = \eta^{(1)} \), where \( \eta_V^{(1)} = \hat{\eta}_V^* \), and \( \eta_r^{(1)} = \hat{\eta}^*_r K_r^{1/2} / J_{\lambda_r} (\| \eta_r^{(0)} \|) \) for \( r \in W \).

We repeat steps 1–5 until convergence; the final \( \eta^{(0)} \) is the regularized estimator. Step 3 essentially solves a group lasso, so we adopt the shooting algorithm of Fu [1998] and Yuan and Lin [2006].

For \( p_n \gg n \), it is likely that \( N_V^T N_V \) in step 4 is singular, so the local linear approximation algorithm is inapplicable. We modify the concave convex procedure used in Kim et al. [2008]. Let \( J_{\lambda_r} (\| \eta_r \|) = J_{\lambda_r} (\| \eta_r \|) - \lambda_r \| \eta_r \| \) for each \( r \). The convex part of the objective function is \( C_{\text{vex}} (\eta) = (2n)^{-1} \| Y_M - N \eta \|^2 + \sum_{r = 1}^J \lambda_r \| \eta_r \| \), and the concave part is \( C_{\text{cav}} (\eta) = \sum_{r = 1}^J J_{\lambda_r} (\| \eta_r \|) \). Begin with the initial estimator \( \eta^{(0)} = 0 \) and iteratively update the solution until convergence,

\[
\eta^{(1)} = \arg\min_{\eta} \{ C_{\text{vex}} (\eta) + \nabla C_{\text{cav}} (\eta^{(0)})^\top \eta \} = \arg\min_{\eta} \{ M (\eta | \eta^{(0)}) + \sum_{r = 1}^R \lambda_r \| \eta_r \| \},
\]

where \( M (\eta | \eta^{(0)}) = (2n)^{-1} \| Y_M - N \eta \|^2 + \nabla C_{\text{cav}} (\eta^{(0)})^\top \eta \) is quadratic in \( \eta \). The proximal gradient method of Parikh and Boyd [2013] is adopted to solve the above minimization problem.

Two sets of tuning parameters play crucial roles in the penalized procedure (1.4). The parameter \( \lambda_n \) in the smoothly clipped absolute deviation directly controls the sparsity of both
CHAPTER 1. PARTIALLY FUNCTIONAL LINEAR REGRESSION IN HIGH DIMENSIONS

the functional and non-functional predictors. Wang et al. [2007] showed that minimizing the BIC can identify the true model consistently, while generalized cross-validation might lead to over-fitting. The truncation parameters \( s_{nj} \) control the dimensions of the functional spaces to approximate the true function parameters. Previous work mostly chose \( s_{nj} \) based on the functional principal component representation, such as leave-one-curve-out cross-validation [Rice and Silverman, 1991] and the pseudo-AIC [Yao et al., 2005a]. However, a sensible tuning criterion of \( s_{nj} \) in a regression setting should take into account its impact on the response. The process \( X_j \) is infinite-dimensional, and its coefficient function \( \beta_j(t) = \sum_{k=1}^{\infty} b_{jk} \phi_k(t) \) does not have a finite cut-off. This is similar to the situation that the true model does not lie in the space formed by finite-dimensional candidate models. Therefore we propose a hybrid tuning procedure, which in principle combines BIC for tuning \( \lambda_n \) and AIC for choosing \( s_{nj} \) due to the infinite-dimensional parameter spaces of \( \{\beta_j; j = 1, \ldots, \infty\} \). In practice it is computationally prohibitive to choose \( \{s_{nj}; j = 1, \ldots, d\} \) simultaneously in the penalized procedure (1.4).

Table 1.1 suggests that, when using a common truncation, the selection of both functional and scalar covariates is accurate and stable for a wide range of \( s_n \). Thus we propose to use a common truncation parameter \( s_n \) when solving (1.4), then refit the selected model with the significant functional and scalar predictors using ordinary least squares, while different truncation parameters \( s_{nj} \) are tuned simultaneously by AIC for the retained functional predictors.

Specifically, for a fixed pair \((s_n, \lambda_n)\), the ABIC criterion is defined as

\[
\text{ABIC}(s_n, \lambda_n) = \log \{\text{RSS}(s_n, \lambda_n)\} + 2g(s_n, \lambda_n)s_n/n + n^{-1} \text{df}(s_n, \lambda_n) \log(n),
\]

where

\[
\text{RSS}(s_n, \lambda_n) = \sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{d} \sum_{k=1}^{s_n} \hat{\xi}_{ijk} \hat{b}_{jk}(s_n, \lambda_n) - z_i^T \hat{\gamma}(s_n, \lambda_n) \right\}^2,
\]

and \( g(s_n, \lambda_n) \) is the number of non-zero estimates of the regression functions, \( g(s_n, \lambda_n) = \sum_{j=1}^{d} I(\hat{\beta}_j; s_n, \lambda_n) \). The degree of freedom \( \text{df}(s_n, \lambda_n) \) equals \( I(\hat{\gamma}; s_n, \lambda_n) + \sum_{j=1}^{d} \sum_{k=1}^{s_n} \hat{w}_{jk} I(\hat{\beta}_j; s_n, \lambda_n) \).
with \( I(\hat{\gamma}; s_n, \lambda_n) \) indicating the number of non-zero elements in \( \hat{\gamma} \). This procedure requires estimation using the whole data only once and is computationally fast.

For the refit step, denote the index set of the selected functional predictors by \( D \subset \{1, \ldots, d\} \), and the index set of the selected scalar covariates by \( S \subset \{1, \ldots, p_n\} \). We minimize

\[
AIC(s_nj : j \in D) = \log \{ \text{RSS}(s_nj : j \in D) \} + 2n^{-1} \sum_{j \in D} s_{nj},
\]

with respect to combinations of \( \{s_nj : j \in D\} \), where

\[
\text{RSS}(s_nj : j \in D) = \sum_{i=1}^{n} \left\{ y_i - \sum_{j \in D} \sum_{k=1}^{s_{nj}} \hat{\xi}_{ijk} \hat{b}_{jk}^*(s_nj) - \sum_{l \in S} z_{ij} \hat{\gamma}_l^*(s_nj) \right\}^2,
\]

and \( \hat{b}_{jk}^*(s_nj) \) and \( \hat{\gamma}_l^*(s_nj) \) are the refitted values using ordinary least squares.

### 1.3 Asymptotic Properties

Denote the true values of \( b^{(1)} \) and \( \gamma \) by \( b_0^{(1)} \) and \( \gamma_0 \), and similarly for the rest of parameters. Recall that the boundedness of the covariance functions \( K_j(s, t) \) and regression operators implies that \( \sum_{k=1}^{\infty} w_{jk} < \infty \) and \( \sum_{k=1}^{\infty} b_{jk0}^2 < \infty \). We impose mild conditions on the decay rates of eigenvalues \( \{w_{jk}\} \) and regression coefficients \( \{b_{jk0}\} \), similar to those adopted by Hall and Horowitz [2007] and Lei [2014]. We assume that, for \( j = 1, \ldots, d \):

\[
(A1) \quad w_{jk} - w_{j(k+1)} \geq Ck^{-a-1} \text{ for } k \geq 1.
\]

This implies that \( w_{jk} \geq Ck^{-a} \). As the covariance functions \( K_1, \ldots, K_d \) are bounded, one has \( a > 1 \). Regarding the regression function \( \beta_j(\cdot) \), in order to prevent the coefficients \( b_{jk0} \) from decreasing too slowly, we assume that

\[
(A2) \quad |b_{jk0}| \leq Ck^{-b} \text{ for } k > 1.
\]

The decay conditions are needed only to control the tail behaviors for large \( k \), and so are not as restrictive as they appear. Without loss of generality, we use a common truncation parameter...
s_n in theoretical analysis. It is important to control s_n appropriately. On one hand s_n cannot be too large due to increasingly unstable functional principal component estimates,

(A3) \( (s_n^{2a+2} + s_n^{a+4})/n = o(1) \).

On the other hand s_n cannot be too small, so that the covariances between Z and the unobservable \( \{\xi_{jk}: k \geq s_n + 1\} \) are asymptotically negligible,

(A4) \( s_n^{2b-1}/n \to \infty \) as \( n \to \infty \).

Combining (A3) and (A4) entails \( b > \max(a+3/2, a/2+5/2) \) as a sufficient condition for such an \( s_n \) to exist. This implies that the regression function is smoother than the lower bound on the smoothness of \( \hat{K}_j \). Regarding the dimension of scalar covariates, assume that the number of significant covariates satisfies

(A5) \( s_n^{a+2} q_n^2/n = o(1) \).

Such \( q_n = o(n^{1/2} s_n^{-a/2-1}) \) does exist and is allowed to diverge with the sample size given (A3).

The dimension of the candidate set, \( p_n \), is allowed to be ultra-high,

(A6) \( p_n = O\{\exp(n^\alpha)\} \) for some \( \alpha \in (0, 1/2) \).

Lastly we require the following to hold for the tuning parameter \( \lambda_n \) and the sparsity of \( \gamma \) to achieve consistent estimation,

(A7) \( \lambda_n = o(1), \max\{n^{2\alpha-1}, \min_{t=1,\ldots,q_n} |\gamma_t|/\lambda_n \} \to \infty \).

We defer to section 1.6 the standard conditions (B1)–(B5) on the underlying processes \( x_{ij} \), how the data are sampled and smoothed, and the moments of scalar covariates, followed by the auxiliary lemmas and proofs.

To facilitate theoretical analysis, we re-parameterize by writing \( \tilde{b}_{jk} = w_{jk}^{1/2} b_{jk} \), so that the functional principal component scores serving as predictor variables are on a common scale of variabilities. This re-parameterization is only used for technical derivations, and does not appear in the estimation procedure. Let \( \tilde{\eta} = (\tilde{b}_{1}^{(1)^T}, \gamma^{(1)^T})^T \), where \( \tilde{b}_{1}^{(1)} = (\tilde{b}_{1}^{(1)^T}, \ldots, \tilde{b}_{d}^{(1)^T})^T \),
\( \tilde{b}_j^{(1)} = A_j b_j^{(1)} \), and \( A_j \) is the \( s_n \times s_n \) diagonal matrix with \( A_j(k,k) = w_{jk}^{1/2} \). Then, the minimization of (1.4) is equivalent to minimizing

\[
Q_n(\tilde{\eta}) = \sum_{i=1}^{n} \{ y_i - \sum_{j=1}^{d} \sum_{k=1}^{s_n} (\tilde{\xi}_{ijk} w_{jk}^{-1/2}) \tilde{b}_{jk} - z_i^T \gamma \}^2 + 2n \sum_{l=1}^{p_n} J_{\lambda_n}(\|\gamma_l\|) + 2n \sum_{j=1}^{d} J_{\lambda_{j_n}}(\|b_j^{(1)}\|).
\]

Theorem 1 establishes the estimation and selection consistency for both the functional and scalar regression parameters. For a random variable \( \varepsilon \) with mean zero, define \( \varepsilon \) as a subgaussian random variable if there exists some positive constant \( C_1 > 0 \) such that \( \Pr(|\varepsilon| > t) \leq \exp(-2^{-1}C_1 t^2) \) for \( t \geq 0 \). Let \( \tilde{b}^{(1)} \) denote the estimate of \( b^{(1)} \).

**Theorem 1** If \( \varepsilon_1, \ldots, \varepsilon_n \) are independent and identically distributed subgaussian random variables, then under conditions (A1)–(A7) and (B1)–(B5), there exists a local minimizer \( \tilde{\eta} = (\tilde{b}^{(1)\tau}, \tilde{\gamma}^{(1)\tau}) \) of \( Q_n(\tilde{\eta}) \) such that \( \|\tilde{\eta} - \tilde{\eta}_0\| = O_p\{(q_n + s_n)/n\}^{1/2} \) and \( \Pr(\tilde{\gamma}_2 = 0, \tilde{b}^{(1)} = 0, j = g + 1, \ldots, d) \to 1 \).

The estimation consistency result is expressed in terms of \( \tilde{b}^{(1)} \), not the original parameter \( b^{(1)} = (b_1^{(1)\tau}, \ldots, b_d^{(1)\tau}) \). For estimation, given \( \tilde{b}^{(1)} = A_j^{-1} \tilde{b}^{(1)} \), it is easy to deduce that

\[
\|\hat{\beta}_j - \beta_{j0}\|_{L^2} = O_p\{s_n(q_n + s_n)/n\}, \text{ where } \hat{\beta}_j = \sum_{k=1}^{s_n} \hat{b}_{jk} \hat{\phi}_{jk} \text{ and } \beta_{j0} = \sum_{k=1}^{\infty} b_{jk0} \phi_{jk}.
\]

**Theorem 2** Establishes the asymptotic normality for the \( q_n \)-dimensional vector \( \tilde{\gamma}^{(1)} \). Write \( \Sigma_1 = E(z_i^{(1)} z_i^{(1)\tau}) \), and \( \hat{\Sigma}_1 = n^{-1} \sum_{i=1}^{n} z_i^{(1)} z_i^{(1)\tau} \) with \( z_i^{(1)} = (z_{i1}, \ldots, z_{iq_n})^\tau \).

**Theorem 2** If \( \varepsilon_1, \ldots, \varepsilon_n \) are independent and identically distributed subgaussian random variables and \( q_n = o(n^{1/3}) \), then under conditions (A1)–(A7) and (B1)–(B5), for the local minimizer in Theorem 1, \( n^{1/2} A_n \hat{\Sigma}_1 (\tilde{\gamma}^{(1)} - \gamma_0^{(1)}) \to N(0, \sigma^2 H^* + B^*) \) in distribution, for any \( r \times q_n \) matrix \( A_n \) such that \( G = \lim_{n \to \infty} A_n A_n^\tau \) is positive definite, where \( \sigma^2 = \text{var}(\varepsilon) \), \( H^* = \lim_{n \to \infty} A_n \Sigma_1 A_n^\tau \), \( B^* = \lim_{n \to \infty} A_n B_n A_n^\tau \) with

\[
B_n = \text{cov}\left\{ \sum_{j=1}^{g} \sum_{k=1}^{s_n} \sum_{\ell \neq k} b_{jk0}(w_{jk} - w_{j\ell})^{-1} \langle \Xi_j, \phi_{j\ell} \rangle \int (x_{ij} \otimes x_{ij}) \phi_{jk} \phi_{j\ell} \right\},
\]

where \( \Xi_j = (\Xi_{j1}, \ldots, \Xi_{jq_n})^\tau \), \( E\{X_j(t) Z_l\} = \Xi_{jl}(t) \), and \( (x_{ij} \otimes x_{ij})(s,t) = x_{ij}(s)x_{ij}(t) \).
The asymptotic covariance is inflated by estimating the unobservable functional principal component scores. The inflation is quantified by a convergent sequence $B_n$ associated with the truncation size $s_n$.

### 1.4 Simulation Studies

The simulated data $\{y_i, i = 1, \ldots, n\}$ are generated from the model

$$y_i = \sum_{j=1}^{d} \int_0^1 \beta_j(t) x_{ij}(t) dt + z_i^T \gamma + \epsilon_i = \sum_{j=1}^{d} \sum_{k} b_{jk} \xi_{ijk} + z_i^T \gamma + \epsilon_i,$$

with $d = 4$ functional predictors, $p_n$ scalar covariates, the errors $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed from $N(0, \sigma^2)$, and $\gamma$ is the vector of scalar coefficients. The functional predictors have mean zero and covariance function derived from the Fourier basis

$$\phi_{2\ell-1} = 2^{-1/2} \cos \{(2\ell - 1)\pi t\} \quad \text{and} \quad \phi_{2\ell} = 2^{-1/2} \sin \{(2\ell - 1)\pi t\}, \quad \ell = 1, \ldots, 25, \quad \text{and} \quad t \in T = [0, 1].$$

The underlying regression function is $\beta_j(t) = \sum_{k=1}^{50} b_{jk} \phi_k(t)$, a linear combination of the eigenbasis. The scalar covariates $z_i = (z_{i1}, \ldots, z_{ip_n})^T$ are jointly normal with mean zero, unit variance and AR(0.5) correlation structure. Next, we describe how to generate the $d = 4$ functional predictors $x_{ij}(t)$. For $j = 1, \ldots, 4$, define $V_{ij}(t) = \sum_{k=1}^{50} \xi_{ijk} \phi_k(t)$, where $\{\xi_{ijk}, i = 1, \ldots, n\}$ follow independent and identically distributed $N(0, 16k^{-2})$ for different $i$ and $j$. The four functional predictors are then defined through the linear transformations

$$x_{i1} = V_{i1} + 0.5(V_{i2} + V_{i3}), \quad x_{i2} = V_{i2} + 0.5(V_{i1} + V_{i3}), \quad x_{i3} = V_{i3} + 0.5(V_{i1} + V_{i2}), \quad x_{i4} = V_{i4}.$$ 

Here the first three functional predictors are correlated with each other. To be more realistic, we allow a moderate correlation between $V_{i1}$ and $z_i$ for $i = 1, \ldots, n$, by setting that $\tilde{\xi} = (\tilde{\xi}_{i11}, \tilde{\xi}_{i12}, \tilde{\xi}_{i13}, \tilde{\xi}_{i14})^T$ and $z_i = (z_{i1}, \ldots, z_{ip_n})^T$ have a correlation structure specified by $\text{corr}(\tilde{\xi}_{i1k}, z_{ij}) = r^{|k-l|+1}$, $(k = 1, \ldots, 4; \ l = 1, \ldots, p_n)$, with $r = 0\cdot2$. For the actual obser-
vations, we assume they are realizations of \( \{x_{ij}(\cdot), j = 1, 2, 3, 4\} \) at 100 equally spaced times \( \{t_{ijl}, l = 1, \ldots, 100\} \in T \) with independent and identically distributed noise \( \varepsilon_{ijl} \sim N(0, 1) \).

We use 200 Monte Carlo runs for model assessment. Since inferences on both the parametric component \( \gamma \) and the functional components \( \beta_j \) are of interest, we report the Monte Carlo averages for the numbers of false nonzero and false zero functional predictors, and the functional mean squared error 

\[
\text{MSE}_f = \sum_{d=1}^{d} E(\|\hat{\beta}_j - \beta_j\|^2_{L_2}).
\]

For the scalar covariates, we report the Monte Carlo averages for the numbers of false nonzero and false zero scalar covariates, and the scalar mean squared error 

\[
\text{MSE}_s = E(\|\hat{\gamma} - \gamma\|^2).
\]

The prediction error is assessed using an independent test set of size \( N = 1000 \) for each Monte Carlo repetition, and

\[
\text{PE} = N^{-1} \sum_{i=1}^{N} (y_i^*-\hat{y}_i^*)^2 - \sigma^2,
\]

where \( \{x_i^*, z_i^*, y_i^*, j = 1, \ldots, 4\} \) are the testing data generated from the same model, and the predictions are \( \hat{y}_i^* = \sum_j \sum_k \hat{\xi}_{ijk} \hat{b}_{jk} + z_i^* \hat{\gamma} \) by plugging in estimates from the corresponding training sample.

Design I is for a moderate number of scalar covariates with sample size \( n = 200 \) and error variance \( \sigma^2 = 1 \). Specifically, for \( j = 1, 2, b_{j1} = 1, b_{j2} = 0.8, b_{j3} = 0.6, b_{j4} = 0.5 \) and \( b_{jk} = 8(k-2)^{-4} (k = 5, \ldots, 50) \), \( \beta_3 = \beta_4 = 0 \), and \( \gamma = (1_T^T, 0_{95}^T)^T \). Thus \( p_n = 20, q_n = 5 \). To illustrate the impact of the choice of \( s_n \), we inspect the results for \( s_n \) ranging from 1 to 16 with \( \lambda_n \) chosen by \text{BIC} in Table 1.1. The selection of functional and scalar predictors is quite accurate and stable for a wide range of \( s_n \), but with a very small number of false nonzero scalars. For functional predictors, the functional mean square error improves until \( s_n \) reaches an optimal level, then deteriorates as \( s_n \) continues to increase. For the scalar covariates, the mean square error and prediction error appear more stable after the optimal level. We then use \text{ABIC} with a common \( s_n \) to select both \( s_n \) and \( \lambda_n \). It yields similar results to those at optimal mean square and prediction errors, selecting an average \( \hat{s}_n = 4.30 \) with standard error \( 0.050 \). Refitting the selected model using ordinary least squares with jointly tuned \( s_{nj} \) via \text{AIC} improves the estimation of the functional coefficients and the overall prediction.

Design II illustrates the situation with ultra-high dimension of scalar covariates \( \gamma = (1_T^T, 0_{995}^T)^T \) with \( p_n = 1000 \), and other settings the same as in Design I. The \text{ABIC} yields results similar
to those giving the optimal estimation and prediction. The number of false nonzero scalar covariates, the scalar mean square error and prediction error in Step 1 become larger than those in Design I, mainly due to the ultra-high number of insignificant scalar covariates. The functional mean square error is also higher, as the correlation between functional predictors and scalar covariates becomes greater for a larger $p_n$. To improve the estimation and prediction, for each Monte Carlo run, after obtaining the estimates based on ABIC in Step 1, we generate an additional sample of size 200 and implement the penalized procedure using the significant variables and $s_n$ selected in Step 1. The results summarized in Step 2 are dramatically improved and become comparable to those for Design I. This hints at a promising two-step procedure via sample splitting in when $p_n$ is ultra-high, in a similar spirit to Fan et al. [2012]. A further improvement can be achieved by refitting the selected model with jointly tuned $s_{nj}$ using ordinary least squares. In particular, Design III illustrates the situation that the response does not necessarily depend on the leading principal components and the regression coefficients may decay more slowly than theoretically required. Specifically, we use $b_{j1} = 0, b_{j2} = 1, b_{j3} = 0.5$, $b_{jk} = (k - 2)^{-3}$ for $k = 4, \ldots, 50, j = 1, 2$, and other settings are the same as Design I. The results across $s_n$ follow a similar pattern, while the automated ABIC captures the regression relationship adaptively and resembles the optimal estimation and prediction. The refitting step using least squares with jointly tuned $s_{nj}$ behaves similarly as in Design I. Designs IV contains ultra-high numbers of scalar covariates $\gamma = (1, 0)^T_{5, 995}$ with $p_n = 1000$, and other settings the same as Design II. The results exhibit similar phenomenon as those in Design II. Moreover, Table 1.3 includes the results obtained from the same settings as Design I–IV except for a larger sample size $n = 400$. For Table 1.4, the only setting difference is that the regression error $\epsilon_i^*$ is generated from $N(0, 2)$ instead of $N(0, 1)$. As expected, a larger sample size reduced the estimation and prediction errors, while a higher noise level increased such errors.
<table>
<thead>
<tr>
<th>Design</th>
<th>$s_n$</th>
<th>$FZ_f$</th>
<th>$FN_f$</th>
<th>$MSE_f$</th>
<th>$FZ_s$</th>
<th>$FN_s$</th>
<th>$MSE_s$</th>
<th>PE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1</td>
<td>0.95</td>
<td>0</td>
<td>4.1 (0.03)</td>
<td>0.39</td>
<td>4.7 (0.15)</td>
<td>7.1</td>
<td>27.9 (0.6)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.35</td>
<td>0</td>
<td>2.2 (0.10)</td>
<td>0.05</td>
<td>3.2</td>
<td>1.1 (0.06)</td>
<td>7.9 (0.3)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.0</td>
<td>0</td>
<td>0.6 (0.01)</td>
<td>0.91</td>
<td>0.22 (0.013)</td>
<td>1.5</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.0</td>
<td>0</td>
<td>0.12 (0.005)</td>
<td>0.36</td>
<td>0.067 (0.005)</td>
<td>0.21</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.0</td>
<td>0</td>
<td>0.14 (0.005)</td>
<td>0.38</td>
<td>0.069 (0.004)</td>
<td>0.19</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.0</td>
<td>0</td>
<td>0.19 (0.007)</td>
<td>0.44</td>
<td>0.072 (0.004)</td>
<td>0.19</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.0</td>
<td>0</td>
<td>0.65 (0.03)</td>
<td>0.38</td>
<td>0.073 (0.004)</td>
<td>0.22</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.03</td>
<td>0.08</td>
<td>3.1 (0.13)</td>
<td>0.11</td>
<td>0.074 (0.006)</td>
<td>0.54</td>
<td>0.1</td>
</tr>
<tr>
<td>ABIC</td>
<td>0.0</td>
<td>0.13</td>
<td>0.007</td>
<td>0.34</td>
<td>0.067 (0.004)</td>
<td>0.20</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.09</td>
<td>0.003</td>
<td>0.28</td>
<td>0.065 (0.004)</td>
<td>0.18</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td>TUNE</td>
<td>0.0</td>
<td>0.14</td>
<td>0.005</td>
<td>0.10</td>
<td>0.047 (0.003)</td>
<td>0.17</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.095</td>
<td>0.005</td>
<td>0.10</td>
<td>0.047 (0.003)</td>
<td>0.17</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.076</td>
<td>0.003</td>
<td>0.09</td>
<td>0.046 (0.003)</td>
<td>0.16</td>
<td>0.004</td>
<td></td>
</tr>
</tbody>
</table>

Table 1.1: Simulation results with sample size $n = 200$ based on 200 Monte Carlo replicates for Designs I and II. Shown are the Monte Carlo averages (standard errors in parentheses) for the number of false zero functional predictors ($FZ_f$), the number of the false nonzero functional predictors ($FN_f$), the functional mean squared error ($MSE_f$), the number of false zero scalar covariates ($FZ_s$), the number of false nonzero scalar covariates ($FN_s$), the scalar mean squared error ($MSE_s$), and the prediction error (PE). We first use ABIC to choose the tuning parameter $\lambda_n$ and a common truncation $s_n$, then tune $s_{nj}$ jointly with AIC by refitting the selected model using ordinary least squares. In Design II, Step 1 results are based on the original sample in each Monte Carlo run, while Step 2 contains the improved results by fitting the penalized procedure to the selected model in Step 1 with an additional sample of $n = 200$. 
Table 1.2: Simulation results with sample size $n = 200$ based on 200 Monte Carlo replicates for Designs III and IV. Shown are the Monte Carlo averages (standard errors in parentheses) for the number of false zero functional predictors ($FZ_f$), the number of the false nonzero functional predictors ($FN_f$), the functional mean squared error ($MSE_f$), the number of false zero scalar covariates ($FZ_s$), the number of false nonzero scalar covariates ($FN_s$), the scalar mean squared error ($MSE_s$), and the prediction error ($PE$). We first use ABIC to choose the tuning parameter $\lambda_n$ and a common truncation $s_n$, then tune $s_{nj}$ jointly with AIC by refitting the selected model using ordinary least squares. In Design IV, Step 1 results are based on the original sample in each Monte Carlo run, while Step 2 contains the improved results by fitting the penalized procedure to the selected model in Step 1 with an additional sample of $n = 200$. 

<table>
<thead>
<tr>
<th>Design</th>
<th>$s_n$</th>
<th>$FZ_f$</th>
<th>$FN_f$</th>
<th>$MSE_f$</th>
<th>$FZ_s$</th>
<th>$FN_s$</th>
<th>$MSE_s$</th>
<th>$PE$</th>
</tr>
</thead>
<tbody>
<tr>
<td>III $p_n = 20$</td>
<td>1</td>
<td>2.0</td>
<td>0</td>
<td>2.5 (0)</td>
<td>1.5</td>
<td>1.1</td>
<td>4.2 (0.15)</td>
<td>26.0 (0.1)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.70</td>
<td>0</td>
<td>1.7 (0.06)</td>
<td>0.02</td>
<td>1.2</td>
<td>0.49 (0.028)</td>
<td>3.9 (0.1)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0.076 (0.004)</td>
<td>0</td>
<td>0.38</td>
<td>0.069 (0.004)</td>
<td>0.21 (0.009)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0.069 (0.003)</td>
<td>0</td>
<td>0.38</td>
<td>0.065 (0.004)</td>
<td>0.13 (0.005)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0.11 (0.005)</td>
<td>0</td>
<td>0.40</td>
<td>0.067 (0.004)</td>
<td>0.14 (0.005)</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0.15 (0.006)</td>
<td>0</td>
<td>0.41</td>
<td>0.068 (0.004)</td>
<td>0.15 (0.005)</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0</td>
<td>0.01</td>
<td>0.60 (0.02)</td>
<td>0</td>
<td>0.37</td>
<td>0.071 (0.004)</td>
<td>0.19 (0.006)</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.17</td>
<td>0.21</td>
<td>3.4 (0.1)</td>
<td>0</td>
<td>0.05</td>
<td>0.089 (0.007)</td>
<td>0.65 (0.05)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Design</th>
<th>$s_n$</th>
<th>$FZ_f$</th>
<th>$FN_f$</th>
<th>$MSE_f$</th>
<th>$FZ_s$</th>
<th>$FN_s$</th>
<th>$MSE_s$</th>
<th>$PE$</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV $p_n = 1000$</td>
<td>0.09</td>
<td>0.10 (0.003)</td>
<td>0</td>
<td>6.5</td>
<td>0.41 (0.02)</td>
<td>0.81 (0.021)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.059 (0.003)</td>
<td>0</td>
<td>1.3</td>
<td>0.48 (0.003)</td>
<td>0.13 (0.005)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TUNE $s_{nj}$</td>
<td>0.042 (0.002)</td>
<td>0</td>
<td>0.09</td>
<td>0.045 (0.003)</td>
<td>0.11 (0.004)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Design</td>
<td>$s_n$</td>
<td>$FZ_f$</td>
<td>$FN_f$</td>
<td>$MSE_f$</td>
<td>$FZ_s$</td>
<td>$FN_s$</td>
<td>$MSE_s$</td>
<td>PE</td>
</tr>
<tr>
<td>--------</td>
<td>-------</td>
<td>--------</td>
<td>--------</td>
<td>---------</td>
<td>--------</td>
<td>--------</td>
<td>---------</td>
<td>----</td>
</tr>
</tbody>
</table>
| I  
$p_n = 20$ | 1     | 0.90   | 0      | 4-1 (0.04) | 0.15   | 5-1    | 2.3 (0.08)     | 25.8 (0.2) |
|       | 2     | 0      | 0      | 1.3 (0.009) | 0      | 1.4    | 0.34 (0.01)     | 4.6 (0.04) |
|       | 3     | 0      | 0      | 0.58 (0.008) | 0      | 0.23   | 0.11 (0.004)    | 1.5 (0.02) |
|       | 4     | 0      | 0      | 0.063 (0.002) | 0      | 0.23   | 0.027 (0.002)   | 0.13 (0.004) |
|       | 5     | 0      | 0      | 0.071 (0.003) | 0      | 0.27   | 0.028 (0.002)   | 0.12 (0.004) |
|       | 6     | 0      | 0      | 0.095 (0.003) | 0      | 0.24   | 0.028 (0.002)   | 0.12 (0.004) |
|       | 10    | 0      | 0      | 0.32 (0.01)  | 0      | 0.31   | 0.029 (0.002)   | 0.13 (0.004) |
|       | 16    | 0      | 0      | 1.2 (0.04)   | 0      | 0.10   | 0.026 (0.002)   | 0.15 (0.004) |
|       | ABIC  | 0      | 0      | 0.067 (0.003) | 0      | 0.23   | 0.027 (0.002)   | 0.13 (0.004) |
|       | TUNE $s_{nj}$ | 0      | 0      | 0.049 (0.002) | 0      | 0.20   | 0.026 (0.002)   | 0.12 (0.004) |
| II  
$p_n = 1000$ | 1     | 0      | 0      | 0.11 (0.004) | 0      | 4.6    | 0.13 (0.007)    | 0.51 (0.01) |
|       | 2     | 0      | 0      | 0.052 (0.002) | 0      | 0.07   | 0.025 (0.001)   | 0.11 (0.004) |
|       | 3     | 0      | 0      | 0.028 (0.01)  | 0      | 0.25   | 0.028 (0.002)   | 0.099 (0.004) |
|       | 4     | 0      | 0      | 0.033 (0.001) | 0      | 0.23   | 0.026 (0.001)   | 0.073 (0.004) |
|       | 5     | 0      | 0      | 0.048 (0.002) | 0      | 0.26   | 0.027 (0.002)   | 0.075 (0.004) |
|       | 6     | 0      | 0      | 0.072 (0.003) | 0      | 0.25   | 0.027 (0.002)   | 0.078 (0.004) |
|       | 10    | 0      | 0      | 0.28 (0.01)   | 0      | 0.25   | 0.028 (0.002)   | 0.099 (0.004) |
|       | 16    | 0      | 0      | 1.2 (0.04)    | 0      | 0.02   | 0.024 (0.001)   | 0.13 (0.004) |
|       | ABIC  | 0      | 0      | 0.034 (0.002) | 0      | 0.23   | 0.026 (0.001)   | 0.075 (0.004) |
|       | TUNE $s_{nj}$ | 0      | 0      | 0.026 (0.001) | 0      | 0.22   | 0.026 (0.001)   | 0.070 (0.004) |
| III  
$p_n = 20$ | 1     | 2.0    | 0      | 2.5 (0)      | 1.0    | 0.43   | 2.7 (0.1)       | 25.2 (0.09) |
|       | 2     | 1.10   | 0      | 0.72 (0.04)  | 0      | 0.43   | 0.19 (0.01)     | 2.8 (0.06) |
|       | 3     | 0      | 0      | 0.059 (0.002) | 0      | 0.25   | 0.032 (0.002)   | 0.16 (0.006) |
|       | 4     | 0      | 0      | 0.033 (0.001) | 0      | 0.23   | 0.026 (0.001)   | 0.073 (0.004) |
|       | 5     | 0      | 0      | 0.048 (0.002) | 0      | 0.26   | 0.027 (0.002)   | 0.075 (0.004) |
|       | 6     | 0      | 0      | 0.072 (0.003) | 0      | 0.25   | 0.027 (0.002)   | 0.078 (0.004) |
|       | 10    | 0      | 0      | 0.28 (0.01)   | 0      | 0.25   | 0.028 (0.002)   | 0.099 (0.004) |
|       | 16    | 0      | 0      | 1.2 (0.04)    | 0      | 0.02   | 0.024 (0.001)   | 0.13 (0.004) |
|       | ABIC  | 0      | 0      | 0.034 (0.002) | 0      | 0.23   | 0.026 (0.001)   | 0.075 (0.004) |
|       | TUNE $s_{nj}$ | 0      | 0      | 0.026 (0.001) | 0      | 0.22   | 0.026 (0.001)   | 0.070 (0.004) |
| IV   
$p_n = 1000$ | 1     | 0      | 0      | 0.047 (0.001) | 0      | 3.8    | 0.17 (0.009)    | 0.35 (0.008) |
|       | 2     | 0      | 0      | 0.031 (0.002) | 0      | 0.08   | 0.024 (0.001)   | 0.073 (0.004) |
|       | 3     | 0      | 0      | 0.023 (0.001) | 0      | 0.07   | 0.024 (0.001)   | 0.058 (0.004) |

Table 1.3: Simulation results using the same settings as Design I & II in Section 4 of the paper and Design III & IV as above, based on 200 Monte Carlo replicates, except the sample size $n = 400$. 
<table>
<thead>
<tr>
<th>Design</th>
<th>$s_n$</th>
<th>FZ$_f$</th>
<th>FN$_f$</th>
<th>MSE$_f$</th>
<th>FZ$_s$</th>
<th>FN$_s$</th>
<th>MSE$_s$</th>
<th>PE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1</td>
<td>0.94</td>
<td>0</td>
<td>4.1 (0.03)</td>
<td>0.41</td>
<td>7.3</td>
<td>4.9 (0.2)</td>
<td>28.0 (0.6)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.56</td>
<td>0</td>
<td>2.8 (0.1)</td>
<td>0.08</td>
<td>3.3</td>
<td>1.4 (0.07)</td>
<td>9.9 (0.3)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.01</td>
<td>0</td>
<td>0.63 (0.02)</td>
<td>0</td>
<td>1.7</td>
<td>0.38 (0.03)</td>
<td>1.7 (0.07)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0.17 (0.007)</td>
<td>0</td>
<td>0.58</td>
<td>0.15 (0.009)</td>
<td>0.31 (0.01)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0.23 (0.1)</td>
<td>0</td>
<td>0.58</td>
<td>0.15 (0.009)</td>
<td>0.30 (0.01)</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0.33 (0.01)</td>
<td>0</td>
<td>0.63</td>
<td>0.15 (0.009)</td>
<td>0.32 (0.01)</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.01</td>
<td>0</td>
<td>1.3 (0.05)</td>
<td>0</td>
<td>0.49</td>
<td>0.15 (0.009)</td>
<td>0.44 (0.05)</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.31</td>
<td>0.18</td>
<td>7.7 (0.3)</td>
<td>0.08</td>
<td>0.20</td>
<td>0.36 (0.03)</td>
<td>3.6 (0.3)</td>
</tr>
</tbody>
</table>

| ABIC   | $s_n=4.12$ (0.026) | $s_n=4.04$ (0.021) |
| TUNE $s_{nj}$ | $s_{n1}=4.59$ (0.069), $s_{n2}=4.61$ (0.060) | $s_{n1}=4.53$ (0.056), $s_{n2}=4.46$ (0.050) |
| II     | $p_n=1000$ |
| STEP 1 | 0.04   | 0.29 (0.008) | 7.3    | 0.55 (0.03) | 1.9 (0.07) |
| STEP 2 | 0      | 0.15 (0.008) | 0.16   | 0.096 (0.005) | 0.26 (0.02) |
| TUNE $s_{nj}$ | 0      | 0.12 (0.006) | 0.15   | 0.095 (0.005) | 0.24 (0.008) |

| III    | $p_n=20$ |
| 1      | 2.0    | 2.5 (0) | 1.5    | 1.2    | 4.3 (0.15) | 26.1 (0.1) |
| 2      | 0.33   | 1.1 (0.06) | 0.02   | 0.69   | 0.43 (0.03) | 3.3 (0.1)  |
| 3      | 0      | 0.10 (0.005) | 0      | 0.39   | 0.13 (0.008) | 0.29 (0.01) |
| 4      | 0      | 0.12 (0.006) | 0      | 0.39   | 0.13 (0.008) | 0.22 (0.01) |
| 5      | 0      | 0.20 (0.009) | 0      | 0.41   | 0.13 (0.008) | 0.24 (0.01) |
| 6      | 0      | 0.29 (0.01)  | 0      | 0.41   | 0.14 (0.008) | 0.26 (0.01) |
| 10     | 0.03   | 0.07    | 1.4    | 0.06   | 0.17   | 0.12 (0.007) | 0.41 (0.03) |
| 16     | 0.02   | 1.3     | 11.8   | 0.4    | 0.02   | 0.39   | 0.18 (0.02) | 1.0 (0.03) |

| ABIC   | $s_n=3.76$ (0.065) | $s_n=3.33$ (0.040) |
| TUNE $s_{nj}$ | $s_{n1}=3.76$ (0.064), $s_{n2}=3.79$ (0.066) | $s_{n1}=3.96$ (0.060), $s_{n2}=3.73$ (0.048) |
| IV     | $p_n=1000$ |
| STEP 1 | 0.06   | 0.18 (0.006) | 0.02   | 0.63   | 0.63 (0.03) | 1.5 (0.04) |
| STEP 2 | 0.091  | 0.005   | 0.12   | 0.099  | 0.005  | 0.23 (0.009) |
| TUNE $s_{nj}$ | 0.005  | 0.065 (0.003) | 0.10   | 0.095  | 0.005  | 0.18 (0.009) |

Table 1.4: Simulation results using the same settings as Design I & II in Section 4 of the paper and Design III & IV as above, based on 200 Monte Carlo replicates, except the variance of the regression error $\sigma^2 = 2$. 
1.5 Application

We applied our method to a dataset from the National Mortality, Morbidity, and Air Pollution Study that contains air pollution measurements and mortality counts for U.S. cities with the U.S. census information for year 2000. A main goal of the study is to investigate the impact of air pollution on the non-accidental mortality rate across different cities in the United States, when we take into account climate patterns and inform the U.S. census. Previous studies conducted a two-stage analysis: first modelling the short-term effect of certain air pollutants on the mortality count for each city, then combining the estimates across cities [Peng et al., 2005, 2006]. By contrast, we apply the partially functional linear regression model to the data for different cities. In particular, we are interested in studying the effect of particulate matter with an aerodynamic diameter of less than 2.5\(\mu m\), abbreviated PM\(2.5\) and measured in \(\mu gm^{-3}\), as its negative effect on health, revealed by recent toxicological and epidemiological studies, has brought it to the public’s attention. Other studies [Samoli et al., 2013, Pascal et al., 2014] have shown that PM\(2.5\) has a larger effect on mortality in warm weather, so we focused on the daily concentration measurements of PM\(25\) from April 1, 2000 to August 31, 2000, which along with the daily observations of temperature and humidity were treated as the functional predictors. After removing the cities which have more than ten consecutive missing measurements of PM\(2.5\), we included a total of 69 cities in our analysis. The response of interest is the log-transformed total non-accidental mortality rate in the following month, September 2000, of the population of age 65 and older, which accounts for the majority of non-accidental deaths. The scalar covariates available from the U.S. census for each city are land area per individual, water area per individual, proportion of the urban population, proportion of the population with at least a high school diploma, proportion of the population with at least a university degree, proportion of the population below poverty line, and proportion of household owners.

The A\(BIC\) was used to first choose significant predictors with a common truncation, followed by a least squares refitting using A\(IC\) to tune \(s_{nj}\) jointly. Among scalar covariates, our analysis shows that only the proportion of household owners has a negative effect \(-1.80\) with
the standard error of $0.41$, indicating that household owners tend to incur a lower mortality rate, where the standard error was obtained based on 1000 bootstrap samples by fitting the selected model using ordinary least squares. Our method also selected two significant functional predictors, PM2·5 and temperature. The least squares refitting chose the truncation numbers $\hat{s}_{n1} = 2$ and $\hat{s}_{n2} = 2$. The estimated regression parameter functions with their 95\% bootstrap confidence bands are shown in Figure 1.1. We observe that higher PM2·5 concentrations during the summer, especially July and August, can lead to an increased mortality during the immediately following period, which coincides with the findings in Samoli et al. [2013] and Pascal et al. [2014]. This may be partially explained by the proximity of the pollution period to the time of death. Higher temperatures in the summer, contrasted with lower temperatures in April, may also increase the mortality rate, which agrees with Curriero et al. [2002]. To better understand the effects of functional predictors, we fitted a linear regression using only the selected scalar covariate, giving $R^2 = 0.15$. Including temperature leads to $R^2 = 0.25$, and including both temperature and PM2·5 yields $R^2 = 0.38$. A heuristic $F$ test for the significance of two principal component scores of temperature gives a $p$-value of $0.01$, and adding additional two principal component scores of PM2·5 gives a $p$-value of $0.0008$. For comparison, we also fitted the marginal models containing only PM2·5 or temperature using classical functional linear regression. The marginal $F$ tests for temperature and PM2·5 yield $p$-values of 0.0001 and 0.004. The regression parameter functions show similar patterns and are omitted. We conclude that, after adjusting for temperature and household ownership, summer PM2·5 concentrations have a significant impact on the near-future mortality rate of elder citizens in the U.S..

1.6 Proof of Lemmas and Main Theorems

1.6.1 Regularity Conditions

Without loss of generality, we assume that $\{X_j, j = 1, \ldots, d\}$, $Y$ and $Z$ have been centred to have mean zero. With $W_{ijk} = x_{ij}(t_{ijk}) + \epsilon_{ijk}$, for definiteness, we consider the local lin-
Chapter 1. Partially Functional Linear Regression in High Dimensions

Figure 1.1: The estimated regression parameter functions and 95% confidence bands based on 1000 bootstrap samples for PM$_{2.5}$ and temperature with $\hat{s}_{n1} = 2$ and $\hat{s}_{n2} = 2$, respectively. The solid line denotes the estimated regression parameter functions and the dashed lines denote the 95% bootstrap confidence bands. The left and right panels are for the PM$_{2.5}$ and temperature, respectively.

ear smoother for each set of subjects using bandwidths $\{h_{ij}, j = 1, \ldots, d\}$, and denote the smoothed trajectories by $\hat{x}_{ij}$. Denote the minimum and maximum eigenvalues of a symmetric matrix $A$ by $\lambda_{\text{min}}(A)$ and $\lambda_{\text{max}}(A)$. Recall that the first $g$ functional predictors are significant, while the rest are not. Define the $(g s_n + q_n) \times 1$ vector $\tilde{N} = (\xi_{11} w_{11}^{-1/2}, \ldots, \xi_{1s_n} w_{1s_n}^{-1/2}, \ldots, \xi_{g1} w_{g1}^{-1/2}, \ldots, \xi_{gs_n} w_{gs_n}^{-1/2}, Z_1, \ldots, Z_{q_n})^T$ to combine all functional and scalar predictors.

Condition (B1) consists of regularity assumptions for functional data, for example, a Gaussian process with Hölder continuous sample paths satisfies (B1), see Hall and Hosseini-Nasab [2006]. Condition (B2) is standard for local linear smoothers, (B3)–(B4) concern how the functional predictors are sampled and smoothed, while (B5) is for the moments of non-functional covariates $Z = (Z_1, \ldots, Z_{p_n})^T$.

(B1) For $j = 1, \ldots, d$, for any $C > 0$ there exists an $\epsilon > 0$ such that

$$\sup_{t \in T} [E\{|X_j(t)|^C\}] < \infty, \quad \sup_{s, t \in T} (E[\{|s - t|^{-\epsilon}|X_j(s) - X_j(t)|\}^C]) < \infty.$$ 

For each integer $r \geq 1$, $w_{jk}^{-r} E(\xi_{jk}^{2r})$ is bounded uniformly in $k$.

(B2) For $j = 1, \ldots, d$, $X_j$ is twice continuously differentiable on $T$ with probability 1, and
\[ \int E\{X_j^{(2)}(t)\}^4 dt < \infty, \] where \( X_j^{(2)}(\cdot) \) denotes the second derivative of \( X_j(\cdot) \).

The following condition concerns the design on which \( x_{ij} \) is observed and the local linear smoother \( \hat{x}_{ij} \). When a function is said to be smooth, we mean that it is continuously differentiable to an adequate order.

(B3) For \( j = 1, \ldots, d \), \( \{t_{ijl}, l = 1, \ldots, m_{ij}\} \) are considered deterministic and ordered increasingly for \( i = 1, \ldots, n \). There exist densities \( g_{ij} \) uniformly smooth over \( i \), satisfying
\[ \int_T g_{ij}(t) dt = 1 \text{ and } 0 < c_1 < \inf_i \{\inf_{t \in T} g_{ij}(t)\} < \sup_i \{\sup_{t \in T} g_{ij}(t)\} < c_2 < \infty \]
for all \( t_{ijl} \) according to \( t_{ijl} = G_{ij}^{-1}(l - 1)/(m_{ij} - 1) \), where \( G_{ij}^{-1} \) is the inverse of \( G_{ij}(t) = \int_{-\infty}^t g_{ij}(s) ds \). For each \( j = 1, \ldots, d \), there exist a common sequence of bandwidths \( h_j \) such that
\[ 0 < c_1 < \inf_i h_{ij}/h_j < \sup_i h_{ij}/h_j < c_2 < \infty, \]
where \( h_{ij} \) is the bandwidth for \( \hat{x}_{ij} \). The kernel density function is smooth and compactly supported.

Let \( T = [a_0, b_0], t_{ij0} = a_0, t_{ij,m_{ij}+1} = b_0 \), let \( \Delta_{ij} = \sup \{t_{ij,l+1} - t_{ij,l}, l = 0, \ldots, m_{ij}\} \) and \( m_j = m_j(n) = \inf_i \{1, \ldots, n \} m_{ij} \). The condition below is to let the smooth estimate \( \hat{x}_{ij} \) serve as well as the true \( x_{ij} \) in the asymptotic analysis, denoting \( 0 < \lim a_n/b_n < \infty \) by \( a_n \sim b_n \).

(B4) For \( j = 1, \ldots, d \), \( \sup_i \Delta_{ij} = O(m_j^{-1}), h_j \sim m_j^{-1/5}, m_j n^{-5/4} \rightarrow \infty \).

The condition for the scalar covariates is given below.

(B5) \( Z_1, \ldots, Z_{pn} \) are subgaussian random variables such that \( pr(|Z_l| > t) \leq \exp(-2^{-1}Ct^2) \)
for any \( t \geq 0 \) and some \( C > 0 \) that does not depend on \( l \), \( E(Z_l^4) \) is uniformly bounded for \( l = 1, \ldots, p_n \), \( \lambda_{\max}(\Sigma) \leq c_1 < \infty \) and \( 0 < c_2 \leq \lambda_{\min}(U_1) \leq \lambda_{\max}(U_1) \leq c_3 < \infty \)
for all \( n \), where \( \Sigma = E(ZZ^T) \) and \( U_1 = E(\tilde{N}\tilde{N}^T) \).

### 1.6.2 Auxiliary Lemmas

For each \( j = 1, \ldots, d \), given the estimated covariances \( \hat{K}_j(s,t) = n^{-1} \sum_{i=1}^n \hat{x}_{ij}(s)\hat{x}_{ij}(t) \), the eigenvalues/functions and functional principal component scores are estimated by
\[ \int_T \hat{K}_j(s,t)\hat{\phi}_{jk}(s)ds = \hat{\lambda}_{jk}\hat{\phi}_{jk}(t), \quad \hat{\xi}_{ijk} = \int_T \hat{x}_{ij}(t)\hat{\phi}_{jk}(t)dt, \quad (1.5) \]
subject to $\int_T \dot{\phi}_j^2(k)(t)dt = 1$ and $\int_T \dot{\phi}_j(k_1)(t)\dot{\phi}_j(k_2)(t)dt = 0$ for $k_1 \neq k_2$. Denote $\hat{\Delta}_j^2 = \|\hat{K}_j - K_j\|^2 = \int_T \int_T \{\hat{K}_j(s, t) - K_j(s, t)\}^2dsdt$, $R_j(s, t) = n^{1/2}\{\hat{K}_j(s, t) - K_j(s, t)\}$, $(x_{ij}(s,t) = x_{ij}(s)x_{ij}(t)$, the second derivative of $x_{ij}$ by $x_{ij}^{(2)}$, the minimum and maximum eigenvalues of a symmetric matrix $A$ by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$. Let the $n \times p_n$ matrix $Z_M = (z_1, \ldots, z_n)^T = (Z_M^{(1)}, Z_M^{(2)})$ with $Z_M^{(1)}$ containing the first $q_n$ columns of $Z_M$. Define $M_j$ as the $n \times s_n$ matrix with $(i,k$)th element $\hat{x}_{ijk}$ and $M = (M_1, \ldots, M_d)$, $M^{(1)} = (M_1, \ldots, M_q)$ and $M^{(2)} = (M_{g+1}, \ldots, M_d)$. Similarly, we define $\hat{M}_j$, $\hat{M}^{(1)}$ and $\hat{M}^{(2)}$, where $\hat{x}_{ijk}$ is replaced by $\hat{x}_{ijk}w_{jk}^{-1/2}$. Moreover, we denote $\hat{M}_j$ as the $n \times s_n$ matrix with $(i,k$)th element $\hat{\phi}_{ijk}$, and $\hat{M} = (\hat{M}_1, \ldots, \hat{M}_d)$, $\hat{M}^{(1)} = (\hat{M}_1, \ldots, \hat{M}_q)$ and $\hat{M}^{(2)} = (\hat{M}_{g+1}, \ldots, \hat{M}_d)$. Similarly, we define $\hat{M}_j$, $\hat{M}^{(1)}$ and $\hat{M}^{(2)}$, where $\hat{x}_{ijk}$ is replaced by $\hat{x}_{ijk}w_{jk}^{-1/2}$. Combining the functional and scalar covariates, we have $\hat{N}_1 = (\hat{M}^{(1)}, Z_M^{(1)}) = (\hat{N}_1, \ldots, \hat{N}_n)^T$, where $\hat{N}_i = (\hat{N}_1, \ldots, \hat{N}_g, \hat{x}_{n+1})^T, i = 1, \ldots, n$ are independently and identically distributed as $\hat{N}$. We further denote $N_1 = (M^{(1)}, Z_M^{(1)}) = (N_1, \ldots, N_n)^T$, $\hat{N}^{(1)} = (\hat{M}^{(1)}, Z_M^{(1)})$, $\hat{N}^{(1)} = (\hat{M}^{(1)}, Z_M^{(1)})$, and let $E(\hat{N}_1 \hat{N}_1^T) = U_1$. Recall that $\hat{b}^{(1)}$ denotes the estimate of $\hat{b}^{(1)}$, $\hat{\eta} = (\hat{b}^{(1)^T}, \hat{\eta}_T)^T$, and we further denote $\hat{b}_j^{(1)}$ the estimate of $\hat{b}_j^{(1)}$. Suppose $\alpha(\cdot), \beta(\cdot)$ and $G(\cdot, \cdot)$ are square-integrable functions on $T$ and $T \times T$. Write $\|\alpha\|$, $\int \alpha(\beta)$ (or $\langle \alpha, \beta \rangle$) and $\int G(\alpha, \beta)$ for $\{\int_T \alpha^2(t)dt\}^{1/2}$, $\int_T \alpha(t)\beta(t)dt$ and $\int \int_T G(s, t)\alpha(s)\beta(t)dsdt$.

Lemma 1 provides results for the estimates obtained by functional principal component analysis. We first quantify the smoothing error of $\hat{x}_{ij}$ and its influence carried over to the covariance and functional principal component estimates, $\hat{x}_{ijk} - x_{ijk} = \hat{x}_{ijk} - \hat{x}_{ijk} + \hat{x}_{ijk} - \hat{x}_{ijk} = \int x_{ij}(\hat{\phi}_{jk} - \hat{\phi}_{jk} + \int (\hat{x}_{ij} - x_{ij})\hat{\phi}_{jk}$. Then we obtain upper bounds for the differences between the estimated and true eigenfunctions in terms of covariance perturbation and eigenvalue spacings for quantifying the increased estimation error of the higher order terms. For each $j = 1, \ldots, d$, denote

$$
\delta_{jk} = \min_{l=1,\ldots,k} (w_{jl} - w_{j,l+1}), \quad J_n = \{k = 1, \ldots, \infty : w_{jk} - w_{jk+1} > 2\hat{\Delta}_j\}.
$$
where \( \hat{\Delta}_j = \| \hat{K}_j - K_j \| \), that is, consider \( k \in J_{jn} \) for which the distance of \( w_{jk} \) to the nearest other eigenvalues does not fall below \( 2\hat{\Delta}_j \). It is known that \( \phi_{jk} \) can be consistently estimated for \( k \in J_{jn} \) [Theorem 1, Hall and Hosseini-Nasab, 2006], \( \| \hat{\phi}_{jk} - \phi_{jk} \| = O_p(\delta_{jk}\hat{\Delta}_j) = O_p(k^{\alpha+1}n^{-1/2}) \), implying \( \sup J_{jn} = o\{n^{1/(2\alpha+2)}\} \). However, for our theoretical analysis, we need a sharper bound without compromising the number of eigenfunctions considered. Define the set of realizations such that, for sample size \( n \), some \( C \) and any \( \tau < 1 \),

\[
\mathcal{F}_{s_{n,j}} = \left\{ (\hat{w}_{jk1} - \hat{w}_{jk2})^{-2} \leq 2(w_{jk1} - w_{jk2})^{-2} \leq Cn^\tau, k_1, k_2 = 1, \ldots, s_n, k_1 \neq k_2 \right\}.
\]

**Lemma 1**  
(a) Under conditions (B1)-(B4), for each \( i = 1, \ldots, n \) and each \( j = 1, \ldots, d \), we have \( E(\| \hat{x}_{ij} - x_{ij} \|^2) = o(n^{-1}) \), \( E\left\{ \int (\hat{x}_{ij} - x_{ij})^4 \right\} = o(n^{-2}) \). For each \( j = 1, \ldots, d \), we have \( E(\| \hat{K}_j - K_j \|^2) = O(n^{-1}) \).

(b) Under conditions (A1), (A3), (B1)-(B4), for each \( j = 1, \ldots, d \), we have \( s_n \in J_{jn} \) for large \( n \), \( P(\mathcal{F}_{s_{n,j}}) \to 1 \) as \( n \to \infty \) for any \( \tau < 1 \). Moreover, for each \( k = 1, \ldots, s_n \),

\[
\| \hat{\phi}_{jk} - \phi_{jk} \| = O_p(kn^{-1/2}), \text{ where } O_p(\cdot) \text{ is uniform in } k = 1, \ldots, s_n.
\]

(c) Under conditions (A1), (A3), (B1)-(B4), for each \( j = 1, \ldots, d \),

\[
\hat{\phi}_{jk}(t) - \phi_{jk}(t) = n^{-1/2} \sum_{v:v \neq k} (w_{jk} - w_{jv})^{-1} \phi_{jv}(t) \int \mathbb{R}_j \phi_{jk} \phi_{jv} + \alpha_{jk}(t),
\]

where \( \| \alpha_{jk} \| = O_p(k^{\alpha+2}n^{-1}) \), \( \mathbb{R}_j = n^{1/2}(\hat{K}_j - K_j) \), and \( O_p(\cdot) \) is uniform in \( k = 1, \ldots, s_n \).

(d) Under conditions (A1), (B1)-(B4) and \( n^{-1/2}s_n = o_p(1) \), for any \( j = 1, \ldots, d \), we have \( c_1\lambda_n \leq \lambda_{jn} \leq c_2\lambda_n \) for some positive constants \( c_1, c_2 \).

The next lemma quantifies the asymptotic orders of several important types of expressions that will be encountered in the proofs of our lemmas and main theorems. For convenience, define
the following notations, \( l, l_1, l_2 = 1, \ldots, p_n, k, k_1, k_2 = 1, \ldots, s_n, j, j_1, j_2 = 1, \ldots, d, \)

\[
\begin{align*}
\theta^{(1)}_{jk} &= \sum_{i=1}^{n} (\hat{\xi}_{ijk} - \xi_{ijk})^2 w_{jk}^{-1}, \\
\theta^{(2)}_{j_1 j_2} &= n^{-1} \sum_{i=1}^{n} (\hat{\xi}_{ij_1 k_1} \hat{\xi}_{ijk} - \xi_{ijk} \xi_{ijk}) (w_{j_1 k_1} w_{j_2 k_2})^{-1/2}, \\
\theta^{(3)}_{k_1 k_2} &= n^{-1} \sum_{i=1}^{n} \{\xi_{ijk} \xi_{ijk} - E(\xi_{ijk} \xi_{jk_1 k_2})\} (w_{j_1 k_1} w_{j_2 k_2})^{-1/2}, \\
\theta^{(4)}_{j_1 j_2} &= \theta^{(2)}_{j_1 j_2} + \theta^{(3)}_{k_1 k_2}, \\
\theta^{(5)}_{jkl} &= n^{-1} \sum_{i=1}^{n} (\hat{\xi}_{ijk} - \xi_{ijk}) z_{ij} w_{jk}^{-1/2}, \\
\theta^{(6)}_{jkl} &= n^{-1} \sum_{i=1}^{n} \{\xi_{ijk} z_{ij} - E(\xi_{ijk} z_{ij})\} w_{jk}^{-1/2}, \\
\vartheta^{(7)}_{jkl} &= \theta^{(5)}_{jkl} + \theta^{(6)}_{jkl}, \\
\vartheta^{(8)}_{l_1 l_2} &= n^{-1} \sum_{i=1}^{n} \{z_{il_1} z_{il_2} - E(z_{il_1} z_{il_2})\}, \\
\gamma^{(1)}_{jl} &= \sum_{i=1}^{n} \sum_{k=0}^{s_n} \xi_{ijk} b_{jk0} z_{ij}, \\
\gamma^{(2)}_{jl} &= \sum_{i=1}^{n} \sum_{k=1}^{s_n} (\hat{\xi}_{ijk} - \xi_{ijk}) (\hat{b}_{jk} - b_{jk0}) z_{ij}, \\
\gamma^{(3)}_{jl} &= \sum_{i=1}^{n} \sum_{k=1}^{s_n} (\hat{\xi}_{ijk} - \xi_{ijk}) b_{jk0} z_{ij}, \\
\gamma^{(4)}_{jl} &= \sum_{i=1}^{n} \sum_{k=1}^{s_n} (\hat{\xi}_{ijk} - \xi_{ijk}) b_{jk0} z_{ij}, \\
\gamma^{(5)}_{jl} &= \sum_{k=1}^{s_n} b_{jk0} \int (\sum_{j=1}^{n} x_{ij} z_{ij} - n \Xi_{jl}) (\hat{\phi}_{jl} - \phi_{jl}), \\
\gamma^{(6)}_{jl} &= \sum_{k=1}^{s_n} b_{jk0} (\Xi_{jl}, \alpha_{jk}), \\
\gamma^{(7)}_{jl} &= n^{1/2} \sum_{k=1}^{s_n} \sum_{v \neq k} b_{jk0} (w_{jk} - w_{jv})^{-1} (\Xi_{jl}, \phi_{jv}) \int (\mathcal{R}_j - \mathcal{R}_{jv}) \phi_{jk} \phi_{jv}, \\
\gamma^{(8)}_{jl} &= n^{1/2} \sum_{k=1}^{s_n} \sum_{v \neq k} b_{jk0} (w_{jk} - w_{jv})^{-1} (\Xi_{jl}, \phi_{jv}) \int \mathcal{R}_{jv} \phi_{jk} \phi_{jv},
\end{align*}
\]

where \( \vartheta^{(m)}_{j} = (\vartheta^{(m)}_{j_1}, \ldots, \vartheta^{(m)}_{j_n})^\top \) denote corresponding vectors and \( \mathcal{R}_{j} = n^{1/2}(\tilde{K}_j - K_j) \), where \( \tilde{K}_j = n^{-1} \sum_{i=1}^{n} x_{ij} \otimes x_{ij} \).

**Lemma 2**  
(a) Under conditions (A1), (A3), (B1)—(B5), we have

\[
\begin{align*}
\theta^{(1)}_{jk} &= O_p(k^{a+2}), \\
\theta^{(2)}_{j_1 j_2} &= O_p(k_1^{a/2+1} n^{-1/2} + k_2^{a/2+1} n^{-1/2}), \\
\theta^{(3)}_{k_1 k_2} &= O_p(n^{-1/2}), \\
\theta^{(4)}_{j_1 j_2} &= O_p(k_1^{a/2+1} n^{-1/2} + k_2^{a/2+1} n^{-1/2}), \\
\theta^{(5)}_{jkl} &= O_p(k^{a/2+1} n^{-1/2}), \\
\theta^{(6)}_{jkl} &= O_p(n^{-1/2}), \\
\theta^{(7)}_{jkl} &= O_p(k^{a/2+1} n^{-1/2}), \\
\theta^{(8)}_{l_1 l_2} &= O_p(n^{-1/2}),
\end{align*}
\]
where the $O_p(\cdot)$ and $o_p(\cdot)$ terms are uniform for $k, k_1, k_2 = 1, \ldots, s_n$ and $l, l_1, l_2 = 1, \ldots, p_n$.

(b) Under conditions (A1)–(A7), (B1)–(B5), uniformly for $l = 1, \ldots, p_n$,

\[
\vartheta_{jl}(1) = \vartheta_{jl}(2) = \vartheta_{jl}(3) = \vartheta_{jl}(4) = \vartheta_{jl}(5) + \vartheta_{jl}(6) + \vartheta_{jl}(7) + \vartheta_{jl}(8).
\]

Lemma 3 characterizes the eigenvalues of the design matrices $\tilde{N}_1$, and Lemma 4 concerns the asymptotic order of $\Lambda_1 = P_{\tilde{N}_1}(Y_M - \tilde{N}_1\tilde{\eta}_{10})$, where $P_{\tilde{N}_1} = \tilde{N}_1(\tilde{N}_1^T\tilde{N}_1)^{-1}\tilde{N}_1^T$, $Y_M = (y_1, \ldots, y_n)^T$, and $\tilde{\eta}_{10}$ is the true parameter of $\tilde{\eta}_1$.

**Lemma 3** Under conditions (A1), (A3), (A5), (B1)–(B5), we have $|\lambda_{\min}(\tilde{N}_1^T\tilde{N}_1/n) - \lambda_{\min}(U_1)| = o_p(1)$,

| $\lambda_{\max}(\tilde{N}_1^T\tilde{N}_1/n) - \lambda_{\max}(U_1)$ | = o_p(1).

**Lemma 4** Under conditions (A1)–(A5), (B1)–(B5), we have $\|\Lambda_1\|^2 = O_p(r_n^2)$, where $r_n^2 = q_n + s_n$.

### 1.6.3 Proof of Lemmas

**Proof of Lemma 1.** We shall show Lemma 1 for any fixed $j = 1, \ldots, d$, we suppress the subscript $j$ in this proof for convenience. For part (a), recall that $W_{il} = x_i(t_{il}) + \varepsilon_{il}$, where the error $\varepsilon_{il}$ are independent of $x_i$. Thus one can factor the probability space $\Omega = \Omega_1 \times \Omega_2$, where $\Omega_1$ is for $\{x_i\}_{i=1,\ldots,n}$ and $\Omega_2$ for $\{\varepsilon_{il}\}_{l=1,\ldots,m_i}$. Given the data for a single subject, a specific realization $x_i$ corresponds to fixing a value $\omega_i \in \Omega_1$, that is $x_i(\cdot) = x_i(\cdot, \omega_i)$. We write $E_e$ for expectations with regard to the probability measure on $\Omega_2$, and $E_X$ with respect to $\Omega_1$. For each fixed $\omega_i$, the error $\{\varepsilon_{il}, i = 1, \ldots, n\}$ are independent and identically distributed for different $l$, with $E_e(\varepsilon_{il}) = 0$ and $E_e(\varepsilon_{il}^2) = \sigma_i^2$. Given condition (B3) together with a local linear smoother, one has $E(\|\hat{x}_i - x_i\|^2) = E_X[E_e\{\int(\hat{x}_i - x_i)^2\}]$. Using standard Taylor expansion argument, together with the dominant convergence theorem given $E(\|x_i^{(2)}\|^4) < \infty$,
it is easy to verify $E_X\left[E_\varepsilon\left\{\int (\hat{x}_i - x_i)^2\right\}\right] \leq C \int \left[E_X\left\{\left(x_i^{(2)}\right)^2\right\}\right] h^4 + (mh)^{-1}$ under (B3), yielding $E(\|\hat{x}_i - x_i\|^2) = o(n)$ by (B4). Similar arguments will also lead to $E_X \left[E_\varepsilon\left\{\int (\hat{x}_i - x_i)^4\right\}\right] \leq C \int \left[E_X\left\{\left(x_i^{(2)}\right)^4\right\}\right] h^8 + (mh)^{-2}$, thus $E\left\{\int (\hat{x}_i - x_i)^4\right\} = o(n^{-2})$. Regarding $\hat{K} = n^{-1} \sum_{i=1}^n \hat{x}_i \otimes \hat{x}_i$, notice that $\hat{K} - K = n^{-1} \sum_{i=1}^n \{(\hat{x}_i - x_i) \otimes x_i + (\hat{x}_i - x_i) \otimes (\hat{x}_i - x_i)\} + (n^{-1} \sum_{i=1}^n x_i \otimes x_i - K)$. It is easy to see that the first term on the right hand side is dominated by $n^{-1} \sum_{i=1}^n \|(\hat{x}_i - x_i) \otimes x_i\|$. Since $\{\hat{x}_i, x_i\}$ is independent of $\{\hat{x}_{i'}, x_{i'}\}$ for $i \neq i'$, we have $E\left\{n^{-1} \sum_{i=1}^n \|(\hat{x}_i - x_i) \otimes x_i\|^2\right\} = \{E\|(\hat{x}_i - x_i) \otimes x_i\|^4\} + n^{-1} \text{var}(\|(\hat{x}_i - x_i) \otimes x_i\|) \leq E(\|\hat{x}_i - x_i\|^2)(E\|x_i\|^4) + n^{-1} \{E(\|(\hat{x}_i - x_i)\|^4)(E\|x_i\|^4)\}^{1/2} = o(n^{-1})$. The second term $E(\|(n^{-1} \sum_{i=1}^n x_i \otimes x_i - K\|^2) = O(n^{-1})$ by Lemma 3.3 of Hall and Hosseini-Nasab [2009]. This leads to $E(\|\hat{K} - K\|^2) = O(n^{-1})$.

We next obtain the bound in (b) for $k = 1, \ldots, s_n$. First notice that $F_n$ can be implied by $\min_{k=1,\ldots,s_n}(w_k - w_{k+1}) \geq s_n^{-a-1} \geq C n^{-\tau/2}$ for some $C$ and any $\tau < 1$, which is fulfilled by $s_n^{-1} n^{-1/2} \rightarrow 0$ in (A3), leading to $P(F_n) \rightarrow 1$ as $n \rightarrow \infty$. It is easy to see that for $k = 1, \ldots, s_n$, $w_k - w_{k+1} \geq s_n^{-a-1} > 2\Delta$ as $n \rightarrow \infty$, that is $s_n \in J_n$ for large $n$. Now we will bound $\|\hat{\phi}_k - \phi_k\|^2$ for $k = 1, \ldots, s_n$. By (5.16) in Hall and Horowitz [2007], one has $\|\hat{\phi}_k - \phi_k\|^2 \leq 2\hat{u}_k^2$, where $\hat{u}_k^2 = \sum_{v \neq k} (\hat{w}_k - w_v)^{-2} \left\{\int (\hat{K} - K)\hat{\phi}_k\phi_v\right\}^2$. Also $P(F_n) \rightarrow 1$ implies that $(\hat{w}_k - w_v)^{-2} \leq C n^\tau$ with probability tending to 1. Then we have

\[
\hat{u}_k^2 \leq \sum_{v \neq k} (\hat{w}_k - w_v)^{-2} \left[2 \left\{\int (\hat{K} - K)\phi_j\phi_v\right\}^2 + 2 \left\{\int (\hat{K} - K)(\hat{\phi}_k - \phi_k)\phi_v\right\}^2\right] \\
\leq 2 \sum_{v \neq k} (\hat{w}_k - w_v)^{-2} \left\{\int (\hat{K} - K)\phi_k\phi_v\right\}^2 + 2Cn^\tau \sum_{v \neq k} \left\{\int (\hat{K} - K)(\hat{\phi}_k - \phi_k)\phi_v\right\}^2 \\
\leq 4 \sum_{v \neq k} (w_k - w_v)^{-2} \left\{\int (\hat{K} - K)\phi_k\phi_v\right\}^2 + 2Cn^\tau \hat{\Delta}^2 \|\hat{\phi}_k - \phi_k\|^2.
\]
Plugging this into $\|\hat{\phi}_k - \phi_k\|^2 \leq 2\hat{u}_k^2$, one has

$$(1 - 2Cn^\tau \hat{\Delta}^2)\|\hat{\phi}_k - \phi_k\|^2 \leq 4 \sum_{v:v \neq k} (w_k - w_v)^{-2}\left\{\int (\hat{K} - K)\phi_k \phi_v \right\}^2.$$ 

As $\hat{\Delta} = O_p(n^{-1/2})$ and $\tau < 1$, we have $n^\tau \hat{\Delta}^2 = o_p(1)$, and $\|\hat{\phi}_k - \phi_k\|^2 \leq C \sum_{v:v \neq k} (w_k - w_v)^{-2}\left\{\int (\hat{K} - K)\phi_k \phi_v \right\}^2$. Given the results in (a), by analogy to (5.22) in Hall and Horowitz [2007], $nE[\sum_{v:v \neq k} (w_k - w_v)^{-2}\left\{\int (\hat{K} - K)\phi_k \phi_v \right\}^2] = O(k^2)$ still holds, the result follows by Chebyshev’s inequality.

To obtain (c), using Lemma 5.1 of Hall and Horowitz [2007] with $\psi_k = \hat{\phi}_k$, $\lambda_k = \hat{w}_k$ and $L = \hat{K}$, since $\inf_{\{v:v \neq k\}} |\hat{w}_k - w_v| > 0$, we have

$$\hat{\phi}_k - \phi_k = \sum_{v:v \neq k} (\hat{w}_k - w_v)^{-1} \phi_v \int (\hat{K} - K)\hat{\phi}_k \phi_v + \phi_k \int (\hat{\phi}_k - \phi_k)\phi_k$$

$$= \sum_{v:v \neq k} (w_k - w_v)^{-1} \phi_v \int (\hat{K} - K)\phi_k \phi_v + \phi_k \int (\hat{\phi}_k - \phi_k)\phi_k$$

$$+ \sum_{v:v \neq k} (w_k - w_v)^{-1} \phi_v \int (\hat{K} - K)(\hat{\phi}_k - \phi_k)\phi_v$$

$$- \sum_{v:v \neq k} (\hat{w}_k - w_v)(w_k - w_v) (\hat{w}_k - w_v)^{-1} \phi_v \int (\hat{K} - K)\hat{\phi}_k \phi_v.$$ 

Denote the last three terms by $\alpha_k \equiv \alpha_{1k} + \alpha_{2k} + \alpha_{3k}$. For $\alpha_{1k}$, one has

$$\|\hat{\phi}_k - \phi_k\|^2 = 2 - 2 \int \hat{\phi}_k \phi_k = 2(\int \hat{\phi}_k^2 - \int \hat{\phi}_k \phi_k) = -2 \int (\hat{\phi}_k - \phi_k)\phi_k,$$

that is $\alpha_{1k} = -\|\hat{\phi}_k - \phi_k\|^2/2$. From part (b), one has $\|\hat{\phi}_k - \phi_k\| = O_p(kn^{-1/2})$ uniformly in $k = 1, \ldots, s_n$, then $\|\alpha_{1k}\| = O_p(k^2n^{-1})$. To bound $\alpha_{2k}$, noticing $\|\sum_v c_v \phi_v\|^2 = \sum_v c_v^2$ due to orthonormal $\{\phi_k\}$,

$$\|\alpha_{2k}\| \leq \left[\sum_{v:v \neq k} (w_k - w_v)^{-2}\left\{\int (\hat{K} - K)(\hat{\phi}_k - \phi_k)\phi_v \right\}^2\right]^{1/2}$$

$$\leq \delta_k^{-1} \hat{\Delta} \|\hat{\phi}_k - \phi_k\| = O_p(k^{a+2}n^{-1}).$$
For $\alpha_{3k}$, noticing that $\sup_{k=1,\ldots,\infty} |\hat{w}_k - w_k| = O_p(\hat{\Delta})$, we can bound $\|\alpha_{3k}\|$ as follows,

$$2|\hat{w}_k - w_k| \left[ \sum_{v:v \neq j} (w_k - w_v)^{-4} \left\{ \int (\hat{K} - K) \hat{\phi}_k \phi_v \right\}^2 \right]^{1/2} \leq C \hat{\Delta} \left[ \sum_{v:v \neq k} (w_k - w_v)^{-4} \left\{ \int (\hat{K} - K) \hat{\phi}_k \phi_v \right\}^2 \right. + \left. \sum_{v:v \neq k} (w_k - w_v)^{-4} \left\{ \int (\hat{K} - K) (\hat{\phi}_k - \phi_k) \phi_v \right\}^2 \right]^{1/2} = C \hat{\Delta} (A_1 + A_2)^{1/2},$$

where $A_1 = \sum_{v:v \neq k} (w_k - w_v)^{-4} \left\{ \int (\hat{K} - K) \hat{\phi}_k \phi_v \right\}^2$ and $A_2 = \sum_{v:v \neq k} (w_k - w_v)^{-4} \left\{ \int (\hat{K} - K)(\hat{\phi}_k - \phi_k) \phi_v \right\}^2$. By the derivations of part (b), we have

$$A_1 \leq \delta_k^{-2} \sum_{v:v \neq k} (w_k - w_v)^{-2} \left\{ \int (\hat{K} - K) \phi_k \phi_v \right\}^2 = O_p(k^2 \delta_k^{-2} n^{-1}),$$

$$A_2 \leq \delta_k^{-4} \hat{\Delta}^2 \|\hat{\phi}_k - \phi_k\|^2 = O_p(\delta_k^{-4} j^2 \hat{\Delta}^4).$$

Notice that $\delta_k^{-2} = O\{k^{2(a+1)}\} = o(n)$ by (A3), which indicates $A_2 = o_p(k^2 \delta_k^{-2} n^{-1})$. Thus $\|\alpha_{3k}\| = O_p(k \delta_k^{-1} n^{-1})$. It leads to $\|\alpha_k\| = O_p(k^{a+2} n^{-1})$.

For part (d), since $\sum_{k=1}^{s_n} w_k < \infty$ and $\sup_{k=1,\ldots,s_n} |\hat{w}_k - w_k| \leq \||\hat{K} - K|| = O_p(n^{-1/2})$ by Theorem 1 of Hall and Hosseini-Nasab [2006] and part (a), we can see that $(\sum_{k=1}^{s_n} \hat{w}_k)^{1/2} = O_p(1)$ given $n^{-1/2} s_n = o(1)$, which completes the proof.
Proof of Lemma 2. For \( \theta_{jk}^{(1)} \) and any fixed \( j \), we have

\[
\theta_{jk}^{(1)} = \sum_{i=1}^{n} (\hat{\xi}_{ijk} - \xi_{ijk})^2 w_{jk}^{-1} = \sum_{i=1}^{n} (\hat{\xi}_{ijk} - \xi_{ijk} + \tilde{\xi}_{ijk} - \xi_{ijk})^2 w_{jk}^{-1}
\]

\[
\leq 2 \sum_{i=1}^{n} (\hat{\xi}_{ijk} - \xi_{ijk})^2 w_{jk}^{-1} + 2 \sum_{i=1}^{n} (\xi_{ijk} - \xi_{ijk})^2 w_{jk}^{-1}
\]

\[
= 2 \sum_{i=1}^{n} \left\{ \int \hat{x}_{ij} (\hat{\phi}_{jk} - \phi_{jk}) \right\}^2 w_{jk}^{-1} + 2 \sum_{i=1}^{n} \left\{ \int (\hat{x}_{ij} - x_{ij}) \hat{\phi}_{jk} \right\}^2 w_{jk}^{-1}
\]

\[
\leq 2 \sum_{i=1}^{n} (\|\hat{x}_{ij}\|^2 \|\hat{\phi}_{jk} - \phi_{jk}\|^2 + \|\phi_{jk}\|^2 \|\hat{x}_{ij} - x_{ij}\|^2) w_{jk}^{-1}.
\]

Given Lemma 1 and (B1), we know

\[
E(\|\hat{x}_{ij}\|^2) \leq 2E(\|\hat{x}_{ij} - x_{ij}\|^2) + 2E(\|x_{ij}\|^2) = O(1),
\]

hence, \( \|\hat{x}_{ij}\|^2 = O_p(1) \). From Lemma 1, we have

\[
\theta_{jk}^{(1)} = \sum_{i=1}^{n} \left\{ O_p(1)O_p(k^2 n^{-1}) + o_p(n^{-1}) \right\} k^a = O_p(k^{a+2}),
\]

uniformly for \( k = 1, \ldots, s_n \).

For \( \theta_{k_1k_2}^{(2)} \), it is obvious that

\[
|\theta_{k_1k_2}^{(2)}| = \left| n^{-1} \sum_{i=1}^{n} \left( \hat{\xi}_{ij1k_1} \hat{\xi}_{ij2k_2} - \xi_{ij1k_1} \xi_{ij2k_2} \right) (w_{j1k_1} w_{j2k_2})^{-1/2} \right|
\]

\[
= n^{-1} \sum_{i=1}^{n} \left( \hat{\xi}_{ij1k_1} (\hat{\xi}_{ij2k_2} - \xi_{ij2k_2}) \right) (w_{j1k_1} w_{j2k_2})^{-1/2} + \sum_{i=1}^{n} \xi_{ij2k_2} (\hat{\xi}_{ij1k_1} - \xi_{ij1k_1}) (w_{j1k_1} w_{j2k_2})^{-1/2}
\]

\[
\leq n^{-1} \left( \sum_{i=1}^{n} \hat{\xi}_{ij1k_1} w_{j1k_1}^{-1} \right)^{1/2} \left( \sum_{i=1}^{n} (\hat{\xi}_{ij2k_2} - \xi_{ij2k_2})^2 w_{j2k_2}^{-1} \right)^{1/2} +
\]

\[
n^{-1} \left( \sum_{i=1}^{n} \xi_{ij2k_2} w_{j2k_2}^{-1} \right)^{1/2} \left( \sum_{i=1}^{n} (\hat{\xi}_{ij1k_1} - \xi_{ij1k_1})^2 w_{j1k_1}^{-1} \right)^{1/2}.
\]

Since \( E(\sum_{i=1}^{n} \xi_{ij2k_2} w_{j2k_2}^{-1}) = n \) for any \( k_2 = 1, \ldots, s_n \), we have \( \sum_{i=1}^{n} \xi_{ij2k_2} w_{j2k_2}^{-1} = O_p(n) \),
uniformly for \( k_2 = 1, \ldots, s_n \). Moreover,

\[
\sum_{i=1}^{n} \xi_{ij}^2 w_{j1}^{-1} w_{j2}^{-1} \leq 2 \sum_{i=1}^{n} (\xi_{ij1} - \xi_{ij2})^2 w_{j1}^{-1} + 2 \sum_{i=1}^{n} \xi_{ij2}^2 w_{j2}^{-1} = O_p(k_1^{a/2} + n) = O_p(n),
\]

uniformly for \( k_1 = 1, \ldots, s_n \). In conclusion, we have

\[
|\tilde{\theta}_{k_1,k_2}^{(2)}| = n^{-1}O_p(n^{1/2})O_p(k_2^{a/2+1}) + n^{-1}O_p(n^{1/2})O_p(k_1^{a/2+1}) = O_p\left(k_1^{a/2+1}n^{-1/2} + k_2^{a/2+1}n^{-1/2}\right),
\]

uniformly for \( k_1, k_2 = 1, \ldots, s_n \).

For \( \theta_{k_1,k_2}^{(3)} \),

\[
E(\theta_{k_1,k_2}^{(3)})^2 \leq n^{-1}\left\{ E(\xi_{ij1}^4 w_{j1}^{-2}) E(\xi_{ij2}^4 w_{j2}^{-2}) \right\}^{1/2} \leq c_1n^{-1},
\]

uniformly for \( k_1, k_2 = 1, \ldots, s_n \) by (B1). It follows that \( \theta_{k_1,k_2}^{(3)} = O_p(n^{-1/2}) \), uniformly for \( k_1, k_2 = 1, \ldots, s_n \). Moreover, it is trivial that

\[
\theta_{k_1,k_2}^{(4)} = \theta_{k_1,k_2}^{(2)} + \theta_{k_1,k_2}^{(3)} = O_p(n^{-1/2}k_1^{a/2+1} + n^{-1/2}k_2^{a/2+1}),
\]

uniformly for \( k_1, k_2 = 1, \ldots, s_n \).

Based on (B5), we conclude that \( z_{ij}^4 = O_p(1) \), uniformly for \( l = 1, \ldots, p_n \). For \( \theta_{jkl}^{(5)} \), we have

\[
|\theta_{jkl}^{(5)}| \leq n^{-1}\left\{ \sum_{i=1}^{n} (\xi_{ijk} - \xi_{ijkl})^2 w_{jk}^{-1} \right\}^{1/2} \left( \sum_{i=1}^{n} z_{ijl}^2 \right)^{1/2} = n^{-1}O_p(k_1^{a/2+1})O_p(n^{1/2}) = O_p\left(k_1^{a/2+1}n^{-1/2}\right),
\]

uniformly for \( k = 1, \ldots, s_n, l = 1, \ldots, p_n \). Furthermore,

\[
E(\theta_{jkl}^{(6)})^2 \leq n^{-1}\left\{ E(\xi_{ijlk}^4 w_{jk}^{-2}) E(z_{ijkl}^4) \right\}^{1/2} = O\left(n^{-1}\right),
\]

uniformly for \( k = 1, \ldots, s_n, l = 1, \ldots, p_n \). Thus \( \theta_{jkl}^{(6)} = O_p(n^{-1/2}) \), uniformly for \( k = 1, \ldots, s_n, l = 1, \ldots, p_n \). Then it follows that \( \theta_{jkl}^{(7)} = \theta_{jkl}^{(5)} + \theta_{jkl}^{(6)} = O_p(k_1^{a/2+1}n^{-1/2}) \), uniformly for \( k = 1, \ldots, s_n, l = 1, \ldots, p_n \). For \( \theta_{l1l2}^{(8)} \), we have

\[
E(\theta_{l1l2}^{(8)})^2 \leq n^{-1}\left\{ E(z_{il1}^4 E(z_{l1l2}^4) \right\}^{1/2} =
\]
Since it follows that we have $\|O_\vartheta H\|$ uniformly for $C$, $l_2 = 1, \ldots, p_n$, and this entails that $\theta_{l_1l_2}^{(8)} = O_p(n^{-1/2})$, uniformly for $l_1 = 1, \ldots, p_n, l_2 = 1, \ldots, p_n$.

For $\vartheta_{j1}^{(1)}$, given (A4), we have

$$E|\vartheta_{j1}^{(1)}| \leq \sum_{i=1}^{n} \sum_{k=s_{n}+1}^{\infty} |b_{jk0}| E|\xi_{ij}z_{ij}| \leq \sum_{i=1}^{n} \sum_{k=s_{n}+1}^{\infty} |b_{jk0}| \{E(\xi_{ij}^2)E(z_{ij}^2)\}^{1/2} \leq c_1 \sum_{i=1}^{n} \sum_{k=s_{n}+1}^{\infty} k^{-b}k^{-1/2} = O(n s_n^{-b+1/2}) = o(n^{1/2}).$$

Hence $\vartheta_{j1}^{(1)} = o_p(n^{1/2})$, uniformly for $l = 1, \ldots, p_n$. For $\vartheta_{j1}^{(2)}$, we have

$$\vartheta_{j1}^{(2)} = \sum_{i=1}^{n} \sum_{k=1}^{s_n} \{\int \hat{x}_{ij}(\phi_{jk} - \phi_{jk}) + \int (\hat{x}_{ij} - x_{ij})\phi_{jk}\} (\hat{b}_{jk} - b_{jk0}) z_{ij}$$

$$= \int \{\sum_{i=1}^{n} \hat{x}_{ij} z_{ij}\} \{\sum_{k=1}^{s_n} (\phi_{jk} - \phi_{jk})(\hat{b}_{jk} - b_{jk0})\}$$

$$+ \int \{\sum_{i=1}^{n} (\hat{x}_{ij} - x_{ij})z_{ij}\} \{\sum_{k=1}^{s_n} \phi_{jk} (\hat{b}_{jk} - b_{jk0})\}.$$

It follows that

$$(\vartheta_{j1}^{(2)})^2 \leq 2 \int \sum_{i=1}^{n} \hat{x}_{ij} z_{ij}^2 \sum_{k=1}^{s_n} (\phi_{jk} - \phi_{jk})(\hat{b}_{jk} - b_{jk0})^2$$

$$+ 2 \int \sum_{i=1}^{n} (\hat{x}_{ij} - x_{ij})z_{ij}^2 \sum_{k=1}^{s_n} \phi_{jk} (\hat{b}_{jk} - b_{jk0})^2.$$

Since

$$E\|\sum_{i=1}^{n} \hat{x}_{ij} z_{ij}\|^2 \leq E\|\sum_{i=1}^{n} (\hat{x}_{ij} z_{ij})\|^2 \leq n^2\{E(\|\hat{x}_{1j}\| z_{1j})\}^2 + n\{E(\|\hat{x}_{1j}\|^4)E(z_{1j}^4)\}^{1/2} = O(n^2),$$

we have $\sum_{i=1}^{n} \hat{x}_{ij} z_{ij}^2 = O_p(n^2)$. Similarly,

$$E\|\sum_{i=1}^{n} (\hat{x}_{ij} - x_{ij})z_{ij}\|^2 \leq n^2(E\|\hat{x}_{ij} - x_{ij}\| z_{ij})^2 + n\{E(\|\hat{x}_{ij} - x_{ij}\|^4)E(z_{ij}^4)\}^{1/2} = o(n),$$
lead to \( \| \sum_{i=1}^{n} (\hat{x}_{ij} - x_{ij}) z_{ij} \|^2 = o_p(n) \). Moreover,

\[
\begin{align*}
\| \sum_{k=1}^{s_n} (\hat{\phi}_{jk} - \phi_{jk})(\hat{b}_{jk} - b_{jk0}) \|^2 & \leq \| \hat{b}_{j0}^{(1)} - \hat{b}_{j0}^{(1)} \|^2 \sum_{k=1}^{s_n} \| \hat{\phi}_{jk} - \phi_{jk} \|^2 w_{jk}^{-1} \\
& = O_p(r_n^2/n) O_p \{ \sum_{k=1}^{s_n} (k^2/n) k^a \} \\
& = O_p(r_n^2 s_n^{a+3} / n^2),
\end{align*}
\]

and

\[
\begin{align*}
\| \sum_{k=1}^{s_n} \phi_{jk}(\hat{b}_{jk} - b_{jk0}) \|^2 & \leq \left( \sum_{k=1}^{s_n} |b_{jk} - b_{jk0}| \right)^2 \leq \| \hat{b}_{j0}^{(1)} - \hat{b}_{j0}^{(1)} \|^2 \sum_{k=1}^{s_n} w_{jk}^{-1} = O_p(r_n^2 s_n^{a+1} / n),
\end{align*}
\]

by Theorem 1. In summary, we get \((\hat{\psi}_{jl}^{(2)})^2 = o_p(r_n^2 s_n^{a+3}) = o_p(q_n s_n^{a+3} + s_n^{a+4}),\) uniformly for \(l = 1, \ldots, p_n\). Under conditions (A3) and (A5), \(\hat{\psi}_{jl}^{(2)} = o_p(n^{1/2}),\) uniformly for \(l = 1, \ldots, p_n\).

For \(\hat{\psi}_{jl}^{(3)}\), we have

\[
(\hat{\psi}_{jl}^{(3)})^2 = \left\| \left\{ \sum_{i=1}^{n} (\hat{x}_{ij} - x_{ij}) z_{ij} \right\} \left( \sum_{k=1}^{s_n} \phi_{jk} b_{jk0} \right) \right\|^2 \leq \left\| \sum_{i=1}^{n} (\hat{x}_{ij} - x_{ij}) z_{ij} \right\|^2 \left\| \sum_{k=1}^{s_n} \phi_{jk} b_{jk0} \right\|^2
\]

\[
\leq \left( \sum_{k=1}^{s_n} \left| b_{jk0} \right| \right)^2 o_p(n) = o_p(n),
\]

hence \(\hat{\psi}_{jl}^{(3)} = o_p(n^{1/2})\) uniformly for \(l = 1, \ldots, p_n\).

Next, we show that \(\hat{\psi}_{jl}^{(4)} = \hat{\psi}_{jl}^{(5)} + \hat{\psi}_{jl}^{(6)} + \hat{\psi}_{jl}^{(7)} + \hat{\psi}_{jl}^{(8)}\). Recall that \(\Xi_j = (\Xi_{j1}, \ldots, \Xi_{jn})^\top\) with \(E\{x_{ij}(t)z_{ij}\} = \Xi_{jl}(t)\). By lemma 1, we have the expression

\[
\hat{\psi}_{jk}(t) - \phi_{jk}(t) = n^{-1/2} \sum_{v:v \neq k} (w_{jk} - w_{jv})^{-1} \phi_{jv}(t) \int \mathcal{R}_j \phi_{jk} \phi_{jv} + \alpha_{jk}(t),
\]

where \(||\alpha_{jk}|| = O_p(k^{a+2} n^{-1}), \mathcal{R}_j = n^{1/2}(\hat{K}_j - K_j),\) and \(O_p(\cdot)\) is uniform in \(k = 1, \ldots, s_n\).

Since

\[
\hat{\psi}_{jl}^{(4)} = \sum_{i=1}^{n} \sum_{k=1}^{s_n} (\hat{\xi}_{ijk} - \xi_{ijk}) b_{jk0} z_{ij} = \int \left\{ \sum_{i=1}^{n} x_{ij} z_{ij} \right\} \left\{ \sum_{k=1}^{s_n} (\hat{\phi}_{jk} - \phi_{jk}) b_{jk0} \right\},
\]
substitute the expression for $\hat{\phi}_{jk}(t) - \phi_{jk}(t)$ into the above equation, we immediately get $\vartheta_{jl}^{(4)} = \vartheta_{jl}^{(5)} + \vartheta_{jl}^{(6)} + \vartheta_{jl}^{(7)} + \vartheta_{jl}^{(8)}$.

For $\vartheta_{jl}^{(5)}$, it is obvious that

$$\left(\vartheta_{jl}^{(5)}\right)^2 \leq \left\| \sum_{k=1}^{s_n} (\hat{\phi}_{jk} - \phi_{jk}) b_{jk0} \right\|^2 \left\| \sum_{i=1}^{n} (x_{ij} z_{ij} - \Xi_{jl}) \right\|^2,$$

where

$$E\left\| \sum_{i=1}^{n} (x_{ij} z_{ij} - \Xi_{jl}) \right\|^2 = \int E\left\{ \sum_{i=1}^{n} (x_{ij} z_{ij} - \Xi_{jl}) \right\}^2 = n \int \text{var}(x_{ij} z_{ij}) \leq n \int \{E(x_{ij}^4) E(z_{ij}^4)\}^{1/2} = O(n),$$

and

$$\left\| \sum_{k=1}^{s_n} (\hat{\phi}_{jk} - \phi_{jk}) b_{jk0} \right\| \leq \sum_{k=1}^{s_n} |b_{jk0}||\hat{\phi}_{jk} - \phi_{jk}| = O_p\left(\sum_{k=1}^{s_n} k^{-b} k^{n-1/2}\right) = O_p(n^{-1/2}),$$

thus $\left(\vartheta_{jl}^{(5)}\right)^2 = O_p(1) = o_p(n)$, and it follows that $\vartheta_{jl}^{(5)} = o_p(n^{1/2})$ uniformly for $l = 1, \ldots, p_n$.

For $\vartheta_{jl}^{(6)}$, we have

$$\left(\vartheta_{jl}^{(6)}\right)^2 \leq n^2 \int (E x_{ij} z_{ij})^2 \left\| \sum_{k=1}^{s_n} b_{jk0} \alpha_{jk} \right\|^2 \leq n^2 \int \{E(x_{ij}^4) E(z_{ij}^4)\}^{1/2} \left\| \sum_{k=1}^{s_n} |b_{jk0}| \alpha_{jk} \right\|^2$$

$$= n^2 O_p\left\{ \left(\sum_{k=1}^{s_n} k^{-b} k^{a+2} n^{-1}\right)^2 \right\} = o_p(n),$$

hence $\vartheta_{jl}^{(6)} = o_p(n^{1/2})$ uniformly for $l = 1, \ldots, p_n$.

From the proof of lemma 1, it is easy to see that $\mathcal{R}_j - \mathcal{R}_j^* = n^{1/2}(\hat{K}_j - \tilde{K}_j) = o_p(1)$, which follows that $\vartheta_{jl}^{(7)} = o_p(n^{1/2})$, uniformly for $l = 1, \ldots, p_n$.

Proof of Lemma 3. First, it is obvious that $|\lambda_{\min}(\hat{N}_1^T \hat{N}_1/n) - \lambda_{\min}(U_1)| \leq ||\hat{N}_1^T \hat{N}_1/n - U_1||_1$,
where \(||.||_1\) is the \(L_1\) norm for matrix. Since

\[||\tilde{N}_1^T \tilde{N}_1/n - U_1||_1 \leq O_p\left(\sum_{k_1=1}^{s_n} |\theta_{k_1}^{(4)}| + \sum_{l=1}^{q_n} |\theta_{l \in s_n}^{(7)}| + \sum_{k_1=1}^{s_n} |\theta_{k_1 q_n}^{(7)}| + \sum_{l_1=1}^{q_n} |\theta_{l_1 q_n}^{(8)}|\right)\]

by Lemma 2 (a), hence we have

\[|\lambda_{\min}(\tilde{N}_1^T \tilde{N}_1/n) - \lambda_{\min}(U_1)| = O_p(s_n^{a/2+2} n^{-1/2} + q_n s_n^{a/2+1} n^{-1/2}).\]

Under conditions (A3) and (A5), it is obvious that

\[|\lambda_{\min}(\tilde{N}_1^T \tilde{N}_1/n) - \lambda_{\min}(U_1)| = o_p(1).\]

Similarly, \(\lambda_{\max}(\tilde{N}_1^T \tilde{N}_1/n) - \lambda_{\max}(U_1)\) is invertible, hence \(P_{\tilde{N}_1}\) exists.

**Proof of Lemma 4.** By Lemma 3 and (B5), we know that \(\tilde{N}_1^T \tilde{N}_1\) is invertible, hence \(P_{\tilde{N}_1}\) exists.

For \(A_1\), we have

\[A_1 = P_{\tilde{N}_1}(Y_M - \tilde{N}_1 \tilde{\eta}_{10}) = P_{\tilde{N}_1}\{Y_M - \tilde{N}_1 \tilde{\eta}_{10} - \nu + (\tilde{N}_1 - \tilde{N}_1) \tilde{\eta}_{10}\}\]

\[= P_{\tilde{N}_1}\{\epsilon + \nu + (\tilde{N}_1 - \tilde{N}_1) \tilde{\eta}_{10}\},\]

where \(\epsilon = Y_M - \tilde{N}_1 \tilde{\eta}_{10} - \nu = (\epsilon_1, ..., \epsilon_n)^T\), \(\nu = (\nu_1, ..., \nu_n)^T\) with \(\nu_i = \sum_{j=1}^g \sum_{k=s_n+1}^{\infty} \xi_{ijk} b_{jk} \).

For \(P_{\tilde{N}_1} \epsilon\), we have

\[E\|P_{\tilde{N}_1} \epsilon\|^2 = E(\epsilon^T P_{\tilde{N}_1} \epsilon) = E\{E(\epsilon^T P_{\tilde{N}_1} \epsilon | \epsilon)\} = E[\text{tr}(P_{\tilde{N}_1} E(\epsilon \epsilon^T))]\]

\[= \sigma^2 \text{tr}(P_{\tilde{N}_1}) = \sigma^2 (q_n + g s_n) = O(q_n + s_n),\]

hence \(\|P_{\tilde{N}_1} \epsilon\|^2 = O_p(q_n + s_n).\)
For $P_{\tilde{N}_1}(\tilde{N}_1 - \tilde{N}_1)\tilde{\eta}_{10}$, we have

$$
\|P_{\tilde{N}_1}(\tilde{N}_1 - \tilde{N}_1)\tilde{\eta}_{10}\|^2 \leq \|(\tilde{N}_1 - \tilde{N}_1)\tilde{\eta}_{10}\|^2 \leq O\left[\sum_{i=1}^{n} \sum_{j=1}^{g} \sum_{k=1}^{s_n} (\tilde{\xi}_{ijk} - \xi_{ijk})b_{jk0}\right]^2
$$

\[
\leq O\left[2g \sum_{i=1}^{n} \sum_{j=1}^{g} \sum_{k=1}^{s_n} (\tilde{\xi}_{ijk} - \xi_{ijk})b_{jk0}^2 + 2g \sum_{i=1}^{n} \sum_{j=1}^{g} \sum_{k=1}^{s_n} (\tilde{\xi}_{ijk} - \xi_{ijk})b_{jk0}^2\right]
\leq O\left[\sum_{j=1}^{g} \sum_{i=1}^{n} \|x_{ij}\|^2 O_p\left(\sum_{k=1}^{s_n} k^{-b} kn^{-1/2}\right)\right] + O_p\left[\sum_{j=1}^{g} \sum_{i=1}^{n} \|\hat{x}_{ij} - x_{ij}\|^2 \sum_{k=1}^{s_n} k^{-b}\right]
\]

\[= O_p(1),\]

since $b > 2$ implies that $\sum_{j=1}^{g} \sum_{i=1}^{n} \|x_{ij}\|^2 O_p\left(\sum_{k=1}^{s_n} k^{-b} kn^{-1/2}\right) = O_p(1)$, and Lemma 1 entails that $\sum_{j=1}^{g} \sum_{i=1}^{n} \|\hat{x}_{ij} - x_{ij}\|^2 \sum_{k=1}^{s_n} k^{-b} = O_p(1)$. It then follows that $\|P_{\tilde{N}_1}(\tilde{N}_1 - \tilde{N}_1)\tilde{\eta}_{10}\|^2 = O_p(1)$.

For $P_{\tilde{N}_1}\nu$, it is obvious that $\|P_{\tilde{N}_1}\nu\|^2 \leq \|\nu\|^2$. For $i = 1, \ldots, n$, we have

$$
E(\nu_i)^2 = O\left(\sum_{j=1}^{g} \sum_{k=s_n+1}^{\infty} \text{var} (\sum_{k=s_n+1}^{\infty} \xi_{ijk} b_{jk0})\right) = O\left(\sum_{j=1}^{g} \sum_{k=s_n+1}^{\infty} b_{jk0}^2 w_{jk}\right)
$$

\[= O\left(\sum_{j=1}^{g} \sum_{k=s_n+1}^{\infty} k^{-2b} k_{-1}\right) = O(s_n^{-2b}).\]

It follows that $\|P_{\tilde{N}_1}\nu\|^2 = O_p(ns_n^{-2b})$. In summary,

$$
\|\Lambda_1\|^2 \leq O(\|P_{\tilde{N}_1}\epsilon\|^2 + \|P_{\tilde{N}_1}(\tilde{N}_1 - \tilde{N}_1)\tilde{\eta}_{10}\|^2 + \|P_{\tilde{N}_1}\nu\|^2)
\]

\[= O_p(q_n + s_n + 1 + ns_n^{-2b} = O_p(q_n + s_n) = O_p(r_n^2),\]

where $r_n^2 = q_n + s_n$. This completes the proof.

### 1.6.4 Proofs of Main Theorems

**Proof of Theorem 1.** First, we constrain $Q_n(\tilde{\eta})$ on the subspace, where the true zero parameters are set as 0, that is $\{\tilde{\eta} \in R^{ds_n+p_n} : \tilde{b}_{jk}^{(1)} = 0, k = g + 1, \ldots, d, \gamma^{(2)} = 0\}$, and prove consistency
in the \((g s_n + q_n)\)-dimensional space. Define the constrained penalized function

\[
\tilde{Q}_n(\tilde{\eta}_1) = \sum_{i=1}^{n} \{y_i - \sum_{j=1}^{g} \sum_{k=1}^{s_n} (\hat{\xi}_{ijk} w_j^{-1/2} \tilde{b}_{jk} - z_i(1)^T \gamma(1)) \}^2 + 2n \sum_{l=1}^{q_n} J_{\lambda_n}(|\gamma_l|) + 2n \sum_{j=1}^{g} J_{\lambda_{jn}}(||b_j^{(1)}||),
\]

where \(\tilde{\eta}_1 = (\tilde{b}_1^{(1)^T}, \ldots, \tilde{b}_g^{(1)^T}, \gamma(1)^T)^T\) and \(z_i(1) = (z_{i1}, \ldots, z_{iq_n})^T\). We now show there exists a local minimizer \(\tilde{\eta}_1\) such that \(\|\tilde{\eta}_1 - \tilde{\eta}_{10}\| = O_p(r_n n^{-1/2})\) with \(r_n = (q_n + r_n)^{1/2}\). Let \(\alpha_n = r_n n^{-1/2}\), and

\[
\begin{align*}
    \tilde{Q}_n(\tilde{\eta}_1 + \alpha_n u) &\geq \|N_1 \alpha_n u\|^2 - 2 \Lambda_1^T N_1 \alpha_n u + 2n \sum_{l=1}^{q_n} \{J_{\lambda_n}(|\gamma_{10} + \alpha_n u_l^\gamma|) - J_{\lambda_n}(|\gamma_{10}|) \} \\
    &\quad + \sum_{j=1}^{g} \{J_{\lambda_{jn}}(||b_j^{(1)}|| + \alpha_n A_j^{-1} u_j^{(1)}||) - J_{\lambda_{jn}}(||b_j^{(1)}||) \} \\
    &\geq n \lambda_{min} (N_1^T N_1 / n) \alpha_n^2 || u ||^2 - 2n^{1/2} \Lambda_1 ||2 \lambda_{max} (N_1^T N_1 / n) \alpha_n || u || \\
    &\quad + 2n \sum_{l=1}^{q_n} \{J'_{\lambda_n}(|\gamma_{10}|) \text{sgn}(\gamma_{10}) \alpha_n u_l^\gamma + J''_{\lambda_n}(|\gamma_{10}|) \alpha_n^2 (u_l^\gamma)^2 (1 + o(1)) \} \\
    &\quad + \sum_{j=1}^{g} \{J'_{\lambda_{jn}}(||b_j^{(1)}||) \alpha_n || A_j^{-1} u_j^{(1)}|| + J''_{\lambda_{jn}}(||b_j^{(1)}||) \alpha_n^2 || A_j^{-1} u_j^{(1)}||^2 (1 + o(1)) \}.
\end{align*}
\]

The Taylor expansion in the second inequality holds because \(\alpha_n w_{j,s_n}^{-1/2} \leq r_n s_n^{a/2} n^{-1/2} = o(1)\) by condition (A5). As for the smoothly clipped absolute deviation penalty, we have

\[
J'_{\lambda_n}(|\gamma_{10}|) = J''_{\lambda_n}(|\gamma_{10}|) = 0 \text{ for all } l = 1, \ldots, q_n \text{ under condition (A7)}, \text{ and } J'_{\lambda_{jn}}(||b_j^{(1)}||) = J''_{\lambda_{jn}}(||b_j^{(1)}||) = 0 \text{ for all } j = 1, \ldots, g \text{ since } ||b_j^{(1)}|| \geq C_1 \text{ for some constant } C_1. \text{ Thus, we can}
see that

\[
\hat{Q}_n(\tilde{\eta}_{10} + \alpha_n u) - \hat{Q}_n(\tilde{\eta}_{10}) \\
\geq n\lambda_{\min}(\tilde{N}_1^T \tilde{N}_1/n)c_n^2\|u\|_2^2 - 2n^{1/2}\|\Lambda_1/2\|\lambda_{\max}(\tilde{N}_1^T \tilde{N}_1/n)\alpha_n\|u\|_2 \\
\geq c_3n\alpha_n^2\|u\|_2^2 - c_4n^{1/2}r_n\alpha_n\|u\| = c_3r_n^2\|u\|_2^2 - c_4r_n^2\|u\|,
\]

where \(c_3, c_4\) are some positive constants, and the second inequality is by Lemma 3 and (B5).

When \(C\) is large enough, we have \(\hat{Q}_n(\tilde{\eta}_{10} + \alpha_n u) - \hat{Q}_n(\tilde{\eta}_{10}) > 0\), which implies that there exists a local minimizer \(\tilde{\eta}_1\) of \(\hat{Q}_n(\tilde{\eta})\) such that \(\|\tilde{\eta}_1 - \tilde{\eta}_{10}\| = O_p(\alpha_n)\).

Next, we denote \(\tilde{\eta} = (\tilde{\eta}_1^{(1)T}, ..., \tilde{\eta}_g^{(1)T}, 0^T, ..., 0^T, \tilde{\gamma}^{(1)T}, 0^T)^T\), where \(\tilde{\eta}_l = (\tilde{\eta}_1^{(1)T}, ..., \tilde{\eta}_g^{(1)T}, \tilde{\gamma}^{(1)T})^T\), and our second goal is to show that \(\tilde{\eta}\) is a local minimizer of \(\hat{Q}_n(\tilde{\eta})\) over the whole space \(R^{ds_n + p_n}\). Denote \(S_l(\tilde{\eta}) = \partial\{(2n)^{-1}\|Y_M - \hat{M}b^{(1)} - Z_M\gamma\|^2\}/\partial\gamma_l\) and \(S_j^*(\tilde{\eta}) = \partial\{(2n)^{-1}\|Y_M - \hat{M}b^{(1)} - Z_M\gamma\|^2\}/\partial b_j^{(1)}\). By Karush-Kuhn-Tucker condition, it suffices to show that \(\tilde{\eta}\) satisfying the following conditions,

\[
S_l(\tilde{\eta}) = 0, \text{ and } |\gamma_l| \geq a\lambda_n \text{ for } l = 1, \ldots, q_n, \\
|S_l(\tilde{\eta})| \leq \lambda_n, \text{ and } |\gamma_l| < \lambda_n \text{ for } l = q_n + 1, \ldots, p_n, \\
S_j^*(\tilde{\eta}) = 0, \text{ and } \|b_j^{(1)}\| \geq a\lambda_n \text{ for } j = 1, \ldots, g, \\
\|S_j^*(\tilde{\eta})\| \leq \lambda_n, \text{ and } \|b_j^{(1)}\| < \lambda_n \text{ for } j = g + 1, \ldots, d,
\]

so that \(\tilde{\eta}\) is a local minimizer of \(\hat{Q}_n(\tilde{\eta})\). When \(l = 1, \ldots, q_n\), since \(\min_{t=1, \ldots, q_n}|\gamma_t| \geq \min_{t=1, \ldots, q_n}|\gamma_{10}| - \|\tilde{\gamma}^{(1)} - \gamma_0^{(1)}\|\) and \(\|\tilde{\gamma}^{(1)} - \gamma_0^{(1)}\| = o_p(\lambda_n)\), hence under (A7), we have \(\min_{t=1, \ldots, q_n}|\gamma_t|/\lambda_n \rightarrow \infty\) in probability. It follows that

\[
P(|\gamma_l| \geq a\lambda_n \text{ for } l = 1, \ldots, q_n) \rightarrow 1. \quad (1.6)
\]

When \(j = 1, \ldots, g\), since \(\|\hat{b}_j^{(1)}\| \geq \|b_j^{(1)}\| - \|\tilde{b}_j^{(1)} - b_j^{(1)}\|, \|\hat{b}_j^{(1)} - b_j^{(1)}\| = o_p(s_n^{1/2}\alpha_n) = o_p(1)\) and \(\min_{j=1, \ldots, g}\|\hat{b}_j^{(1)}\| > C_1\) for some positive constant \(C_1\), hence, we have \(\min_{j=1, \ldots, g}\|\hat{b}_j^{(1)}\|/\lambda_n \rightarrow 0\).
\(\infty\) in probability. It follows that

\[
P(\|\hat{b}_j^{(1)}\| \geq a\lambda_{jn} \text{ for } j = 1, \ldots, g) \to 1. \tag{1.7}
\]

When \(l = 1, \ldots, q_n\) and \(j = 1, \ldots, g\), \(S_l(\tilde{\eta}) = 0\) and \(S_j^*(\tilde{\eta}) = 0\) hold trivially since (1.6), (1.7) hold and \(\tilde{\eta}_1\) is a local minimizer of \(\hat{Q}_n(\tilde{\eta}_1)\).

Next, we show that \(\tilde{\eta}\) satisfy

\[|S_l(\tilde{\eta})| \leq \lambda_n, \text{ and } |\gamma_l| < \lambda_n \text{ for } l = q_n + 1, \ldots, p_n\]

and

\[||S_j^*(\tilde{\eta})|| \leq \lambda_{jn}, \text{ and } ||\hat{b}_j^{(1)}|| < \lambda_{jn} \text{ for } j = g + 1, \ldots, d.\]

Since \(\hat{\gamma}_l = 0\) for \(l = q_n + 1, \ldots, p_n\) and \(\hat{b}_j^{(1)} = 0\) for \(j = g + 1, \ldots, d\), it suffices to show that

\[
P(|S_l(\tilde{\eta})| > \lambda_n \text{ for some } l = q_n + 1, \ldots, p_n) \to 0, \tag{1.8}
\]

and

\[
P(||S_j^*(\tilde{\eta})|| > \lambda_{jn} \text{ for some } j = g + 1, \ldots, d) \to 0. \tag{1.9}
\]

Denote \(d_n = (S_{q_n+1}(\tilde{\eta}), \ldots, S_{p_n}(\tilde{\eta}))^\top = -n^{-1}\{Z_M^{(2)}\}^\top(Y_M - \tilde{N}_1\tilde{\eta}_1)\), where

\[Y_M - \tilde{N}_1\tilde{\eta}_1 = \epsilon + \nu + (\tilde{N}_1 - \tilde{\eta}_1)\tilde{\eta}_{10} + \tilde{N}_1(\tilde{\eta}_{10} - \tilde{\eta}_1),\]

and \(\nu = (\nu_1, \ldots, \nu_n)^\top\) with \(\nu_i = \sum_{j=1}^{g} \sum_{k=s_n+1}^{\infty} \xi_{ijk}b_{jk0}\). From the proof of Lemma 4 and previous derivations, we have

\[\|\nu + (\tilde{N}_1 - \tilde{\eta}_1)\tilde{\eta}_{10} + \tilde{N}_1(\tilde{\eta}_{10} - \tilde{\eta}_1)\|^2 = O_p(r_n^2) = o_p(n\lambda_n^2).\]
Moreover, we have $\max_{l=q_n+1,\ldots,p_n} \sum_{i=1}^{n} z_{ij}^2 = O_p(n)$ by (B5). Thus

$$
\|n^{-1} \{Z_M^{(2)}\}^T (\nu + (\tilde{N}_1 - \tilde{N}_1)\tilde{\eta}_{10} + \tilde{N}_1 (\tilde{\eta}_{10} - \tilde{\eta}_1))\|_\infty \\
\leq n^{-1} \left( \max_{l=q_n+1,\ldots,p_n} \sum_{i=1}^{n} z_{ij}^2 \right)^{1/2} \|\nu + (\tilde{N}_1 - \tilde{N}_1)\tilde{\eta}_{10} + \tilde{N}_1 (\tilde{\eta}_{10} - \tilde{\eta}_1)\|= o_p(\lambda_n).
$$

Next, we denote $n^{-1/2} \{Z_M^{(2)}\}^T \epsilon = (\delta_{q_n+1}, \ldots, \delta_{p_n})^T$ with $\delta_l = n^{-1/2} \sum_{i=1}^{n} z_{ij} \epsilon_i$ for $l = q_n + 1, \ldots, p_n$. As $\{\epsilon_i, i = 1, \ldots, n\}$ and $\{z_{il}, i = 1, \ldots, n\}$ are subgaussian random variables and independent of each other, given (B5), hence $\delta_l$ is subexponential random variables with uniformly bounded second moments for $l = q_n + 1, \ldots, p_n$. Thus, for any constant $C > 0$, there exists positive constants $C_1$ and $C_2$ such that uniformly for $l = q_n + 1, \ldots, p_n$,

$$
P(|\delta_l| > Cn^{1/2} \lambda_n) \leq C_2 \exp(-2^{-1} C_1 n^{1/2} \lambda_n),
$$

and it follows that

$$
P(\|n^{-1/2} \{Z_M^{(2)}\}^T \epsilon\|_\infty > Cn^{1/2} \lambda_n) \leq \sum_{l=q_n+1}^{p_n} P(|\delta_l| > Cn^{1/2} \lambda_n) \\
\leq O\{p_n \exp(-2^{-1} C_1 n^{1/2} \lambda_n)\} = O\{\exp(n^n - 2^{-1} C_1 n^{1/2} \lambda_n)\} = o(1),
$$

where the last two equalities are by (A6) and (A7). Hence, it is obvious that $\|n^{-1} \{Z_M^{(2)}\}^T \epsilon\|_\infty = o_p(\lambda_n)$. In conclusion, we have $\|d_n\|_\infty \leq \|n^{-1} \{Z_M^{(2)}\}^T \epsilon\|_\infty + \|n^{-1} \{Z_M^{(2)}\}^T (\nu + (\tilde{N}_1 - \tilde{N}_1)\tilde{\eta}_{10} + \tilde{N}_1 (\tilde{\eta}_{10} - \tilde{\eta}_1))\|_\infty = o_p(\lambda_n)$ and this implies that (1.8) holds.

Next, we start to show (1.9). For $j = g + 1, \ldots, d$, we have $S_j^*(\tilde{\eta}) = -n^{-1} \tilde{M}_j^T (Y_M - \tilde{N}_1 \tilde{\eta}_1) = (S_{j1}, \ldots, S_{j, s_n})^T$ such that $\sum_{k=1}^{s_n} S_{jk}^2$ is bounded by

$$
2 \sum_{k=1}^{s_n} (n^{-1} \sum_{i=1}^{n} \tilde{\xi}_{ijk} \epsilon_i)^2 + 2n^{-2} \sum_{k=1}^{s_n} \sum_{i=1}^{n} \tilde{\xi}_{ijk}^2 \|\nu + (\tilde{N}_1 - \tilde{N}_1)\tilde{\eta}_{10} + \tilde{N}_1 (\tilde{\eta}_{10} - \tilde{\eta}_1)\|^2.
$$
For \( \sum_{k=1}^{s_n} (n^{-1} \sum_{i=1}^{n} \hat{\xi}_{ijk} \epsilon_i)^2 \), we have

\[
E\{\sum_{k=1}^{s_n} (n^{-1} \sum_{i=1}^{n} \hat{\xi}_{ijk} \epsilon_i)^2\} = n^{-2} \sum_{k=1}^{s_n} \sum_{i=1}^{n} \sigma^2 E(\hat{\xi}_{ijk}^2) \leq n^{-2} \sum_{k=1}^{s_n} \sum_{i=1}^{n} \sigma^2 E(\|\hat{x}_{ij}\|^2) = O(s_n n^{-1}).
\]

It is easy to see that \( \sum_{k=1}^{s_n} \sum_{i=1}^{n} \hat{\xi}_{ijk}^2 = O_p(n \sum_{k=1}^{s_n} w_{jk}) = O_p(n) \). Thus

\[
\sum_{k=1}^{s_n} S_{jk}^2 = O_p(s_n n^{-1}) + n^{-2} O_p(n) O_p(r_n^2) = O_p((q_n + s_n)n^{-1}),
\]

and it follows that

\[
\|S_j^* (\hat{\eta})\|^2 = \sum_{k=1}^{s_n} S_{jk}^2 = O_p\{n^{-1}(q_n + s_n)\} = o_p(\lambda_n^2),
\]

under (A7), which entails (1.9) immediately. Hence, we conclude that \( \hat{\eta} \) is a local minimizer of \( Q_n(\hat{\eta}) \) such that \( \|\hat{\gamma} - \gamma_0\| = O_p(r_n n^{-1/2}) \), \( ||\hat{b}^{(1)} - \tilde{b}_0^{(1)}|| = O_p(r_n n^{-1/2}) \), \( ||\hat{b}^{(1)} - \hat{b}_0^{(1)}|| = O_p(r_n s_n^{-1/2} n^{-1/2}) \), and \( P(\hat{b}_{j}^{(1)} = 0, j = g + 1, \ldots, d, \hat{\gamma}_2 = 0) \to 1 \). This completes the proof.
Proof of Theorem 2. For convenience, define the following,

\[ L_n(\eta) = \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{d} \sum_{k=1}^{\infty} \xi_{ij}k b_{jk} - z_i^T \gamma \right)^2, \]

\[ T_{1n}(\eta) = -\sum_{i=1}^{n} \left( \sum_{j=1}^{d} \sum_{k=1}^{\infty} \xi_{ij}k b_{jk} \right)^2 + 2 \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{d} \sum_{k=1}^{s_n} \xi_{ij}k b_{jk} - z_i^T \gamma \right) \left( \sum_{j=1}^{d} \sum_{k=1}^{s_n+1} \xi_{ij}k b_{jk} \right), \]

\[ T_{2n}(\eta) = -2 \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{d} \sum_{k=1}^{s_n} \xi_{ij}k b_{jk} - z_i^T \gamma \right) \left\{ \sum_{j=1}^{d} \sum_{k=1}^{s_n} (\hat{\xi}_{ijk} - \xi_{ijk}) b_{jk} \right\}, \]

\[ T_{3n}(\eta) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{d} \sum_{k=1}^{s_n} (\hat{\xi}_{ijk} - \xi_{ijk}) b_{jk} \right\}^2, \]

\[ T_{4n}(\eta) = 2n \left\{ \sum_{i=1}^{p_n} J_{\lambda_n}(|\gamma_l|) + \sum_{j=1}^{d} J_{\lambda_n}(\|b_{j}^{(1)}\|) \right\}, \]

\[ P_{n}(\eta) = L_n(\eta) + T_{1n}(\eta) = \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{d} \sum_{k=1}^{s_n} \xi_{ijk} b_{jk} - z_i^T \gamma \right)^2. \]

Then we have \( Q_n(\eta) = L_n(\eta) + T_{1n}(\eta) + T_{2n}(\eta) + T_{3n}(\eta) + T_{4n}(\eta), \) and denote \( \nabla Q_n(\eta) = \partial Q_n(\eta)/\partial \gamma = (\partial Q_n(\eta)/\partial \gamma_1, \ldots, \partial Q_n(\eta)/\partial \gamma_{qn})^T \) and \( \nabla^2 Q_n(\eta) = \partial^2 Q_n(\eta)/\partial \gamma^{(1)T} \partial \gamma^{(1)} \).

Recall that \( \eta \) is a local minimizer of \( Q_n(\eta) \), hence, we have \( \nabla Q_n(\eta) \neq 0 \) and \( \nabla^2 Q_n(\eta) \) and \( \nabla^2 Q_n(\eta) \neq 0 \) for some \( \eta \). Moreover, by properties of \( \eta \) derived before and (A7), we have that uniformly for \( l = 1, \ldots, q_n \), \( \partial T_{4n}(\eta)/\partial \gamma_l = 2n J'_{\lambda_n}(|\gamma_l|) \text{sgn}(\gamma_l) = 0 \) with probability tending to one, thus \( \nabla T_{4n}(\eta) = 0 \). For \( \nabla P_{n}(\eta) \), by
Taylor expansion we have,

\[
\nabla P_n(\hat{\eta}) = \nabla P_n(\tilde{\eta}_0) + \nabla^2 P_n(\tilde{\eta}_0)(\hat{\gamma}^{(1)} - \gamma_0^{(1)})
\]

\[
= \nabla L_n(\tilde{\eta}_0) + \nabla T_{1n}(\tilde{\eta}_0) + \nabla^2 L_n(\tilde{\eta}_0)(\hat{\gamma}^{(1)} - \gamma_0^{(1)})
\]

\[
= (-2 \sum_{i=1}^{n} \epsilon_i z_i^{(1)}) + (-2 \sum_{j=1}^{d} \vartheta_j^{(1)}) + 2 \sum_{i=1}^{n} z_i^{(1)} z_i^{(1)^T}(\hat{\gamma}^{(1)} - \gamma_0^{(1)})
\]

where the last equality is by Lemma 2 (b), and \(1_{q_n}\) is the vector of ones with dimension \(q_n\). For \(\nabla T_{2n}(\hat{\eta})\),

\[
\nabla T_{2n}(\hat{\eta}) = 2 \sum_{j=1}^{d} (\vartheta_j^{(2)} + \vartheta_j^{(3)} + \vartheta_j^{(4)})
\]

\[
= 2 \sum_{j=1}^{d} (\vartheta_j^{(2)} + \vartheta_j^{(3)} + \vartheta_j^{(5)} + \vartheta_j^{(6)} + \vartheta_j^{(7)} + \vartheta_j^{(8)})
\]

\[
= o_p(n^{1/2})1_{q_n} + 2 \sum_{j=1}^{d} \vartheta_j^{(8)},
\]

where the last two equalities are by Lemma 2 (b). Notice that \(\vartheta_j^{(8)} = 0\) for \(j = g + 1, \ldots, d\),

\[
\nabla Q_n(\hat{\eta}) = \nabla L_n(\hat{\eta}) + \nabla T_{1n}(\hat{\eta}) + \nabla T_{2n}(\hat{\eta}) + \nabla T_{3n}(\hat{\eta}) + \nabla T_{4n}(\hat{\eta})
\]

\[
= o_p(n^{1/2})1_{q_n} + 2 \sum_{j=1}^{g} \vartheta_j^{(8)} + (-2 \sum_{i=1}^{n} \epsilon_i z_i^{(1)})
\]

\[
+ 2 \sum_{i=1}^{n} z_i^{(1)} z_i^{(1)^T}(\hat{\gamma}^{(1)} - \gamma_0^{(1)}) = 0,
\]

and it follows that

\[
n^{-1/2} \sum_{i=1}^{n} z_i^{(1)} z_i^{(1)^T}(\hat{\gamma}^{(1)} - \gamma_0^{(1)}) = n^{-1/2} \sum_{i=1}^{n} \epsilon_i z_i^{(1)} - n^{-1/2} \sum_{j=1}^{g} \vartheta_j^{(8)} + o_p(1)1_{q_n}
\]

\[
= \sum_{i=1}^{n} W_{in} + o_p(1)1_{q_n},
\]
with $W_{in} = n^{-1/2} \epsilon_i z_i^{(1)} - n^{-1/2} \sum_{j=1}^g \sum_{k=1}^n \sum_{\ell \neq k} b_{jk\ell} (w_{jk} - w_{j\ell})^{-1} \langle z_j, \phi_{j\ell} \rangle \int (x_{ij} \otimes x_{ij} - K_j) \phi_{j\ell}$, it is obvious that $\{W_{in}, i = 1, \ldots, n\}$ are independently and identically distributed. With zero mean and $\text{cov}(W_{in}) = n^{-1}(\sigma^2 \Sigma_1 + B_n)$. It satisfies to show that $V_n^{-1/2} A_n \sum_{i=1}^n W_{in}$ converges to $N(0, I_r)$ in distribution, where $V_n = A_n(\sigma^2 \Sigma_1 + B_n) A_n^T V_n = \sigma^2 H^* + B^*$, and $V^*$ is positive definite. We start to show this by Linderberg Feller theorem.

First, we denote $\Pi_{in} = V_n^{-2} A_n W_{in}$, and it is trivial that $\{\Pi_{in}, i = 1, \ldots, n\}$ are independently and identically distributed centered random vectors with $\text{cov}(\sum_{i=1}^n \Pi_{in}) = I_r$. Second, for any $\zeta > 0$, we have

$$\sum_{i=1}^n E[\|\Pi_{in}\|^2 1_{kn} \{\|\Pi_{in}\| > \zeta\}] \leq nE(\|\Pi_{1n}\|^4)^{1/2} \{P(\|\Pi_{1n}\| > \zeta)\}^{1/2},$$

where

$$\|\Pi_{1n}\|^2 = \Pi_{1n}^T \Pi_{1n} = W_{1n}^T A_n^T V_n^{-1} A_n W_{1n} \leq \lambda_{\text{max}}(A_n^T V_n^{-1} A_n) \|W_{1n}\|^2 \leq \lambda_{\text{max}}(V_n^{-1}) \lambda_{\text{max}}(A_n^T A_n) \|W_{1n}\|^2 = O(1) \|W_{1n}\|^2.$$

It follows that $E(\|\Pi_{1n}\|^4) = O(1) E(\|W_{1n}\|^4) = O(1) q_n^2 / n^2$. Moreover,

$$P(\|\Pi_{1n}\| > \zeta) = P(\|V_n^{-1/2} A_n W_{1n}\| > \zeta) \leq E(\|V_n^{-1/2} A_n W_{1n}\|^2) / \zeta^2 \leq \lambda_{\text{max}}(A_n^T V_n^{-1} A_n) E(\|W_{1n}\|^2) / \zeta^2 = O(1) E(\|W_{1n}\|^2) = O(1) q_n / n.$$

By previous derivations and assumptions, we know

$$\sum_{i=1}^n E[\|\Pi_{in}\|^2 1_{kn} \{\|\Pi_{in}\| > \zeta\}] = nO(q_n / n) O(q_n n^{-1/2}) = O(q_3^3 n^{-1/2}) = o(1),$$

since $q_n = o(n^{1/3})$. By Linderberg Feller theorem, we conclude that $V_n^{-1/2} A_n \sum_{i=1}^n W_{in}$ con-
verges to $N(0_r, I_r)$ in distribution, which completes the proof.
Chapter 2

Testing General Hypothesis in Large-scale FLR

2.1 Introduction

The classical functional linear regression (FLR) has been widely adopted for modelling the linear relationship between a scalar response \( Y \) and a functional predictor that is often assumed to be sampled from an \( L^2(T) \) random process \( X(t) \) defined on a compact interval \( T \subseteq \mathbb{R} \). Specifically, given \( n \) independent and identically distributed (i.i.d.) pairs \( \{Y_i, X_i(\cdot)\} \), the classical FLR is formulated as

\[
Y_i = \int_T X_i(t) \beta(t) dt + \epsilon_i, \quad i = 1, \ldots, n
\]  

(2.1)

where both \( Y_i \) and \( X_i \) are centred without loss of generality, i.e., \( EY_i = 0 \) and \( EX_i(t) = 0 \) for \( t \in T \), the unknown regression parameter function \( \beta(t) \) is square-integrable, i.e., \( \beta \in L^2(T) \), and the i.i.d. regression error \( \epsilon_i \) is independent of \( X_i \) with mean zero and variance \( 0 < \sigma^2 < \infty \). This model has been extensively researched in the literature of functional data analysis [Ramsay and Dalzell, 1991, Cardot et al., 1999, Fan and Zhang, 2000, Yuan and Cai, 2010, among others], including theoretical considerations [Cai and Hall, 2006, Hall and Horowitz,
2007, Cai and Yuan, 2012] and statistical inference [Cardot et al., 2003, Lei, 2014, Hilgert et al., 2013], and we refer readers to Ramsay and Silverman [2005] for an overview and numerical examples. There has also been numerous work that extended the classical FLR to various settings, such as the functional response [Faraway, 1997, Cuevas et al., 2002, Yao et al., 2005b], the generalized FLR [Escabias et al., 2004, James, 2002, Müller and Stadtmüller, 2005, Shang and Cheng, 2015], the partially FLR [Lian, 2011, Kong et al., 2016], the additive regression [Müller and Yao, 2008, Zhu et al., 2014, Fan et al., 2015], among many others.

In modern scientific experiments, the response $Y$ can be potentially associated with multiple or even a large number of functional predictors, e.g., Lian [2013] proposed a FLR involving a fixed number of functional predictors and Kong et al. [2016] considered regularized estimation and variable selection for a partially FLR that contains high-dimensional scalar covariates in addition to a finite number of functional predictors. However, in large-scale data analysis where a FLR is exploited, the number of potential functional predictors $p_n$ can be possibly much larger than the sample size $n$, despite the significant ones of size $q_n$ are usually assumed to be sparse or at a fraction polynomial order of $n$. Examples can be found in neuroimage analysis, where the relationship between certain disease marker and a number of brain regions of interest (ROI) scanned over time is of interest. This consideration motivates, namely, a large-scale FLR model as follows,

$$Y_i = \sum_{j=1}^{p_n} \int_T X_{ij}(t) \beta_j(t) dt + \epsilon_i, \quad i = 1, \ldots, n$$ \hspace{1cm} (2.2)

where $p_n$ is allowed to grow exponentially in sample size $n$, without loss of generality, the first $q_n$ regression parameter functions $\{\beta_j : j = 1 \ldots, q_n\}$ are assumed nonzero (i.e., important ones), while the rest are zero, and the i.i.d. error $\epsilon_i$ is independent of $\{X_{ij} : j = 1, \ldots, p_n\}$ with mean zero and variance $\sigma^2$. It is common to use a set of pre-fixed (i.e., B-splines, wavelets) or data-driven (i.e., eigenfunctions) basis to represent the underlying process $X_j$ of each predictor $\{X_{ij} : i = 1, \ldots, n\}$. Since the data-driven bases, such as eigenfunctions, have to be estimated
from $p_n$ separate functional principal component analysis procedures, which is computationally prohibited when $p_n \gg n$ and does not necessarily produces efficient representations for the sake of regression/prediction of $Y$, hence we adopt a common pre-fixed basis $\{b_k : k \geq 1\}$ that is complete and orthonormal in $L^2(t)$ for all processes $X_j$ for $j = 1, \ldots, p_n$, and do not further pursue other complicated basis-seeking procedures, e.g., functional partial least squares [Reiss and Ogden, 2007].

To be specific, functional predictors and the associated regression functions can be expressed as linear combinations of the complete orthonormal basis $\{b_k : k \geq 1\}$, such that $\beta_j = \sum_{k=1}^{\infty} \eta_{jk} b_k$, $X_{ij} = \sum_{k=1}^{\infty} \theta_{ijk} b_k$, where the coefficients $\theta_{ijk} = \int_T X_{ij}(t) \beta_{k}(t) dt$ coinciding with the projections are mean-zero random variables whose variances are $E(\theta_{ijk}^2) = \omega_{jk} > 0$. As a result, model (2.2) can be reformulated as

$$Y_i = \sum_{j=1}^{p_n} \sum_{k=1}^{\infty} \theta_{ijk} \eta_{jk} + \epsilon_i. \quad (2.3)$$

To perform estimation and inference on the regression functions of primary interest, a direct minimization of the square loss with respect to the infinite sequences of unknown coefficients $\eta_{jk}$ is not feasible. A common practice is to truncate up to the first $s_n$ leading terms that is allowed to grow with $n$, where $s_n$ controls the complexity of $\beta_j$ as a whole function and balances the bias-variance tradeoff in a similar spirit of classical nonparametric regression.

To our knowledge, this type of large-scale FLR first appeared in Fan et al. [2015] that considered a penalized procedure for model estimation and selection, while our primary interest is the hypothesis testing. A careful inspection on Condition 1(A) in Fan et al. [2015, Appendix B] which requires $\sum_{k=1}^{\infty} \theta_{ijk}^2 k^4 < C^2$ for a universal constant $C$ for $i = 1, \ldots, n, j = 1, \ldots, p_n$, reveals that all random processes $X_j$ are bounded in $L^2(T)$ uniformly in $j$, thus excludes even the case of Gaussian processes. Moreover, Condition 2(D) assumes that the minimal eigenvalues of $n^{-1} \sum_{j=1}^{q_n} \Theta_j' \Theta_j \geq c_0$, bounded from below by a constant $c_0$ uniformly in $1 \leq j \leq q_n$ (i.e., the important ones), where $\Theta_j = (\theta_{ijk})_{1 \leq i \leq n, 1 \leq k \leq s_n}$ is the $n \times s_n$ design matrix induced by $X_j$. 

\[ \text{CHAPTER 2. TESTING GENERAL HYPOTHESIS IN LARGE-SCALE FLR} \]
in Fan et al. [2015] that used \( q_n \) to denote the truncation size with a notation abuse. Unfortunately, this crucial condition is not valid for infinite-dimensional \( L^2 \) process, as the minimal eigenvalues necessarily approach 0 when \( s_n \) diverges, with a typical illustration given by the Karhunen-Loève expansion. By contrast, we do not make these stringent assumptions so that the predictor processes are genuinely functional in the large-scale FLR (2.2).

The main contribution of this paper is to develop a rigorous testing procedure for the general hypothesis on an arbitrary subset of regression functions \( \{ \beta_k : k = 1, \ldots, p_n \} \). The challenge arises from not only the ultrahigh-dimensionality in \( p_n \) that can be as exponentially large as \( n \) but also the intrinsic infinite-dimensionality of each \( X_j, j = 1, \ldots, p_n \). Although the FLR (and its variants) has been well studied, there are only a few contributions on inference procedures, e.g., Hilgert et al. [2013] and Lei [2014] considered adaptive tests of a single regression function in classical FLR and Shang and Cheng [2015] for the generalized FLR. In the current exposition, we adopt a general class of nonconvex penalty functions [Loh and Wainwright, 2015] which include LASSO [Tibshirani, 1996], SCAD [Fan and Li, 2001] and MCP [Zhang, 2010] penalties as special cases and whose theoretical properties in high-dimensional linear regression have been extensively studied [Meinshausen and Bühlmann, 2006, van de Geer, 2008, Meinshausen and Yu, 2009, Bickel et al., 2009, Zhang, 2009, Fan and Lv, 2011, Wang et al., 2013, 2014, Fan et al., 2014, Loh and Wainwright, 2015, among many others].

Recently the research on inference of high-dimensional linear regression has flourished, especially for LASSO-type convex penalty [Tibshirani, 1996], such as Wasserman and Roeder [2009], Meinshausen and Bühlmann [2010], Shah and Samworth [2013] based on sample splitting or subsampling, Zhang and Zhang [2014], van de Geer et al. [2014] on bias correction methods, Lockhart et al. [2014], Taylor et al. [2014] for and conditional inference on the even that some covariates have been selected, among others.

In this article, we are inspired by the unconditional inference based on decorrelated score function proposed in Ning and Liu [2016] owing to its generality, no need of data splitting or strong minimal signal conditions. We first exploit a penalized least squares procedure treating
the truncated coefficients of each $\beta_k$ altogether in a group fashion, and obtain estimation consistency with no need of oracle properties under weaker minimal signal conditions that allow for a wider class of suitable settings. Then we devise the decorrelated score function in the context of large-scale FLR for testing general null hypothesis on any subset of $\{\beta_k : k \geq 1\}$. Unlike testing null hypothesis on a single parameter in high-dimensional linear regression, the limiting distribution for such a general null hypothesis is intractable. Finally we adopt the multiplier bootstrap method to approximate the limiting distribution of the score test statistic under null hypothesis and provide theoretical guarantees for all possible sizes in a uniform manner.

The rest of the article is organized as follows. In section 2.2, we describe the general hypothesis we want to test and the way to construct the proposed score test based on a penalized estimator and a well defined decorrelated score function. Section 2.3 is devoted to introducing the theoretical properties of the penalized estimator and the score test defined in section 2.2 under some regularity conditions. Section 2.4 proposes the simulation results to back up the theoretical properties. The algorithm to obtain the penalized estimator, technical conditions on penalty functions, notations used throughout the paper and the proofs of Lemmas and Theorems are deferred to section 2.5.

### 2.2 Proposed Methodology

#### 2.2.1 Regularized estimation by group penalized least squares

Recall the large-scale FLR defined in (2.2), the underlying predictor processes $X_j$ and the corresponding regression functions $\beta_j$ are expressed by a set of complete and orthonormal basis $\{b_k : k \geq 1\}$, leading to an infinite-dimensional representation (2.3). Since the square loss function $\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p_n} \sum_{k=1}^{\infty} \theta_{ijk} \eta_{jk})^2$ cannot be directly minimized with respect to $\eta_{jk}$ due to infinite-dimensionality, we control the complexity of each $\beta_k$ as a whole function and adopt the commonly-used truncation for the purpose of regularization, rather than viewing its basis terms as separate predictors. Here this truncation parameter is allowed to grow with $n$,
and plays a role of smoothing parameter balancing the trade-off between approximation bias and sampling variability.

Remark. Ideally one can use different truncation sizes for each $\beta_k$, however, selection of different truncation sizes for a large number of functional predictors is computationally infeasible. Therefore, in practice, we may adopt the strategy suggested by Kong et al. [2016], i.e., use a common $s_n$ to perform regularized estimation, then amend a refit step using ordinary least squares for the retained variables and choose different truncations using a $K$-fold cross-validation, saying $K = 5$. Nonetheless, the case of common $s_n$ suffices for methodological development and theoretical analysis.

Besides the truncation, it is essential to impose suitable penalty for each regression function as a whole by utilizing a functional version of group regularization [Yuan and Lin, 2006]. To regularize all predictors on a comparable scale, one often standardizes scalar predictors in linear regression [Fan and Li, 2001]. For functional predictors $X_j$, we choose to account the variability of the grouped projection coefficients $\theta_{ijk}$ in the $n \times s_n$ design matrix $\Theta_j = (\theta_{ijk})_{1 \leq i \leq n; 1 \leq k \leq s_n}$. Hence the penalty imposed on $n^{-1/2}||\Theta_j \eta_j||_2$ invokes a group penalty that shrinks the unimportant regression function to zero, where $|| \cdot ||_2$ is the Euclidean or $\ell_2$ norm (if an infinite sequence). For technical convenience, we scale up the penalty parameter $\lambda_n$ by $s_n^{1/2}$, which does not affect the relative weighting of penalties given the common group size $s_n$.

Therefore, our objective is to minimize the penalized square loss function as follows, denoting $\eta = (\eta_1', \ldots, \eta_{pn}')'$ with vectors $\eta_j = (\eta_{j1}, \ldots, \eta_{js_n})'$, and $|| \cdot ||_1$ the $\ell_1$ norm,

$$
\min_{\eta: ||\eta||_1 \leq R_n} \left[ (2n)^{-1} \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{pn} \sum_{k=1}^{s_n} \theta_{ijk} \eta_{jk})^2 + \frac{1}{n} \sum_{j=1}^{pn} \rho_{\lambda_n s_n^{1/2}} (n^{-1/2} ||\Theta_j \eta_j||_2) \right]
$$

where $\rho_\lambda(\cdot)$ with the tuning parameter $\lambda$ belongs to a general class of nonconvex penalty functions satisfying conditions (P1)–(P5) in section 2.5, which includes the mostly used penalties,
such as LASSO, SCAD and MCP and so on. [Loh and Wainwright, 2015]. The positive constraint $R_n$ should be chosen carefully to make the true value $\eta^*$ a feasible point such that $\|\eta^*\|_1 \leq R_n$. Upon solving the optimization problem (2.4), which is guaranteed to have a global minimum by Weierstrass extreme value theorem if $\rho_\lambda(\cdot)$ is continuous, the regularized estimator for each $\beta_j(t)$ can be expressed as $\hat{\beta}_j(t) = \sum_{k=1}^{s_n} \hat{\eta}_{jk} b_k(t)$, where $\hat{\eta}$ is obtained from solving (2.4). The implementation by a coordinate descent algorithm, similar to Ravikumar et al. [2008] with slight modification, is presented in section 2.5, while the tuning parameters $\lambda_n$ and $s_n$ are chosen by a $K$-fold cross-validation, saying $K = 5$ or 10. It is worth mentioning that, for the purpose of general hypothesis testing, it is sufficient to obtain consistent estimation of $\eta$ from (2.4) in both $\ell^1$ and $\ell^2$ sense stated in Theorem 3, while the selection consistency or the oracle properties is not necessary.

2.2.2 General hypothesis and score decorrelation

Our goal is to test a class of hypotheses that is of full generality in the framework of large-scale FLR. Denote $\mathcal{P}_n = \{1, \ldots, p_n\}$ as the index set of all functional predictors, and let $\mathcal{H}_n \subseteq \mathcal{P}_n$ be an arbitrary nonempty subset of $\mathcal{P}_n$ with the cardinality $|\mathcal{H}_n| = h_n \leq p_n$, and the complement of $\mathcal{H}_n$ is denoted by $\mathcal{H}^c = \mathcal{P}_n \setminus \mathcal{H}_n$. Then the general hypothesis can be expressed as

$$H_0 : \|\beta_j\|_{L^2} = 0 \text{ for all } j \in \mathcal{H}_n \quad \text{v.s.} \quad H_a : \|\beta_j\|_{L^2} > 0 \text{ for at least one } j \in \mathcal{H}_n, \quad (2.5)$$

which is asymptotically equivalent to the hypothesis

$$H_0 : \eta_j = 0 \text{ for all } j \in \mathcal{H}_n \quad \text{v.s.} \quad H_a : \eta_j \neq 0 \text{ for at least one } j \in \mathcal{H}_n,$$

with each $\eta_j = (\eta_{j1}, \ldots, \eta_{js_n})'$, as $s_n \to \infty$ given $n \to \infty$. It is worth noting that the cardinality $h_n$ can be as large as the cardinality of the full predictor set $p_n$, allowing for any hypothesis on $\{\beta_k : k = 1, \ldots, p_n\}$ of interest.

For testing the general null hypothesis in (2.5), the building block originates from the
CHAPTER 2. TESTING GENERAL HYPOTHESIS IN LARGE-SCALE FLR

54

combination of consistently estimated regression functions and a new type of score function. As illustrated in Ning and Liu [2016], the motivation of considering a decorrelated score function is the high-dimensionality of the nuisance parameter space \( \mathcal{H}_n^c = \mathcal{P}_n \setminus \mathcal{H}_n \) that makes the limiting distribution of the estimated nuisance parameter constrained by null hypothesis intractable [Fu and Knight, 2000]. Hence the key is to decorrelate the score function of the primary parameter in \( \mathcal{H}_n \) from that of the nuisance parameter in \( \mathcal{H}_n^c \) in order to control the variability induced by high-dimensional nuisance parameter. As a result, the decorrelation operation is a natural extension of profile score to high-dimensional case and leads to a test that is asymptotically equivalent to the classical Rao’s score test in low-dimensional case [Cox and Hinkley, 1979, Ning and Liu, 2016].

We first introduce some notation for adopting the score function decorrelation in the proposed large-scale FLR. Recall that \( \omega_{jk} \) is the variance of the i.i.d. projection coefficient \( \{ \theta_{ijk} = \int_T X_i(t)b_k(t)dt : i = 1, \ldots, n \} \). Denote \( \Lambda_j = \text{diag}\{\omega_{j1}^{1/2}, \ldots, \omega_{js_n}^{1/2}\} \) for \( j \leq p_n \), the block diagonal matrix \( \Lambda_{\mathcal{H}_n} = \text{diag}\{\Lambda_j : j \in \mathcal{H}_n\} \), and similarly for \( \Lambda_{\mathcal{P}_n} \equiv \Lambda \). Let \( \Theta = (G_1', \ldots, G_n') = (\Theta_{\mathcal{H}_n}, \Theta_{\mathcal{H}_n^c}) \), \( \Theta_{\mathcal{H}_n} = (E_1', \ldots, E_n') \), \( \Theta_{\mathcal{H}_n^c} = (F_1', \ldots, F_n') \), where \( G_i, E_i \) and \( F_i \) are the vectors containing coefficients \( \theta_{ijk} \) from corresponding functional predictors for the \( i \)th subject, respectively, \( \Theta_{\mathcal{H}_n} \) is formed by concatenating \( \{\Theta_j : j \in \mathcal{H}_n\} \) in a row, and similarly for \( \Theta_{\mathcal{H}_n^c} \). Here we view the least squares \( \ell(\eta) = \ell(\eta_{\mathcal{H}_n}, \eta_{\mathcal{H}_n^c}) = n^{-1}(Y - \Theta\eta)'(Y - \Theta\eta) \) as the likelihood function of \( \eta \) without introducing extra notations. Further we define

\[
w = I_{\mathcal{H}_n}^{-1} I_{\mathcal{H}_n \setminus \mathcal{H}_n} (I_{\mathcal{H}_n \setminus \mathcal{H}_n} E_{\mathcal{H}_n^c} I_{\mathcal{H}_n^c} \mathcal{H}_n), \quad \text{where } I_{\mathcal{H}_n \setminus \mathcal{H}_n} = E(F_iE_i'), \ I_{\mathcal{H}_n \setminus \mathcal{H}_n} E_{\mathcal{H}_n^c} = E(F_iF_i').
\]

For the purpose of decorrelation, a new score function with respect to the primary parameter...
\( \eta_{H_n} \), denoted by \( S(\eta) \), is defined in the context of our large-scale FLR as follows,

\[
S(\eta) = S(\eta_{H_n}, \eta_{H^n_m}) = n^{-1}\Lambda_{H_n}^{-1}(w'\Theta'H_n - \Theta'H_n)(Y - \Theta'H_n\eta_{H_n} - \Theta'H_n\eta_{H^n_m})
\]

\[
= n^{-1}\sum_{i=1}^{n}\Lambda_{H_n}^{-1}(w'F_i - E_i)(Y_i - E'_i\eta_{H_n} - F'_i\eta_{H^n_m}).
\]  

(2.6)

It is easy to verify that this new score function with respect to the primary parameter \( \eta_{H_n} \) is uncorrelated with the traditional score function with respect to the nuisance parameter \( \eta_{H^n_m} \), i.e., \( E\{S(\eta)\nabla_{\eta_{H^n_m}}\ell(\eta)\} = 0 \), as in Ning and Liu [2016], where \( \nabla_\gamma \) denotes the gradient vector taken with respect to \( \gamma \).

2.2.3 Bootstrapped score test for general hypothesis in large-scale FLR

Given the consistent estimation of regression functions and the decorrelated score function, we are ready to construct the proposed score test for the general hypothesis (2.5) in large-scale FLR. It is noted that the decorrelated score function \( S(\eta) \) defined in (2.6) cannot be calculated from observed data directly due to the unknown quantities \( w = I_{H^n_m}^{-1}H_n'H_n \) and \( \Lambda_{H_n} \). It is straightforward to estimate \( \Lambda_{H_n} \) by plugging in \( \hat{\omega}_{jk} = \sum_{i=1}^{n}\theta_{ijk}^2 \), denoted by \( \hat{\Lambda}_{H_n} \), and similarly for \( \hat{\Lambda}_{H^n_m} \). To estimate \( w \), a natural choice is the moment estimator \( \hat{w} = \hat{I}_{H^n_m}^{-1}H_n'H_n \), where \( \hat{I}_{H^n_m}^{-1}H_n'H_n = n^{-1}\Theta_{H^n_m}'\Theta_{H^n_m} \) for \( I_{H^n_m}^{-1}H_n'H_n = E(F_iE_i') \) and \( \hat{I}_{H^n_m}^{-1}H_n'H_n = n^{-1}\Theta_{H^n_m}'\Theta_{H^n_m} \) for \( I_{H^n_m}^{-1}H_n'H_n = E(F_iF_i') \). However, this estimator may not exist, since the matrix \( \hat{I}_{H^n_m}^{-1}H_n'H_n \) can be singular in high-dimensional settings. We follow the suggestion by Ning and Liu [2016] to adopt the Dantizig selection [Candes and Tao, 2007] to estimate the \((p_n - h_n) \times h_n s_n \) unknown matrix \( w \) by column, while alternative procedures can also be used but not pursued here for brevity. Specifically, for each \( l = 1, \ldots, h_n s_n \), we solve

\[
\hat{w}_l \in \arg\min_{w_l} ||w_l||_1 \quad \text{s.t.} \quad ||n^{-1}\sum_{i=1}^{n}E_{il}F_i' - w_l'n^{-1}\sum_{i=1}^{n}F_iF_i'||_\infty \leq \tau_n,
\]

(2.7)
where $\tau_n$ as a common tuning parameter chosen by a $K$-fold cross-validation in practice, giving
the resulting estimator $\hat{w}$. It then follows the estimated decorrelated score function

$$
\hat{S}(\eta) = \hat{S}(\eta_{\mathcal{H}_n}, \eta_{\mathcal{H}_n}) = n^{-1} \hat{\Lambda}_n^{-1}(\hat{w}'\Theta'_{\mathcal{H}_n} - \Theta'_{\mathcal{H}_n})(Y - \Theta_{\mathcal{H}_n} \hat{\eta}_{\mathcal{H}_n} - \Theta_{\mathcal{H}_n} \hat{\eta}_{\mathcal{H}_n})
= n^{-1} \sum_{i=1}^{n} \hat{\Lambda}_n^{-1}(\hat{w}'F_i - E_i)(Y_i - E_i'\hat{\eta}_{\mathcal{H}_n} - F_i'\hat{\eta}_{\mathcal{H}_n}).
$$

Then we plug in the estimator $\hat{\eta}$ obtained from minimizing (2.4) to construct the decorrelated
score test statistic under the null hypothesis of $\hat{\eta}_{\mathcal{H}_n} = 0$, leading to

$$
\hat{T} = n^{1/2} \hat{S}(0, \hat{\eta}_{\mathcal{H}_n}) = n^{-1/2} \sum_{i=1}^{n} \hat{S}_i, \quad \text{where} \quad \hat{S}_i = \hat{\Lambda}_n^{-1}(\hat{w}'F_i - E_i)(Y_i - F_i'\hat{\eta}_{\mathcal{H}_n}).
$$

Note that the null hypothesis (2.5) is of full generality with the dimension $h_n s_n$, where $s_n$ grows with $n$ (often at a factional polynomial order) to approximate the infinite-dimensional functional spaces and $h_n$ can be as large as $p_n$. Unlike testing a finite dimensional null hypothesis, it is rather intricate to find a limiting distribution that is likely to be intractable, even for the case of $h_n = 1$. Hence, we use its infinity norm $||\hat{T}||_{\infty} = \max\{|\hat{T}_l| : l = 1, \ldots, h_n s_n\}$ to test against the null hypothesis (2.5), and adopt a computationally efficient and theoretically guaranteed bootstrap method to approximate the limiting distribution of $||\hat{T}||_{\infty}$. Since a standard bootstrap is computationally intensive by repeatedly estimating $\eta$ and $w$, we consider a multiplier bootstrap method which was studied by Chernozhukov et al. [2014]. Specifically, denote $\hat{T}_e = n^{-1/2} \sum_{i=1}^{n} e_i \hat{S}_i$, where $\{e_1, \ldots, e_n\}$ is a set of i.i.d. standard normal random variables independent of the data, and define $c_B(\alpha) = \inf\{t \in \mathbb{R} : P_e(||\hat{T}_e||_{\infty} \geq t) \leq \alpha\}$ as the $100(1 - \alpha)$th percentile of $||\hat{T}_e||_{\infty}$, where $P_e(\cdot)$ denotes the probability with respect to $\{e_1, \ldots, e_n\}$. Based on this critical value, we reject the null hypothesis at the significance level $\alpha$ provided that $||\hat{T}||_{\infty} \geq c_B(\alpha)$. In Section 2.3, we show that under the null hypothesis and some mild conditions, the Kolmogorov distance between the distributions of $||\hat{T}||_{\infty}$ and $||\hat{T}_e||_{\infty}$ converges to zero as sample size grows, which provides theoretical guarantees for the proposed score test.
Remark. Based on the decorrelated score function (2.6), in principle one can construct the counterparts of other classical tests, such as the Wald and likelihood ratio tests, for a variety of high-dimensional models, for instance, the Cox proportional hazard model [Fang et al., 2014]. We expect that these asymptotically equivalent tests may also hold for the large-scale FLR model, which deserves further investigation.

2.3 Asymptotic Properties

Before introducing the main Theorems, we impose some mild conditions on the covariance structure of functional predictors, and the orders of various tuning parameters. First, recall that we have $\omega_{jk} = E(\theta_{ijk}^2) > 0$. Accordingly, we assume that the expected norm of the functional predictors are uniformly bounded such that

\[(A1) \sup_{j \leq p_n} \sum_{k=1}^{\infty} \omega_{jk} < \infty.\]

Notice that (A1) implies that $\lambda_{\text{max}}(\Lambda) = O(1)$. Before discussing the next condition which is devoted to describing the distribution of several subgaussian random variables, we introduce some useful notations as follows. For any sub-gaussian random variable $X$, define $||X||_{\phi_1} = \sup_{q \geq 1} q^{-1/2}(E|X|^q)^{1/q}$ as the sub-gaussian norm, and for any sub-exponential random variable $X$, define $||X||_{\phi_2} = \sup_{q \geq 1} q^{-1}(E|X|^q)^{1/q}$ as the sub-exponential norm, similar to those adopted in Ning and Liu [2016].

\[(A2) \text{ Assume that } \epsilon_i, \omega_{jk}^{-1/2}\theta_{ijk}, (w_i'F_i - E_{il})[E(E_{il}^2)]^{-1/2} \text{ are centred sub-gaussian random variables satisfying } ||\epsilon_i||_{\phi_1} \leq c, ||\omega_{jk}^{-1/2}\theta_{ijk}||_{\phi_1} \leq c, ||(w_i'F_i - E_{il})[E(E_{il}^2)]^{-1/2}||_{\phi_1} \leq c \text{ for some positive constant } c, \text{ uniformly in } i = 1, \ldots, n, j = 1, \ldots, p_n, k = 1, \ldots, \infty, l = 1, \ldots, h_ns_n.\]

Notice that (A1) together with (A2) implies that $\theta_{ijk}, w_i'(F_i - E_{il})$ are also centred sub-gaussian random variables satisfying $||\theta_{ijk}||_{\phi_1} \leq c_1, ||w_i'F_i - E_{il}||_{\phi_1} \leq c_1$ for some positive constant $c_1$, uniformly in $1 \leq i \leq n, 1 \leq j \leq p_n, 1 \leq l \leq h_ns_n$ and $k \geq 1$. Before introducing
the subsequent conditions, we denote the information matrix and partial information matrix as \( I = E(G_i G_i') \) and \( I_{H_n[H_n]} = I_{H_n[H_n]} - w' I_{H_n[H_n]} \) respectively. Similarly, the standardized version of the information matrix and partial information matrix are denoted as \( \bar{I} = \Lambda^{-1} I \Lambda^{-1} \) and \( \bar{I}_{H_n[H_n]} = \Lambda^{-1} I_{H_n[H_n]} \Lambda^{-1} \) respectively. Next, we assume that

(A3) \( \lambda_{\min}(\Lambda) \geq c_s^{-a/2} \) for some constants \( c > 0 \) and \( a > 1 \).

(A4) \( 0 < m_0 \leq \lambda_{\min}(\bar{I}) \leq \lambda_{\max}(\bar{I}) < \infty \) for some constants \( m_1 > m_0 > 0 \), satisfying \( m_0 > 2^{-1} m_1 \mu \).

Note that \( \mu > 0 \) is a constant such that \( \rho_{\lambda,\mu}(t) \) is convex in \( t \), which can be referred to Appendix.

Combining (A3) and (A4), it is easy to see that \( \lambda_{\min}(I) = \lambda_{\min}(\Lambda \bar{I} \Lambda) \geq c_s^{-a} \), for some \( c > 0 \).

In addition, we need to control the correlation between the functional part for testing and the rest of the functional predictors, so we assume

(A5) \( 0 < c_1 \leq \lambda_{\min}(\bar{I}_{H_n[H_n]}) \leq \lambda_{\max}(\bar{I}_{H_n[H_n]}) \leq c_2 < \infty \) for some constants \( c_2 > c_1 > 0 \).

Next, the overall number of functional predictors is permitted to grow exponentially in sample size and for any two sequences \( a_n \) and \( b_n \), \( a_n \sim b_n \) if and only if \( c_1 \leq \lim_{n \to \infty} |a_n/b_n| \leq c_2 \) for some positive constants \( c_1, c_2 \). Then we assume

(A6) \( p_n \sim \exp(n^\beta) \) for some \( \beta \in (0, 9^{-1}) \).

Regarding the first \( q_n \) nonzero regression functions, we assume that they belong to a sobolev ball as follows

(A7) \( \sup_{j \leq q_n} \sum_{k=1}^{\infty} \eta_{jk}^2 k^{2\delta} < c \) for some positive constants \( \delta \) and \( c \).

Next, in (A8), we quantify the relationship among the four parameters \( q_n, \rho_n, s_n, R_n \), and the sample size \( n \). \( q_n \) is the number of significant predictors. \( \rho_n \) is equal to \( \sup_{l \leq h_n, s_n} ||w_l||_0 \), where the vector norm \( || \cdot ||_0 \) denotes the number of nonzeros in the vector. \( R_n \) is a general parameter such that \( ||\eta^*||_1 \leq R_n \), where \( \eta^* \) represents the true value of \( \eta \). \( s_n \) is the truncation size for each functional predictor. We assume
Note that (A8) is sufficient for the existence of such \( \lambda \) and (A8) also implies that both \( \rho_n \) and \( q_n \) are small in sample size reflecting the sparseness of the model. Next, for a concrete example, if \( s_n, R_n, \rho_n \) and \( q_n \) are on the same order, (A8) can be rewritten as

\[
\max \{n^{3\beta/2} \rho_n q_n s_n^{3\alpha/2} - \delta \log s_n, \ n^\delta q_n^2 s_n^{a+1-\delta}, \ n^2 \beta q_n^2 s_n^{a+1-\delta}, \ n^{5\beta/2-1/2} \rho_n q_n s_n^{2a+1-\delta}, \ n^{2\beta-1/2} (\log n)^{1/2} \rho_n s_n^{3\alpha/2}, \ n^{5\beta/2-1/2} R_n q_n s_n^{a+1}, \ n^{3\beta-1} R_n \rho_n q_n s_n^{2a+1}, \n\}
\]

\[= o(1). \tag{2.10} \]

It is easy to verify that there exists \( s_n \) satisfying (2.10) if and only if

\[
\min \{ \frac{2\delta - 3a - 4}{3\beta}, \frac{\delta - a - 3}{\beta}, \frac{2\delta - 2}{2\beta + 1} \} > \max \{ \frac{4a + 8 - 2\delta}{1 - 5\beta}, \frac{3a + 2}{1 - 4\beta}, \frac{a + 6}{1 - 5\beta}, \frac{2a + 6}{1 - 3\beta} \}, \tag{2.11} \]

and it is easy to show that the setting \( \{a = 2, \delta = 7, \beta < 28^{-1}\} \) satisfies (2.11) for instance.

Next, in (A9), we impose some conditions on the tuning parameter \( \lambda \) for the regularizer in (3) to ensure estimation consistency.

\[
\begin{align*}
(A9) \quad & n^\delta q_n s_n^{a+1} = o(\lambda_n^{-1}), \quad n^\delta q_n s_n^{a+1} = o(\lambda_n^{-1}), \quad n^{5\beta/2 - 1/2} \rho_n q_n s_n^{2a+1} = o(\lambda_n^{-1}), \quad n^{3/2 - 1/2} R_n = o(\lambda_n), \quad q_n s_n^{-\delta} = o(\lambda_n). 
\end{align*}
\]

Note that (A8) is sufficient for the existence of such \( \lambda \) which satisfy (A9). Lastly, in (A10), we assume the tuning parameter \( \tau_n \) for the Dantzig selector method to satisfy

\[
(A10) \quad \tau_n \sim \left[ \log(p_n s_n)/n \right]^{1/2}.
\]

Combining (A6), (A8) with (A10), we have \( \tau_n \sim n^{3/2 - 1/2} \). In the following, we establish the main theorems. Several auxiliary lemmas and the proofs of those lemmas and the main theorems are deferred to Appendix. Theorem 3 establishes the estimation consistency for the
estimator obtained from (2.4) under some mild conditions, where the conditions (P1)–(P5) are delegated to Appendix.

**Theorem 3** Under conditions (A1)–(A4), (A6)–(A9) and (P1)–(P5), every local minimizer \( \hat{\eta} \) of \( Q_n(\eta) \) obtained from (2.4) satisfies

1) \[ ||\hat{\eta} - \eta||_2 \leq c_0 \lambda_n s_n^{a/2+1/2} q_n^{1/2}, \text{ with probability tending to one, for some constant } c_0 > 0 \]

2) \[ ||\hat{\eta} - \eta||_1 \leq c_1 \lambda_n s_n^{a/2+1} q_n, \text{ with probability tending to one, for some constant } c_1 > 0. \]

In particular, regarding the consistency property of the estimated regression curves \( \hat{\beta}_j(t) = \sum_{k=1}^{s_n} \hat{\eta}_{jk} b_k(t) \), it is straightforward to see that

\[
\sup_{j \leq p_n} ||\hat{\beta}_j - \beta_j||_{L^2} \leq \sup_{j \leq p_n} ||\hat{\eta}_j - \eta_j||_2 + s_n^{-\delta} \sup_{j \leq q_n} \left( \sum_{k=s_n+1}^{\infty} \eta_{jk} k^{2\delta} \right)^{1/2} = O(\lambda_n s_n^{a/2+1/2} q_n^{1/2} + s_n^{-\delta}).
\]

Next, we demonstrate Theorem 4, which lay the foundation for hypothesis testing through the multiplier bootstrap method.

**Theorem 4** Under conditions (A1)–(A10) and (P1)–(P5), by using the local minimizer \( \hat{\eta} \) from Theorem 3, then under \( H_0 : ||\beta_j||_{L^2} = 0 \) for all \( j \in \mathcal{H}_n \), the Kolmogorov distance between the distributions of \( ||\hat{T}||_{\infty} \) and \( ||\hat{T}_e||_{\infty} \) satisfies

\[
\lim_{n \to \infty} \sup_{t \geq 0} \left| P(||\hat{T}||_{\infty} \leq t) - P_e(||\hat{T}_e||_{\infty} \leq t) \right| = 0,
\]

and consequently,

\[
\lim_{n \to \infty} \sup_{\alpha \in (0,1)} \left| P(||\hat{T}||_{\infty} \leq c_B(\alpha)) - \alpha \right| = 0.
\]
2.4 Simulation Studies

The simulated data \( \{y_i, i = 1, \ldots, n\} \) are generated from the model

\[
y_i = \sum_{j=1}^{p_n} \int_0^1 \beta_j(t) x_{ij}(t) \, dt = \sum_{j=1}^{p_n} \sum_k \eta_{ijk} \theta_{ijk} + \epsilon_i,
\]

with \( p_n \) functional predictors, the errors \( \epsilon_1, \ldots, \epsilon_n \) are independent and identically distributed from \( N(0, \sigma^2) \), and the functional predictors have mean zero and covariance function derived from the Fourier basis \( \phi_1 = 1, \phi_{2\ell} = 2^{1/2} \cos\{\ell \pi (2t - 1)\}, \ell = 1, \ldots, 25 \) and \( \phi_{2\ell-1} = 2^{1/2} \sin\{((\ell - 1) \pi (2t - 1)\}, \ell = 2, \ldots, 25 \), and \( t \in T = [0, 1] \). The underlying regression function is \( \beta_j(t) = \sum_{k=1}^{50} \eta_{jk} \phi_k(t) \) for \( j \leq q_n \), where \( \eta_{jk} = c_j(1.2 - 0.2k) \) for \( k \leq 4 \) and \( \eta_{jk} = 0.4c_j(k - 3)^{-8} \) for \( 5 \leq k \leq 50 \), with some constants \( \{c_j : j \leq q_n\} \) that are chosen for different settings, and the remaining \( \beta_j(t) = 0 \). Next, we describe how to generate the \( p_n \) functional predictors \( x_{ij}(t) \). For \( j = 1, \ldots, p_n \), define \( V_{ij}(t) = \sum_{k=1}^{50} \tilde{\theta}_{ijk} \phi_k(t) \), where \( \{\tilde{\theta}_{ijk}\} \) follow independent distributed \( N(0, k^{-2}) \) for different \( i, j \) and \( k \). The \( p_n \) functional predictors are then defined through the AR(\( \rho \)) structure linear transformations

\[
x_{ij}(t) = \sum_{j'=1}^{p_n} \rho^{ij-j'} V_{ij'}(t) = \sum_{k=1}^{50} \sum_{j'=1}^{p_n} \rho^{ij-j'} \tilde{\theta}_{ij'k} \phi_k(t) = \sum_{k=1}^{50} \theta_{ijk} \phi_k(t),
\]

with \( \theta_{ijk} = \sum_{j'=1}^{p_n} \rho^{ij-j'} \tilde{\theta}_{ij'k} \), and the constant \( \rho \in (0, 1) \) controls the correlation among the functional predictors where we set \( \rho = 0.3 \) in our simulation study. For the actual observations, we assume they are realizations of \( \{x_{ij}(\cdot), j = 1, \ldots, p_n\} \) at 100 equally spaced times \( \{t_{ij}, l = 1, \ldots, 100\} \in T \). To be more realistic, we adopt a known orthonormal cubic spline basis containing \( m = 30 \) basis functions to fit each observed functional data to obtain the corresponding estimated coefficients rather than using the Fourier basis. After obtaining the estimated \( x_{ij}(t) \), we apply the algorithm in section 2.5.1 to get the estimated \( \beta_j(t) \) under SCAD penalty function. Then, we construct the test statistic and its associated \( \alpha = 5\% \) empirical quantile through wild bootstrap by using \( N = 1000 \) bootstrap samples to test on several alter-
CHAPATER 2. TESTING GENERAL HYPOTHESIS IN LARGE-SCALE FLR

native hypothesis. We summarize and compare the finite sample performance of several null hypothesis under various settings of regression curves specified by \( \{c_j : j \leq q_n\} \) based on the rejection proportion over 300 Monte Carlos in the table below.

### Table 2.1: Results for the several different combinations of null hypothesis and settings of \( \beta_j(t) \) that were measured over 300 simulations

<table>
<thead>
<tr>
<th>Settings of ( \beta_j(t) )</th>
<th>Null hypothesis ( H_0 )</th>
<th>Rejection proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n = 100 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_n = 200 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( q_n = 3 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \sigma^2 = 1 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( c_1 = 0, c_2 = 0, c_3 = 0 )</td>
<td>( H_0 : \beta_1(t) = 0 )</td>
<td>.0467</td>
</tr>
<tr>
<td>( c_1 = .1, c_2 = 0, c_3 = 0 )</td>
<td>( H_0 : \beta_1(t) = 0 )</td>
<td>.1067</td>
</tr>
<tr>
<td>( c_1 = .2, c_2 = 0, c_3 = 0 )</td>
<td>( H_0 : \beta_1(t) = 0 )</td>
<td>.2400</td>
</tr>
<tr>
<td>( c_1 = .6, c_2 = 0, c_3 = 0 )</td>
<td>( H_0 : \beta_1(t) = 0 )</td>
<td>.9133</td>
</tr>
<tr>
<td>( c_1 = 1 )</td>
<td>( H_0 : \beta_1(t) = 0 )</td>
<td>.9967</td>
</tr>
<tr>
<td>( c_2 = 1 )</td>
<td>( H_0 : \beta_1(t) = \cdots = \beta_5(t) = 0 )</td>
<td>1</td>
</tr>
<tr>
<td>( c_3 = 1 )</td>
<td>( H_0 : \beta_1(t) = \cdots = \beta_20(t) = 0 )</td>
<td>1</td>
</tr>
<tr>
<td>( H_0 : \beta_5(t) = \cdots = \beta_20(t) = 0 )</td>
<td>0.0533</td>
<td></td>
</tr>
</tbody>
</table>

From Table 2.1, we find that the rejection proportions of the first four null hypothesis, where \( c_2 \) and \( c_3 \) are held to be zero, are in ascending order as the signal of the first predictor increases, which is consistent with the shape of a power curve. In addition, we also find that the rejection proportion of the first and the last null hypothesis are close to the pre-specified significance level \( \alpha = 5\% \), which is to be expected since the null hypothesis is true based on the underlying model and the rejection proportion represents the average Type I error of the multiplier decorrelated score test. In particular, when the minimal signal of the significant functional predictors is strong enough to be detected (i.e. \( c_1 = c_2 = c_3 = 1 \)), we observe that the rejection proportions of the associated null hypothesis are close to one, which is to be expected. In conclusion, we summarize that the multiplier decorrelated score test performs well in various settings and this justify the validity of the proposed score test.


2.5 Proof of Lemmas and Main Theorems

2.5.1 Algorithm and parameter tuning

First, without loss of generality, we assume that the data are centered so that we have

\[ n^{-1} \sum_{i=1}^{n} Y_i = 0, \quad n^{-1} \sum_{i=1}^{n} \theta_{ijk} = 0, \]

for any \( j = 1, \ldots, p_n, k = 1, \ldots, s_n \). In addition, for each \( j = 1, \ldots, p_n \), we denote \( \hat{f}_j = \Theta_j \hat{\eta}_j \), where \( \hat{\eta}_j \) is an estimator for \( \eta_j \), and \( U_j = \Theta_j (\Theta_j' \Theta_j)^{-1} \Theta_j' \). The optimization of (2.4) can be achieved by adopting the coordinate descent method similar to those used in Ravikumar et al. [2008] and Fan et al. [2015] with a little modification, where \( \rho_{\lambda_n} (\cdot) \) is replaced by \( \rho_{\lambda_n s_n^{1/2}} (\cdot) \). In the following, we propose an algorithm corresponding to general penalty functions satisfying the technical conditions (P1)–(P5) in Appendix.

---

Algorithm for general \( \rho_{\lambda} (\cdot) \)

1. Start with the initial estimator \( \hat{f}_j = 0 \), for each \( j = 1, \ldots, p_n \).
2. Calculate the residual \( R_j = Y - \sum_{k \neq j} \hat{f}_k \), while fixing the values of \( \{ \hat{f}_k : k \neq j \} \).
3. Calculate the \( \hat{P}_j = U_j R_j \).
4. Let \( \hat{f}_j = \max \left\{ 1 - \rho'_{\lambda_n s_n^{1/2}} (n^{-1/2} ||\hat{f}_j||_2) n^{1/2} / ||\hat{P}_j||_2, 0 \right\} \hat{P}_j \).
5. Let \( \hat{f}_j = \hat{f}_j - n^{-1} 1_n' \hat{f}_j 1_n \), where \( 1_n \) denotes the \( n \times 1 \) vector of ones.
6. Repeat (ii) to (v) for \( j = 1, \ldots, p_n \) and iterate until convergence to obtain the final estimates \( \hat{f}_j \), for \( j = 1, \ldots, p_n \).
7. Compute \( \hat{\eta}_j = (\Theta_j' \Theta_j)^{-1} \Theta_j' \hat{f}_j \) by using the final estimates \( \hat{f}_j \) from step (vi) to get the final estimates \( \hat{\eta}_j \), for \( j = 1, \ldots, p_n \).

---

In addition, several tuning parameters play critical roles in the penalized procedure. The optimal values of the parameters \( \lambda_n, s_n \) will be automatically chosen by the method of cross validation.
2.5.2 Technical Conditions on Regularizers

We impose some conditions on the nonconvex penalty function $\rho_\lambda$, similar to those adopted in Loh and Wainwright [2015].

(P1) $\rho_\lambda$ is an even function, and $\rho_\lambda(0) = 0$.

(P2) For $t \geq 0$, $\rho_\lambda(t)$ is nondecreasing in $t$.

(P3) The function $g_\lambda(t) = \rho_\lambda(t)/t$ is nonincreasing in $t$, for $t > 0$.

(P4) The function $\rho_\lambda(t)$ is differentiable everywhere except at $t = 0$. At $t = 0$, we have

$$\lim_{t \to 0^+} \rho_\lambda'(t) = \lambda L,$$

for some positive constant $L$.

(P5) The function $\rho_{\lambda, \mu}(t)$ is convex in $t$, for some positive constant $\mu$, where $\rho_{\lambda, \mu}(t) = \rho_\lambda(t) + 2^{-1} \mu t^2$.

From Loh and Wainwright [2015], it is obvious that most nonconvex regularizers such as LASSO, SCAD and MCP meet those conditions. Note that Lemma 5 is devoted to summarizing the general properties of those penalty functions.

2.5.3 Notations

We summarize most of the notations that will be used throughout the paper as follows. First, for any vector $u = (u_1, \ldots, u_s)'$, we let $\|u\|_q = (\sum_{l=1}^s |u_l|^q)^{1/q}$ for $q \geq 1$, $\|u\|_\infty = \max_{1 \leq s} |u_l|$, and $\|u\|_0$ represents the cardinality of supp$(u)$, where supp$(u) = \{l : u_l \neq 0\}$. Moreover, if $S = \text{supp}(u)$, we let $u_S$ to be the vector containing only nonzeros of $u$, while $u_{S^c}$ contains only zero elements. For any matrix $C = [c_{ij}]$, we denote $\|C\|_\infty = \max_{i,j} |c_{ij}|$, if $C$ is symmetric, $\lambda_{\min}(C)$ and $\lambda_{\max}(C)$ are the minimum and maximum eigenvalues. For sequences $a_n$ and $b_n$, $a_n \sim b_n$ if and only if $c_1 \leq \lim_{n \to \infty} |a_n/b_n| \leq c_2$ for some positive constants $c_1$, $c_2$. For any sub-gaussian random variable $X$, define $\|X\|_{\phi_1} = \sup_{q \geq 1} q^{-1/2}(E|X|^q)^{1/q}$ as the sub-gaussian
norm, and for any sub-exponential random variable $X$, define $||X||_{\phi_2} = \sup_{q \geq 1} q^{-1}(E|X|^q)^{1/q}$ as the sub-exponential norm, similar to those adopted in Ning and Liu [2016].

Second, recall that $\mathcal{P}_n = \{1, \ldots, p_n\}$ is the index set representing all functional predictors, while $\mathcal{H}_n \subseteq \mathcal{P}_n$ is any nonempty subset of $\mathcal{P}_n$, with cardinality $|\mathcal{H}_n| = h_n \leq p_n$, and we also denote $\mathcal{H}^c_n = \mathcal{P}_n \setminus \mathcal{H}_n$. For each $j \leq p_n$, we let $\Lambda_j = \text{diag}\{\omega_{j1}^{1/2}, \ldots, \omega_{j\tilde{s}_n}^{1/2}\}$. Moreover, we also let $\Lambda = \Lambda_{\mathcal{P}_n}$ be the block diagonal matrix with $\{\Lambda_j : j \leq p_n\}$ as its diagonal submatrices and let $\Lambda_{\mathcal{H}_n}$ be the block diagonal matrix with $\{\Lambda_j : j \in \mathcal{H}_n\}$ as its diagonal submatrices. Similarly, we define $\hat{\Lambda}_j = \text{diag}\{\hat{\omega}_{j1}^{1/2}, \ldots, \hat{\omega}_{j\tilde{s}_n}^{1/2}\}$, $\hat{\Lambda} = \hat{\Lambda}_{\mathcal{P}_n}$ and $\hat{\Lambda}_{\mathcal{H}_n}$. Furthermore, we denote $\hat{\eta} = \Lambda \eta = (\hat{\eta}_1, \ldots, \hat{\eta}_{p_n})'$, with each $\hat{\eta}_j = \Lambda_j \eta_j = (\eta_{j1} \omega_{j1}^{1/2}, \ldots, \eta_{j\tilde{s}_n} \omega_{j\tilde{s}_n}^{1/2})'$. Analogously, we denote $\nu = \eta - \hat{\eta}$ and $\hat{\nu} = \hat{\eta} - \hat{\eta}^*$ = $\Lambda \nu$, where $\eta^*$ and $\hat{\eta}^*$ are true versions of $\eta$ and $\hat{\eta}$ respectively. For any $\mathcal{H}_n \subseteq \mathcal{P}_n$, $\hat{\nu}_{\mathcal{H}_n}$ is constructed by stacking the vectors $\{\hat{\nu}_j = \Lambda_j (\eta_j - \eta_j^*) : j \in \mathcal{H}_n\}$ in a column.

Third, without loss of generality, we let $\Theta = (G_1, \ldots, G_n)' = [\Theta_{\mathcal{H}_n}, \Theta_{\mathcal{H}_n^c}]$, $\Theta_{\mathcal{H}_n} = (E_1, \ldots, E_n)'$, $\Theta_{\mathcal{H}_n^c} = (F_1, \ldots, F_n)'$, where $\Theta_{\mathcal{H}_n}$ is constructed by stacking $\{\Theta_j : j \in \mathcal{H}_n\}$ in a row. Similarly, we denote $\check{\Theta} = (\check{G}_1, \ldots, \check{G}_n)' = [\check{\Theta}_{\mathcal{H}_n}, \check{\Theta}_{\mathcal{H}_n^c}] = \Theta \Lambda^{-1} = [\Theta_{\mathcal{H}_n} \Lambda^{-1}_{\mathcal{H}_n}, \Theta_{\mathcal{H}_n^c} \Lambda^{-1}_{\mathcal{H}_n^c}]$, $\check{\Theta}_{\mathcal{H}_n} = (\check{E}_1, \ldots, \check{E}_n)'$, $\check{\Theta}_{\mathcal{H}_n^c} = (\check{F}_1, \ldots, \check{F}_n)'$, where $\check{\Theta}_{\mathcal{H}_n}$ is constructed by stacking $\{\check{\Theta}_j = \Theta_j \Lambda^{-1}_j : j \in \mathcal{H}_n\}$ in a row.

Fourth, we denote a series of terms that can be constructed from the information matrix $I$.

\[
I = E(G_i G_i'), \quad I_{\mathcal{H}_n \mathcal{H}_n} = E(E_i E_i'), \quad I_{\mathcal{H}_n \mathcal{H}_n^c} = E(E_i F_i'), \quad I_{\mathcal{H}_n^c \mathcal{H}_n^c} = E(F_i F_i'),
\]

\[
w = I^{-1}_{\mathcal{H}_n \mathcal{H}_n} I_{\mathcal{H}_n \mathcal{H}_n^c} = (w_1, \ldots, w_{\tilde{s}_n \tilde{s}_n}), \quad \check{I} = E(\check{G}_i \check{G}_i') = \Lambda^{-1} I \Lambda^{-1},
\]

\[
\check{I}_{\mathcal{H}_n \mathcal{H}_n} = E(\check{E}_i \check{E}_i') = \Lambda^{-1}_{\mathcal{H}_n} I_{\mathcal{H}_n \mathcal{H}_n} \Lambda^{-1}_{\mathcal{H}_n}, \quad \check{I}_{\mathcal{H}_n \mathcal{H}_n^c} = E(\check{E}_i \check{F}_i') = \Lambda^{-1}_{\mathcal{H}_n} I_{\mathcal{H}_n \mathcal{H}_n^c} \Lambda^{-1}_{\mathcal{H}_n^c},
\]

\[
\check{I}_{\mathcal{H}_n^c \mathcal{H}_n^c} = E(\check{F}_i \check{F}_i') = \Lambda^{-1}_{\mathcal{H}_n^c} I_{\mathcal{H}_n^c \mathcal{H}_n^c} \Lambda^{-1}_{\mathcal{H}_n^c}, \quad \check{I}_{\mathcal{H}_n \mathcal{H}_n} = \Lambda^{-1}_{\mathcal{H}_n} I_{\mathcal{H}_n \mathcal{H}_n} \Lambda^{-1}_{\mathcal{H}_n^c}, \quad \check{I}_{\mathcal{H}_n \mathcal{H}_n^c} = \Lambda^{-1}_{\mathcal{H}_n} I_{\mathcal{H}_n \mathcal{H}_n^c} \Lambda^{-1}_{\mathcal{H}_n^c}, \quad \check{I}_{\mathcal{H}_n^c \mathcal{H}_n^c} = \Lambda^{-1}_{\mathcal{H}_n^c} I_{\mathcal{H}_n^c \mathcal{H}_n^c} \Lambda^{-1}_{\mathcal{H}_n^c},
\]

\[
\rho_n = \max_{t \leq \tilde{s}_n} ||w_t||_0 = \max_{t \leq \tilde{s}_n} \rho_{nt}, \quad \rho_{nl} = ||w_l||_0.
\]
Finally, by using the \( \hat{\eta} \) generated from (2.4) under the conditions of Theorem 3, we further denote

\[
\hat{T} = n^{-1/2} \sum_{i=1}^{n} \tilde{\Lambda}^{-1}_{\mathcal{H}_n}(\hat{w}'F_i - E_i)(Y_i - F_i'\hat{\eta}_{\mathcal{H}_n}) = n^{-1/2} \sum_{i=1}^{n} \hat{S}_i,
\]

\[
T^* = n^{-1/2} \sum_{i=1}^{n} \Lambda^{-1}_{\mathcal{H}_n}(w'F_i - E_i)\varepsilon_i = n^{-1/2} \sum_{i=1}^{n} S_i,
\]

\[
\hat{T}_e = n^{-1/2} \sum_{i=1}^{n} e_i \hat{S}_i, \quad T^*_e = n^{-1/2} \sum_{i=1}^{n} e_i S_i,
\]

\[
\phi'(\alpha) = \inf\{t \in \mathbb{R} : P_e(||\hat{T}_e||_{\infty} \leq t) \geq \alpha\}, \quad \alpha \in (0, 1),
\]

\[
\phi^*(\alpha) = \inf\{t \in \mathbb{R} : P_e(||T^*_e||_{\infty} \leq t) \geq \alpha\}, \quad \alpha \in (0, 1),
\]

\[
\hat{S}_i = (\hat{S}_{i1}, \ldots, \hat{S}_{i,h_n,s_n}) = \tilde{\Lambda}^{-1}_{\mathcal{H}_n}(\hat{w}'F_i - E_i)(Y_i - F_i'\hat{\eta}_{\mathcal{H}_n}),
\]

\[
S_i = (S_{i1}, \ldots, S_{i,h_n,s_n}) = \Lambda^{-1}_{\mathcal{H}_n}(w'F_i - E_i)\varepsilon_i,
\]

where \( \{e_i : i \leq n\} \) is a set of independent and identically distributed standard normal random variables independent of the data, \( P_e(\cdot) \) means the probability with respect to \( e \), \( \phi'(\alpha) \) denotes the \( \alpha \) quantile of \( ||\hat{T}_e||_{\infty} \), and \( \phi^*(\alpha) \) denotes the \( \alpha \) quantile of \( ||T^*_e||_{\infty} \). It is also worth to notice that \( \{S_i : i \leq n\} \) are iid centered random vectors such that \( E(S_iS_i') = \sigma^2 I_{\mathcal{H}_n|\mathcal{H}_n} \).

### 2.5.4 Auxiliary Lemmas

**Lemma 5**  
(a) Under conditions (P1)–(P4), we have \( |\rho_{\lambda}(t_1) - \rho_{\lambda}(t_2)| \leq \lambda L|t_1 - t_2| \), for all \( t_1, t_2 \in \mathbb{R} \).

(b) Under conditions (P1)–(P4), we have \( |\rho_{\lambda}'(t)| \leq \lambda L \), for all \( t \neq 0 \).

(c) Under conditions (P1)–(P5), we have \( \lambda L|t| \leq \rho_{\lambda}(t) + 2^{-1} \mu t^2 \), for all \( t \in \mathbb{R} \).

**Lemma 6**  
Under conditions (P1)–(P4), if \( P_{\lambda_n}(\eta^*) - P_{\lambda_n}(\eta) \geq 0 \), where
Under conditions (A2), (A6), (A8), we have $\eta^*$ is the true version of $\eta$, then we have

$$0 \leq P_{\lambda_n}(\eta^*) - P_{\lambda_n}(\eta) \leq \lambda_n s_n^{1/2} \left( \sum_{j \in A_n} n^{-1/2} \| \Theta_j (\eta_j - \eta_j^*) \|_2 - \sum_{j \in A_n^c} n^{-1/2} \| \Theta_j (\eta_j - \eta_j^*) \|_2 \right),$$

where $A_n \subseteq \mathcal{P}_n$ is the index set corresponding to the largest $q_n$ elements of $\{n^{-1/2} \| \Theta_j \eta_j \|_2 : j \leq p_n \}$ in magnitude, while $A_n^c = \mathcal{P}_n \setminus A_n$.

**Lemma 7** Under conditions (A2), (A6), (A8), we have

1) $\| \hat{\Lambda} - I \|_\infty \leq c_1 n^{\beta/2 - 1/2}$, with probability tending to one, for some constant $c_1 > 0$

2) $\| \hat{\Lambda} - I \|_\infty \leq c_2 n^{\beta/2 - 1/2}$, with probability tending to one, for some constant $c_2 > 0$,

where $I$ denotes the $p_n s_n \times p_n s_n$ identity matrix.

**Lemma 8** Under conditions (A1), (A2), (A6), (A8), we have

1) $\| n^{-1} \Theta' \Theta - E(G_t G_t') \|_\infty \leq c_0 n^{\beta/2 - 1/2}$, with probability tending to one, for some constant $c_0 > 0$

2) $\| n^{-1} \tilde{\Theta}' \tilde{\Theta} - E(\tilde{G}_t \tilde{G}_t') \|_\infty \leq c_1 n^{\beta/2 - 1/2}$, with probability tending to one, for some constant $c_1 > 0$

3) $\| n^{-1} \tilde{\Theta}' \tilde{\epsilon} \|_\infty \leq c_2 n^{\beta/2 - 1/2}$, with probability tending to one, for some constant $c_2 > 0$

4) $\| n^{-1} \tilde{\Theta}' \tilde{\epsilon} \|_\infty \leq c_3 n^{\beta/2 - 1/2}$, with probability tending to one, for some constant $c_3 > 0$

5) $\max_{t \leq h_n s_n} \| n^{-1} \sum_{i=1}^n (w_i F_i - E_{it}) F_i' \|_\infty \leq c_4 n^{\beta/2 - 1/2}$, with probability tending to one, for some constant $c_4 > 0$

6) $\max_{t \leq h_n s_n} \| n^{-1} \sum_{i=1}^n (w_i F_i - E_{it}) [E(E_{it}^2)]^{-1/2} F_i' \|_\infty \leq c_5 n^{\beta/2 - 1/2}$, with probability tending to one, for some constant $c_5 > 0$

7) $\max_{t \leq h_n s_n} \| n^{-1} \sum_{i=1}^n (w_i F_i - E_{it}) \epsilon_i \|_\infty \leq c_6 n^{\beta/2 - 1/2}$, with probability tending to one, for some constant $c_6 > 0$
8) $\max_{l \leq h, s_0} \left| n^{-1} \sum_{i=1}^{n} (w_i' F_i - E_i) \left[ E_i^2 \right]^{-1/2} \epsilon_i \right| \leq c_7 n^{\beta/2-1/2}$, with probability tending to one, for some constant $c_7 > 0$.

where $\epsilon$ denotes the $n \times 1$ random error vector.

**Lemma 9** Under conditions (A1), (A2), (A6)-(A8), and $H_0 : \beta_j(t) = 0$ for all $j \in \mathcal{H}_n$, we have

1) $\| n^{-1} \hat{\Theta}' \mathcal{H}_n (Y - \Theta \mathcal{H}_n \eta \mathcal{H}_n) \|_\infty \leq c_0 (n^{\beta/2-1/2} + q_n s_n^{-\delta})$, with probability tending to one, for some constant $c_0 > 0$

2) $\| n^{-1} \Theta \mathcal{H}_n ^t (Y - \Theta \mathcal{H}_n \eta \mathcal{H}_n) \|_\infty \leq c_1 (n^{\beta/2-1/2} + q_n s_n^{-\delta})$, with probability tending to one, for some constant $c_1 > 0$.

**Lemma 10** Under conditions (A1), (A2), (A4), (A6), (A8), we have

1) $n^{-1} \sum_{j=1}^{p_n} \| \Theta_j (\eta_j - \eta^*_j) \|^2 \|_2 \leq m_1 [1 + o(1)] \| \hat{\nu} \|_2^2$, with probability tending to one, where the constant $m_1$ is defined in (A4)

2) $\| \hat{\nu} \|_1 \leq c_0 s_n^{1/2} \sum_{j=1}^{p_n} n^{-1/2} \| \Theta_j (\eta_j - \eta^*_j) \|_2$, with probability tending to one, for some constant $c_0 > 0$

3) $\lambda_n \| \hat{\nu} \|_1 \leq c_1 [P_{\lambda_n}(\eta^*) + P_{\lambda_n}(\eta) + \| \hat{\nu} \|_2^2]$, with probability tending to one, for some constant $c_1 > 0$.

Recall that $\hat{\nu} = \Lambda \nu = \hat{\eta} - \hat{\eta}^* = \Lambda (\eta - \eta^*)$.

**Lemma 11** Under conditions (A1)–(A4), (A6), (A8), (A10), we have

1) $\max_{l \leq h, s_0} \left| n^{-1} \sum_{i=1}^{n} F_i F_i' \hat{w}_l - w_l \right| \|_\infty \leq c_0 n^{\beta/2-1/2}$, with probability tending to one, for some constant $c_0 > 0$

2) $\max_{l \leq h, s_0} \left| \hat{w}_l - w_l \right|_2 \leq c_1 s_n^{1/2} n^{\beta/2-1/2}$, with probability tending to one, for some constant $c_1 > 0$.
3) \[ \max_{t \leq h_n} ||\hat{w}_t - w_t||_1 \leq c_2 \rho_n s_n \beta/2 - 1/2, \] with probability tending to one, for some constant \( c_2 > 0 \).

Recall that \( \rho_n = \max_{t \leq h_n} ||w_t||_0 = \max_{t \leq h_n} \rho_{nl}, \quad \rho_{nl} = ||w_t||_0. \)

Lemma 12 Under conditions (A1)–(A4), (A6)–(A10), and (P1)–(P5), we have

1) \[ \max_{t \leq h_n} \left| n^{-1/2} \sum_{i=1}^{n} (\hat{S}_{it} - S_{it}) \right| \leq c_0 (n^{-1/2} \rho_n s_n^{3\beta/2} + \lambda_n n^{\beta/2} q_n s_n^{\alpha+1} + n^{\beta/2+1/2} q_n s_n^{\delta} \log s_n + n^{\beta} \rho_n q_n s_n^{3\alpha/2-\delta} \log s_n), \] with probability tending to one, for some constant \( c_0 > 0 \)

2) \[ \max_{t \leq h_n} \left[ n^{-1} \sum_{i=1}^{n} (\hat{S}_{it} - S_{it})^2 \right]^{1/2} \leq c_1 \left[ \lambda_n n^{\beta} q_n s_n^{\alpha+1} + n^{\beta/2} q_n s_n^{\delta} \log s_n + \rho_n s_n^{3\alpha/2} n^{\beta-1/2} (\log n)^{1/2} + \lambda_n \rho_n q_n s_n^{\alpha+1} n^{3\beta/2-1/2} + n^{\beta-1/2} \rho_n q_n s_n^{3\alpha/2-\delta} \log s_n \right], \] with probability tending to one, for some constant \( c_1 > 0 \).

Recall that we denote

\[ \hat{S}_i = (\hat{S}_{i1}, \ldots, \hat{S}_{ih_n s_n}) = \hat{\Lambda}_n^{-1}(\hat{w}'F_i - E_i)(Y_i - F_i'\hat{\eta}_{H_i}), \]
\[ S_i = (S_{i1}, \ldots, S_{ih_n s_n}) = \Lambda_n^{-1}(w'F_i - E_i)\epsilon_i. \]

2.5.5 Proof of Lemmas

Proof of Lemma 5. For part (a), if \( |t_1| = |t_2| \), then by condition (P1), it is easy to see that

\[ |\rho_\lambda(t_1) - \rho_\lambda(t_2)| = |\rho_\lambda(|t_1|) - \rho_\lambda(|t_2|)| = 0 \leq \lambda L |t_1 - t_2|. \] If \( |t_1| \neq |t_2| \), then, without loss of generality, we assume \( |t_1| > |t_2| \). By conditions (P1)–(P3), we have

\[ |\rho_\lambda(t_1) - \rho_\lambda(t_2)| \leq (|t_1| - |t_2|)\{\rho_\lambda(|t_1| - |t_2|)/(|t_1| - |t_2|)\} \leq |t_1 - t_2| \lim_{t \to 0^+} \{\rho_\lambda(t)/t\}. \]

Given conditions (P1), (P3) and (P4), it is easy to see that \( \lim_{t \to 0^+} \{-\rho_\lambda(t)/t\} = \lambda L. \) It follows that \( |\rho_\lambda(t_1) - \rho_\lambda(t_2)| \leq \lambda L |t_1 - t_2| \), which completes the proof of part (a).

For part (b), by definition, for all \( t \neq 0 \), we have \( |\rho_\lambda'(t)| = |\lim_{\Delta \to 0} ((\rho_\lambda(t + \Delta) - \rho_\lambda(t))/\Delta)|. \) By Lemma 5(a), we have \( |\rho_\lambda'(t)| \leq \lambda L, \) which completes the proof of part (b).
For part (c), if \( t = 0 \), the inequality holds trivially. If \( t \neq 0 \), without loss of generality, we let \( t > 0 \). Since \( \rho_{\lambda,\mu}(s) = \rho_{\lambda}(s) + 2^{-1}\mu s^2 \) is convex for \( s \in [0,t] \), and is differentiable for \( s \in (0,t] \), it is easy to see that \( \rho_{\lambda,\mu}(t) - \rho_{\lambda,\mu}(0) \geq t\lim_{s_1 \to 0^+} \{ \rho_{\lambda}(s_1) + \mu s_1 \} = t\lambda L \), which completes the proof of part (c).

**Proof of Lemma 6.** First, we define \( f_n(t) = t/\rho_{\lambda_n,s_n^{1/2}}(t) \) for \( t > 0 \), and \( f_n(t) = (\lambda_n s_n^{1/2} L)^{-1} \) for \( t = 0 \). Under conditions (P1)–(P4), it is easy to see that \( f_n(t) \) is nondecreasing in \( t \) for \( t \geq 0 \). Then, it follows that

\[
\sum_{j \in A_n} \rho_{\lambda_n,s_n^{1/2}}^{-1}(n^{-1/2}\|\Theta_j(\eta_j - \eta_j^*)\|_2) = \sum_{j \in A_n} n^{-1/2}\|\Theta_j(\eta_j - \eta_j^*)\|_2 \left[ f_n(n^{-1/2}\|\Theta_j(\eta_j - \eta_j^*)\|_2) \right]^{-1}
\]

\[
\geq \left[ f_n(\max_{j \in A_n} n^{-1/2}\|\Theta_j(\eta_j - \eta_j^*)\|_2) \right]^{-1} \sum_{j \in A_n} n^{-1/2}\|\Theta_j(\eta_j - \eta_j^*)\|_2. \tag{2.12}
\]

Moreover, we also have

\[
\sum_{j \in A_n} \rho_{\lambda_n,s_n^{1/2}}^{-1}(n^{-1/2}\|\Theta_j(\eta_j - \eta_j^*)\|_2) = \sum_{j \in A_n} n^{-1/2}\|\Theta_j(\eta_j - \eta_j^*)\|_2 \left[ f_n(n^{-1/2}\|\Theta_j(\eta_j - \eta_j^*)\|_2) \right]^{-1}
\]

\[
\leq \left[ f_n(\max_{j \in A_n} n^{-1/2}\|\Theta_j(\eta_j - \eta_j^*)\|_2) \right]^{-1} \sum_{j \in A_n} n^{-1/2}\|\Theta_j(\eta_j - \eta_j^*)\|_2. \tag{2.13}
\]
Combining (2.12), (2.13) and $P_{n}(\eta^{*}) - P_{n}(\eta) \geq 0$, we have

$$0 \leq P_{n}(\eta^{*}) - P_{n}(\eta) = \sum_{j=1}^{p_{n}} \rho_{n}s_{n}^{-1/2} (n^{-1/2} ||\Theta_{j} \eta_{j}^{*}||_{2}) - \sum_{j=1}^{p_{n}} \rho_{n}s_{n}^{-1/2} (n^{-1/2} ||\Theta_{j} \eta_{j}||_{2})$$

$$= \sum_{j=1}^{q_{n}} \rho_{n}s_{n}^{-1/2} (n^{-1/2} ||\Theta_{j} \eta_{j}^{*}||_{2}) - \sum_{j=1}^{p_{n}} \rho_{n}s_{n}^{-1/2} (n^{-1/2} ||\Theta_{j} \eta_{j}||_{2})$$

$$\leq \sum_{j=1}^{q_{n}} \rho_{n}s_{n}^{-1/2} (n^{-1/2} ||\Theta_{j} (\eta_{j} - \eta_{j}^{*})||_{2}) + \sum_{j=1}^{q_{n}} \rho_{n}s_{n}^{-1/2} (n^{-1/2} ||\Theta_{j} \eta_{j}||_{2}) - \sum_{j=1}^{p_{n}} \rho_{n}s_{n}^{-1/2} (n^{-1/2} ||\Theta_{j} \eta_{j}||_{2})$$

$$= \sum_{j=1}^{q_{n}} \rho_{n}s_{n}^{-1/2} (n^{-1/2} ||\Theta_{j} (\eta_{j} - \eta_{j}^{*})||_{2}) - \sum_{j=q_{n}+1}^{p_{n}} \rho_{n}s_{n}^{-1/2} (n^{-1/2} ||\Theta_{j} \eta_{j}||_{2})$$

$$\leq \sum_{j \in A_{n}} \rho_{n}s_{n}^{-1/2} (n^{-1/2} ||\Theta_{j} (\eta_{j} - \eta_{j}^{*})||_{2}) - \sum_{j \in A_{n}} \rho_{n}s_{n}^{-1/2} (n^{-1/2} ||\Theta_{j} (\eta_{j} - \eta_{j}^{*})||_{2})$$

$$\leq \left[ \left( \sum_{j \in A_{n}} n^{-1/2} ||\Theta_{j} (\eta_{j} - \eta_{j}^{*})||_{2} \right) \right]^{-1} \left( \sum_{j \in A_{n}} n^{-1/2} ||\Theta_{j} \eta_{j}||_{2} - \sum_{j \in A_{n}} n^{-1/2} ||\Theta_{j} \eta_{j}^{*}||_{2} \right)$$

$$\leq \lambda_{n}s_{n}^{-1/2} \left( \sum_{j \in A_{n}} n^{-1/2} ||\Theta_{j} (\eta_{j} - \eta_{j}^{*})||_{2} - \sum_{j \in A_{n}} n^{-1/2} ||\Theta_{j} (\eta_{j} - \eta_{j}^{*})||_{2} \right),$$

which completes the proof.

**Proof of Lemma 7.** First, under condition (A2) and by Bernstein inequality, it is easy to see that for any $t > 0$,

$$P\left(|n^{-1} \sum_{i=1}^{n} (\omega_{jk}^{-1} \theta_{ijk}^{2} - 1)| \geq t\right) \leq 2 \exp \left( -c_{0}\min\{c_{1}^{-2}t^{2}, c_{1}^{-1}t\}n \right),$$

for some universal constants $c_{0}, c_{1} > 0$, uniformly in $j = 1, \ldots, p_{n}$, $k = 1, \ldots, s_{n}$. It then follows from union bound inequality that

$$P\left( \max_{j \leq p_{n}} \max_{k \leq s_{n}} |n^{-1} \sum_{i=1}^{n} (\omega_{jk}^{-1} \theta_{ijk}^{2} - 1)| \geq t\right) \leq 2p_{n}s_{n} \exp \left( -c_{0}\min\{c_{1}^{-2}t^{2}, c_{1}^{-1}t\}n \right),$$

by choosing $t = c_{2} [\log(p_{n}s_{n})/n]^{1/2}$, for some sufficiently large $c_{2} > 0$, it is easy to see that

$$\max_{j \leq p_{n}} \max_{k \leq s_{n}} |n^{-1} \sum_{i=1}^{n} (\omega_{jk}^{-1} \theta_{ijk}^{2} - 1)| \leq c_{2} [\log(p_{n}s_{n})/n]^{1/2},$$
with probability tending to one. Accordingly, it is easy to see that

\[ ||\hat{\Lambda}^{-1} - I||_\infty = \max_{j \leq p_n} \max_{k \leq s_n} |(n^{-1} \sum_{i=1}^{n} \omega_{jk}^{-1} \theta_{ijk}^2)^{1/2} - 1| \]

\[ \leq \max_{j \leq p_n} \max_{k \leq s_n} n^{-1} \sum_{i=1}^{n} |(\omega_{jk}^{-1} \theta_{ijk}^2 - 1)| \leq c_2 [\log(p_n s_n)/n]^{1/2}, \]

with probability tending to one. Under (A6) and (A8), it is obvious that \( ||\hat{\Lambda}^{-1} - I||_\infty \leq c_3 n^{\beta/2 - 1/2} \), with probability tending to one, for some constant \( c_3 > 0 \). Moreover, we have

\[ ||\hat{\Lambda}^{-1} - I||_\infty = \max_{j \leq p_n} \max_{k \leq s_n} |\omega_{jk}^{1/2} (n^{-1} \sum_{i=1}^{n} \theta_{ijk}^2)^{-1/2} - 1| \]

\[ \leq [\max_{j \leq p_n} \max_{k \leq s_n} \omega_{jk}^{1/2} (n^{-1} \sum_{i=1}^{n} \theta_{ijk}^2)^{-1/2}] ||\hat{\Lambda}^{-1} - I||_\infty \]

\[ \leq (||\hat{\Lambda}^{-1} - I||_\infty + 1) ||\hat{\Lambda}^{-1} - I||_\infty, \]

which further implies that

\[ ||\hat{\Lambda}^{-1} - I||_\infty \leq ||\hat{\Lambda}^{-1} - I||_\infty / (1 - ||\hat{\Lambda}^{-1} - I||_\infty) \]

\[ \leq c_3 n^{\beta/2 - 1/2} / (1 - c_3 n^{\beta/2 - 1/2}) \]

\[ \leq 2c_3 n^{\beta/2 - 1/2}, \]

with probability tending to one, which completes the proof.

**Proof of Lemma 8.** By using Bernstein inequality and union bound inequality repeatedly, the Lemma holds trivially.
Proof of Lemma 9. First, under \( H_0 : \beta_j(t) = 0 \) for all \( j \in \mathcal{H}_n \), it is easy to see that

\[
||n^{-1} \hat{\Theta}'_{\mathcal{H}_n} (Y - \Theta_{\mathcal{H}_n} \eta_{\mathcal{H}_n})||_\infty = ||n^{-1} \sum_{i=1}^{n} \tilde{F}_i (Y_i - F_i' \eta_{\mathcal{H}_n})||_\infty
\]

\[
= ||n^{-1} \sum_{i=1}^{n} \tilde{F}_i \epsilon_i + n^{-1} \sum_{i=1}^{n} \tilde{F}_i \left( \sum_{j \in \mathcal{H}_n} \sum_{k = s_n + 1}^{\infty} \theta_{ijk} \eta_{jk} \right)||_\infty
\]

\[
= ||n^{-1} \sum_{i=1}^{n} \tilde{F}_i \epsilon_i + n^{-1} \sum_{i=1}^{n} \sum_{j \in \mathcal{H}_n} \sum_{k = s_n + 1}^{\infty} \left[ \tilde{F}_i \theta_{ijk} - E(\tilde{F}_i \theta_{ijk}) \right] \eta_{jk} + n^{-1} \sum_{i=1}^{n} \sum_{j \in \mathcal{H}_n} \sum_{k = s_n + 1}^{\infty} E(\tilde{F}_i \theta_{ijk}) \eta_{jk}||_\infty
\]

\[
\leq ||n^{-1} \sum_{i=1}^{n} \tilde{F}_i \epsilon_i + n^{-1} \sum_{i=1}^{n} \sum_{j \in \mathcal{H}_n} \sum_{k = s_n + 1}^{\infty} \left[ \tilde{F}_i \theta_{ijk} - E(\tilde{F}_i \theta_{ijk}) \right] \eta_{jk}||_\infty + n^{-1} \sum_{i=1}^{n} \sum_{j \in \mathcal{H}_n} \sum_{k = s_n + 1}^{\infty} E(\tilde{F}_i \theta_{ijk}) \eta_{jk}||_\infty.
\]

(2.14)

In addition, for each \( l = 1, \ldots, (p_n - h_n)s_n \), we have

\[
E[(n^{-1} \sum_{i=1}^{n} \tilde{F}_{il} \epsilon_i)^2] = n^{-1} \sigma^2 \sim n^{-1}.
\]

(2.15)
Moreover, for each \( l = 1, \ldots, (p_n - h_n)s_n \), we have

\[
E\left( n^{-1} \sum_{i=1}^{n} \sum_{j \in H_n^l} \sum_{k=s_n+1}^{\infty} \left[ \tilde{F}_{il} \theta_{ijk} - E(\tilde{F}_{il} \theta_{ijk}) \right] \eta_{jk} \right)^2
\]

\[
=n^{-2} \sum_{i=1}^{n} E\left( \sum_{j \in H_n^l} \sum_{k=s_n+1}^{\infty} \left[ \tilde{F}_{il} \theta_{ijk} - E(\tilde{F}_{il} \theta_{ijk}) \right] \eta_{jk} \right)^2
\]

\[
\leq n^{-2} \sum_{i=1}^{n} E\left( \sum_{j=1}^{q_n} \sum_{k=s_n+1}^{\infty} \left[ \tilde{F}_{il} \theta_{ijk} - E(\tilde{F}_{il} \theta_{ijk}) \right] \eta_{jk} \right)^2
\]

\[
\leq n^{-2} q_n \sum_{i=1}^{n} q_n \sum_{j=1}^{\infty} \left( \sum_{k=s_n+1}^{\infty} \eta_{jk}^2 k^{2\delta} \right) \left( \sum_{k_1=s_n+1}^{\infty} \text{Var}(\tilde{F}_{il} \theta_{ij1k_1}) k_1^{-2\delta} \right)
\]

\[
\leq n^{-2} q_n \sum_{i=1}^{n} \sum_{j=1}^{\infty} \sum_{k=s_n+1}^{\infty} \left( \sum_{k_1=s_n+1}^{\infty} \eta_{jk}^2 k^{2\delta} \right) \left( \sum_{k_1=s_n+1}^{\infty} E(\tilde{F}_{il}^2 \theta_{ij1k_1}^2) k_1^{-2\delta} \right)
\]

\[
= n^{-2} q_n \sum_{i=1}^{n} \sum_{j=1}^{q_n} \left( \sum_{k=s_n+1}^{\infty} \eta_{jk}^2 k^{2\delta} \right) \left( \sum_{k_1=s_n+1}^{\infty} E(\tilde{F}_{il}^2 \theta_{ij1k_1}^2) \omega_{jk1} k_1^{-2\delta} \right)
\]

\[
\leq cn^{-1} q_n^2 s_n^{-2\delta} = o(n^{-1}), \quad (2.16)
\]

where the last inequality is by (A1), (A2), (A7) and (A8). Therefore, by combining (2.15) with (2.16), it is easy to see that for any \( l = 1, \ldots, (p_n - h_n)s_n \), there exists a universal constant \( c_0 > 0 \) such that

\[
\lim_{n \to \infty} P\left( |n^{-1} \sum_{i=1}^{n} \sum_{j \in H_n^l} \sum_{k=s_n+1}^{\infty} \left[ \tilde{F}_{il} \theta_{ijk} - E(\tilde{F}_{il} \theta_{ijk}) \right] \eta_{jk} | \leq c_0 |n^{-1} \sum_{i=1}^{n} \tilde{F}_{il} \epsilon_i | \right) = 1,
\]
which implies that for any \( t > 0 \),

\[
\lim_{n \to \infty} P\left( |n^{-1} \sum_{i=1}^{n} \tilde{F}_{it} \epsilon_i + n^{-1} \sum_{i=1}^{n} \sum_{j \in \mathcal{H}_n^k} \sum_{k=s_n+1}^{\infty} \left[ \tilde{F}_{it} \theta_{ijk} - E(\tilde{F}_{it} \theta_{ijk}) \right] \eta_{jk} | \geq t \right)
\]

\[
\leq \lim_{n \to \infty} P\left( |n^{-1} \sum_{i=1}^{n} \tilde{F}_{it} \epsilon_i | \geq t \right) + \lim_{n \to \infty} P\left( |n^{-1} \sum_{i=1}^{n} \sum_{j \in \mathcal{H}_n^k} \sum_{k=s_n+1}^{\infty} \left[ \tilde{F}_{it} \theta_{ijk} - E(\tilde{F}_{it} \theta_{ijk}) \right] \eta_{jk} | \geq t \right)
\]

\[
\leq \lim_{n \to \infty} P\left( |n^{-1} \sum_{i=1}^{n} \tilde{F}_{it} \epsilon_i | \geq t \right) + \lim_{n \to \infty} P\left( |n^{-1} \sum_{i=1}^{n} \tilde{F}_{it} \epsilon_i | \geq c_0^{-1} t \right)
\]

\[
\leq 2 \lim_{n \to \infty} P\left( |n^{-1} \sum_{i=1}^{n} \tilde{F}_{it} \epsilon_i | \geq c_1 t \right),
\] (2.17)

where \( c_1 = \min\{1, c_0^{-1}\} > 0 \). Moreover, under (A2) and by Bernstein inequality, we have that for any \( l = 1, \ldots, (p_n - h_n)s_n \), and any \( t > 0 \),

\[
\lim_{n \to \infty} P\left( |n^{-1} \sum_{i=1}^{n} \tilde{F}_{it} \epsilon_i | \geq c_1 t \right) \leq \lim_{n \to \infty} 2 \exp \left( - c_2 \min\{c_3^{-2} t^2, c_3^{-1} t\} n \right),
\] (2.18)

where \( c_2 \) and \( c_3 \) are some universal positive constants. By combining (2.17) with (2.18), we have that for any \( l = 1, \ldots, (p_n - h_n)s_n \), and any \( t > 0 \),

\[
\lim_{n \to \infty} P\left( |n^{-1} \sum_{i=1}^{n} \tilde{F}_{it} \epsilon_i + n^{-1} \sum_{i=1}^{n} \sum_{j \in \mathcal{H}_n^k} \sum_{k=s_n+1}^{\infty} \left[ \tilde{F}_{it} \theta_{ijk} - E(\tilde{F}_{it} \theta_{ijk}) \right] \eta_{jk} | \geq t \right)
\]

\[
\leq 2 \lim_{n \to \infty} P\left( |n^{-1} \sum_{i=1}^{n} \tilde{F}_{it} \epsilon_i | \geq c_1 t \right) \leq \lim_{n \to \infty} 4 \exp \left( - c_2 \min\{c_3^{-2} t^2, c_3^{-1} t\} n \right),
\]

invoking union bound inequality, we have that for any \( t > 0 \),

\[
\lim_{n \to \infty} P\left( ||n^{-1} \sum_{i=1}^{n} \tilde{F}_{it} \epsilon_i + n^{-1} \sum_{i=1}^{n} \sum_{j \in \mathcal{H}_n^k} \sum_{k=s_n+1}^{\infty} \left[ \tilde{F}_{it} \theta_{ijk} - E(\tilde{F}_{it} \theta_{ijk}) \right] \eta_{jk} ||_{\infty} \geq t \right)
\]

\[
\leq \lim_{n \to \infty} 4p_n s_n \exp \left( - c_2 \min\{c_3^{-2} t^2, c_3^{-1} t\} n \right),
\]
by choosing \( t = c_4 \log(p_n s_n)/n \)^{1/2}, for some sufficiently large \( c_4 > 0 \), it is easy to see that

\[
\| n^{-1} \sum_{i=1}^{n} \tilde{F}_i \epsilon_i + n^{-1} \sum_{i=1}^{n} \sum_{j \in H_n} \sum_{k = s_n+1}^{\infty} [\tilde{F}_i \theta_{ijk} - E(\tilde{F}_i \theta_{ijk})] \eta_{jk} \|_\infty \\
\leq c_4 \log(p_n s_n)/n \)^{1/2} \leq c_5 n^{\beta/2-1/2}
\]

(2.19)

with probability tending to one, for some \( c_5 > 0 \). Furthermore, we have

\[
\| n^{-1} \sum_{i=1}^{n} \sum_{j \in H_n} \sum_{k = s_n+1}^{\infty} E(\tilde{F}_i \theta_{ijk}) \eta_{jk} \|_\infty \leq \max_{t \leq (p_n - h_n)s_n} \| n^{-1} \sum_{i=1}^{n} \sum_{j \in H_n} \sum_{k = s_n+1}^{\infty} E(\tilde{F}_i \theta_{ijk}) \eta_{jk} \| \\
\leq \max_{t \leq (p_n - h_n)s_n} n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{q_n} \sum_{k = s_n+1}^{\infty} E(\tilde{F}_i \theta_{ijk}) \eta_{jk} \\
\leq \max_{t \leq (p_n - h_n)s_n} n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{q_n} \left( \sum_{k = s_n+1}^{\infty} \left[ E(\tilde{F}_i \theta_{ijk}) \right]^2 k^{-2\delta} \right)^{1/2} \left( \sum_{k = s_n+1}^{\infty} \eta_{jk}^2 k_1^{2\delta} \right)^{1/2} \\
\leq n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{q_n} \left( \sum_{k = s_n+1}^{\infty} \omega_{jk} k^{-2\delta} \right)^{1/2} \left( \sum_{k = s_n+1}^{\infty} \eta_{jk}^2 k_1^{2\delta} \right)^{1/2} \leq c_6 q_n s_n^{-\delta}
\]

(2.20)

for some constant \( c_6 > 0 \), where the last inequality is by (A1) and (A7). By combining (2.12), (2.19) with (2.20), we conclude that \( \| n^{-1} \tilde{\Theta}'_{H_n}(Y - \Theta_{H_n} \eta_{H_n}) \|_\infty \leq c_7 (n^{\beta/2-1/2} + q_n s_n^{-\delta}) \), with probability tending to one, for some constant \( c_7 > 0 \), which completes the proof of part 1). For part 2), it is easy to observe that

\[
\| n^{-1} \Theta'_{H_n}(Y - \Theta_{H_n} \eta_{H_n}) \|_\infty = \| n^{-1} \Lambda_{H_n} \tilde{\Theta}'_{H_n}(Y - \Theta_{H_n} \eta_{H_n}) \|_\infty \\
\leq \lambda_{\max}(\Lambda_{H_n}) \| n^{-1} \tilde{\Theta}'_{H_n}(Y - \Theta_{H_n} \eta_{H_n}) \|_\infty \\
\leq c_8 (n^{\beta/2-1/2} + q_n s_n^{-\delta})
\]

with probability tending to one, for some constant \( c_8 > 0 \), which completes the proof.
Proof of Lemma 10. First, it is easy to see that

\[ n^{-1} \sum_{j=1}^{p_n} \| \Theta_j (\eta_j - \eta_j^*) \|^2 = n^{-1} \sum_{j=1}^{p_n} \| \tilde{\Theta}_j (\tilde{\eta}_j - \tilde{\eta}_j^*) \|^2 = \sum_{j=1}^{p_n} (\tilde{\eta}_j - \tilde{\eta}_j^*)' (n^{-1} \tilde{\Theta}_j' \tilde{\Theta}_j) (\tilde{\eta}_j - \tilde{\eta}_j^*) \]

\[ \leq \lambda_{\text{max}}(\tilde{I}) \| \tilde{\nu} \|^2 + \sum_{j=1}^{p_n} (\tilde{\eta}_j - \tilde{\eta}_j^*)' [n^{-1} \tilde{\Theta}_j' \tilde{\Theta}_j - E(n^{-1} \tilde{\Theta}_j' \tilde{\Theta}_j)] (\tilde{\eta}_j - \tilde{\eta}_j^*) \]

\[ \leq \lambda_{\text{max}}(\tilde{I}) \| \tilde{\nu} \|^2 + \| n^{-1} \tilde{\Theta}' \tilde{\Theta} - E(n^{-1} \tilde{\Theta}' \tilde{\Theta}) \|_{\infty} \sum_{j=1}^{p_n} \| \tilde{\eta}_j - \tilde{\eta}_j^* \|^2 \]

\[ \leq [\lambda_{\text{max}}(\tilde{I}) + s_n \| n^{-1} \tilde{\Theta}' \tilde{\Theta} - E(n^{-1} \tilde{\Theta}' \tilde{\Theta}) \|_{\infty}] \| \tilde{\nu} \|^2 \]

\[ \leq (m_1 + c_2 s_n n^{\beta/2 - 1/2}) \| \tilde{\nu} \|^2, \]

for some constant $c_2 > 0$, where the last inequality is by (A4) and Lemma 8. Since $s_n n^{\beta/2 - 1/2} = o(1)$, it follows that $n^{-1} \sum_{j=1}^{p_n} \| \Theta_j (\eta_j - \eta_j^*) \|^2 \leq m_1 [1 + o(1)] \| \tilde{\nu} \|^2$, with probability tending
to one, which completes the proof of part 1). Moreover, we have

\[
\|\hat{\nu}\|_1 = \sum_{j=1}^{p_n} \sum_{k=1}^{s_n} |\hat{\nu}_{jk} - \hat{\nu}_{jk}^*| \leq s_n^{1/2} \sum_{j=1}^{p_n} \|\hat{\nu}_j - \hat{\nu}_j^*\|_2
\]

\[
= \left[\lambda_{\min}(\tilde{I})\right]^{-1/2} s_n^{1/2} \sum_{j=1}^{p_n} \left[\lambda_{\min}(\tilde{I})\right]^{1/2} \|\hat{\nu}_j - \hat{\nu}_j^*\|_2
\]

\[
\leq m_0^{-1/2} s_n^{1/2} \sum_{j=1}^{p_n} \left[\lambda_{\min}(\tilde{I})\right]^{1/2} \|\hat{\nu}_j - \hat{\nu}_j^*\|_2 \leq m_0^{-1/2} s_n^{1/2} \sum_{j=1}^{p_n} \left[(\hat{\nu}_j - \hat{\nu}_j^*)' E(n^{-1} \hat{\Theta}_j' \hat{\Theta}_j) (\hat{\nu}_j - \hat{\nu}_j^*)\right]^{1/2}
\]

\[
\leq m_0^{-1/2} s_n^{1/2} \sum_{j=1}^{p_n} \left(n^{-1} \|\Theta_j(\eta_j - \eta_j^*)\|_2^2 + \|E(n^{-1} \hat{\Theta}_j' \hat{\Theta}_j) - n^{-1} \hat{\Theta}_j' \hat{\Theta}_j\|_\infty \|\hat{\nu}_j - \hat{\nu}_j^*\|_1^2\right)^{1/2}
\]

\[
\leq m_0^{-1/2} s_n^{1/2} \sum_{j=1}^{p_n} \left(n^{-1/2} \|\Theta_j(\eta_j - \eta_j^*)\|_2 + \|E(n^{-1} \hat{\Theta}_j' \hat{\Theta}_j) - n^{-1} \hat{\Theta}_j' \hat{\Theta}_j\|_\infty \|\hat{\nu}_j - \hat{\nu}_j^*\|_1\right)
\]

\[
= \left(m_0^{-1/2} s_n^{1/2} \sum_{j=1}^{p_n} n^{-1/2} \|\Theta_j(\eta_j - \eta_j^*)\|_2 + \left(m_0^{-1/2} s_n^{1/2} \|E(n^{-1} \hat{\Theta}_j' \hat{\Theta}_j) - n^{-1} \hat{\Theta}_j' \hat{\Theta}_j\|_\infty \|\hat{\nu}\|_1\right)\right)
\]

\[
\leq \left(m_0^{-1/2} s_n^{1/2} \sum_{j=1}^{p_n} n^{-1/2} \|\Theta_j(\eta_j - \eta_j^*)\|_2 + \left(c_3 s_n^{1/2} n^{3/4} - 1/4 \|\hat{\nu}\|_1\right)\right)
\]

for some constant \(c_3 > 0\), where the last inequality is by Lemma 8. Invoking (A8), it is easy to see that \(\|\hat{\nu}\|_1 \leq c_4 s_n^{1/2} \sum_{j=1}^{p_n} n^{-1/2} \|\Theta_j(\eta_j - \eta_j^*)\|_2\), with probability tending to one, for some constant \(c_4 > 0\), which completes the proof of part 2). For part 3), it is easy to observe that by combining part 2) with Lemma 5(c), we have

\[
\lambda_n \|\hat{\nu}\|_1 \leq c_4 \lambda_n s_n^{1/2} \sum_{j=1}^{p_n} n^{-1/2} \|\Theta_j(\eta_j - \eta_j^*)\|_2
\]

\[
=c_4 L^{-1} \sum_{j=1}^{p_n} \lambda_n s_n^{1/2} L n^{-1/2} \|\Theta_j(\eta_j - \eta_j^*)\|_2
\]

\[
\leq c_4 L^{-1} \sum_{j=1}^{p_n} \left[\rho \lambda_n s_n^{1/2} (n^{-1/2} \|\Theta_j(\eta_j - \eta_j^*)\|_2 + 2^{-1} \mu n^{-1} \|\Theta_j(\eta_j - \eta_j^*)\|_2^2\right]
\]

\[
\leq c_4 L^{-1} \left[P_{\lambda_n}(\eta^*) + P_{\lambda_n}(\eta) + 2^{-1} \mu n^{-1} \sum_{j=1}^{p_n} \|\Theta_j(\eta_j - \eta_j^*)\|_2^2\right].
\]
By combining (2.21) with part 1), it is easy to see that $\lambda_n \|\hat{\nu}\|_1 \leq c_5 \left[ P_{\lambda_n}(\eta^*) + P_{\lambda_n}(\eta) + \|\hat{\nu}\|_2^2 \right]$, with probability tending to one, for some constant $c_5 > 0$, which completes the proof of part 3).

**Proof of Lemma 11.** First, by Lemma 8, we have

$$\max_{l \leq h_n s_n} \|n^{-1} \sum_{i=1}^n (w_l' F_i - E_{il}) F_i'\|_\infty \leq c_0 n^{\beta/2 - 1/2},$$

with probability tending to one, for some constant $c_0 > 0$. Moreover, under (A6), (A8), and (A10), we have $\tau_n \sim n^{\beta/2 - 1/2}$. By choosing $\tau_n = c_0 n^{\beta/2 - 1/2}$, then it is easy to see that

$$\max_{l \leq h_n s_n} \|n^{-1} \sum_{i=1}^n (w_l' F_i - E_{il}) F_i'\|_\infty \leq \tau_n = c_0 n^{\beta/2 - 1/2}. \quad (2.22)$$

Under (2.7), we have

$$\max_{l \leq h_n s_n} \|n^{-1} \sum_{i=1}^n (\hat{w}_l' F_i - E_{il}) F_i'\|_\infty \leq \tau_n = c_0 n^{\beta/2 - 1/2}. \quad (2.23)$$

Combining (2.22) with (2.23), it is easy to see that

$$\max_{l \leq h_n s_n} \left\| n^{-1} \sum_{i=1}^n F_i F_i' (\hat{w}_l - w_l) \right\|_\infty$$

$$\leq \max_{l \leq h_n s_n} \left\| n^{-1} \sum_{i=1}^n (w_l' F_i - E_{il}) F_i' \right\|_\infty + \max_{l \leq h_n s_n} \left\| n^{-1} \sum_{i=1}^n (\hat{w}_l' F_i - E_{il}) F_i' \right\|_\infty$$

$$\leq 2 \tau_n = 2 c_0 n^{\beta/2 - 1/2}, \quad (2.24)$$

which completes the proof of part 1). In addition, by the definition of the Dantzig selector method, it is easy to see that

$$P \left( \bigcap_{l=1}^{h_n s_n} \{ ||\hat{w}_l||_1 \leq ||w_l||_1 \} \right) \to 1, \quad \text{as} \quad n \to \infty. \quad (2.25)$$
Denote $S_l = \{ j : w_{lj} \neq 0 \}$ as the support set for each vector $w_l$, then (2.25) implies that

\[
\max_{l \leq h_n s_n} ||\hat{w}_l - w_l||_1 \leq \max_{l \leq h_n s_n} ||(\hat{w}_l - w_l)S_l||_1 + \max_{l \leq h_n s_n} ||(\hat{w}_l)_{S_l^c}||_1 \\
\leq \max_{l \leq h_n s_n} ||(\hat{w}_l - w_l)S_l||_1 + \max_{l \leq h_n s_n} \left[ ||(w_l)_{S_l}||_1 - ||(\hat{w}_l)_{S_l}||_1 \right] \\
\leq 2 \max_{l \leq h_n s_n} ||(\hat{w}_l - w_l)S_l||_1 \leq 2 \max_{l \leq h_n s_n} \rho_n^{1/2} ||(\hat{w}_l - w_l)S_l||_2 \\
\leq 2 \rho_n^{1/2} \max_{l \leq h_n s_n} ||\hat{w}_l - w_l||_2,
\]  

(2.26)

with probability tending to one. Furthermore, we have

\[
\max_{l \leq h_n s_n} (\hat{w}_l - w_l)'(n^{-1} \sum_{i=1}^{n} F_i F_i')(\hat{w}_l - w_l) \\
\leq \max_{l \leq h_n s_n} ||\hat{w}_l - w_l||_1 ||n^{-1} \sum_{i=1}^{n} F_i F_i'(w_l - w_l)||_\infty \\
\leq 2 c_0 n \beta/2 - 1/2 \max_{l \leq h_n s_n} ||\hat{w}_l - w_l||_1,
\]  

(2.27)

with probability tending to one, where the last inequality is by (2.24). In addition, we also have

\[
\max_{l \leq h_n s_n} (\hat{w}_l - w_l)'(n^{-1} \sum_{i=1}^{n} F_i F_i')(\hat{w}_l - w_l) \\
= \max_{l \leq h_n s_n} \left( (\hat{w}_l - w_l)'E(F_i F_i')(\hat{w}_l - w_l) - (\hat{w}_l - w_l)'[E(F_i F_i') - n^{-1} \sum_{i=1}^{n} F_i F_i'](\hat{w}_l - w_l) \right) \\
\geq \max_{l \leq h_n s_n} \left( \lambda_{\text{min}}(I)||\hat{w}_l - w_l||_2^2 - ||\hat{w}_l - w_l||_1^2 ||E(F_i F_i') - n^{-1} \sum_{i=1}^{n} F_i F_i'||_\infty \right) \\
\geq \max_{l \leq h_n s_n} \left( \lambda_{\text{min}}(I)||\hat{w}_l - w_l||_2^2 - ||\hat{w}_l - w_l||_1^2 ||E(G_i G_i') - n^{-1} \Theta' \Theta||_\infty \right) \\
\geq \lambda_{\text{min}}(I) \left( \max_{l \leq h_n s_n} ||\hat{w}_l - w_l||_2^2 \right) - ||E(G_i G_i') - n^{-1} \Theta' \Theta||_\infty \left( \max_{l \leq h_n s_n} ||\hat{w}_l - w_l||_1^2 \right) \\
\geq c_1 s_n^{-a} \left( \max_{l \leq h_n s_n} ||\hat{w}_l - w_l||_2^2 \right) - c_2 n^{\beta/2 - 1/2} \left( \max_{l \leq h_n s_n} ||\hat{w}_l - w_l||_1^2 \right) \\
\geq (c_1 s_n^{-a} - c_3 n^{\beta/2 - 1/2}) \left( \max_{l \leq h_n s_n} ||\hat{w}_l - w_l||_2^2 \right),
\]  

(2.28)

with probability tending to one, where the second last inequality is by (A3), (A4), and Lemma 8,
while the last inequality is by (2.26). Combining (2.26), (2.27), (2.28) with (A8) entails that

$$\max_{l \leq h_n s_n} ||\hat{w}_l - w_l||_2 \leq c_4 \rho_n^{1/2} s_n^a n^{3/2-1/2},$$

(2.29)

with probability tending to one, for some constant $c_4 > 0$, which completes the proof of part 2). Lastly, combining (2.26) with (2.29) entails that

$$\max_{l \leq h_n s_n} ||\hat{w}_l - w_l||_1 \leq c_5 \rho_n s_n^a n^{3/2-1/2},$$

with probability tending to one, for some constant $c_5 > 0$, which completes the proof of part 3).
Proof of Lemma 12. First, it is easy to see that for each \( l = 1, \ldots, h_n s_n \), we have

\[
\hat{S}_{il} - S_{il}
\]

\[
= [n^{-1} \sum_{i=1}^{n} E_{i,l}^2]^{-1/2} (\hat{w}_l' F_i - E_{il})(Y_i - F_i' \hat{\eta}_{\mathcal{H}_n}) - [E(E_{il}^2)]^{-1/2} (w_l' F_i - E_{il}) \epsilon_i
\]

\[
= [E(E_{il}^2)/(n^{-1} \sum_{i=1}^{n} E_{i,l}^2)]^{1/2} [E(E_{il}^2)]^{-1/2} (\hat{w}_l' F_i - E_{il})(Y_i - F_i' \hat{\eta}_{\mathcal{H}_n})
\]

\[
- [E(E_{il}^2)]^{-1/2} (w_l' F_i - E_{il}) \epsilon_i
\]

\[
= [E(E_{il}^2)/(n^{-1} \sum_{i=1}^{n} E_{i,l}^2)]^{1/2} [E(E_{il}^2)]^{-1/2} (\hat{w}_l' F_i - E_{il}) \epsilon_i + F_i' (\eta_{\mathcal{H}_n} - \hat{\eta}_{\mathcal{H}_n}) +
\]

\[
\sum_{j \in \mathcal{H}_n} \sum_{k = s_n + 1}^{\infty} \theta_{ijh} \eta_{jk} - [E(E_{il}^2)]^{-1/2} (w_l' F_i - E_{il}) \epsilon_i
\]

\[
= [E(E_{il}^2)/(n^{-1} \sum_{i=1}^{n} E_{i,l}^2)]^{1/2} [(w_l' F_i - E_{il}) + (\hat{w}_l - w_l)' F_i]
\]

\[
[\epsilon_i + F_i' (\eta_{\mathcal{H}_n} - \hat{\eta}_{\mathcal{H}_n}) + \sum_{j \in \mathcal{H}_n} \sum_{k = s_n + 1}^{\infty} \theta_{ijh} \eta_{jk}] - [E(E_{il}^2)]^{-1/2} (w_l' F_i - E_{il}) \epsilon_i
\]

\[
= \left( [E(E_{il}^2)/(n^{-1} \sum_{i=1}^{n} E_{i,l}^2)]^{1/2} - 1 \right) \left( [E(E_{il}^2)]^{-1/2} (w_l' F_i - E_{il}) \epsilon_i \right)
\]

\[
+ \left( [E(E_{il}^2)/(n^{-1} \sum_{i=1}^{n} E_{i,l}^2)]^{1/2} \right) \left( [E(E_{il}^2)]^{-1/2} (w_l' F_i - E_{il}) F_i' (\eta_{\mathcal{H}_n} - \hat{\eta}_{\mathcal{H}_n}) \right)
\]

\[
+ \left( [E(E_{il}^2)/(n^{-1} \sum_{i=1}^{n} E_{i,l}^2)]^{1/2} \right) \left( [E(E_{il}^2)]^{-1/2} (w_l' F_i - E_{il}) (\sum_{j \in \mathcal{H}_n} \sum_{k = s_n + 1}^{\infty} \theta_{ijh} \eta_{jk}) \right)
\]

\[
+ \left( [E(E_{il}^2)/(n^{-1} \sum_{i=1}^{n} E_{i,l}^2)]^{1/2} \right) \left( [E(E_{il}^2)]^{-1/2} (\hat{w}_l - w_l)' F_i \epsilon_i \right)
\]

\[
+ \left( [E(E_{il}^2)/(n^{-1} \sum_{i=1}^{n} E_{i,l}^2)]^{1/2} \right) \left( [E(E_{il}^2)]^{-1/2} (\hat{w}_l - w_l)' F_i (\eta_{\mathcal{H}_n} - \hat{\eta}_{\mathcal{H}_n}) \right)
\]

\[
+ \left( [E(E_{il}^2)/(n^{-1} \sum_{i=1}^{n} E_{i,l}^2)]^{1/2} \right) \left( [E(E_{il}^2)]^{-1/2} (\hat{w}_l - w_l)' F_i (\sum_{j \in \mathcal{H}_n} \sum_{k = s_n + 1}^{\infty} \theta_{ijh} \eta_{jk}) \right)
\]

\[
= \Delta_{1il} + \Delta_{2il} + \Delta_{3il} + \Delta_{4il} + \Delta_{5il} + \Delta_{6il},
\]
where we denote

\[
\begin{align*}
\Delta_{1il} &= \left( [E(E_{il}^2)/(n^{-1} \sum_{i_1=1}^{n} E_{i_1 il}^2)]^{1/2} - 1 \right) \left( [E(E_{il}^2)]^{-1/2} (w_i' F_i - E_{il}) \epsilon_i \right), \\
\Delta_{2il} &= \left( [E(E_{il}^2)/(n^{-1} \sum_{i_1=1}^{n} E_{i_1 il}^2)]^{1/2} \right) \left( [E(E_{il}^2)]^{-1/2} (w_i' F_i - E_{il}) F_i' (\eta_{H_n} - \hat{\eta}_{H_n}) \right), \\
\Delta_{3il} &= \left( [E(E_{il}^2)/(n^{-1} \sum_{i_1=1}^{n} E_{i_1 il}^2)]^{1/2} \right) \left( [E(E_{il}^2)]^{-1/2} (w_i' F_i - E_{il}) \left( \sum_{j \in H_n} \sum_{k=s_n+1}^{\infty} \theta_{ijk} \eta_{jk} \right) \right), \\
\Delta_{4il} &= \left( [E(E_{il}^2)/(n^{-1} \sum_{i_1=1}^{n} E_{i_1 il}^2)]^{1/2} \right) \left( [E(E_{il}^2)]^{-1/2} (\hat{w}_i - w_i)' F_i \epsilon_i \right), \\
\Delta_{5il} &= \left( [E(E_{il}^2)/(n^{-1} \sum_{i_1=1}^{n} E_{i_1 il}^2)]^{1/2} \right) \left( [E(E_{il}^2)]^{-1/2} (\hat{w}_i - w_i)' F_i F_i' (\eta_{H_n} - \hat{\eta}_{H_n}) \right), \\
\Delta_{6il} &= \left( [E(E_{il}^2)/(n^{-1} \sum_{i_1=1}^{n} E_{i_1 il}^2)]^{1/2} \right) \left( [E(E_{il}^2)]^{-1/2} (\hat{w}_i - w_i)' F_i \left( \sum_{j \in H_n} \sum_{k=s_n+1}^{\infty} \theta_{ijk} \eta_{jk} \right) \right).
\end{align*}
\]

Accordingly, it is easy to observe that

\[
\begin{align*}
\max_{l \leq h_n s_n} \left| n^{-1/2} \sum_{i=1}^{n} (\hat{S}_{il} - S_{il}) \right| &\leq \max_{l \leq h_n s_n} \left| n^{-1/2} \sum_{i=1}^{n} \Delta_{1il} \right| + \max_{l \leq h_n s_n} \left| n^{-1/2} \sum_{i=1}^{n} \Delta_{2il} \right| \\
&+ \max_{l \leq h_n s_n} \left| n^{-1/2} \sum_{i=1}^{n} \Delta_{3il} \right| + \max_{l \leq h_n s_n} \left| n^{-1/2} \sum_{i=1}^{n} \Delta_{4il} \right| \\
&+ \max_{l \leq h_n s_n} \left| n^{-1/2} \sum_{i=1}^{n} \Delta_{5il} \right| + \max_{l \leq h_n s_n} \left| n^{-1/2} \sum_{i=1}^{n} \Delta_{6il} \right|.
\end{align*}
\]

(2.30)
For $\max_{t \leq h_n s_n} |n^{-1/2} \sum_{i=1}^{n} \Delta_{2il}|$, we have

$$
\max_{t \leq h_n s_n} |n^{-1/2} \sum_{i=1}^{n} \Delta_{2il}|
= \max_{t \leq h_n s_n} |n^{-1/2} \sum_{i=1}^{n} \left( [E\left(E_{il}^2\right)/\left(n^{-1} \sum_{i=1}^{n} E_{i1}^2\right)]^{1/2} \left( [E\left(E_{il}^2\right)]^{-1/2} (w_i'F_i - E_{il}) \hat{H}_{il} \right) \right) |
\leq \|\Delta \hat{\Lambda}^{-1} - I\|_\infty \max_{t \leq h_n s_n} |n^{-1/2} \sum_{i=1}^{n} \left( [E\left(E_{il}^2\right)]^{-1/2} (w_i'F_i - E_{il}) \hat{H}_{il} \right) |
\leq n^{1/2} \left( 1 + \|\Delta \hat{\Lambda}^{-1} - I\|_\infty \right) \|\hat{H}_{il} - \hat{\eta}_{il}\| \max_{t \leq h_n s_n} |n^{-1/2} \sum_{i=1}^{n} (w_i'F_i - E_{il}) [E\left(E_{il}^2\right)]^{-1/2} F_i' \|_\infty
\leq n^{1/2} \left( 1 + \|\Delta \hat{\Lambda}^{-1} - I\|_\infty \right) \|\hat{\eta} - \eta\| \max_{t \leq h_n s_n} |n^{-1/2} \sum_{i=1}^{n} (w_i'F_i - E_{il}) [E\left(E_{il}^2\right)]^{-1/2} F_i' \|_\infty
\leq c_1 \lambda_n n^{3/2} q_n s_n^{a/2 + 1},
$$

(2.32)

with probability tending to one, for some constant $c_1 > 0$, where the last inequality is by Lemma 7, Lemma 8 and Theorem 3.
For \( \max_{l \leq h \cdot s \cdot n} |n^{-1/2} \sum_{i=1}^{n} \Delta_{3li}| \), we have

\[
\max_{l \leq h \cdot s \cdot n} |n^{-1/2} \sum_{i=1}^{n} \Delta_{3li}|
= \max_{l \leq h \cdot s \cdot n} |n^{-1/2} \sum_{i=1}^{n} \left( \left[E(E_{il}^2) / (n^{-1} \sum_{i=1}^{n} E_{i1}^2) \right]^{1/2} \left[E(E_{il}^2)\right]^{-1/2} (w_i' F_i - E_{il}) \left( \sum_{j \in \mathcal{H}_n} \sum_{k=s+1}^{\infty} \theta_{ijk} \eta_{jk} \right) \right) |
\leq (1 + \|\Lambda \hat{\Lambda}^{-1} - I\|_\infty) \max_{l \leq h \cdot s \cdot n} |n^{-1/2} \sum_{i=1}^{n} \left( \left[E(E_{il}^2)\right]^{-1/2} (w_i' F_i - E_{il}) (\sum_{j \in \mathcal{H}_n} \sum_{k=s+1}^{\infty} \theta_{ijk} \eta_{jk}) \right) |
\leq n^{1/2} \left(1 + \|\Lambda \hat{\Lambda}^{-1} - I\|_\infty\right) \left( \max_{l \leq h \cdot s \cdot n} \max_{i \leq n} \left| \left[E(E_{il}^2)\right]^{-1/2} (w_i' F_i - E_{il}) \right| \right)
\cdot \left( n^{-1} \sum_{i=1}^{n} \sum_{j \in \mathcal{H}_n} \sum_{k=s+1}^{\infty} \theta_{ijk} \eta_{jk} \right).
\tag{2.33}
\]

Regarding \( \max_{l \leq h \cdot s \cdot n} \max_{i \leq n} \left| \left[E(E_{il}^2)\right]^{-1/2} (w_i' F_i - E_{il}) \right| \), it is easily to be seen from (A2), (A6) and (A8) that we have

\[
\max_{l \leq h \cdot s \cdot n} \max_{i \leq n} \left| \left[E(E_{il}^2)\right]^{-1/2} (w_i' F_i - E_{il}) \right|
\leq c_1 [\log(n p_{n \cdot s \cdot n})]^{1/2}
\leq c_2 n^{\beta/2},
\tag{2.34}
\]

with probability tending to one, for some constants \( c_1, c_2 > 0 \).
Regarding $n^{-1} \sum_{i=1}^{n} \left| \sum_{j \in H_n^c} \sum_{k=s_n+1}^{\infty} \theta_{ijk} \eta_{jk} \right|$, we have

$$E\left(n^{-1} \sum_{i=1}^{n} \left| \sum_{j \in H_n^c} \sum_{k=s_n+1}^{\infty} \theta_{ijk} \eta_{jk} \right|\right) \leq E\left(n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{q_n} \sum_{k=s_n+1}^{\infty} \left| \theta_{ijk} \eta_{jk} \right|\right)$$

$$\leq n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{q_n} \left( \sum_{k=s_n+1}^{\infty} \theta_{ijk}^2 k^{-2\delta} \right)^{1/2} \left( \sum_{k=s_n+1}^{\infty} \eta_{jk}^2 k_1^{2\delta} \right)^{1/2}$$

$$\leq n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{q_n} \left( \sum_{k_1=s_n+1}^{\infty} \eta_{j k_1}^2 k_1^{2\delta} \right)^{1/2} \left[ E\left( \sum_{k=s_n+1}^{\infty} \theta_{ijk}^2 k^{-2\delta} \right)^{1/2} \right]$$

$$= n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{q_n} \left( \sum_{k_1=s_n+1}^{\infty} \eta_{j k_1}^2 k_1^{2\delta} \right)^{1/2} \left( \sum_{k=s_n+1}^{\infty} \omega_{jk} k^{-2\delta} \right)^{1/2}$$

$$\leq O(q_n s_n^{-\delta}), \quad (2.35)$$

where the last inequality is by (A1) and (A7). Hence, by combining (2.33), (2.34) with (2.35), it is easy to see that

$$\max_{l \leq h_n s_n} \left| n^{-1/2} \sum_{i=1}^{n} \Delta_{3il} \right| \leq O_p\left(n^{\beta/2+1/2} q_n s_n^{-\delta}\right). \quad (2.36)$$
For \( \max_{t \leq h_n s_n} |n^{-1/2} \sum_{i=1}^{n} \Delta_{4il}| \), we have

\[
\max_{t \leq h_n s_n} |n^{-1/2} \sum_{i=1}^{n} \Delta_{4il}|
= \max_{t \leq h_n s_n} \left| n^{-1/2} \sum_{i=1}^{n} \left( \left[ E(E_{il}^2) / (n^{-1} \sum_{i=1}^{n} E_{iil}^2) \right]^{1/2} \left[ E(E_{il}^2) \right]^{-1/2} (\hat{w}_t - w_t) F_i \epsilon_i \right) \right|
\leq (1 + \| \Lambda \hat{\Lambda}^{-1} - I \|_\infty) \max_{t \leq h_n s_n} \left| n^{-1/2} \sum_{i=1}^{n} \left[ E(E_{il}^2) \right]^{-1/2} (\hat{w}_t - w_t) F_i \epsilon_i \right|
\leq n^{1/2} \left( 1 + \| \Lambda \hat{\Lambda}^{-1} - I \|_\infty \right) \left( \max_{t \leq h_n s_n} \left| E(E_{il}^2) \right|^{-1/2} \right)
\cdot \left( \max_{t \leq h_n s_n} \| \hat{w}_t - w_t \|_1 \right) \left( \| n^{-1} \Theta' \epsilon \|_\infty \right)
\leq n^{1/2} \left( 1 + \| \Lambda \hat{\Lambda}^{-1} - I \|_\infty \right) \left( \left[ \lambda_{\min}(\Lambda) \right]^{-1} \right)
\cdot \left( \max_{t \leq h_n s_n} \| \hat{w}_t - w_t \|_1 \right) \left( \| n^{-1} \Theta' \epsilon \|_\infty \right)
\leq c_3 n^{\beta - 1/2} p_n^{-3\alpha/2},
\]  
(2.37)

with probability tending to one, for some constant \( c_3 > 0 \), where the last inequality is by Lemma 7, Lemma 8, Lemma 11 and (A3).

For \( \max_{t \leq h_n s_n} |n^{-1/2} \sum_{i=1}^{n} \Delta_{5il}| \), we have

\[
\max_{t \leq h_n s_n} \left| n^{-1/2} \sum_{i=1}^{n} \Delta_{5il} \right|
= \max_{t \leq h_n s_n} \left| n^{-1/2} \sum_{i=1}^{n} \left( \left[ E(E_{il}^2) / (n^{-1} \sum_{i=1}^{n} E_{iil}^2) \right]^{1/2} \left[ E(E_{il}^2) \right]^{-1/2} (\hat{w}_t - w_t) F_i F_i' (\eta_{H_n} - \hat{\eta}_{H_n}) \right) \right|
\leq (1 + \| \Lambda \hat{\Lambda}^{-1} - I \|_\infty) \max_{t \leq h_n s_n} \left| n^{-1/2} \sum_{i=1}^{n} \left[ E(E_{il}^2) \right]^{-1/2} (\hat{w}_t - w_t) F_i F_i' (\eta_{H_n} - \hat{\eta}_{H_n}) \right|
\leq n^{1/2} \left( 1 + \| \Lambda \hat{\Lambda}^{-1} - I \|_\infty \right) \left( \left[ \lambda_{\min}(\Lambda) \right]^{-1} \right) \left( \| \eta - \hat{\eta} \|_1 \right) \left( \max_{t \leq h_n s_n} \| n^{-1} \sum_{i=1}^{n} F_i F_i' (\hat{w}_t - w_t) \|_\infty \right)
\leq c_4 \lambda_n n^{\beta/2} q_n^{-\alpha+1},
\]  
(2.38)

with probability tending to one, for some constant \( c_4 > 0 \), where the last inequality is by Lemma 7, Lemma 11, Theorem 3 and (A3).
For \( \max_{l \leq h_n s_n} |n^{-1/2} \sum_{i=1}^{n} \Delta_{6il}| \), we have

\[
\max_{l \leq h_n s_n} |n^{-1/2} \sum_{i=1}^{n} \Delta_{6il}|
= \max_{l \leq h_n s_n} \left| n^{-1/2} \sum_{i=1}^{n} \left( [E(E_i^2)] / (n^{-1} \sum_{i=1}^{n} E_i^2) \right)^{1/2} \left( [E(E_i^2)]^{-1/2} (\hat{\omega}_l - \omega_l)' F_i \left( \sum_{j \in H_n^c} \sum_{k = s_n + 1}^{\infty} \theta_{ijk} \eta_{j} \right) \right) \right|
\leq (1 + \| \Lambda \hat{\Lambda}^{-1} - I \|_\infty) \max_{l \leq h_n s_n} \left| n^{-1/2} \sum_{i=1}^{n} \left( [E(E_i^2)]^{-1/2} (\hat{\omega}_l - \omega_l)' F_i \left( \sum_{j \in H_n^c} \sum_{k = s_n + 1}^{\infty} \theta_{ijk} \eta_{j} \right) \right) \right|
\leq n^{1/2} \left( 1 + \| \Lambda \hat{\Lambda}^{-1} - I \|_\infty \right) \left( \max_{l \leq h_n s_n} [E(E_i^2)]^{-1/2} \right) \left( \max_{l \leq h_n s_n} \| \hat{\omega}_l - \omega_l \|_1 \right)
\cdot \left( \| n^{-1} \sum_{i=1}^{n} F_i \left( \sum_{j \in H_n^c} \sum_{k = s_n + 1}^{\infty} \theta_{ijk} \eta_{j} \right) \|_\infty \right)
\leq n^{1/2} \left( 1 + \| \Lambda \hat{\Lambda}^{-1} - I \|_\infty \right) \left( \left[ \lambda_{\min}(\Lambda) \right]^{-1} \right) \left( \max_{l \leq h_n s_n} \| \hat{\omega}_l - \omega_l \|_1 \right)
\cdot \left( \max_{l \leq (p_n - h_n) s_n} \left| n^{-1} \sum_{i=1}^{n} F_{il} \left( \sum_{j \in H_n^c} \sum_{k = s_n + 1}^{\infty} \theta_{ijk} \eta_{j} \right) \right| \right)
\leq n^{1/2} \left( 1 + \| \Lambda \hat{\Lambda}^{-1} - I \|_\infty \right) \left( \left[ \lambda_{\min}(\Lambda) \right]^{-1} \right) \left( \max_{l \leq h_n s_n} \| \hat{\omega}_l - \omega_l \|_1 \right)
\cdot \left( \max_{l \leq (p_n - h_n) s_n} \max_{i \leq n} | F_{il} | \right) \left( n^{-1} \sum_{i=1}^{n} \left| \sum_{j \in H_n^c} \sum_{k = s_n + 1}^{\infty} \theta_{ijk} \eta_{j} \right| \right)
\leq O_p(n^\beta \rho_n q_n \delta_n^{3/2-\delta}),
\]

where the last inequality is by Lemma 7, Lemma 11, (A2), (A3) and (2.35).

In summary, by combining (2.30), (2.31), (2.32), (2.36), (2.37), (2.38) with (2.39), it is
easy to see that

$$\max_{l \leq h, s_n} |n^{-1/2} \sum_{i=1}^{n} (\hat{S}_{il} - S_{il})| \leq \max_{l \leq h, s_n} |n^{-1/2} \sum_{i=1}^{n} \Delta_{1il}| + \max_{l \leq h, s_n} |n^{-1/2} \sum_{i=1}^{n} \Delta_{2il}|$$

$$+ \max_{l \leq h, s_n} |n^{-1/2} \sum_{i=1}^{n} \Delta_{3il}| + \max_{l \leq h, s_n} |n^{-1/2} \sum_{i=1}^{n} \Delta_{4il}|$$

$$+ \max_{l \leq h, s_n} |n^{-1/2} \sum_{i=1}^{n} \Delta_{5il}| + \max_{l \leq h, s_n} |n^{-1/2} \sum_{i=1}^{n} \Delta_{6il}|$$

$$\leq c_0 n^{\beta-1/2} + c_1 \lambda_n n^{\beta/2} q_n s_n^{\alpha/2 + 1} + O_p(n^{\beta/2+1/2} q_n s_n^{-\delta})$$

$$+ c_3 n^{\beta-1/2} \rho_n s_n^{3\alpha/2} + c_4 \lambda_n n^{\beta/2} q_n s_n^{\alpha+1} + O_p(n^{\beta/2+1/2} q_n s_n^{-\delta})$$

$$\leq c_5 (n^{\beta-1/2} \rho_n s_n^{3\alpha/2} + \lambda_n n^{\beta/2} q_n s_n^{\alpha+1} + n^{\beta/2+1/2} q_n s_n^{-\delta} \log s_n)$$

$$+ n^{\beta} \rho_n q_n s_n^{3\alpha/2 - \delta} \log s_n),$$

with probability tending to one, for some constant $c_5 > 0$, which completes the proof of part 1).

Next, we start to prove part 2). First, it is easy to observe that

$$\max_{l \leq h, s_n} \left[ n^{-1} \sum_{i=1}^{n} (\hat{S}_{il} - S_{il})^2 \right]$$

$$\leq 100 \left( \max_{l \leq h, s_n} n^{-1} \sum_{i=1}^{n} \Delta_{1il}^2 + \max_{l \leq h, s_n} n^{-1} \sum_{i=1}^{n} \Delta_{2il}^2 + \max_{l \leq h, s_n} n^{-1} \sum_{i=1}^{n} \Delta_{3il}^2 \right.$$}

$$\left. + \max_{l \leq h, s_n} n^{-1} \sum_{i=1}^{n} \Delta_{4il}^2 + \max_{l \leq h, s_n} n^{-1} \sum_{i=1}^{n} \Delta_{5il}^2 + \max_{l \leq h, s_n} n^{-1} \sum_{i=1}^{n} \Delta_{6il}^2 \right). \quad (2.40)$$
Lemma 7, Theorem 3 and (A2).

For \( \max_{l \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \Delta_{il}^2 \), we have

\[
\max_{l \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \Delta_{il}^2 \\
= \max_{l \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \left( \left[ E(E_i^2)/(n^{-1} \sum_{i_1=1}^{n} E_{i_1}^2) \right]^{1/2} \right)^2 \left( \left[ E(E_i^2) \right]^{-1/2}(w_i' F_i - E_{il}) \epsilon_i \right)^2 \\
\leq \left( ||\hat{\Lambda} - I||_\infty \right)^2 \max_{l \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \left( \left[ E(E_i^2) \right]^{-1/2}(w_i' F_i - E_{il}) \epsilon_i \right)^2 \\
\leq \left( ||\hat{\Lambda} - I||_\infty \right)^2 \max_{l \leq h_n s_n} \max_{i \leq n} \left( \left[ E(E_i^2) \right]^{-1/2}(w_i' F_i - E_{il}) \epsilon_i \right)^2 \\
\leq c_6 n^{2\beta - 1} \log n, \tag{2.41}
\]

with probability tending to one, for some constant \( c_6 > 0 \), where the last inequality is by Lemma 7 and (A2).

For \( \max_{l \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \Delta_{il}^2 \), we have

\[
\max_{l \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \Delta_{il}^2 \\
= \max_{l \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \left( \left[ E(E_i^2)/(n^{-1} \sum_{i_1=1}^{n} E_{i_1}^2) \right]^{1/2} \right)^2 \left( \left[ E(E_i^2) \right]^{-1/2}(w_i' F_i - E_{il}) F_i'(\eta_{h_n} - \hat{\eta}_{h_n}) \right)^2 \\
\leq \left( 1 + ||\hat{\Lambda} - I||_\infty \right)^2 \max_{l \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \left( \left[ E(E_i^2) \right]^{-1/2}(w_i' F_i - E_{il}) F_i'(\eta_{h_n} - \hat{\eta}_{h_n}) \right)^2 \\
\leq \left( 1 + ||\hat{\Lambda} - I||_\infty \right)^2 \left( ||\hat{\eta} - \eta||_1 \right)^2 \max_{l \leq h_n s_n} \max_{i \leq n} \left( \left[ E(E_i^2) \right]^{-1/2}(w_i' F_i - E_{il}) F_i'(\eta_{h_n} - \hat{\eta}_{h_n}) \right)^2 \\
\leq \left( 1 + ||\hat{\Lambda} - I||_\infty \right)^2 \left( ||\hat{\eta} - \eta||_1 \right)^2 \max_{l \leq h_n s_n} \max_{i \leq n} \max_{l_1 \leq (p_n - h_n)s_n} \left( \left[ E(E_i^2) \right]^{-1/2}(w_i' F_i - E_{il}) F_{il_1} \right)^2 \\
\leq c_7 \lambda_n^2 n^{2\beta} q_n^2 s_n^{a+2}, \tag{2.42}
\]

with probability tending to one, for some constant \( c_7 > 0 \), where the last inequality is by Lemma 7, Theorem 3 and (A2).
For \( \max_{l \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \Delta^2_{3il} \), we have

\[
\max_{l \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \Delta^2_{3il} = \max_{l \leq h_n s_n} n^{-1} \left( \left[ E(E^2_{il})/(n^{-1} \sum_{i=1}^{n} E^2_{il}) \right]^{1/2} \right)^2 \left( [E(E^2_{il})]^{-1/2} (w_i' F_i - E_{il}) \left( \sum_{j \in H_n} \sum_{k=s_n+1}^{\infty} \theta_{ijk} \eta_{jk} \right) \right)^2 \\
\leq \left( 1 + ||\hat{\Lambda}^{-1} - I||_{\infty} \right)^2 \max_{l \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \left( [E(E^2_{il})]^{-1/2} (w_i' F_i - E_{il}) \left( \sum_{j \in H_n} \sum_{k=s_n+1}^{\infty} \theta_{ijk} \eta_{jk} \right) \right)^2 \\
\leq \left( 1 + ||\hat{\Lambda}^{-1} - I||_{\infty} \right)^2 \left( \max_{l \leq h_n s_n} \max_{i \leq n} \left[ E(E^2_{il}) \right]^{-1/2} (w_i' F_i - E_{il}) \right)^2 \\
\cdot \left( n^{-1} \sum_{i=1}^{n} \left( \sum_{j \in H_n} \sum_{k=s_n+1}^{\infty} \theta_{ijk} \eta_{jk} \right)^2 \right).
\]

(2.43)

Regarding \( \max_{l \leq h_n s_n} \max_{i \leq n} \left[ E(E^2_{il}) \right]^{-1/2} (w_i' F_i - E_{il}) \), it is easy to see that

\[
\max_{l \leq h_n s_n} \max_{i \leq n} \left[ E(E^2_{il}) \right]^{-1/2} (w_i' F_i - E_{il})^2 \\
\leq c_8 \log(nh_n s_n) \leq c_8 \log(np_n s_n) \\
\leq c_9 n^{\beta},
\]

(2.44)

with probability tending to one, for some constants \( c_8, c_9 > 0 \), where the last inequality is by (A2), (A6) and (A8).
Regarding $n^{-1} \sum_{i=1}^{n} (\sum_{j \in \mathcal{H}_n^c} \sum_{k=s_n+1}^{\infty} \theta_{ijk} \eta_{jk})^2$, it is easy to see that

$$E\left[n^{-1} \sum_{i=1}^{n} \left( \sum_{j \in \mathcal{H}_n^c} \sum_{k=s_n+1}^{\infty} \theta_{ijk} \eta_{jk} \right)^2 \right] \leq E\left[n^{-1} \sum_{i=1}^{n} \left( \sum_{j \in \mathcal{H}_n^c} \sum_{k=s_n+1}^{\infty} \theta_{ijk} \eta_{jk} \right)^2 \right]$$

$$\leq n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{q_n} \left( \sum_{k=s_n+1}^{\infty} \theta_{ijk}^2 k^{-2\delta} \left( \sum_{k_1=s_n+1}^{\infty} \eta_{jk_1}^2 k_1^{2\delta} \right) \right)$$

$$= n^{-1} q_n \sum_{i=1}^{n} \sum_{j=1}^{q_n} \left( \sum_{k=s_n+1}^{\infty} \omega_{jk} k^{-2\delta} \left( \sum_{k_1=s_n+1}^{\infty} \eta_{jk_1}^2 k_1^{2\delta} \right) \right)$$

$$= O(q_n^2 s_n^{-2\delta}). \quad (2.45)$$

Hence, by combining (2.43), (2.44), (2.45) with Lemma 7, it is easy to show that

$$\max_{t \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \Delta_{3it}^2 \leq O_p(n^\beta q_n^2 s_n^{-2\delta}). \quad (2.46)$$

For $\max_{t \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \Delta_{4it}^2$, we have

$$\max_{t \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \Delta_{4it}^2$$

$$= \max_{t \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \left( \left[ E(E_{it}^2) / (n^{-1} \sum_{i=1}^{n} E_{it}^2) \right]^{1/2} \right)^2 \left( \left[ E(E_{it}^2) \right]^{-1/2} (\hat{w}_t - w_t)' F_i \epsilon_i \right)^2$$

$$\leq \left( 1 + \|\Lambda \hat{\Lambda}^{-1} - I\|_\infty \right)^2 \left( [\lambda_{\min}(\Lambda)]^{-1} \right)^2 \left( \max_{t \leq h_n s_n} \|\hat{w}_t - w_t\|_1 \right)^2 \left( n^{-1} \sum_{i=1}^{n} \|F_i \epsilon_i\|_\infty^2 \right)$$

$$\leq \left( 1 + \|\Lambda \hat{\Lambda}^{-1} - I\|_\infty \right)^2 \left( [\lambda_{\min}(\Lambda)]^{-1} \right)^2 \left( \max_{t \leq h_n s_n} \|\hat{w}_t - w_t\|_1 \right)^2 \left( \max_{t \leq n} \max_{t \leq (p_n - h_n) s_n} \|F_{it} \epsilon_i\|_\infty \right)^2$$

$$\leq c_{10} \rho_n^2 s_n^3 \epsilon_n^{2\beta-1} \log n, \quad (2.47)$$

with probability tending to one, for some constant $c_{10} > 0$, where the last inequality is by Lemma 7, Lemma 11, (A1), (A2) and (A3).
For \( \max_{l \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \Delta_{5il}^2 \), we have

\[
\max_{l \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \Delta_{5il}^2 \leq \left( 1 + \| \Lambda \hat{\Lambda}^{-1} - I \|_\infty \right)^2 \left( \left[ \lambda_{\min}(\Lambda) \right]^{-1} \right)^2 \left( \max_{l \leq h_n s_n} \| \hat{w}_l - w_l \|_1 \right)^2 \left( \| \hat{\eta} - \eta \|_1 \right)^2 \left( \max_{i \leq n} \| F_i \|_\infty^4 \right)
\]

\[
\leq c_1 \lambda_n^2 \rho_n^2 q_n^2 s_n^{4a+2} n^{3\beta-1},
\]

(2.48)

with probability tending to one, for some constant \( c_1 > 0 \), where the last inequality is by Lemma 7, Lemma 11, Theorem 3, (A1), (A2) and (A3).

For \( \max_{l \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \Delta_{6il}^2 \), we have

\[
\max_{l \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \Delta_{6il}^2 = \max_{l \leq h_n s_n} n^{-1} \sum_{i=1}^{n} \left( \left[ E(E_{il}^2) / (n^{-1} \sum_{i=1}^{n} E_{il}^2) \right]^{1/2} \right)^2 \left( \left[ E(E_{il}^2) \right]^{-1/2} (\hat{w}_l - w_l)' F_i (\sum_{j \in \mathcal{H}_n^c} \sum_{k=s_n+1} \theta_{ijk} \eta_{jk}) \right)^2
\]

\[
\leq \left( 1 + \| \Lambda \hat{\Lambda}^{-1} - I \|_\infty \right)^2 \left( \left[ \lambda_{\min}(\Lambda) \right]^{-1} \right)^2 \left( \max_{l \leq h_n s_n} \| \hat{w}_l - w_l \|_1 \right)^2
\]

\[
\cdot \left( n^{-1} \sum_{i=1}^{n} \left\| F_i (\sum_{j \in \mathcal{H}_n^c} \sum_{k=s_n+1} \theta_{ijk} \eta_{jk}) \right\|_\infty^2 \right)
\]

\[
= \left( 1 + \| \Lambda \hat{\Lambda}^{-1} - I \|_\infty \right)^2 \left( \left[ \lambda_{\min}(\Lambda) \right]^{-1} \right)^2 \left( \max_{l \leq h_n s_n} \| \hat{w}_l - w_l \|_1 \right)^2
\]

\[
\cdot \left( n^{-1} \sum_{i=1}^{n} \left\| F_i \right\|_\infty^2 \left( \sum_{j \in \mathcal{H}_n^c} \sum_{k=s_n+1} \theta_{ijk} \eta_{jk} \right)^2 \right)
\]

\[
= \left( 1 + \| \Lambda \hat{\Lambda}^{-1} - I \|_\infty \right)^2 \left( \left[ \lambda_{\min}(\Lambda) \right]^{-1} \right)^2 \left( \max_{i \leq n} \| \hat{F}_i \|_\infty^2 \right)
\]

\[
\cdot \left( n^{-1} \sum_{i=1}^{n} \left( \sum_{j \in \mathcal{H}_n^c} \sum_{k=s_n+1} \theta_{ijk} \eta_{jk} \right)^2 \right)
\]

\[
\leq O_p \left( \rho_n^2 q_n^2 s_n^{3a-2\delta} n^{2\beta-1} \right),
\]

(2.49)

where the last inequality is by Lemma 7, Lemma 11, (2.45), (A1), (A2) and (A3).

In summary, by combining (2.41), (2.42), (2.46), (2.47), (2.48), (2.49) with (2.40), it is
easy to observe that

\[
\max_{l \leq h_n s_n} \left[ n^{-1} \sum_{i=1}^{n} (\hat{S}_i - S_i)^2 \right] \\
\leq c_{12} \left[ \lambda_n^2 n^{2\beta} q_n^2 s_n^{a+2} + n^{\beta} q_n^2 s_n^{-2\delta} (\log s_n)^2 + \rho_n^2 s_n^{3a} n^{-2\beta - 1} \log n \right. \\
+ \left. \lambda_n^2 \rho_n^2 q_n^2 s_n^{4a+2} n^{\beta - 1} + n^{\beta - 1} \rho_n^2 q_n^2 s_n^{3a - 2\delta} (\log s_n)^2 \right],
\]

(2.50)

with probability tending to one, for some constant \(c_{12} > 0\), which further implies that

\[
\max_{l \leq h_n s_n} \left[ n^{-1} \sum_{i=1}^{n} (\hat{S}_i - S_i)^2 \right]^{1/2} \\
\leq c_{13} \left[ \lambda_n n^{\beta} q_n^{2/2 + 1} + n^{\beta/2} q_n s_n^{-\delta} \log s_n + \rho_n s_n^{3a/2} n^{-1/2} (\log n)^{1/2} \right. \\
+ \left. \lambda_n \rho_n q_n s_n^{2a + 1} n^{3\beta/2 - 1/2} + n^{\beta - 1/2} \rho_n q_n s_n^{3a/2 - \delta} \log s_n \right],
\]

(2.51)

with probability tending to one, for some constant \(c_{13} > 0\), which completes the proof of part 2).

### 2.5.6 Proofs of Main Theorems

**Proof of Theorem 3.** Recall that we denote \(\nu = \eta - \eta^*\), and \(\bar{\nu} = \bar{\eta} - \bar{\eta}^* = \Lambda \nu\). By first order necessary condition of the optimization theory, any local minimizer \(\hat{\eta}\) of \(Q_n(\eta)\) from (2.4) must satisfy \(\hat{\eta} \in \{\eta : \langle \nabla L_n(\eta) + \nabla P_{\lambda_n}(\eta), -\nu \rangle \geq 0, ||\nu||_1 \leq R_n\}\), where \(\langle a, b \rangle = a'b\) for any vectors \(a, b\). Hence, in order to prove Theorem 3, it is sufficient to show that for any \(\eta \in \{\eta : \langle \nabla L_n(\eta) + \nabla P_{\lambda_n}(\eta), -\nu \rangle \geq 0, ||\nu||_1 \leq R_n\}\), parts 1) and 2) of Theorem 3 hold.

Next, we begin the proof. First, we assume

\[
\eta \in \{\eta : \langle \nabla L_n(\eta) + \nabla P_{\lambda_n}(\eta), -\nu \rangle \geq 0, ||\nu||_1 \leq R_n\}.
\]

(2.52)
Moreover, we have

\[
\langle \nabla L_n(\eta) - \nabla L_n(\eta^*), \nu \rangle = \tilde{\nu}'(n^{-1}\tilde{\Theta}'\tilde{\Theta})\tilde{\nu}
\]

\[
= \tilde{\nu}'E(n^{-1}\tilde{\Theta}'\tilde{\Theta})\tilde{\nu} - \tilde{\nu}'[E(n^{-1}\tilde{\Theta}'\tilde{\Theta}) - n^{-1}\tilde{\Theta}'\tilde{\Theta}]\tilde{\nu}
\]

\[
\geq \lambda_m(\tilde{I})||\tilde{\nu}||_2^2 - ||E(n^{-1}\tilde{\Theta}'\tilde{\Theta}) - n^{-1}\tilde{\Theta}'\tilde{\Theta}||_1||\tilde{\nu}||_1^2
\]

\[
\geq m_0||\tilde{\nu}||_2^2 - c_0n^{\beta/2-1/2}||\tilde{\nu}||_1^2;
\]

with probability tending to one, for some constant \(c_0 > 0\), where the last inequality is by (A4) and Lemma 8. In addition, with a little abuse of notation, denote 

\[
P_{\lambda_n, \mu}(\eta) = P_{\lambda_n}(\eta) + 2^{-1}\mu n^{-1}\sum_{j=1}^{p_n}\|\Theta_j\eta_j\|_2^2 = \sum_{j=1}^{p_n}\left(\rho_{\lambda_n}s_n^{1/2}(n^{-1/2}\|\Theta_j\eta_j\|_2) + 2^{-1}\mu n^{-1}\|\Theta_j\eta_j\|_2^2\right).
\]

Under (P5), (A3) and (A4), it is easy to see that 

\[
P_{\lambda_n, \mu}(\eta)\]

is convex in \(\eta\), which implies that

\[
P_{\lambda_n, \mu}(\eta^*) - P_{\lambda_n, \mu}(\eta) \geq \langle \nabla P_{\lambda_n, \mu}(\eta), -\nu \rangle,
\]

which further implies that

\[
\langle \nabla P_{\lambda_n}(\eta), -\nu \rangle \leq P_{\lambda_n}(\eta^*) - P_{\lambda_n}(\eta) + 2^{-1}\mu n^{-1}\sum_{j=1}^{p_n}\|\Theta_j(\eta_j - \eta_j^*)\|_2^2.
\]

By combining (2.52), (2.53) with (2.54), we have

\[
m_0||\tilde{\nu}||_2^2 - c_0n^{\beta/2-1/2}||\tilde{\nu}||_1^2 \leq \langle \nabla L_n(\eta) - \nabla L_n(\eta^*), \nu \rangle
\]

\[
= \langle \nabla L_n(\eta) + \nabla P_{\lambda_n}(\eta), \nu \rangle + \langle \nabla P_{\lambda_n}(\eta), -\nu \rangle + \langle -\nabla L_n(\eta^*), \nu \rangle
\]

\[
\leq P_{\lambda_n}(\eta^*) - P_{\lambda_n}(\eta) + 2^{-1}\mu n^{-1}\sum_{j=1}^{p_n}\|\Theta_j(\eta_j - \eta_j^*)\|_2^2
\]

\[
+ ||n^{-1}\tilde{\Theta}'(Y - \Theta\eta^*)||_\infty||\tilde{\nu}||_1.
\]

(2.55)
By combining (2.52), (2.55) with Lemma 9, we have

\[ m_0 ||\hat{\nu}||^2 \leq P_{\lambda_n}(\eta^*) - P_{\lambda_n}(\eta) + 2^{-1} \mu n^{-1} \sum_{j=1}^{p_n} ||\Theta_j(\eta_j - \eta_j^*)||^2 + c_1 (n^{\beta/2-1/2} R_n + q_n s_n^{-\delta}) ||\hat{\nu}||_1, \]  

(2.56)

with probability tending to one, for some constant \(c_1 > 0\). By combining (2.56) with Lemma 10 and (A9), we have

\[ [m_0 - 2^{-1} m_1 \mu (1 + o(1))] ||\hat{\nu}||^2 \leq (1 + o(1)) [P_{\lambda_n}(\eta^*) - P_{\lambda_n}(\eta)]. \]  

(2.57)

By combining (2.57) with Lemma 6 and (A4), we have

\[ 0 \leq [m_0 - 2^{-1} m_1 \mu (1 + o(1))] ||\hat{\nu}||^2 \leq c_2 \lambda_n s_n^{1/2} \left( \sum_{j \in A_n} n^{-1/2} ||\Theta_j(\eta_j - \eta_j^*)||_2 - \sum_{j \in A_n^c} n^{-1/2} ||\Theta_j(\eta_j - \eta_j^*)||_2 \right), \]  

(2.58)

with probability tending to one, for some constant \(c_2 > 0\). On one hand, (2.58) implies that

\[ ||\hat{\nu}||^2 \leq c_3 \lambda_n s_n^{1/2} \sum_{j \in A_n} n^{-1/2} ||\Theta_j(\eta_j - \eta_j^*)||_2 \leq c_3 \lambda_n s_n^{1/2} q_n^{1/2} \left( \sum_{j \in A_n} n^{-1} ||\Theta_j(\eta_j - \eta_j^*)||_2^2 \right)^{1/2} \leq c_3 \lambda_n s_n^{1/2} q_n^{1/2} \left( \sum_{j=1}^{p_n} n^{-1} ||\Theta_j(\eta_j - \eta_j^*)||_2^2 \right)^{1/2} \leq c_4 \lambda_n s_n^{1/2} q_n^{1/2} ||\hat{\nu}||_2, \]

with probability tending to one, for some constants \(c_3, c_4 > 0\), where the last inequality is by Lemma 10. Then, it follows that we have \( ||\hat{\nu}||_2 \leq c_4 \lambda_n s_n^{1/2} q_n^{1/2} \), which further entails that

\[ ||\nu||_2 = ||\Lambda^{-1}\hat{\nu}||_2 \leq \lambda_{\text{max}}(\Lambda^{-1}) ||\hat{\nu}||_2 = \left[ \lambda_{\text{min}}(\Lambda) \right]^{-1} ||\hat{\nu}||_2 \leq c_5 \lambda_n s_n^{n/2+1/2} q_n^{1/2}, \]
with probability tending to one, for some constant \(c_5 > 0\), which completes the proof of part
1). On the other hand, (2.58) also implies that

\[
\sum_{j \in A} n^{-1/2} ||\Theta_j(\eta_j - \eta_j^*)||_2 \geq \sum_{j \in \mathcal{A}_n} n^{-1/2} ||\Theta_j(\eta_j - \eta_j^*)||_2.
\]  

(2.59)

Under Lemma 10 and (2.59), we have

\[
||\tilde{\nu}||_1 \leq c_6 s_n^{1/2} \sum_{j=1}^{p_n} n^{-1/2} ||\Theta_j(\eta_j - \eta_j^*)||_2 \leq 2c_5 s_n^{1/2} \sum_{j \in \mathcal{A}_n} n^{-1/2} ||\Theta_j(\eta_j - \eta_j^*)||_2
\]

\[
\leq 2c_5 s_n^{1/2} q_n^{1/2} \left( \sum_{j=1}^{p_n} n^{-1} ||\Theta_j(\eta_j - \eta_j^*)||_2^2 \right)^{1/2} \leq c_6 s_n^{1/2} q_n^{1/2} ||\hat{\nu}||_2
\]

\[
\leq c_7 \lambda_n s_n q_n,
\]

with probability tending to one, for some constants \(c_5, c_6, c_7 > 0\), which further entails that

\[
||\nu||_1 = ||\Lambda^{-1} \hat{\nu}||_1 \leq \lambda_{\max}(\Lambda^{-1}) ||\hat{\nu}||_1 = [\lambda_{\min}(\Lambda)]^{-1} ||\hat{\nu}||_1 \leq c_8 \lambda_n s_n^{3/2+1} q_n,
\]

with probability tending to one, for some constant \(c_8 > 0\), which completes the proof of part 2).

**Proof of Theorem 4.** First, we have

\[
||\hat{T}||_\infty - ||T^*||_\infty \leq ||\hat{T} - T^*||_\infty
\]

\[
= ||n^{-1/2} \sum_{i=1}^{n} (\hat{S}_i - S_i)||_\infty = \max_{l \leq h_n s_n} ||n^{-1/2} \sum_{i=1}^{n} (\hat{S}_{il} - S_{il})||
\]

\[
\leq c_0 g(n),
\]  

(2.60)

with probability tending to one, for some constant \(c_0 > 0\), where \(g(n) = n^{\beta - 1/2} \rho_n s_n^{3\alpha/2} + \lambda_n n^{\beta/2} q_n s_n^{\alpha+1} + n^{\beta/2+1/2} q_n s_n^{-\delta} \log s_n + n^{\beta} \rho_n q_n s_n^{3\alpha/2-\delta} \log s_n\), and the last inequality is by Lemma 12.
Second, we have

\[ ||\hat{T}_e||_\infty - ||T^*_e||_\infty \leq ||\hat{T}_e - T^*_e||_\infty \]

\[ = |n^{-1/2} \sum_{i=1}^{n} e_i(\hat{S}_i - S_i)|_\infty = \max_{l\leq hns_n} |n^{-1/2} \sum_{i=1}^{n} e_i(\hat{S}_{il} - S_{il})|. \quad (2.61) \]

Since \( \{e_i : i \leq n\} \) is a set of independent and identically distributed standard normal random variables independent of the data, by the Hoeffding inequality, we have that for any \( l = 1, \ldots, hns_n \) and \( t > 0 \),

\[ P_e(|n^{-1/2} \sum_{i=1}^{n} e_i(\hat{S}_i - S_i)| \geq t) \leq 2 \exp\left(-\frac{t^2}{2n^{-1} \sum_{i=1}^{n} (\hat{S}_i - S_i)^2}\right), \]

where \( P_e(\cdot) \) means the probability with respect to \( e \). Then, the union bound inequality yields,

\[ P_e(\max_{l\leq hns_n} |n^{-1/2} \sum_{i=1}^{n} e_i(\hat{S}_i - S_i)| \geq t) \leq 2hns_n \exp\left(-\frac{t^2}{2 \max_{l\leq hns_n} \left[ n^{-1} \sum_{i=1}^{n} (\hat{S}_i - S_i)^2\right]}\right). \]

Together with Lemma 12, it is easy to see that

\[ P_e(\max_{l\leq hns_n} |n^{-1/2} \sum_{i=1}^{n} e_i(\hat{S}_i - S_i)| \leq c_1 f(n)) \to 1, \]

with probability tending to one, for some constant \( c_1 > 0 \), where

\[ f(n) = \lambda_n n^{3\beta/2} q_ns_n^{-\alpha/2+1} + n^\beta q_n s_n^{-\delta} \log s_n + \rho_n s_n^{3\alpha/2} n^{3\beta/2-1/2} (\log n)^{1/2} \]

\[ + \lambda_n \rho_n q_n s_n^{2\alpha+1} n^{2\beta-1/2} + n^{3\beta/2-1/2} \rho_n q_n s_n^{3\alpha/2-\delta} \log s_n. \]

Together with (2.61), we obtain

\[ P_e(||\hat{T}_e||_\infty - ||T^*_e||_\infty | \leq c_1 f(n)) \to 1. \quad (2.62) \]
Moreover, we have

\[ g(n)[\log(p_n s_n)]^{1/2} \sim g(n)n^{3/2} \]
\[ = n^{3\beta/2 - 1/2}\rho_n s_n^{3\alpha/2} + \lambda_n n^\beta q_n s_n^{a+1} + \rho_n s_n^{-\delta} n^{3\beta/2 - 1/2} \log s_n \]
\[ + n^{3\beta/2} \rho_n q_n s_n^{3a/2 - \delta} \log s_n = o(1), \tag{2.63} \]

under (A6), (A8) and (A9). In addition, we also have

\[ f(n)[\log(p_n s_n)]^{1/2} \sim f(n)n^{3/2} \]
\[ = \lambda_n n^{2\beta} q_n s_n^{a+1} + n^{3\beta/2} q_n s_n^{-\delta} \log s_n + \rho_n s_n^{3\alpha/2} n^{2\beta - 1/2} (\log n)^{1/2} \]
\[ + \lambda_n \rho_n q_n s_n^{2a+1} n^{5\beta/2 - 1/2} + n^{2\beta - 1/2} \rho_n q_n s_n^{3a/2 - \delta} \log s_n = o(1), \tag{2.64} \]

under (A6), (A8) and (A9). Furthermore, (A5) implies that

\[ E(S_{il}^2) \geq c_2, \tag{2.65} \]

for some universal constant \( c_2 > 0 \), and (A2) implies that

\[ ||S_{il}||_{\phi_2} \leq c_3, \tag{2.66} \]

for some universal constant \( c_3 > 0 \). Hence, by combining (2.60), (2.62), (2.63), (2.64), (2.65), (2.66) with (A6), we have

\[ \lim_{n \to \infty} \sup_{\alpha \in (0,1)} \left| P(||\hat{T}||_\infty \leq \phi'(\alpha)) - \alpha \right| = 0, \]

by Lemma H.7 in Ning and Liu [2016], which completes the proof.
Bibliography


