Are Latent Factor Regression and Sparse Regression Adequate?

Jianqing Fan Zhipeng Lou Mengxin Yu

Abstract

We propose the **F**actor **A**ugmented (sparse linear) **R**egression **M**odel (FARM) that not only admits both the latent factor regression and sparse linear regression as special cases but also bridges dimension reduction and sparse regression together. We provide theoretical guarantees for the estimation of our model under the existence of sub-Gaussian and heavy-tailed noises (with bounded $(1+\vartheta)$-th moment, for all $\vartheta > 0$) respectively. In addition, the existing works on supervised learning often assume the latent factor regression or sparse linear regression is the true underlying model without justifying its adequacy. To fill in such an important gap on high-dimensional inference, we also leverage our model as the alternative model to test the sufficiency of the latent factor regression and the sparse linear regression models. To accomplish these goals, we propose the **F**actor-**A**djusted de**Bi**ased **Test** (FabTest) and a two-stage ANOVA type test respectively. We also conduct large-scale numerical experiments including both synthetic and FRED macroeconomics data to corroborate the theoretical properties of our methods. Numerical results illustrate the robustness and effectiveness of our model against latent factor regression and sparse linear regression models.

Keyword: Factor model, High-dimensional Inference, Hypothesis, Sparse linear regression, Robustness

[1]Jianqing Fan is Frederick L. Moore '18 Professor of Finance, Professor of Statistics, and Professor of Operations Research and Financial Engineering at the Princeton University. Zhipeng Lou is a Postdoctoral Researcher at Department of Operations Research and Financial Engineering, Princeton University. Mengxin Yu is a Ph.D. student at Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA. Emails: {jqfan, zlou, mengxiny}@princeton.edu. The research is supported in part by the ONR grant N00014-22-1-2340, NSF grants DMS-2210833, DMS-2053832, DMS-2052926 and NIH grant 2R01-GM072611-16

1 Introduction

Over the past two decades, along with the development of technology, datasets with high-dimensionality in various fields such as biology, genomics, neuroscience and finance have been collected. One stylized feature of the high-dimensional data is the high dependence across features that give rises to near collinearity. A common structure to characterize the dependence across features is the approximate factor model [Bai, 2003, Fan et al., 2013], in which the variables are correlated with each other through several common latent factors. More specifically, we assume the observed d-dimensional covariate vector $\boldsymbol{x}$ follows from the model

$$\boldsymbol{x} = \boldsymbol{B}\boldsymbol{f} + \boldsymbol{u}, \tag{1.1}$$

where $\boldsymbol{f}$ is a K-dimensional vector of latent factors, $\boldsymbol{B} \in \mathbb{R}^{d\times K}$ is the corresponding factor loading matrix, and $\boldsymbol{u}$ is a d-dimensional vector of idiosyncratic component which is uncorrelated with $\boldsymbol{f}$.

To tackle the high-dimensionality of datasets, various methods have been proposed. Among these, dimensionality reduction and sparse regression are two popularly used ones to circumvent the curse of dimensionality. They also serve as the backbones for many emerging statistical methods.

In terms of dimension reduction, the factor regression model is one of the most popular methods and has been widely used [Stock and Watson, 2002, Bai and Ng, 2006, Bair et al., 2006, Bai and Ng, 2008, Fan et al., 2017b, Bing et al., 2019, Bunea et al., 2020, Bing et al., 2021]. It assumes that the latent factors drive both dependent and independent variables as follows:

$$\begin{aligned} Y &= \boldsymbol{f}^\top\boldsymbol{\gamma} + \varepsilon, \\ \boldsymbol{x} &= \boldsymbol{B}\boldsymbol{f} + \boldsymbol{u}. \end{aligned} \tag{1.2}$$

Here Y is the response variable and $\varepsilon \in \mathbb{R}$ is the random noise which is independent with the factor $\boldsymbol{f}$. When the factors are unobserved, one usually learns the latent factors based on observed $\boldsymbol{x}$ and substitutes the sample version into the regression model (1.2). There are several methods for estimating latent factors such as Principal Component Analysis (PCA) [Bai, 2003, Fan et al., 2013], maximum likelihood estimation [Bai and Li, 2012], and random projections [Fan and Liao, 2020]. In particular,

when the leading Principal Components are used as an estimator for $\boldsymbol{f}$, the sample version of (1.2) reduces to the classical Principal Component Regression (PCR) [Hotelling, 1933].

As for sparse regression, a commonly used model is the following (sparse) linear regression:

$$Y = \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon. \tag{1.3}$$

In the high dimensional regime where the dimension d can be much larger than the sample size n, it is commonly assumed that the population parameter vector $\boldsymbol{\beta} \in \mathbb{R}^d$ is sparse. Over the last two decades, various regularized methods, which incorporate this notion of sparsity, have been proposed. See, for instance, LASSO [Tibshirani, 1996], SCAD [Fan and Li, 2001], Least Angle Regression [Efron et al., 2004], Dantzig selector [Candes and Tao, 2007], Adaptive LASSO [Zou, 2006], MCP [Zhang, 2010] and many others. For more details, please refer to Fan et al. [2020b] for a comprehensive account.

In this paper, we introduce the **F**actor **A**ugmented (sparse linear) **R**egression **M**odel (FARM) (1.4), which incorporates both the latent factor and the idiosyncratic component into the covariates,

$$\begin{aligned} Y &= \boldsymbol{f}^\top \boldsymbol{\gamma}^\star + \boldsymbol{u}^\top \boldsymbol{\beta}^\star + \varepsilon, \\ \boldsymbol{x} &= \boldsymbol{B}\boldsymbol{f} + \boldsymbol{u}, \end{aligned} \tag{1.4}$$

where $\boldsymbol{\gamma}^\star \in \mathbb{R}^K$ and $\boldsymbol{\beta}^\star \in \mathbb{R}^d$ are population parameter vectors quantifying the contribution of the latent factor $\boldsymbol{f}$ and the idiosyncratic component $\boldsymbol{u}$, respectively. Obviously, the factor regression model (1.2) is a special case of (2.1) in which $\boldsymbol{\beta}^\star = 0$. To better illustrate the difference between model (1.4) and the sparse linear model (1.3), our model can be written in an equivalent form,

$$Y = \boldsymbol{f}^\top \boldsymbol{\varphi}^\star + \boldsymbol{x}^\top \boldsymbol{\beta}^\star + \varepsilon, \qquad \boldsymbol{x} = \boldsymbol{B}\boldsymbol{f} + \boldsymbol{u}, \tag{1.5}$$

where $\boldsymbol{\varphi}^\star = \boldsymbol{\gamma}^\star - \boldsymbol{B}^\top \boldsymbol{\beta}^\star \in \mathbb{R}^K$ quantifies the extra contribution of the latent factor $\boldsymbol{f}$ beyond the observed predictor $\boldsymbol{x}$. Therefore, FARM expands the space spanned by $\boldsymbol{x}$ into useful directions spanned by $\boldsymbol{f}$. It is clear that the sparse regression model (1.3) is also a special case of (1.4) with $\boldsymbol{\varphi}^\star = 0$. Thus, our model is general enough to bridge the dimensionality reduction and the sparse regression.

The motivation of our factor augmented linear model (1.4) comes from two perspectives.

1. Firstly, it origins from Fan et al. [2020a]. In order to get precise estimation of $\boldsymbol{\beta}^\star$ based on highly correlated variables, they study the sparse regression estimation by substituting (1.1) into (1.3) and obtain

$$Y = (\boldsymbol{B}\boldsymbol{f} + \boldsymbol{u})^\top \boldsymbol{\beta}^\star + \varepsilon = \boldsymbol{f}^\top (\boldsymbol{B}^\top \boldsymbol{\beta}^\star) + \boldsymbol{u}^\top \boldsymbol{\beta}^\star + \varepsilon. \tag{1.6}$$

   We observe from (1.6), when the sparse linear regression is adequate, for a given $\boldsymbol{\beta}^\star$, the regression coefficient on $\boldsymbol{f}$ is fixed at $\boldsymbol{\gamma}^\star = \boldsymbol{B}^\top \boldsymbol{\beta}^\star$. However, in reality, especially when the variables are highly correlated, it is very likely that the leading factors possess extra contributions to the response instead of only a fixed portion $\boldsymbol{B}^\top \boldsymbol{\beta}^\star$. This results in our proposition of model (1.4), where we augment the leading factors into sparse regression that expands the linear space spanned by $\boldsymbol{x}$ into useful directions.

2. Secondly, it origins from the factor regression given in (1.2). In reality, the leading common factors $\boldsymbol{f}$ indeed provides some important contributions to the response, but it is hard to believe that they will have fully explanation power, especially when the effect of the factors is weak. Besides, in real applications, several examples illustrate the poor performance of factor regression model or PCR, see Jolliffe [1982] for more details. Thus, completely ignoring the idiosyncratic component $\boldsymbol{u}$ will harm in model generalization. This also motivates us to propose model (1.4), in which we augment the sparse regression by incorporating the idiosyncratic component $\boldsymbol{u}$ into the original factor regression.

In this paper, we first study the properties of estimated parameters under the proposed model (1.4). Specifically, we assume the factors given in (1.4) are unobserved and leverage PCA to estimate them. Incorporated with penalized least-squares with the ℓ_1-penaly, we derive the ℓ_2-consistency results for parameter vectors $\boldsymbol{\gamma}^\star$ and $\boldsymbol{\beta}^\star$. Going beyond the linear regression model and the least squares estimation, our idea can be naturally extended to more general supervised learning models through different loss functons. For instance, quantile regression [Belloni and Chernozhukov, 2011, Fan et al., 2014], support vector machine [Zhang et al., 2016, Peng et al., 2016], Huber regression [Fan et al., 2017a, Sun et al., 2020], generalized linear model [Van de Geer, 2008, Fan et al., 2020a] and many other variants.

In order to demonstrate the general applicability of our proposed methods, in our paper, we further extend our model settings to robust regression. To be more specific, we only assume the existence of $(1+\vartheta)$-th moment of the noise distribution for some $\vartheta > 0$. We adopt Huber loss together with adaptive tuning parameters and ℓ_1-penalization to derive the consistency results for the parameters of our interest. Besides the aforementioned extensions, it is worth to note that our model is also applicable in the field of causal inference [Imbens and Rubin, 2015, Hernan and Robins, 2019]. To be more specific, the latent factors $\boldsymbol{f}$ given in our model is able to be treated as the unobserved confounding variables which affect both the covariate $\boldsymbol{x}$ and the response Y. From the causal perspective, we provide a methodology to conduct (robust) statistical estimation as well as inference of our model under the existence of latent confounding variables.

The aforementioned works on factor regression and sparse linear regression mainly investigate the theoretical properties based on the assumption that either of them is the true underlying model [Stock and Watson, 2002, Tibshirani, 1996, Fan and Li, 2001, Zou, 2006, Bai and Ng, 2006, Zhang, 2010, Fan et al., 2017b, 2020a, Bing et al., 2021]. However, whether a given model is adequate to explain a given dataset plays a crucial role in the model selection step. This motivates us to fill the gap by leveraging our model as the alternative one to perform hypothesis testing on the adequacy of the factor regression model as well as the sparse linear regression model when covariates admit a factor structure.

For the hypothesis test on the adequacy of the latent factor regression model, we consider testing the hypotheses

$$H_0 : Y = \boldsymbol{f}^\top \boldsymbol{\gamma}^\star + \varepsilon \ \text{ versus } \ H_1 : Y = \boldsymbol{f}^\top \boldsymbol{\gamma}^\star + \boldsymbol{u}^\top \boldsymbol{\beta}^\star + \varepsilon. \tag{1.7}$$

This amounts to testing $H_0 : \boldsymbol{\beta}^\star = 0$ under FARM model. To this end, we propose the **F**actor-**A**djusted de**B**iased **Test** statistic (FabTest) $\widetilde{\boldsymbol{\beta}}_\lambda$ which serves as a de-sparsify version of the estimator $\widehat{\boldsymbol{\beta}}_\lambda$ obtained under ℓ_1-regularization. The asymptotic distribution of the proposed test statistic is derived by leveraging the high-dimensional Gaussian approximation. The critical value controlling the Type-I error is estimated based on the multiplier bootstrap method. As a byproduct, we are also able to conduct entrywise and groupwise hypothesis testing on parameter $\boldsymbol{\beta}^\star$ by following similar de-biasing procedure.

For validating the adequacy of the sparse linear regression model, we consider testing the hypotheses

$$H_0 : Y = \boldsymbol{x}^\top \boldsymbol{\beta}^\star + \varepsilon \ \text{ versus } \ H_1 : Y = \boldsymbol{f}^\top \boldsymbol{\varphi}^\star + \boldsymbol{x}^\top \boldsymbol{\beta}^\star + \varepsilon, \tag{1.8}$$

or $\boldsymbol{\varphi}^\star = 0$ under the FARM model. To tackle the testing problem, we propose a two-stage ANOVA test. In the first stage, we use marginal screening [Fan and Lv, 2008] to pre-select a group of variables which cope well the curse of high dimensionality. In the second stage, we derive the ANOVA-type test statistic. Asymptotic null distribution and the power of the test statistic are derived. In addition, we further extend the aforementioned two-stage ANOVA test to linear multi-modal models [Li and Li, 2021], whose data framework has been well applied in a wide range of scientific fields (e.g multi-omics data in genomics, multimodal neuroimaging data in neuroscience, multimodal electronic health records data in health care).

In summary, our main contributions are as follows:

1. Motivated from the factor regression and sparse regression, we propose the **F**actor **A**ugmented (sparse linear) **R**egression **M**odel (FARM) (1.4) [also (1.5)] and investigate in the parameter estimation properties on $\boldsymbol{\gamma}^\star$ and $\boldsymbol{\beta}^\star$ given in (1.4). Our work serves as an extension of Fan et al. [2020a] to a general setting with weaker assumptions. It augments the sparse linear regression in useful directions of common factors.

2. To further demonstrate the wide applicability of our methods, we extend our model to a more robust setting, where we only assume the existence of $(1 + \vartheta)$-th moment ($\vartheta > 0$) of our noise distribution. Leveraging the ℓ_1-penalized adaptive Huber estimation, we establish statistical estimation results for our parameters of interest. Comparing with those closely related literature [Fan et al., 2020a, 2021a], our assumption on the moment condition of the noise variable is the weakest. Our robustified factor augmented regression also serves as an extension of Sun et al. [2020] to a more general setting.

3. In terms of testing the adequacy of the factor regression, we propose the **FabTest** by incorporating the factor structure into the de-biased estimators [van de Geer et al., 2014, Zhang and

Zhang, 2014, Javanmard and Montanari, 2014]. Accompanied with Gaussian approximation, the asymptotic distribution of our test statistic is derived. As for implementation, we propose the multiplier bootstrap method to estimate the critical value in order to control the Type-I error.

4. For testing the adequacy of sparse linear regression model, we propose a two stage ANOVA-type testing procedure. Asymptotic distribution (under the null) and power (under the alternative) of our constructed test statistic are investigated. In addition, we further extend the methodology to the multi-modal sparse linear regression model [Li and Li, 2021], by testing whether the sparse linear regression for some given modals is adequate.

5. We conduct large scale simulation studies for our proposed methodology using both synthetic data and real data. Simulation results via synthetic data lend further support to our theoretical findings. As for real data, we apply our methodology to the studies of the macroeconomics dataset named FRED-MD [McCracken and Ng, 2016]. The experimental results also illustrate the high efficiency and robustness of our model (FARM) against latent factor regression as well as sparse linear regression.

1.1 Notation

For a vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)^\top \in \mathbb{R}^m$, we denote its ℓ_q norm as $\|\boldsymbol{\gamma}\|_q = (\sum_{\ell=1}^m |\gamma_\ell|^q)^{1/q}$, $1 \leqslant q < \infty$, and write $\|\boldsymbol{\gamma}\|_\infty = \max_{1\leqslant \ell \leqslant m} |\gamma_\ell|$. For any integer m, we denote $[m] = \{1, \ldots, m\}$. The sub-Gaussian norm of a scalar random variable Z is defined as $\|Z\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}\exp(Z^2/t^2) \leqslant 2\}$. For a random vector $\boldsymbol{x} \in \mathbb{R}^m$, we use $\|\boldsymbol{x}\|_{\psi_2} = \sup_{\|\boldsymbol{v}\|_2=1} \|\boldsymbol{v}^\top \boldsymbol{x}\|_{\psi_2}$ to denote its sub-Gaussian norm. Let $\mathbb{I}\{\cdot\}$ denote the indicator function and let $\mathbf{I}_K$ denotes the identity matrix in $\mathbb{R}^{K\times K}$. For a matrix $\boldsymbol{A} = [A_{jk}]$, we define $\|\boldsymbol{A}\|_{\mathbb{F}} = \sqrt{\sum_{jk} A_{jk}^2}$, $\|\boldsymbol{A}\|_{\max} = \max_{jk} |A_{jk}|$ and $\|\boldsymbol{A}\|_\infty = \max_j \sum_k |A_{jk}|$ to be its Frobenius norm, element-wise max-norm and matrix ℓ_∞-norm, respectively. Moreover, we use $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ to denote the minimal and maximal eigenvalues of $\boldsymbol{A}$, respectively. We use $|\mathcal{A}|$ to denote the cardinality of set $\mathcal{A}$. For two positive sequences $\{a_n\}_{n\geqslant 1}$, $\{b_n\}_{n\geqslant 1}$, we write $a_n = O(b_n)$ if there exists a positive constant C such that $a_n \leqslant C \cdot b_n$ and we write $a_n = o(b_n)$ if $a_n/b_n \to 0$. In addition, $a_n = O_{\mathbb{P}}(b_n)$ and $a_n = o_{\mathbb{P}}(b_n)$ have similar meanings as above except that the relationship of a_n/b_n

holds with high probability.

1.2 RoadMap

The rest of this paper is organized as follows. We study the parameter estimation properties of our proposed model (FARM) in section 2, where theoretical results of both regular and robust estimators are analyzed. In section 3, we construct a de-biased test statistic to test the adequacy of latent factor regression model. In addition, in section 4, we construct a two-stage ANOVA test to study the adequacy of sparse linear regression under the setting with highly correlated features. Moreover, to corroborate our theoretical findings, in section 5, we conduct exhaustive simulation studies. Last but not least, we apply our methodology to study the real data FRED-MD in section 5.4.

2 Factor Augmented Regression Model

The primary objective of this section is to propose a regularized estimation method for our factor augmented sparse linear model and investigate the corresponding statistical properties. Suppose we observe n independent and identically distributed (i.i.d.) random samples $\{(\boldsymbol{x}_t, Y_t)\}_{t=1}^n$ from $(\boldsymbol{x}, Y)$, which satisfy that

$$\boldsymbol{x}_t = \boldsymbol{B}\boldsymbol{f}_t + \boldsymbol{u}_t \text{ and } Y_t = \boldsymbol{f}_t^\top \boldsymbol{\gamma}^\star + \boldsymbol{u}_t^\top \boldsymbol{\beta}^\star + \varepsilon_t, \quad t = 1, \ldots, n, \tag{2.1}$$

where $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n \in \mathbb{R}^K$, $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n \in \mathbb{R}^d$ and $\varepsilon_1, \ldots, \varepsilon_n \in \mathbb{R}$ are i.i.d. realizations of $\boldsymbol{f}$, $\boldsymbol{u}$ and ε, respectively. To ease the presentation, we rewrite (2.1) in a more compact matrix form as follows,

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{F}\boldsymbol{B}^\top + \boldsymbol{U}, \\ \boldsymbol{Y} &= \boldsymbol{F}\boldsymbol{\gamma}^\star + \boldsymbol{U}\boldsymbol{\beta}^\star + \mathcal{E}, \end{aligned} \tag{2.2}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$, $\boldsymbol{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)^\top$, $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)^\top$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ and $\mathcal{E} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$. Throughout the whole paper, we assume we only get access to observations $\{(\boldsymbol{x}_t, Y_t)\}_{t=1}^n$. Both the latent factors $\boldsymbol{F}$ and the idiosyncratic components $\boldsymbol{U}$ are unobserved and need

to be estimated from the observed predictors $\boldsymbol{X}$. Thus, in the following, we shall first illustrate how to estimate $\boldsymbol{F}$ and $\boldsymbol{U}$ and then proceed with the regularized estimation for model (2.2).

2.1 Factor Estimation

Since only the predictor vector $\boldsymbol{x}$ is observable, the latent factor $\boldsymbol{f}$ and the corresponding loading matrix $\boldsymbol{B}$ are not identifiable under the factor model (1.1). More specifically, for any non-singular matrix $\boldsymbol{S} \in \mathbb{R}^{K\times K}$, we have $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{f} + \boldsymbol{u} = (\boldsymbol{B}\boldsymbol{S})(\boldsymbol{S}^{-1}\boldsymbol{f}) + \boldsymbol{u}$. To resolve this issue, we impose the following identifiability conditions [Bai, 2003, Fan et al., 2013]:

$$\mathrm{Cov}(\boldsymbol{f}) = \boldsymbol{I}_K \ \text{ and } \ \boldsymbol{B}^\top\boldsymbol{B} \ \text{ is diagonal.}$$

Consequently, the constrained least squares estimator of $(\boldsymbol{F}, \boldsymbol{B})$ based on $\boldsymbol{X}$ is given by

$$(\widehat{\boldsymbol{F}}, \widehat{\boldsymbol{B}}) = \underset{\boldsymbol{F}\in\mathbb{R}^{n\times K}, \boldsymbol{B}\in\mathbb{R}^{d\times K}}{\arg\min} \|\boldsymbol{X} - \boldsymbol{F}\boldsymbol{B}^\top\|_{\mathbb{F}}^2$$
$$\text{subject to } \frac{1}{n}\boldsymbol{F}^\top\boldsymbol{F} = \boldsymbol{I}_K \ \text{ and } \ \boldsymbol{B}^\top\boldsymbol{B} \ \text{ is diagonal.}$$

Elementary manipulation yields that the columns of $\widehat{\boldsymbol{F}}/\sqrt{n}$ are the eigenvectors corresponding to the largest K eigenvalues of the matrix $\boldsymbol{X}\boldsymbol{X}^\top$ and $\widehat{\boldsymbol{B}} = (\widehat{\boldsymbol{F}}^\top\widehat{\boldsymbol{F}})^{-1}\widehat{\boldsymbol{F}}^\top\boldsymbol{X} = n^{-1}\widehat{\boldsymbol{F}}^\top\boldsymbol{X}$. Then the least squares estimator for $\boldsymbol{U}$ is given by $\widehat{\boldsymbol{U}} = \boldsymbol{X} - \widehat{\boldsymbol{F}}\widehat{\boldsymbol{B}}^\top = (\boldsymbol{I}_n - n^{-1}\widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}^\top)\boldsymbol{X}$.

Before presenting the asymptotic properties of the estimators $\{\widehat{\boldsymbol{F}}, \widehat{\boldsymbol{B}}, \widehat{\boldsymbol{U}}\}$, we first impose some regularity conditions.

Assumption 2.1. There exists a positive constant $c_0 < \infty$ such that $\|\boldsymbol{f}\|_{\psi_2} \leqslant c_0$ and $\|\boldsymbol{u}\|_{\psi_2} \leqslant c_0$.

Assumption 2.2. There exists a constant $\tau > 1$ such that $d/\tau \leqslant \lambda_{\min}(\boldsymbol{B}^\top\boldsymbol{B}) \leqslant \lambda_{\max}(\boldsymbol{B}^\top\boldsymbol{B}) \leqslant d\tau$. Moreover, we assume $n\log^2 n = O(d)$.

Assumption 2.3. Let $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{u})$. There exists a constant $\Upsilon > 0$ such that $\|\boldsymbol{B}\|_{\max} \leqslant \Upsilon$ and

$$\mathbb{E}|\boldsymbol{u}^\top\boldsymbol{u} - \mathrm{tr}(\boldsymbol{\Sigma})|^4 \leqslant \Upsilon d^2.$$

Assumption 2.4. There exist a positive constant $\kappa < 1$ such that $\kappa \leqslant \lambda_{\min}(\boldsymbol{\Sigma})$, $\|\boldsymbol{\Sigma}\|_1 \leqslant 1/\kappa$ and $\min_{1\leqslant k,\ell\leqslant d}\mathrm{Var}(u_k u_\ell) \geqslant \kappa$.

Remark 1. Assumptions 2.1–2.4 are standard assumptions in the studies of large dimensional factor model. We refer to Bai [2003], Fan et al. [2013] and Li et al. [2018] for more details. □

We next summarize the theoretical results related to consistent factor estimation in the following proposition which directly follows from Lemmas D.1 and D.2 in Wang and Fan [2017].

Proposition 2.1. *Assume that* $d = o(\exp(n))$. *Let* $\boldsymbol{H} = n^{-1}\boldsymbol{V}^{-1}\widehat{\boldsymbol{F}}^\top \boldsymbol{F}\boldsymbol{B}^\top \boldsymbol{B}$, *where* $\boldsymbol{V} \in \mathbb{R}^{K\times K}$ *is a diagonal matrix consisting of the first* K *largest eigenvalues of the matrix* $n^{-1}\boldsymbol{X}\boldsymbol{X}^\top$. *Then, under Assumptions 2.1–2.4, we have*

1. $\|\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H}^\top\|_{\mathbb{F}}^2 = O_{\mathbb{P}}(n/d + 1/n)$.
2. *For any* $\mathcal{I} \subset \{1, 2, \ldots, d\}$, *we have* $\max_{\ell\in\mathcal{I}} \sum_{t=1}^n |\widehat{u}_{t\ell} - u_{t\ell}|^2 = O_{\mathbb{P}}(\log|\mathcal{I}| + n/d)$.
3. $\|\boldsymbol{H}^\top\boldsymbol{H} - \boldsymbol{I}_K\|_{\mathbb{F}}^2 = O_{\mathbb{P}}(1/n + 1/d)$.
4. $\max_{\ell\in[d]} \|\widehat{\boldsymbol{b}}_\ell - \boldsymbol{H}\boldsymbol{b}_\ell\|_2^2 = O_{\mathbb{P}}\{(\log d)/n\}$.

Remark 2. In practice, the number of latent factors K is typically unknown and it is an important issue to determine K in a data-driven way. There have been various methods proposed in the literature to estimate the number K [Bai and Ng, 2002, Lam and Yao, 2012, Ahn and Horenstein, 2013, Fan et al., 2022]. Our theories always work as long as we replace K by any consistent estimator $\widehat{K}$, i.e. we only require

$$\mathbb{P}(\widehat{K} = K) \to 1, \quad \text{as} \quad n \to \infty.$$

Thus, without loss of generality, we assume the number of factors K is known throughout all the theories developed in this paper. As for the application part, throughout this paper, we utilize the eigenvalue ratio method [Lam and Yao, 2012, Ahn and Horenstein, 2013] to select the number of factors. More specifically, we let $\lambda_k(\boldsymbol{X}\boldsymbol{X}^\top)$ denote the eigenvalues of the Gram matrix $\boldsymbol{X}\boldsymbol{X}^\top$ and the number of factors is given by

$$\widehat{K} = \arg\max_{K\leqslant\mathcal{K}} \frac{\lambda_k(\boldsymbol{X}\boldsymbol{X}^\top)}{\lambda_{k+1}(\boldsymbol{X}\boldsymbol{X}^\top)}, \tag{2.3}$$

where $1 \leqslant \mathcal{K} \leqslant n$ is a prescribed upper bound for K. □

2.2 Regularized Estimation

Under the high dimensional regime where the dimension d can be much larger than the sample size n, it is often assumed that only a small portion of the predictors contribute to the response variable, which amounts to assuming that the true parameter vector $\boldsymbol{\beta}^\star$ is sparse. Then the regularized estimator for the unknown parameter vectors $\boldsymbol{\beta}^\star$ and $\boldsymbol{\gamma}^\star$ of our factor augmented linear model is defined as follows:

$$(\widehat{\boldsymbol{\beta}}_\lambda, \widehat{\boldsymbol{\gamma}}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^d, \boldsymbol{\gamma}\in\mathbb{R}^K}{\arg\min} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \widehat{\boldsymbol{U}}\boldsymbol{\beta} - \widehat{\boldsymbol{F}}\boldsymbol{\gamma}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}, \tag{2.4}$$

where $\lambda > 0$ is a tuning parameter.

We let $\widetilde{\boldsymbol{Y}} = (\boldsymbol{I}_n - \widehat{\boldsymbol{P}})\boldsymbol{Y}$ denote the residuals of the response vector $\boldsymbol{Y}$ after projecting onto the column space of $\widehat{\boldsymbol{F}}$, where $\widehat{\boldsymbol{P}} = n^{-1}\widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}^\top$ is the corresponding projection matrix. Recall that $\widehat{\boldsymbol{U}} = (\boldsymbol{I}_n - \widehat{\boldsymbol{P}})\boldsymbol{X}$. Hence $\widehat{\boldsymbol{F}}^\top\widehat{\boldsymbol{U}} = 0$ and it is straightforward to verify that the solution of (2.4) is equivalent to

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_\lambda &= \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\arg\min} \left\{ \frac{1}{2n} \|\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{U}}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}, \\ \widehat{\boldsymbol{\gamma}} &= (\widehat{\boldsymbol{F}}^\top\widehat{\boldsymbol{F}})^{-1}\widehat{\boldsymbol{F}}^\top\boldsymbol{Y} = \frac{1}{n}\widehat{\boldsymbol{F}}^\top\boldsymbol{Y}. \end{aligned}$$

For any subset $\mathcal{S}$ of $\{1, \ldots, d\}$, we define the convex cone $\mathcal{C}(\mathcal{S}, 3) = \{\boldsymbol{\delta} \in \mathbb{R}^d : \|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 \leqslant 3\|\boldsymbol{\delta}_{\mathcal{S}}\|_1\}$. For simplicity of notation, we write

$$\mathcal{V}_{n,d} = \frac{n}{d} + \sqrt{\frac{\log d}{n}} + \sqrt{\frac{n \log d}{d}}. \tag{2.5}$$

To investigate the consistency property of $(\widehat{\boldsymbol{\beta}}_\lambda, \widehat{\boldsymbol{\gamma}})$, we impose the following moment condition on the random noise ε.

Assumption 2.5. There exists a positive constant $c_1 < \infty$ such that $\|\varepsilon\|_{\psi_2} \leqslant c_1$.

Theorem 2.2. *Recall $\boldsymbol{\varphi}^\star = \boldsymbol{\gamma}^\star - \boldsymbol{B}^\top\boldsymbol{\beta}^\star \in \mathbb{R}^K$. Under Assumptions 2.1–2.5, we have*

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{H}\boldsymbol{\gamma}^\star\|_2 = O_{\mathbb{P}}\left\{ \frac{1}{\sqrt{n}} + \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{d}} \right) \|\boldsymbol{\varphi}^\star\|_2 + \|\boldsymbol{\beta}^\star\|_1 \left(\sqrt{\frac{\log|\mathcal{S}_\star|}{n}} + \frac{1}{\sqrt{d}} \right) \right\},$$

where $\mathcal{S}_\star = \{j \in [d] : \beta_j^\star \neq 0\}$ *and* $|\mathcal{S}_\star|$ *is its cardinality. Furthermore, if* $|\mathcal{S}_\star|\left(\frac{\log d}{n} + \frac{1}{d}\right) = o(1)$*, then, by taking* $\lambda = (\mathcal{I}_0/n)\|\widehat{\boldsymbol{U}}^\top(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{U}}\boldsymbol{\beta}^\star)\|_\infty$ *for some constant* $\mathcal{I}_0 \geqslant 2$*, we have* $\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star \in \mathcal{C}(\mathcal{S}_\star, 3)$ *and*

$$\|\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{|\mathcal{S}_\star|\log d}{n}} + \frac{\mathcal{V}_{n,d}\|\boldsymbol{\varphi}^\star\|_2\sqrt{|\mathcal{S}_\star|}}{n}\right). \tag{2.6}$$

Remark 3. In most of literature investigating the regularized estimation of sparse linear regression model (1.3), it is commonly assumed that the observed covariate vector $\boldsymbol{x}$ is a sub-Gaussian random vector with bounded sub-Gaussian norm $\|\boldsymbol{x}\|_{\psi_2}$. See, for instance, Loh and Wainwright [2012], Nickl and Van De Geer [2013], van de Geer et al. [2014], Zhang and Cheng [2017] and many others. However, such assumption can be unreasonable in the presence of highly correlated covariates. To see this, suppose now both $\boldsymbol{f}$ and $\boldsymbol{u}$ are Gaussian random vectors and the underlying $\boldsymbol{x}$ satisfies the factor model (1.1). Then $\boldsymbol{x}$ is also a Gaussian random vector with $\mathrm{Cov}(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{\Sigma}$. Under the pervasiveness condition (Assumption 2.2) and Assumption 2.4, it is straightforward to verify that $\|\boldsymbol{x}\|_{\psi_2} = \sqrt{8/3}\lambda_{\max}(\boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{\Sigma}) \asymp d$, which violates the assumption on bounded sub-Gaussian norm. In contrast, our model can circumvent such issue because we decompose the covariate $\boldsymbol{x}$ into $(\boldsymbol{f}, \boldsymbol{u})$, and we only need impose sub-Gaussian assumption on $(\boldsymbol{f}, \boldsymbol{u})$. As the sparse linear regression model serves as a special case to our model, our model serves as a more robust choice to conduct parameter estimation comparing with using linear regression directly, even if the sparse linear regression model is adequate. □

Remark 4. Theorem 2.2 substantially generalize the results in Fan et al. [2020a] with weaker assumptions. First, we did not impose the irrepresentable condition on the design matrix $\boldsymbol{U}$, only the lower bound on $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{u})$ is required. In addition, although Fan et al. [2020a] also decompose the covariate $\boldsymbol{x}$ into $(\boldsymbol{f}, \boldsymbol{u})$ in order to get precise estimator for $\boldsymbol{\beta}^\star$, they mainly focus on the linear model $Y = \boldsymbol{x}^\top\boldsymbol{\beta}^\star + \varepsilon$ which corresponds to the special case with $\boldsymbol{\varphi}^\star = 0$ in our results given in Theorem 2.2. □

Remark 5. Our study is different from the related work by Fan et al. [2021a], although they also study one kind of factor augment linear regression model. First of all, they do hypothesis testing for covariance matrix of the idiosyncratic component whereas we focus on testing the adequacy of factor

regression and sparse regression and address also robustness issue. Secondly, their study focuses on panel data and concerning more on prediction rather than the inference. □

2.3 Factor Augmented Robust Linear Regression

In reality, datasets, especially collected from the field of finance, are often contaminated by noises with relatively heavy tails. To resolve such issue, we leverage the adaptive Huber regression to study the parameter of interest in our FARM under the existence of heavy-tailed noise [Sun et al., 2020].

Let $\rho_\omega(\cdot)$ denote the Huber function,

$$\rho_\omega(z) = \begin{cases} z^2/2, & \text{if } |z| \leqslant \omega, \\ \omega z - \omega^2/2, & \text{if } |z| > \omega, \end{cases}$$

where $\omega > 0$ is the robustification parameter which balances robustness and bias. As an robust version of (2.4), our factor augmented adaptive Huber estimator for $(\boldsymbol{\beta}^\star, \boldsymbol{\gamma}^\star)$ is given by

$$(\widehat{\boldsymbol{\beta}}_h, \widehat{\boldsymbol{\gamma}}_h) = \underset{\boldsymbol{\beta}\in\mathbb{R}^d, \boldsymbol{\gamma}\in\mathbb{R}^K}{\arg\min} \left\{ \frac{1}{n}\sum_{t=1}^{n} \rho_\omega(y_t - \widehat{\boldsymbol{u}}_t^\top \boldsymbol{\beta} - \widehat{\boldsymbol{f}}_t^\top \boldsymbol{\gamma}) + \lambda\|\boldsymbol{\beta}\|_1 \right\}, \tag{2.7}$$

where $\lambda > 0$ is a tuning parameter and ω depends on sample size, dimenality, and noise level. For simplicity of notation, we write $\widehat{\boldsymbol{\phi}}_h = (\widehat{\boldsymbol{\beta}}_h^\top, \widehat{\boldsymbol{\gamma}}_h^\top)^\top \in \mathbb{R}^{d+K}$ and $\widetilde{\boldsymbol{\phi}} = (\boldsymbol{\beta}^{\star\top}, \widetilde{\boldsymbol{\gamma}}^\top)^\top \in \mathbb{R}^{d+K}$, where $\widetilde{\boldsymbol{\gamma}} = \widehat{\boldsymbol{B}}^\top\boldsymbol{\beta}^\star + n^{-1}\widehat{\boldsymbol{F}}^\top\boldsymbol{F}\boldsymbol{\varphi}^\star$. The following theorem establishes the statistical consistency of $\widehat{\boldsymbol{\phi}}_h$.

Proposition 2.3. *Assume that $\mathbb{E}|\varepsilon|^{1+\vartheta} < \infty$ for some constant $\vartheta > 0$. Let*

$$\omega \asymp \left(\frac{n}{\log d}\right)^{\frac{1}{1+(\vartheta\wedge 1)}} \quad \text{and} \quad \lambda \asymp \left(\frac{\log d}{n}\right)^{\frac{\vartheta\wedge 1}{1+(\vartheta\wedge 1)}}.$$

Furthermore, we assume that $(|\mathcal{S}_\star| + K)(\log d)^{3/2} = o(n)$,

$$\frac{\log n}{n+\sqrt{d}}\|\boldsymbol{\varphi}^\star\|_2 = o(\omega) \quad \text{and} \quad \mathcal{V}_{n,d}\|\boldsymbol{\varphi}^\star\|_2 = O(\omega\log d). \tag{2.8}$$

Then, under Assumptions 2.1–2.4, we have

$$\|\widehat{\boldsymbol{\phi}}_h - \widetilde{\boldsymbol{\phi}}\|_1 = O_{\mathbb{P}}\left\{(|\mathcal{S}_\star| + K)\left(\frac{\log d}{n}\right)^{\frac{\vartheta\wedge 1}{1+(\vartheta\wedge 1)}}\right\}.$$

We establish the ℓ_1-statistical rate for the parameters in model (1.4)[also (1.5)] by only assuming the existence of $(1+\vartheta)$-th moment of the noise distribution. Specifically, when $\vartheta \geqslant 1$, the results reduce to the same rate as the sub-Gaussian assumption of ε. Our result serves as an extension of Sun et al. [2020], who study the robust estimation for high-dimensional linear regression, to a more general setting by incorporating latent factors.

Remark 6. It is worth noting that the statistical errors we obtained throughout section 2 are non-asymptotic, in the sense that the results always hold as long as n is greater than some fixed constant. As our learned covariates contain statistical errors, novel analysis analysis is required for downstream statistical estimation and inference under both scenarios with light and heavy-tailed noises

3 Is Factor Regression Model Adequate?

The latent factor regression is widely applied in many fields as an efficient dimension reduction method. A natural question arises is whether the model is adequate and FARM (1.4) serves naturally as the alternative model. To be more specific, we consider testing the hypotheses

$$H_0 : \boldsymbol{\beta}^\star = 0 \ \text{ versus } \ H_1 : \boldsymbol{\beta}^\star \neq 0 \tag{3.1}$$

in FARM (1.4). As the penalized least-squares estimator $\widehat{\boldsymbol{\beta}}_\lambda$ is used for estimating $\boldsymbol{\beta}^\star$, it creates biases and make it difficulty for inferences. Thus, we first introduce a de-biased version of $\widehat{\boldsymbol{\beta}}_\lambda$ given in (2.4).

3.1 Bias Correction

We begin with the construction of bias-corrected estimator for $\boldsymbol{\beta}^\star$ following similar idea of Zhang and Zhang [2014], van de Geer et al. [2014] and Javanmard and Montanari [2014]. Specifically, let $\widehat{\boldsymbol{\Theta}} \in \mathbb{R}^{d\times d}$ be an approximation for the inverse of the Gram matrix $\widetilde{\boldsymbol{\Sigma}} = n^{-1}\widehat{\boldsymbol{U}}^\top\widehat{\boldsymbol{U}}$, the de-biased estimator for $\boldsymbol{\beta}^\star$ is then defined as

$$\widetilde{\boldsymbol{\beta}}_\lambda = \widehat{\boldsymbol{\beta}}_\lambda + \frac{1}{n}\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{U}}^\top(\boldsymbol{Y} - \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\beta}}_\lambda). \tag{3.2}$$

The rationale behind such construction is that we are able to decompose estimation error as

$$\widetilde{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star = \frac{1}{n}\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{U}}^\top \mathcal{E} + \frac{1}{n}\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{U}}^\top \boldsymbol{F}\boldsymbol{\varphi}^\star + (\boldsymbol{I}_d - \widehat{\boldsymbol{\Theta}}\widetilde{\boldsymbol{\Sigma}})(\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star), \tag{3.3}$$

after we expand $\boldsymbol{Y}$ according to (2.2) and replace $\boldsymbol{X}$ by $\boldsymbol{X} = \widehat{\boldsymbol{F}}\widehat{\boldsymbol{B}} + \widehat{\boldsymbol{U}}$. The first term on the right hand side of (3.3) quantifies the uncertainty of our estimator $\widetilde{\boldsymbol{\beta}}_\lambda$ and the last two terms are biases which will be shown to be of smaller order.

One observes that constructing the de-biased estimator $\widetilde{\boldsymbol{\beta}}_\lambda$ given above requires an estimator $\widehat{\boldsymbol{\Theta}}$. There are many methods for estimating such precision matrix, for example, the node-wise regression proposed in Zhang and Zhang [2014] and van de Geer et al. [2014], and the CLIME-type estimator given in Cai et al. [2011], Javanmard and Montanari [2014] and Avella-Medina et al. [2018]. In our work, we do not restrict $\widehat{\boldsymbol{\Theta}}$ to be any specific one, but require to satisfy the following general conditions.

Assumption 3.1. Let $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ with $\boldsymbol{\Sigma}$ defined in Assumption 2.3. There exist positive $\Lambda_{\max}$ and Δ_∞ such that

$$\|\boldsymbol{I}_d - \widehat{\boldsymbol{\Theta}}\widetilde{\boldsymbol{\Sigma}}\|_{\max} = O_{\mathbb{P}}(\Lambda_{\max}) \text{ and } \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_\infty = O_{\mathbb{P}}(\Delta_\infty).$$

Without loss of generality, here we assume that $\Delta_\infty \leqslant \|\boldsymbol{\Theta}\|_\infty$.

Remark 7. To give a concrete example, under the mild conditions therein, Assumption 3.1 is satisfied with

$$\Lambda_{\max} = O\left(\sqrt{\frac{\log d}{n} + \frac{1}{d}}\right) \text{ and } \Delta_\infty = O\left(\max_{j\in[d]} |\mathcal{S}_j| \sqrt{\frac{\log d}{n} + \frac{1}{d}}\right),$$

by using node-wise regression [Zhang and Zhang, 2014, van de Geer et al., 2014], where $|\mathcal{S}_j| = \sum_{k=1}^d \mathbb{I}\{\Theta_{jk} \neq 0\}$ quantifies the sparsity of j-th column of the precision matrix $\boldsymbol{\Theta}$ for each $1 \leqslant j \leqslant d$. In Appendix **??**, we will provide a detailed analysis on estimating $\widetilde{\boldsymbol{\Sigma}}^{-1}$ via node-wise regression and establish precise theoretical upper bounds for the statistical rates given in Assumption 3.1. □

3.2 Gaussian Approximation

The goal of this section is to derive the asymptotic distribution of $\|\widetilde{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star\|_\infty$ in the high dimensional setting. To this end, we apply the Gaussian approximation result given in Chernozhukov et al. [2013, 2017, 2020] for high dimensional random vectors. More specifically, we let $\boldsymbol{Z} = (Z_1, \ldots, Z_d)^\top \in \mathbb{R}^d$ be a zero-mean Gaussian random vector with the same covariance matrix as that of $n^{-1/2}\boldsymbol{\Theta}\boldsymbol{U}^\top\mathcal{E}$, that is,

$$\mathrm{Cov}(\boldsymbol{Z}) = \mathrm{Cov}\left(\frac{1}{\sqrt{n}}\boldsymbol{\Theta}\boldsymbol{U}^\top\mathcal{E}\right) = \sigma^2\boldsymbol{\Theta}. \tag{3.4}$$

We next present the theoretical results on Gaussian approximation of our test statistics under some mild conditions.

Theorem 3.1. *Recall $\boldsymbol{\varphi}^\star = \boldsymbol{\gamma}^\star - \boldsymbol{B}^\top\boldsymbol{\beta}^\star \in \mathbb{R}^K$. We assume that $(\log d)^5/n \to 0$,*

$$(\Lambda_{\max}|\mathcal{S}_\star| + \Delta_\infty)\log d \to 0 \ \text{ and } \ \left(\mathcal{V}_{n,d}\|\boldsymbol{\varphi}^\star\|_2 + \sqrt{\frac{n}{d}} + \sqrt{\log d}\right)\|\boldsymbol{\Theta}\|_\infty\sqrt{\frac{\log d}{n}} \to 0, \tag{3.5}$$

with $\mathcal{V}_{n,d}$. Then under Assumption 3.1, we have

$$\sup_{x>0}\left|\mathbb{P}\left(\sqrt{n}\|\widetilde{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star\|_\infty \leqslant x\right) - \mathbb{P}\left(\|\boldsymbol{Z}\|_\infty \leqslant x\right)\right| \to 0.$$

For any $\alpha \in (0,1)$, let $c_{1-\alpha}$ denote the $(1-\alpha)$-th quantile of the distribution of $\|\boldsymbol{Z}\|_\infty$. Theorem 3.1 leads to an approximately level α test for (3.1) as follows:

$$\psi_{\infty,\alpha} = \mathbb{I}\left\{\sqrt{n}\|\widetilde{\boldsymbol{\beta}}_\lambda\|_\infty > c_{1-\alpha}\right\}. \tag{3.6}$$

3.3 Gaussian multiplier bootstrap

The critical value $c_{1-\alpha}$ depends on the unknown σ^2 and $\boldsymbol{\Theta}$, which can be estimated by the following Gaussian multiplier bootstrap.

1. Generate i.i.d. random variables $\xi_1, \ldots, \xi_n \sim N(0,1)$ and compute

$$\widehat{L} = \frac{1}{\sqrt{n}}\|\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{U}}^\top\boldsymbol{\xi}\|_\infty, \ \text{ where } \ \boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_n)^\top.$$

2. Repeat the first step independently for B times and obtain $\widehat{L}_1, \ldots, \widehat{L}_B$. Estimate the critical value $c_{1-\alpha}$ via $1-\alpha$ quantile of the empirical distribution of the bootstrap statistics:

$$\widehat{c}_{1-\alpha} = \inf\{t \geqslant 0 : H_B(t) \geqslant 1-\alpha\}, \text{ where } H_B(t) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left\{\widehat{L}_b \leqslant t\right\}.$$

Reject the null hypothesis H_0 when $\sqrt{n}\|\widetilde{\boldsymbol{\beta}}_\lambda\|_\infty/\widehat{\sigma} > \widehat{c}_{1-\alpha}$, for a given consistent estimator $\widehat{\sigma}$ of σ. To validate the procedure, we need some additional conditions on $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\sigma}$.

Assumption 3.2. There exists a $\Delta_{\max} > 0$ such that $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{\max} = O_{\mathbb{P}}(\Delta_{\max})$.

Assumption 3.3. There exists a $0 < \Delta_\sigma \leqslant 1$ such that $|\widehat{\sigma}/\sigma - 1| = O_{\mathbb{P}}(\Delta_\sigma)$.

Remark 8. The estimation of σ^2 for high dimensional linear regression has been studied in the literature. For example, Fan et al. [2012] proposed refitted cross-validation to construct a consistent estimator with clearly quantified uncertainty of $\widehat{\sigma}$ in ultra-high dimension. In addition, Sun and Zhang [2012] and Yu and Bien [2019] derived scaled-Lasso and organic Lasso respectively for estimating σ. Like our case of estimating $\boldsymbol{\Theta}$, we also do not restrict estimating σ by any fixed method mentioned above, our theory works as long as the general condition of Assumption 3.3 holds. □

Let $\mathbb{P}^\star(\cdot) = \mathbb{P}(\cdot|\boldsymbol{X}, \boldsymbol{Y})$ denote the conditional probability. In the following theorem, we establish the validity of the proposed bootstrap procedure.

Theorem 3.2. *Let Assumptions 3.1–3.3 hold. Assume that*

$$\Lambda_{\max}\|\boldsymbol{\Theta}\|_\infty + \Delta_{\max} + \Delta_\sigma = o\left(\frac{1}{\log d}\right). \tag{3.7}$$

Then, under conditions of Theorem 3.1, we have

$$\sup_{x>0}\left|\mathbb{P}\left(\sqrt{n}\|\widetilde{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star\|_\infty \leqslant x\right) - \mathbb{P}^\star\left(\widehat{L} \leqslant x\right)\right| \xrightarrow{\mathbb{P}} 0.$$

Remark 9. Following the same de-biasing procedure as given in (3.2), we are also able to construct entrywise [Javanmard and Montanari, 2014] and groupwise [Zhang and Cheng, 2017, Dezeure et al.,

2017] simultaneous confidence intervals for $\boldsymbol{\beta}^\star$. For each $1 \leqslant j \leqslant d$, a $(1-\alpha)$-confidence interval for $\beta_j^\star$ is given by

$$\mathcal{CI}_\alpha(\beta_j^\star) = \left\{ \widetilde{\beta}_{j,\lambda} - \widehat{\sigma} z_{1-\alpha/2} \sqrt{\frac{\widehat{\Theta}_{jj}}{n}},\ \widetilde{\beta}_{j,\lambda} - \widehat{\sigma} z_{1-\alpha/2} \sqrt{\frac{\widehat{\Theta}_{jj}}{n}} \right\},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-th quantile of standard normal distribution. By choosing this cut-off value, we obtain a tighter confidence interval comparing with using $c_{1-\alpha}$. For simultaneous groupwise inference of $\boldsymbol{\beta}^\star$, let G be a subset of $\{1, \ldots, d\}$ of interest and consider testing the hypotheses

$$H_{0,G} : \beta_j^\star = \beta_j^\circ \ \text{ for all } \ j \in G \ \text{ versus } \ H_{1,G} : \beta_j^\star \neq \beta_j^\circ \ \text{ for some } \ j \in G.$$

In particular, when $\beta_j^\circ = 0$ for all $j \in G$, this reduces to testing the significance of a group of parameters. We obtain that the asymptotic distribution of $\max_{j \in G} \sqrt{n} |\widetilde{\beta}_{j,\lambda} - \beta_j^\star|$ converges to the distribution of $\max_{j \in G} |Z_j|$ by leveraging the Gaussian approximation. The remaining steps follow directly by conducting the Gaussian multiplier bootstrap. □

4 Is Sparse Linear Model Adequate?

Sparse linear regression, which serves as the backbone of high dimensional statistics, has been widely applied in many areas of science, engineering, and social sciences. However, its adequacy has never been validated. This section focuses on testing the adequacy of the sparse linear model.

4.1 Main Results

As mentioned in introduction, the proposed model (1.5) contains the sparse linear regression model as a special case. Thus, we consider testing the hypotheses

$$H_0 : Y = \boldsymbol{x}^\top \boldsymbol{\beta}^\star + \varepsilon \ \text{ versus } \ H_1 : Y = \boldsymbol{f}^\top \boldsymbol{\varphi}^\star + \boldsymbol{x}^\top \boldsymbol{\beta}^\star + \varepsilon, \tag{4.1}$$

which is equivalent to test whether $\boldsymbol{\varphi}^\star = \boldsymbol{\gamma}^\star - \boldsymbol{B}^\top \boldsymbol{\beta}^\star = 0$. Since $\boldsymbol{B}$ is an unknown dense matrix, simultaneously testing this linear equation will suffer from the curse of dimensionality.

On the other hand, for any set $\mathcal{S} \subset [d]$ with $\mathcal{S}_\star \subset \mathcal{S}$, we have $\boldsymbol{B}_{\mathcal{S}}^\top \boldsymbol{\beta}_{\mathcal{S}}^\star = \boldsymbol{B}^\top \boldsymbol{\beta}^\star$. Hence, it suffices to compare the following two linear models in reduced dimension:

$$H_0 : Y = \boldsymbol{x}_{\mathcal{S}}^\top \boldsymbol{\beta}_{\mathcal{S}}^\star + \varepsilon \ \text{ versus } \ H_1 : Y = \boldsymbol{f}^\top \boldsymbol{\varphi}^\star + \boldsymbol{x}_{\mathcal{S}}^\top \boldsymbol{\beta}_{\mathcal{S}}^\star + \varepsilon. \tag{4.2}$$

This hinges applying a sure screening method to reduce the dimensionality. There exist several methods which lead to the sure screening property. Among those, the commonly used one is the marginal screening method [Fan and Lv, 2008, Fan and Song, 2010, Zhu et al., 2011, Li et al., 2012, Liu et al., 2014, Barut et al., 2016, Chu et al., 2016, Wang and Leng, 2016].

We propose an ANOVA-type test for (4.1) with two stages. In the first stage, the data set is split into two data sets $(\boldsymbol{Y}^{(1)}, \boldsymbol{X}^{(1)})$ and $(\boldsymbol{Y}^{(2)}, \boldsymbol{X}^{(2)})$, with sample sizes m and $n-m$, respectively. We use $(\boldsymbol{Y}^{(1)}, \boldsymbol{X}^{(1)})$ to screen variables. Let $\widehat{\mathcal{S}}_1$ denote the set of variables selected. In the second stage, we leverage the selected $\widehat{\mathcal{S}}_1$ and remaining data $(\boldsymbol{Y}^{(2)}, \boldsymbol{X}^{(2)})$ to perform hypothesis testing based on the ANOVA-type test statistic for low-dimensional model (4.2) with $\mathcal{S}$ replaced by $\widehat{\mathcal{S}}_1$. As the first step is based on marginal screening and is relatively crude, the sample size m is relatively small in comparing with the second step. We impose a general assumption on the set $\widehat{S}_1$.

Assumption 4.1 (Sure screening property)**.** There exists an $s_n > 0$ such that

$$\mathbb{P}\left(|\widehat{\mathcal{S}}_1| \leqslant s_n \ \text{ and } \ \mathcal{S}_\star \subset \widehat{\mathcal{S}}_1\right) \to 1, \ \text{ as } \ n \to \infty.$$

A simple procedure that satisfies the above assumption is the follow factor-adjusted marginal screening based on the data $(\boldsymbol{Y}^{(1)}, \boldsymbol{X}^{(1)})$.

1. Estimation. Compute the latent factor estimator $\widehat{\boldsymbol{F}}^{(1)}$, idiosyncratic component $\widehat{\boldsymbol{U}}^{(1)}$ based on $\boldsymbol{X}^{(1)}$, and $\widetilde{\boldsymbol{Y}}^{(1)} = (\boldsymbol{I}_m - \widehat{\boldsymbol{F}}^{(1)}(\widehat{\boldsymbol{F}}^{(1)\top}\widehat{\boldsymbol{F}}^{(1)})^{-1}\widehat{\boldsymbol{F}}^{(1)\top})\boldsymbol{Y}^{(1)}$, the residual after factor regression.

2. Marginal regression. Compute the least square estimate $\widehat{\beta}_{\ell,M} = \widehat{\boldsymbol{U}}_\ell^{(1)\top}\widetilde{\boldsymbol{Y}}^{(1)}/(\widehat{\boldsymbol{U}}_\ell^{(1)\top}\widehat{\boldsymbol{U}}_\ell^{(1)})$ for each $1 \leqslant \ell \leqslant d$.

3. Screening. Let $\widehat{S}_1 := \widehat{\mathcal{S}}_\phi = \{\ell \in [d] : |\widehat{\beta}_{\ell,M}| > \phi\}$ for some prescribed $\phi > 0$.

Here $\widehat{\boldsymbol{U}}_\ell^{(1)} \in \mathbb{R}^d$ stands for the ℓ-th column of the matrix $\widehat{\boldsymbol{U}}^{(1)}$. We next provide a sufficient condition for the Assumption 4.1 to hold.

Proposition 4.1. *Assume that* $m = o(d \log d)$ *and*

$$\mathcal{O}\left(\frac{\|\psi^*\|_2}{d} + \|\beta^*\|_2\sqrt{\frac{\log d}{m}}\right) \leqslant \phi \leqslant \mathcal{O}\left(\min_{\ell\in[d]} \beta^\star_{\ell,M}\right), \tag{4.3}$$

where $\beta^\star_{\ell,M} := \boldsymbol{\Sigma}_\ell^\top \boldsymbol{\beta}^\star / \Sigma_{\ell\ell}$. *Here* $\boldsymbol{\Sigma}_\ell$ *denotes the* ℓ*-th column of* $\boldsymbol{\Sigma}$. *Then, under the Assumptions 2.1–2.5, we have*

$$\mathbb{P}\left(\mathcal{S}_\star \subset \widehat{\mathcal{S}}_\phi\right) \to 1, \ \text{ as } \ m \to \infty.$$

Furthermore, we assume that $\min_{\ell\in\mathcal{S}_\star} |\beta^\star_{\ell,M}| \geqslant c_\star m^{-\kappa}$ *for some positive constant* $\kappa < 1/2$. *Then for any* $\phi = c_\diamond m^{-\kappa}$ *with* $c_\diamond \leqslant c_\star/(1+\bar{c})$, *we have*

$$\mathbb{P}\left\{|\widehat{\mathcal{S}}_\phi| \leqslant \frac{c_\diamond^2 m^{2\kappa}\|\boldsymbol{\Sigma}\boldsymbol{\beta}^\star\|_2^2}{\lambda^2_{\min}(\boldsymbol{\Sigma})(1-\bar{c})^2}\right\} \to 1 \ \text{ as } \ m \to \infty.$$

Remark 10. From the conclusion of Proposition 4.1, we obtain sure screening property by using our first data set with sample size $m = n^\alpha$ for some $\alpha < 1$ as long as the signal satisfies $\min_{\ell\in\mathcal{S}_\star} |\beta^\star_{\ell,M}| \geqslant c_\star m^{-\kappa}$. Thus, the size of the remaining data set for constructing the test statistic in our second step is $n - n^\alpha \approx n$. It is worth to note that this does not lose any efficiency in terms of the asymptotic power in our hypothesis test when n goes to infinity. □

Remark 11. Fan et al. [2020a] proposed a similar sure screening estimator which is a special case of our Proposition 4.1 with $\boldsymbol{\varphi}^\star = \boldsymbol{\gamma}^\star - \boldsymbol{B}^\top\boldsymbol{\beta}^\star = 0$. Moreover, we also provide an upper bound for the number of selected variables whereas Fan et al. [2020a] only provided a sufficient condition for the sure screening property. □

Next, we proceed to the second stage of our hypothesis testing. In this step, we construct an ANOVA test statistic for (4.2) with $\mathcal{S}$ replaced by $\widehat{\mathcal{S}}_1$, which is given by

$$Q_n^{(2)} = \left\|\left(\boldsymbol{I}_{n-m} - \boldsymbol{P}_{\boldsymbol{X}^{(2)}_{\hat{\mathcal{S}}_1}}\right)\boldsymbol{Y}^{(2)}\right\|_2^2 - \left\|\left(\boldsymbol{I}_{n-m} - \boldsymbol{P}_{\hat{\boldsymbol{F}}^{(2)}} - \boldsymbol{P}_{\hat{\boldsymbol{U}}^{(2)}_{\hat{\mathcal{S}}_1}}\right)\boldsymbol{Y}^{(2)}\right\|_2^2. \tag{4.4}$$

We then summarize our results on the asymptotic behaviors of $Q_n^{(2)}$ in the following Theorem 4.2.

Theorem 4.2. *Let Assumptions 2.1–2.5 and Assumption 4.1 hold with*

$$s_n\left(\frac{\log d}{n}+\frac{1}{d}\right)\to 0 \ \text{ and } \ \Delta_\sigma\to 0.$$

We obtain

$$\sup_{x>0}\left|\mathbb{P}\left(Q_n^{(2)}\leqslant x\widehat{\sigma}^2|H_0\right)-\mathbb{P}(\chi_K^2\leqslant x)\right|\to 0, \ \text{ as } \ n\to\infty.$$

Theorem 4.2 yields a level α test for (4.1) , with critical region $\left\{Q_n^{(2)}>\widehat{\sigma}^2\chi^2_{K,1-\alpha}\right\}$, where $\chi^2_{K,1-\alpha}$ is the $(1-\alpha)$-th quantile of χ_K^2-distribution.

Remark 12. Under stronger conditions such as irrepresentable condition [Zhao and Yu, 2006] or RIP condition [Candes and Tao, 2007], the $\widehat{\mathcal{S}}$ achieved by certain explicit regularization [Zhao and Yu, 2006, Fan and Lv, 2011, Shi et al., 2019, Fan et al., 2020a] or implicit regularization accompanied with early stopping and signal truncation [Zhao et al., 2019, Fan et al., 2021b] enjoys variable selection consistency $\mathbb{P}(\widehat{\mathcal{S}}=\mathcal{S}_\star)\to 1$. In this scenario, we take the test statistic as

$$Q_n=\left\|\left(\boldsymbol{P}_{\widehat{\boldsymbol{F}}}+\boldsymbol{P}_{\widehat{\boldsymbol{U}}_{\hat{\mathcal{S}}}}-\boldsymbol{P}_{\boldsymbol{X}_{\hat{\mathcal{S}}}}\right)\boldsymbol{Y}\right\|_2^2$$

without using sample splitting. Under Assumptions 2.1–2.5, we obtain

$$\sup_{x>0}\left|\mathbb{P}\left(Q_n\leqslant x\widehat{\sigma}^2|H_0\right)-\mathbb{P}(\chi_K^2\leqslant x)\right|\to 0, \tag{4.5}$$

by following similar proof idea with Theorem 4.2. □

We now present the power of the test statistic (4.4).

Theorem 4.3. *Define*

$$\mathcal{D}(\alpha,\theta)=\left\{\boldsymbol{\varphi}\in\mathbb{R}^K:\frac{n\|\boldsymbol{\varphi}\|^2}{1+Ks_n\|\boldsymbol{B}\|^2_{\max}/\lambda_{\min}(\boldsymbol{\Sigma})}\geqslant\sigma^2(2+\delta)(\chi^2_{K,1-\alpha}+\chi^2_{K,1-\theta})\right\},$$

where $\delta>0$ is some constant, s_n is the size of selected set from the first stage and K is the number of factors. In addition, parameter θ is a threshold such that for any $\boldsymbol{\varphi}^\star\in\mathcal{D}(\alpha,\theta)$, the power of the test is larger than $1-\theta$. To be more specific, we assume that

$$\|\boldsymbol{\varphi}^\star\|_2\left(\sqrt{n/d}+1/\sqrt{n}\right)\to 0. \tag{4.6}$$

Then, under the conditions of Theorem 4.2, we have

$$\inf_{\boldsymbol{\varphi}^\star \in \mathcal{D}(\alpha,\theta)} \mathbb{P}(\psi_\alpha = 1 | H_1) \geqslant 1 - \theta,$$

where $\psi_\alpha = \mathbb{I}_{\{Q_n^{(2)} > \hat{\sigma}^2 \chi^2_{K,1-\alpha}\}}$.

Remark 13. Dataset with multiple types are now frequently collected for a common set of experimental subjects. This new data structure is also called multimodal data. It is worth to mention, the above hypothesis test can be further extended to test the adequacy of multi-modal sparse linear regression model [Li and Li, 2021]. Interested readers are referred to Appendix F.4 for more details. □

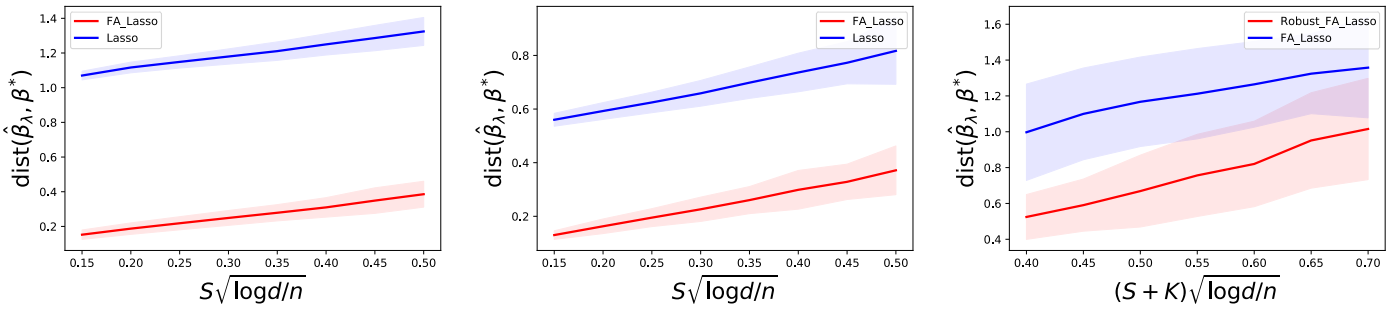
5 Numerical Studies

5.1 Accuracy of Estimation

For data generation, we let number of factors $K = 2$, dimension of covariate $d = 1000$, $\boldsymbol{\gamma}^\star = (0.5, 0.5)$, the first $s = 3$ entries of $\boldsymbol{\beta}^\star$ be 0.5 and remaining $d - s$ entries be 0. Throughout this subsection, we generate every entry of $\boldsymbol{F}, \boldsymbol{U}$ from the standard Gaussian distribution and let every entry of $\boldsymbol{B}$ be generated from the uniform distribution Unif $(-1, 1)$. We choose the noise distribution of ε given in model (2.1) from (i) standard Gaussian, (ii) uniform, and (iii) t_3 distribution respectively.

Distributions (i) and (ii) have sub-Gaussian tails. For these two cases, we select sample size n so that $s\sqrt{\log d/n}$ takes uniform grids in $[0.15, 0.5]$. Then we generate n response variables from model (2.1) and estimate our parameters via (2.4). The results are shown as the red lines in Figure 1. They lend further support to our theoretical findings given in section 2 as the statistical rates there are upper bounded by $O(s\sqrt{\log d/n})$. Moreover, we also show the estimation results by using Lasso directly on measurements $(\boldsymbol{X}, \boldsymbol{Y})$. Results are shown as the blue lines given in the first two figures in Figure 1. Using Lasso directly on $(\boldsymbol{X}, \boldsymbol{Y})$ leads to much worse results due in part to the inadequacy of the model. In addition, as shown in Fan et al. [2020a], even when the sparse regression model is correct, we still have better estimation accuracy using factor adjusted regression.

Distribution (iii) has only the bounded second moment, but no third moment. Likewise, we select corresponding number of observations n so that $(s+K)\sqrt{\log d/n}$ takes uniform grids in $[0.4, 0.7]$. The



(a) (b) (c)

Figure 1: Accuracy for $\widehat{\beta}_\lambda$ with $\text{dist}(\widehat{\beta}_\lambda, \beta^\star) := \|\widehat{\beta}_\lambda - \beta^\star\|_1$ based on 500 replications. The light color regions indicate the standard errors across the simulation. Figure (a) and (b) depict the estimation results of model (1.4) with noise ϵ following the standard Gaussian and uniform distributions respectively. In (a) and (b), the red lines denote the estimation results using the method (2.4) (labeled as FA_Lasso in the figure) and the blue lines represent the results using Lasso with data $(\boldsymbol{X}, \boldsymbol{Y})$. In Figure (c), the noise ϵ follows t_3-distribution. The red line in (c) represents the result of robust factor adjusted regression (Robust_FA_Lasso) via adaptive Huber estimation together with ℓ_1-penalty given in (2.7) and the blue line represents the result achieved by using FA_Lasso.

reduced sample sizes help reduce the computation cost on the regularized adaptive Huber estimation using cross-validation to choose the parameter ω. We compare the results for the robust estimator (2.7) with that of the factor adjusted regression (2.4). The results are shown as the red and blue lines in part (c) of Figure 1 respectively. They provide stark evidence that it is necessary to conduct the robust version of factor adjusted regression (2.7) when noises have heavy tails.

5.2 Adequacy of Factor Regression

Data Generation Processes. We choose $n = 200$, $K = 2$ and d either 200 or 500 and the matrix $\boldsymbol{X} = \boldsymbol{F}\boldsymbol{B}^\top + \boldsymbol{U}$ using the following two models with entries of $\boldsymbol{B}$ generated from $\text{Unif}(-1, 1)$.

1. We generate every row of $\boldsymbol{F} \in \mathbb{R}^{n\times K}, \boldsymbol{U} \in \mathbb{R}^{n\times d}$ from $\text{N}(\boldsymbol{0}, \boldsymbol{I}_K)$ and $\text{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ respectively.

2. We let the t-th row $\boldsymbol{f}_t \in \mathbb{R}^K$ of $\boldsymbol{F} \in \mathbb{R}^{n\times K}$ follow $\boldsymbol{f}_t = \boldsymbol{\Phi}\boldsymbol{f}_{t-1} + \boldsymbol{\xi}_t$ where $\boldsymbol{\Phi} \in \mathbb{R}^{K\times K}$ with

$\boldsymbol{\Phi}_{i,j} = 0.5^{|i-j|+1}, i, j \in [K]$. In addition, $\{\boldsymbol{\xi}_t\}_{t\geqslant 1}$ are drawn independently from $\mathrm{N}(\mathbf{0}, \boldsymbol{I}_K)$. We generate every row of $\boldsymbol{U}$ from $N(\mathbf{0}, \boldsymbol{\Sigma})$ where $\Sigma_{i,j} = 0.6^{|i-j|}, i, j \in [d]$.

The response vector follows $\boldsymbol{Y} = \boldsymbol{F}\boldsymbol{\gamma}^\star + \boldsymbol{U}\boldsymbol{\beta}^\star + \mathcal{E}$ in (4.1) with every entry of $\mathcal{E} \in \mathbb{R}^n$ being generated independently from either from $N(0, 0.5^2)$ or uniform distribution $\mathrm{Unif}\,(-\sqrt{3}/2, \sqrt{3}/2)$. We set $\boldsymbol{\gamma}^\star = (0.5, 0.5)$ and $\boldsymbol{\beta}^\star = (w, w, w, 0, \cdots, 0)$, where $w \geqslant 0$. When $w = 0$, the null hypothesis $\boldsymbol{Y} = \boldsymbol{F}\boldsymbol{\gamma}^\star + \mathcal{E}$ holds and the simulation results correpond to the size of the test. Otherwise, they correspond to the power of the test.

For $n = 200, K = 2$, $d \in \{200, 500\}$ and all $w \in \{0, 0.05, 0.10, 0.15, 0.20\}$, we generate the data from each model and compute the testing results based on procedures in §3 and 2000 simulatons with $\alpha = 0.05$. For every replication, we conduct bootstrap 2000 times to compute the critical value $\widehat{c}_{1-\alpha}$ given in §3. The results are depicted in the Table 1. The column named Gaussian(i), $i \in \{1, 2\}$ represents the simulation results under model i with Gaussian noise. Similar labels applied to the uniform noise distribution.

		Gaussian (1)	Gaussian (2)	Uniform (1)	Uniform (2)
$p = 200$	$w = 0$	0.044	0.047	0.046	0.048
	$w = 0.05$	0.067	0.119	0.065	0.108
	$w = 0.10$	0.326	0.714	0.311	0.653
	$w = 0.15$	0.859	0.989	0.854	0.984
	$w = 0.20$	0.998	1.000	0.996	1.000
$p = 500$	$w = 0$	0.043	0.040	0.048	0.436
	$w = 0.05$	0.067	0.080	0.059	0.071
	$w = 0.10$	0.253	0.632	0.237	0.563
	$w = 0.15$	0.787	0.974	0.780	0.962
	$w = 0.20$	0.993	1.000	0.987	1.000

Table 1: Simulation results of section 3 under different regimes.

Table 1 reveals that our test gives approximately the right size (subject to simulation error; see the rows with $w = 0$). This is consistent with our theoretical findings given in section 3. In addition, when $0 < w < 0.2$, the power of our test increases rapidly to 1 which reveals the efficiency of our test

statistic.

5.3 Adequacy of Sparse Regression

This subsection provides finite-sample validations for the results in section 4. We take the number of data used for screening $m = \lceil n^{0.8} \rceil$, use Iterative Sure Independence Screening method [Fan and Lv, 2008, Saldana and Feng, 2018, Zhang et al., 2019] to select $\widehat{\mathcal{S}}_1$ and apply the refitted cross-validation [Fan et al., 2012] to estimate σ^2. The size and the power of the test are computed based on 2000 simulations.

Data Generation Processes. We let $n = 250$, $K = 3$ and d be either 250 or 600. The noises ε are i.i.d from $N(0, 0.5^2)$ or Unif $(-\sqrt{3}/2, \sqrt{3}/2)$. The covariate $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ follows the factor model $\boldsymbol{X} = \boldsymbol{F}\boldsymbol{B}^\top + \boldsymbol{U}$. We generate $\boldsymbol{F}$, $\boldsymbol{U}$ and $\boldsymbol{B}$ in the same way as those in section 5.2. In addition, the response variable follows $\boldsymbol{Y} = \boldsymbol{F}\boldsymbol{\varphi}^\star + \boldsymbol{X}\boldsymbol{\beta}^\star + \mathcal{E}$ in (4.1) with $\boldsymbol{\beta}^\star = (0.8, 0.8, 0.8, 0.8, 0, \cdots, 0)$ and $\boldsymbol{\varphi}^\star = v \cdot \mathbf{1}_{K \times 1}$ for several different values of $v \geqslant 0$. The case $v = 0$ corresponds to the null hypothesis and it is designed to test the validity of the size.

Results. For $n = 250$, $K = 3$, $d \in \{250, 600\}$ and $v \in \{0, 0.04, 0.08, 0.12, 0.16\}$, we implement the proposed method for every model in section 5.2. The simulation results are depicted in Table 2. The column named Gaussian (or uniform) (i), $i \in \{1, 2\}$ represents the results under model i with Gaussian (or uniform) noise mentioned in section 5.2. When $v = 0$, the null hypothesis holds, our Type-I error is approximately 0.05 which matches with the theoretical value. In addition, when we increase the size of v from $v = 0.04$ to $v = 0.16$, the power of our test statistic increases sharply to 1, which reveals its efficiency.

We next discuss the necessity of using sample splitting. Suppose we do not split samples and use the whole dataset to do sure screening and construct the test statistic. This will result in the high correlation between the selected set $\widehat{\mathcal{S}}$ and covariates when $\widehat{\mathcal{S}}$ is not a consistent estimator of $\mathcal{S}_\star$. In this case, the asymptotic behavior of our test statistic is hard to capture. To demonstrate this point, we simulate the null distribution of the test statistic constructed without using sample splitting and compare it with the asymptotic distribution (χ^2_K) via the quantile-quantile plot in Figure 7 in appendix

		Gaussian (1)	Gaussian (2)	Uniform (1)	Uniform (2)
$p = 250$	$v = 0$	0.051	0.054	0.056	0.053
	$v = 0.04$	0.215	0.278	0.233	0.286
	$v = 0.08$	0.659	0.740	0.655	0.750
	$v = 0.12$	0.965	0.993	0.965	0.996
	$v = 0.16$	1.000	1.000	1.000	1.000
$p = 600$	$v = 0$	0.051	0.052	0.050	0.052
	$v = 0.04$	0.208	0.362	0.197	0.353
	$v = 0.08$	0.624	0.802	0.604	0.785
	$v = 0.12$	0.941	0.994	0.934	0.999
	$v = 0.16$	1.000	1.000	0.999	1.000

Table 2: Simulation results of section 4 under different regimes.

§B.2. Figure 7 reveals that the test statistic constructed without using sample splitting has heavier right tail than that of the χ^2_K distribution. The sizes of the test are much larger than the results in Table 2 when $v = 0$.

We summarize the numerical results as follows. In terms of statistical estimation, our estimated parameters of FARM behave much better than those estimated parameters via sparse linear regression model due to mis-specification. As for prediction, we also conduct additional simulations on comparing FARM with the latent factor regression and sparse linear regression model. Interested readers are referred to §B.1 for more details. For high-dimensional inference, as illustrated in §5.2 and §5.3, when the null hypothesises hold, the size of the test is well-controlled. On the other hand, when the null hypothesises do not hold, the power of our test statistics grow rapidly to 1 even for weak signals.

5.4 Empirical Applications

In this section, we use a macroeconomic dataset named FRED-MD [McCracken and Ng, 2016] to illustrate the performance of our factor augmented regression model (FARM) and investigate whether the latent factor regression model and sparse linear model are adequate.

There are 134 monthly U.S. macroeconomic variables in this dataset. As they measure certain

aspects of economic health, these variables are driven by latent factors and hence correlated. They can be well explained by a few principal components. In our study, we pick out two variables named 'HOUSTNE' and 'GS5' as our responses respectively and let the remaining variables be the covariates. Here 'HOUSTNE' represents the housing starts in the northeast region. Studying the number of housing starts helps one to understand the residents' life condition and economic environment. 'GS5' is correlated with many important variables such as interests rates, inflation and economic growth. It is an important indicator on the financial condition and economics environment of a country.

There exist significant structural breaks for many variables around the year of financial crisis in 2008 which makes our data non-stationary even after performing the suggested transformations. Thus, we analyze the dataset in two separate time periods independently. Specifically, we study the monthly data collected from February 1992 to October 2007 and from August 2010 to February 2020 respectively after examing the missingness and stationarity of the data.

We next compare the performance of our proposed FARM against several benchmarks presented in a few related references which study the same or similar datasets. In specific, we compare the forcasting results of FARM with Lasso (sparse linear regression), PCR (latent factor regression), Ridge (Ridge regression), El-Net (Elastic Net) used in Coulombe et al. [2021b], Smeekes and Wijler [2018], Hall et al. [2018], RF (Random Forest) used in Goulet Coulombe [2020], Coulombe et al. [2021a], Bianchi et al. [2021] and FarmSelect (Factor adjusted Lasso) used in Fan et al. [2020a]. For every given time period and model, we perform the prediction by using the moving window approach with window size 90 months. Indexing the panel data from 1 for each of the two time periods, for all $t > 90$, we use the 90 previous measurements $\{(\boldsymbol{x}_{t-90}, Y_{t-90}), \cdots, (\boldsymbol{x}_{t-1}, Y_{t-1})\}$ to train a model (FARM, sparse linear regression model, latent factor regression model, ridge regression, elastic net, factor adjusted Lasso, random forest), and output a prediction $\widehat{Y}_t$ as well as the in-sample average $\bar{Y}_t := \frac{1}{90}\sum_{i=t-90}^{t-1} Y_i$ (the detailed implementations for different methods are presented in the appendix §B.3). We measure the prediction accuracy by using out-of-sample R^2:

$$R^2 = 1 - \frac{\sum_{t=91}^{T}(Y_t - \widehat{Y}_t)^2}{\sum_{t=91}^{T}(Y_t - \bar{Y}_t)^2},$$

where T denotes the number of total data points in a given time period. Table 3 presents the out-

of-sample R^2 obtained by the aforementioned several models in the two time periods for predicting 'HOUSTNE' and 'GS5'.

From the results, we observe that FARM outperforms all other benchmarks. In specific, in comparison with PCR, our performance is better. This is due to the possibility that the latent factor regression did not adequately explain the data. Additionally, applying traditional penalized regression methods like Lasso or Elastic Net (El-Net) directly will result in an erroneous estimator $\widehat{\beta}$ and worse prediction outcomes when the covariates have a factor structure (highly-correlated). Fan et al. [2020a] propose a factor-adjusted lasso estimator (FarmSelect) to mitigate the impact of latent factors, but they still assume the sparse regression as a sufficient method. From this point of view, their model could be misspecified and performs worse than ours. In terms of the Ridge regression, it assumes the underlying signal β^* is dense instead of sparse. From the outcomes, we conclude that the dense model may not explain this dataset. Finally, our FARM also outperforms the random forest, a well-used model for making predictions via machine learning.

Time period	Data	FARM	Lasso	PCR	Ridge	El-Net	RF	FarmSelect
02.1992-10.2007	HOUSTNE	0.769	0.684	0.372	0.221	0.699	0.497	0.741
	GS5	0.720	0.702	0.056	0.249	0.699	0.557	0.709
08.2010-02.2020	HOUSTNE	0.743	0.374	0.079	0.125	0.348	0.421	0.569
	GS5	0.681	0.650	0.032	0.342	0.653	0.557	0.626

Table 3: Out-of-sample R^2 for predicting 'HOUSTNE' and 'GS5' data using different models in different time periods. In this table, the values in the FARM column denote the prediction results through the factor-augmented linear regression model. In addition, we also compare the prediction results with several benchmarks, Lasso (sparse linear regression), PCR (latent factor regression), Ridge (Ridge regression), El-Net (Elastic Net), RF (Random Forest), and FarmSelect (Factor adjusted Lasso).

We next conduct the hypothesis testing on the adequacy of latent factor regression and sparse linear regression respectively by using FARM as the alternative model. As computing the bootstrap estimate of the null distribution is expensive for testing the adequacy of the factor model, we only conduct the hypothesis testing using the data in the entire two subperiods: 02.1992-10.2007 and 08.2010-02.2020.

The P-values for the tests are given in Table 4. Taking the significant level 0.05, the hypothesis testing results indicate that in most of the cases, the latent factor regression and sparse linear regression, are not sufficient to explain the dataset. These results match well with our prediction results.

Time period	Data	LA_factor	SP_Linear
02.1992-10.2007	HOUSTNE	$< 10^{-3}$	$< 10^{-3}$
	GS5	$1.5 \cdot 10^{-3}$	$4.73 \cdot 10^{-3}$
08.2010-02.2020	HOUSTNE	$< 10^{-3}$	$1.64 \cdot 10^{-1}$
	GS5	$1.98 \cdot 10^{-1}$	$2.94 \cdot 10^{-2}$

Table 4: p-values for testing the adequacy of the latent factor regression and sparse linear regression models to explain 'HOUSTNE' and 'GS5' data in two different time periods. The LA_Factor and SP_Linear have the same meaning as those in Table 3.

6 Conclusion and Discussion

In this paper, we propose a model named Factor Augmented (sparse linear) Regression Model (FARM), which contains the latent factor regression and the sparse linear regression as our special cases. The model expands the space spanned by covariates into useful principal component directions and hence use additional information beyond the linear space spanned by the predictors. We provide theoretical guarantees for our model estimation under the existence of light-tailed and heavy-tailed noises respectively. In addition, we leverage the FARM model as the alternative one to test the adequacy of the latent factor regression model and sparse regression model. We believe that the study is among the first of this kind in high-dimensional inference. The practical performance of our model estimation and our constructed test statistics are proven by extensive simulation studies including both synthetic data and real data. Moreover, it is worth to mention that our model and methodology can be extended to more general supervised learning problems such as nonparametric regression, quantile regression, regression and classification trees, support vector machines, among others where the factor augmentation idea is always useful.

Next, we provide some discussion with several related works. First, we make comparison with Luciani [2014]. It is worth noting although our work shares similar idea with Luciani [2014] in terms of incorporating factors into sparse regression, our framework is more general, systematic and contains theirs as special case. Moreover, our intuition is different. To be more specific, they provide a conceptual idea without specifying any statistical model and do not provide any theoretical guarantees. In contrast, we provide a thorough study of the factor augmented sparse regression model (FARM), from the perspective of (robust) estimation to uncertainty quantification with well established methodology and theoretical results. We further utilize our model to test the goodness-of-fit of two important models, namely, factor regression and sparse regression. More importantly, the starting points of deriving our FARM is different from that in Luciani [2014]. Specifically, our model is intuitive from the inadequacy of factor regression and sparse regression in many situations, whereas they mainly focus on using past factors and idiosyncratic components to forecast time series data.

Next, we discussion our connection with the sparse and dense model proposed by Giannone et al. [2021]. They consider the model

$$Y = \boldsymbol{x}_t^\top \boldsymbol{\beta}^* + \boldsymbol{z}_t^\top \boldsymbol{\varphi}^\star + \epsilon,$$

where $\boldsymbol{z}_t$ is a low-dimensional vectorwhose regression parameter $\boldsymbol{\varphi}^\star$ is considered to be dense and $\boldsymbol{x}_t$ is a high-dimensional covariate whose regression coefficient $\boldsymbol{\beta}^\star$ is considered to be sparse. It is worth noting that their model acts as a special case to our FARM when we assume the factor is observable with $\boldsymbol{z}_t = \boldsymbol{f}_t$ and $\boldsymbol{x}_t$ possess factor structure. Moreover, they aim at identifying explainable regression variables and degree of sparseness using tools from Bayesian statistics and empirical illustration, whereas we provide consistency estimation results and also conduct hypothesis testing on β^* elementwisely or groupwisely via both theories and numerical studies. Last but not least, it is mentioned in their paper, when they analyze the macroeconomic data, although their posterior sparse level is low, the heat map shows high uncertainty on whether certain predictors should be included in the model due to high colinearity of predictors. In such a scenario, conducting regression directly using their method will fail to recover the true support of the covariates [Zhao and Yu, 2006] and thus results in unstable estimation results. To remedy this issue, one needs to decompose covariates into the factor and idiosyncratic component and run factor adjusted regression via FARM. This showcases the neces-

sity of estimating this sparse and dense model using FARM instead of ordinary linear regression under strongly correlated covariate.

It is worth mentioning that, we also discuss the connections with several other related literature, such as Lin and Michailidis [2020], Kneip and Sarda [2011], discuss the model selection consistency, and test the contribution of a particular $\boldsymbol{x}_i$ to Y. Due to the space limit, we put them in the appendix §A.

References

S. C. Ahn and A. R. Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227, 2013.

M. Avella-Medina, H. S. Battey, J. Fan, and Q. Li. Robust estimation of high-dimensional covariance and precision matrices. *Biometrika*, 105(2):271–284, 2018.

J. Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171, 2003.

J. Bai and K. Li. Statistical analysis of factor models of high dimension. *Ann. Statist.*, 40(1):436–465, 2012.

J. Bai and S. Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1): 191–221, 2002.

J. Bai and S. Ng. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150, 2006.

J. Bai and S. Ng. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2): 304–317, 2008. ISSN 0304-4076.

E. Bair, T. Hastie, D. Paul, and R. Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006.

E. Barut, J. Fan, and A. Verhasselt. Conditional sure independence screening. *J. Amer. Statist. Assoc.*, 111(515): 1266–1277, 2016.

A. Belloni and V. Chernozhukov. ℓ_1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.*, 39(1):82–130, 2011.

D. Bianchi, M. Büchner, and A. Tamoni. Bond risk premiums with machine learning. *The Review of Financial Studies*, 34(2):1046–1089, 2021.

X. Bing, F. Bunea, and M. Wegkamp. Inference in latent factor regression with clusterable features. *arXiv:1905.12696*, 2019.

X. Bing, F. Bunea, S. Strimas-Mackey, and M. Wegkamp. Prediction under latent factor regression: Adaptive pcr, interpolating predictors and beyond. *Journal of Machine Learning Research*, 22(177):1–50, 2021.

F. Bunea, S. Strimas-Mackey, and M. Wegkamp. Interpolating predictors in high-dimensional factor regression. *arXiv:2002.02525*, 2020.

T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.*, 106(494):594–607, 2011.

E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, 41(6):2786–2819, 2013.

V. Chernozhukov, D. Chetverikov, and K. Kato. Central limit theorems and bootstrap in high dimensions. *Ann. Probab.*, 45(4):2309–2352, 2017.

V. Chernozhukov, D. Chetverikov, and Y. Koike. Nearly optimal central limit theorem and bootstrap approximations in high dimensions. *arXiv preprint arXiv:2012.09513*, 2020.

W. Chu, R. Li, and M. Reimherr. Feature screening for time-varying coefficient models with ultrahigh dimensional longitudinal data. *Ann. Appl. Stat.*, 10(2):596, 2016.

P. G. Coulombe, M. Leroux, D. Stevanovic, and S. Surprenant. Macroeconomic data transformations matter. *International Journal of Forecasting*, 37(4):1338–1354, 2021a.

P. G. Coulombe, M. Marcellino, and D. Stevanović. Can machine learning catch the covid-19 recession? *National Institute Economic Review*, 256:71–109, 2021b.

R. Dezeure, P. Bühlmann, and C.-H. Zhang. High-dimensional simultaneous inference with the bootstrap. *Test*, 26(4):685–719, 2017.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

J. Fan and Y. Liao. Learning latent factors from diversified projections and its applications to over-estimated and weak factors. *Journal of the American Statistical Association*, 0(0):1–16, 2020.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(5):849–911, 2008.

J. Fan and J. Lv. Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory*, 57(8): 5467–5484, 2011.

J. Fan and R. Song. Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.*, 38(6):3567–3604, 2010.

J. Fan, S. Guo, and N. Hao. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 74(1):37–65, 2012.

J. Fan, Y. Liao, and M. Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 75(4):603–680, 2013. With 33 discussions by 57 authors and a reply by Fan, Liao and Mincheva.

J. Fan, Y. Fan, and E. Barut. Adaptive robust variable selection. *Ann. Statist.*, 42(1):324, 2014.

J. Fan, Q. Li, and Y. Wang. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 79(1):247, 2017a.

J. Fan, L. Xue, and j. Yao. Sufficient forecasting using factor models. *Journal of Econometrics*, 201(2):292–306, 2017b. ISSN 0304-4076.

J. Fan, Y. Ke, and K. Wang. Factor-adjusted regularized model selection. *J. Econometrics*, 216(1):71–85, 2020a.

J. Fan, R. Li, C.-H. Zhang, and H. Zou. Statistical foundations of data science. 2020b.

J. Fan, R. Masini, and M. C. Medeiros. Bridging factor and sparse models. *arXiv:2102.11341*, 2021a.

J. Fan, Z. Yang, and M. Yu. Understanding implicit regularization in over-parameterized single index model. *arXiv:2007.08322v3*, 2021b.

J. Fan, J. Guo, and S. Zheng. Estimating number of factors by adjusted eigenvalues thresholding. *Journal of the American Statistical Association*, 117:852–861, 2022.

D. Giannone, M. Lenza, and G. E. Primiceri. Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89(5):2409–2437, 2021.

P. Goulet Coulombe. The macroeconomy as a random forest. *Available at SSRN 3633110*, 2020.

A. S. Hall et al. Machine learning approaches to macroeconomic forecasting. *The Federal Reserve Bank of Kansas City Economic Review*, 103(63):2, 2018.

M. Hernan and J. Robins. *Causal Inference*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. CRC Press LLC, 2019. ISBN 9781420076165.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

G. Imbens and D. Rubin. Causal inference for statistics, social, and biomedical sciences: An introduction. 2015.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15:2869–2909, 2014.

I. T. Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(3):300–303, 1982.

A. Kneip and P. Sarda. Factor models and variable selection in high-dimensional regression analysis. *The Annals of Statistics*, 39(5):2410–2447, 2011.

C. Lam and Q. Yao. Factor modeling for high-dimensional time series: Inference for the number of factors. *The Annals of Statistics*, 40(2):694 – 726, 2012.

G. Li, H. Peng, J. Zhang, and L. Zhu. Robust rank correlation based screening. *Ann. Statist.*, 40(3):1846–1877, 2012.

Q. Li and L. Li. Integrative factor regression and its inference for multimodal data analysis. *J. Amer. Statist. Assoc.*, 113(521):1–15, 2021.

Q. Li, G. Cheng, J. Fan, and Y. Wang. Embracing the blessing of dimensionality in factor models. *J. Amer. Statist. Assoc.*, 113(521):380–389, 2018.

J. Lin and G. Michailidis. System identification of high-dimensional linear dynamical systems with serially correlated output noise components. *IEEE Transactions on Signal Processing*, 68:5573–5587, 2020.

J. Liu, R. Li, and R. Wu. Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *J. Amer. Statist. Assoc.*, 109(505):266–274, 2014.

P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.*, 40(3):1637–1664, 2012.

M. Luciani. Forecasting with approximate dynamic factor models: the role of non-pervasive shocks. *International Journal of Forecasting*, 30(1):20–29, 2014.

M. W. McCracken and S. Ng. Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589, 2016.

R. Nickl and S. Van De Geer. Confidence sets in sparse regression. *Ann. Statist.*, 41(6):2852–2876, 2013.

B. Peng, L. Wang, and Y. Wu. An error bound for L_1-norm support vector machine coefficients in ultra-high dimension. *J. Mach. Learn. Res.*, 17(1):8279–8304, 2016.

D. F. Saldana and Y. Feng. Sis: An r package for sure independence screening in ultrahigh-dimensional statistical models. *Journal of Statistical Software*, 83(2):1–25, 2018.

C. Shi, R. Song, Z. Chen, and R. Li. Linear hypothesis testing for high dimensional generalized linear models. *Ann. Statist.*, 47(5):2671–2703, 2019.

S. Smeekes and E. Wijler. Macroeconomic forecasting using penalized regression methods. *International journal of forecasting*, 34(3):408–430, 2018.

J. H. Stock and M. W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002.

Q. Sun, W.-X. Zhou, and J. Fan. Adaptive Huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

S. Van de Geer. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2):614–645, 2008.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 2014.

W. Wang and J. Fan. Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Ann. Statist.*, 45(3):1342–1374, 2017.

X. Wang and C. Leng. High dimensional ordinary least squares projection for screening variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, pages 589–611, 2016.

G. Yu and J. Bien. Estimating the error variance in a high-dimensional linear model. *Biometrika*, 106(3): 533–546, 2019.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942, 2010.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(1):217–242, 2014.

N. Zhang, W. Jiang, and Y. Lan. On the sure screening properties of iteratively sure independence screening algorithms. *arXiv:1812.01367*, 2019.

X. Zhang and G. Cheng. Simultaneous inference for high-dimensional linear models. *J. Amer. Statist. Assoc.*, 112(518):757–768, 2017.

X. Zhang, Y. Wu, L. Wang, and R. Li. Variable selection for support vector machines in moderately high dimensions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 78(1):53, 2016.

P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.

P. Zhao, Y. Yang, and Q.-C. He. Implicit regularization via hadamard product over-parametrization in high-dimensional linear regression. *arXiv:1903.09367*, 2019.

L.-P. Zhu, L. Li, R. Li, and L.-X. Zhu. Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.*, 106(496):1464–1475, 2011.

H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006.

Supplement material for "Are Latent Factor Regression and Sparse Regression Adequate?"

A More Discussion

1. Comparison with related work

- Comparison with Kneip and Sarda [2011]. Kneip and Sarda [2011] study a similar problem by incorporating factors into sparse regression under stronger assumptions. Their model requires the idiosyncratic component $\boldsymbol{u}_t$ to be uncorrelated, which is quite a strong assumption in the high dimension, whereas we only require the conditional number of the covariance of $\boldsymbol{u}_t$ to be bounded. In addition, they require the noise distribution to be Gaussian, whereas our framework not only allow for general sub-Gaussian noises but heavy-tailed noises as well. Most importantly, in addition to estimating the parameters of interest, we also focus on high-dimensional inference: testing the goodness of fit of two important models, factor regression and sparse regression which are not considered in Kneip and Sarda [2011].

- Connection with Lin and Michailidis [2020]. Lin and Michailidis [2020] study factor model $\boldsymbol{x}_t = \boldsymbol{B}\boldsymbol{f}_t + \boldsymbol{u}_t$ where factors, idiosyncratic component following certain vector autoregression (VAR) model. In this case, they predict the current covariate $\boldsymbol{x}_t$ by augmenting factors to past covariate $\boldsymbol{x}_{t-1}$. Compared with them, we study different objectives in a sense that we focus on high-dimensional inference for regression problems with general responses. It is worth to note that when our $Y_t = x_{t,i}, i \in [d]$, their model becomes our special case.

2. [Model Selection Consistency] Model selection consistency (recovering the true support of β^*) plays an essential role in model prediction, especially if the dataset is in high dimension but the underlying model is sparse. However, when the features are strongly co-linear, the irrepresentable condition fails, and we can not achieve model selection consistency [Zhao and Yu, 2006]. Just as the case mentioned in Giannone et al. [2021], one may delete the actual important variables but involve useless ones while making predictions. However, it is worth noting that even when sparse regression is adequate, if one decomposes $\boldsymbol{X}$ into $(\boldsymbol{F}, \boldsymbol{U})$, which are two less correlated ones, one will eliminate the high-colinearity across the features and thus obtain model selection consistency [Fan et al., 2020]. Thus, even if the sparse model is adequate, FARM still has many merits in the sense that we can achieve model selection consistency while preserving mean square errors.

3. Contribution of a particular $\boldsymbol{x}_i$ to Y. There are two ways for $\boldsymbol{x}_i$ to influence the response Y.
The first is through idiosyncratic component $\boldsymbol{u}_i$ and the other is indirect effect through $\boldsymbol{f}_t$. To test the first effect, we leverage the debiasing idea given in section 3 to test whether $\beta_i^* = 0$.
For the indirect influence of $\boldsymbol{x}_i$ on Y through $\boldsymbol{f}_t$, this procedure is more involved. This is equivalent to testing whether the i-th row of the $\boldsymbol{B}$ is zero or not. Recall that in our model formulation, we have the eigenspace with respect to spiked eigenvalues of $\mathrm{Cov}(\boldsymbol{x})$ is spanned by columns of $\boldsymbol{B}$. Thus, testing whether the i-th row of $\boldsymbol{B}$ is zero is equivalent to testing whether the i-th row of spiked

eigenvectors $\boldsymbol{V}_{i,\cdot}$ of $\mathrm{Cov}(\boldsymbol{x})$ is zero. To accommodate this issue, we leverage the recent development on doing statistical inference on eigenspace [Yan et al., 2021]. They derive asymptotic distribution for $\widehat{\boldsymbol{V}}_{i,\cdot}\boldsymbol{R} - \boldsymbol{V}_{i,\cdot} \to N(0, \Sigma_{V,i})$. Here $\widehat{\boldsymbol{V}}_{i,\cdot} \in \mathbb{R}^K$ is the i-th row of the estimated spiked eigenvectors of $\boldsymbol{X}$, $\boldsymbol{R} \in \mathbb{R}^{K\times K}$ is a rotation matrix and $\Sigma_{V,i}$ is a covariance matrix that can be estimated via data. Under the null hypothesis, we have $\boldsymbol{V}_{i,\cdot} = 0$. In this scenario, we have $\widehat{\boldsymbol{V}}_{i,\cdot}\boldsymbol{R}\boldsymbol{R}^\top\widehat{\boldsymbol{V}}_{i,\cdot}^\top \to \|\boldsymbol{Z}\|_2^2$ with $\boldsymbol{Z} \sim N(0, \Sigma_{U,i})$. Thus, the hypothesis testing can be done by comparing $\widehat{\boldsymbol{V}}_{i,\cdot}\widehat{\boldsymbol{V}}_{i,\cdot}^\top$ with $1-\alpha$ quantile of the distribution of $\|\boldsymbol{Z}\|_2^2$, which can be estimated via bootstrap.

B Additional Numberical Results

B.1 Model Prediction

In this subsection, we provide a detailed simulation study on the prediction performance of latent factor regression and sparse regression with respective to FARM when our FARM is the true underlying model. First, we demonstrate how latent factor regression and sparse regression behave when the they have different levels of misspecification.

As for data generation, we let $n = 200, K = 5$, dimention of covariate $d = 1000$, $\boldsymbol{\varphi}^\star = 0.8 \cdot \mathbf{1}_K$, $\boldsymbol{\beta}^\star = (v \cdot \mathbf{1}_{20}^\top, \mathbf{0}_{p-20}^\top)^\top$ with the magnitude of v varies uniformly from 0 to 1.20. Throughout this subsection, we generate every entry of $\boldsymbol{F}$ and $\boldsymbol{U}$ from the standard Gaussian distribution and let every entry of $\boldsymbol{B}$ be generated from uniform distribution with $\mathrm{Unif}(-2, 2)$. The noise distribution ϵ is given by standard Gaussian distribution. The experiment is repeated 500 times and we record the out-of-sample MSE. We then illustrate our comparion with factor regression in Table 1. As for our comparison with sparse

$\|\boldsymbol{\beta}^\star\|_\infty$	0.00	0.20	0.40	0.60	0.80	1.00	1.20
FA Reg	0.25	1.02	3.69	8.33	13.42	18.74	28.14
FARM	0.27	0.65	0.69	0.67	0.65	0.74	0.96

Table 1: Out-of-sample MSE of our model (FARM) with factor regression (FA Reg).

regresssion, we adopt the aforementioned setting, except that we fix $\boldsymbol{\beta}^\star = (0.8 \cdot \mathbf{1}_{20}^\top, \mathbf{0}_{p-20}^\top)^\top$ and let $\boldsymbol{\varphi}^\star = v \cdot \mathbf{1}_K$, with v ranges uniformly from 0 to 1.20. We summarize the comparison results in Table 2.

$\|\boldsymbol{\varphi}^\star\|_\infty$	0.00	0.20	0.40	0.60	0.80	1.00	1.20
SP Reg	1.94	2.14	2.15	2.37	2.24	2.40	2.54
FARM	0.55	0.59	0.61	0.62	0.59	0.60	0.61

Table 2: Out-of-sample MSE of our model (FARM) with sparse regression (SP Reg).

Finally, we illustrate the model prediction performances of factor regression and sparse regression

with respective to our FARM when the same size n increases. All settings are the same with those mentioned above, except that we fix $\boldsymbol{\beta}^\star = (0.8 \cdot \mathbf{1}_{20}^\top, \mathbf{0}_{p-20}^\top)^\top$ and let $\boldsymbol{\varphi}^\star = 0.8 \cdot \mathbf{1}_K$. The out-of-sample MSE are recorded below in Table 3. In addition, we also compare FARM with the other two models (Sparse regression and Latent factor regression) using huber loss when there exist heavy-tailed noise. We keep all the aforementioned settings except for changing the noise distribution from standard Gaussian distribution to t_3 distribution. The simulation results are summarized in the following Table 4.

n	200	229	257	286	314	343	371	400
FA Reg	13.59	12.03	14.37	11.81	12.27	13.20	11.79	12.46
SP Reg	2.25	1.86	1.87	1.63	1.59	1.49	1.36	1.38
FARM	0.58	0.48	0.45	0.39	0.35	0.39	0.34	0.31

Table 3: Out-of-sample MSE of our model (FARM) with sparse regression (SP Reg) and factor regression (FA Reg) when n varies.

n	200	229	257	286	314	343	371	400
FA Reg	16.39	18.00	16.60	14.37	15.12	15.97	18.19	14.85
SP Reg	13.73	15.06	14.07	12.63	13.22	13.90	16.67	13.14
FARM	6.67	5.99	4.71	4.04	4.02	3.53	4.77	2.93

Table 4: Out-of-sample MSE of our model (FARM) with sparse regression (SP Reg) and factor regression (FA Reg) when n varies when heavy-tailed error exists.

B.2 Additional Plots

B.3 Implementation Details of Empirical Applications

In this section, we describe the implementation details of different methods used in Empirical Applications. For the implementation of Lasso [Tibshirani, 1996], Ridge, Elastic Net [Zou and Hastie, 2005] we use *glmnet* package in R directly to conduct the numerical studies. The tuning parameters λ and α in the models are chosen via leave-one-out cross-validation (by definition, the parameter α only needs to be tuned in Elastic Net). For the stableness of the algorithms, for FARM, PCR, FarmSelect [Fan et al., 2020], throughout this empirical application, we use PCA to estimate the factors, as suggested in §2, and let the number of factors be the maximum number of (2.3) and 2. For the implementation of FARM, FarmSelect, their first steps of estimating $\widehat{\beta}$ are the same. In specific, we decompose the covariate $\mathbf{X}$ into factor and idiosyncratic component $[\widehat{\boldsymbol{F}}, \widehat{\boldsymbol{U}}]$ and then use Lasso to estimate the loadings $\widehat{\beta}$ as we described in §2. The prediction procedures are different, namely, we use covariate $\boldsymbol{x}_{\text{new}}$ and $[\widehat{\boldsymbol{f}}_{\text{new}}, \widehat{\boldsymbol{u}}_{\text{new}}]$ to make prediction, for FarmSelect and FARM, respectively. In terms of using Random Forest, we use the package *randomForest* in R to implement the model estimation and prediction

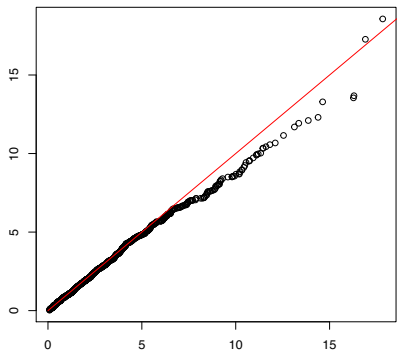


Figure 7: Quantiles of the χ^2_K distribution against those of the test statistic without sample splitting. The x-axis represents the quantiles of the test statistic whereas y-axis is the quantiles of χ^2_K distribution.

directly. For more details of our implementation and reproduction, we put the relevant codes in github path given in the ACC form.

C Technical Lemmas

Recall that $\widetilde{\boldsymbol{\Sigma}} = n^{-1}\widehat{\boldsymbol{U}}^\top\widehat{\boldsymbol{U}}$ and $\mathcal{C}(\mathcal{S}, 3) = \{\boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v}_{\mathcal{S}}\|_1 \leqslant 3\|\boldsymbol{v}_{\mathcal{S}^c}\|_1\}$ for any subset $\mathcal{S} \subset [d]$. First, we introduce the following Lemma C.1. In this Lemma, when the RSC condition is satisfied, we are able to achieve some certain ℓ_2-statistical rates for $\widehat{\boldsymbol{\beta}}_\lambda$ and $\widehat{\boldsymbol{U}}\widehat{\boldsymbol{\beta}}_\lambda$.

Lemma C.1. *Assume that* $\lambda \geqslant (2/n)\|\widehat{\boldsymbol{U}}^\top(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{U}}\boldsymbol{\beta}^\star)\|_\infty$ *and for some positive constant* $\kappa(\mathcal{S}_\star, 3)$,

$$\min_{0\neq\boldsymbol{v}\in\mathcal{C}(\mathcal{S}_\star,3)} \frac{\boldsymbol{v}^\top\widetilde{\boldsymbol{\Sigma}}\boldsymbol{v}}{\|\boldsymbol{v}\|_2^2} \geqslant \kappa(\mathcal{S}_\star, 3).$$

Then we have $\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star \in \mathcal{C}(\mathcal{S}_\star, 3)$,

$$\|\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star\|_2 \leqslant \frac{3\lambda\sqrt{|\mathcal{S}_\star|}}{\kappa(\mathcal{S}_\star, 3)} \text{ and } \|\widehat{\boldsymbol{U}}(\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star)\|_2^2 \leqslant \frac{9n\lambda^2|\mathcal{S}_\star|}{\kappa(\mathcal{S}_\star, 3)}.$$

Proof of Lemma C.1. Write $\boldsymbol{\delta} = \widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star$. Following the proof of Theorem 7.2 in Bickel et al. [2009], we obtain $\boldsymbol{\delta} \in \mathcal{C}(\mathcal{S}_\star, 3)$ and

$$\kappa(\mathcal{S}_\star, 3)\|\boldsymbol{\delta}\|_2^2 \leqslant \boldsymbol{\delta}^\top\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\delta} = \frac{1}{n}\|\widehat{\boldsymbol{U}}\boldsymbol{\delta}\|_2^2 \leqslant 3\lambda\|\boldsymbol{\delta}_{\mathcal{S}_\star}\|_1 \leqslant 3\lambda\sqrt{|\mathcal{S}_\star|}\|\boldsymbol{\delta}\|_2.$$

Then we have

$$\|\boldsymbol{\delta}\|_2 \leqslant \frac{3\lambda\sqrt{|\mathcal{S}_\star|}}{\kappa(\mathcal{S}_\star, 3)} \text{ and } \|\widehat{\boldsymbol{U}}\boldsymbol{\delta}\|_2^2 \leqslant 3n\lambda\sqrt{|\mathcal{S}_\star|}\|\boldsymbol{\delta}\|_2 \leqslant \frac{9n\lambda^2|\mathcal{S}_\star|}{\kappa(\mathcal{S}_\star, 3)}.$$

□

In the next Lemma C.2, we prove that under quite mild conditions, the RSC condition given in Lemma C.1 holds with high probability.

Lemma C.2. *Under Assumptions 2.1–2.4, for any set $\mathcal{S} \subset \{1, \ldots, d\}$ with*

$$|\mathcal{S}|\left(\frac{\log d}{n} + \frac{1}{d}\right) \to 0, \tag{C.1}$$

there exists a constant $\kappa(\mathcal{S}, 3) > 0$ such that

$$\mathbb{P}\left(\min_{0 \neq \boldsymbol{v} \in \mathcal{C}(\mathcal{S}, 3)} \frac{\boldsymbol{v}^\top \widetilde{\boldsymbol{\Sigma}} \boldsymbol{v}}{\|\boldsymbol{v}\|_2^2} \geqslant \kappa(\mathcal{S}, 3)\right) \to 1, \text{ as } n \to \infty.$$

Proof of Lemma C.2. For any vector $\boldsymbol{v} \in \mathbb{R}^d$ such that $\|\boldsymbol{v}_{\mathcal{S}^c}\|_1 \leqslant 3\|\boldsymbol{v}_{\mathcal{S}}\|_1$, we have

$$\|\boldsymbol{v}\|_1^2 \leqslant 16\|\boldsymbol{v}_{\mathcal{S}}\|_1^2 \leqslant 16|\mathcal{S}|\|\boldsymbol{v}_{\mathcal{S}}\|_2^2 \leqslant 16|\mathcal{S}|\|\boldsymbol{v}\|_2^2,$$

which implies that

$$\frac{\boldsymbol{v}^\top \widetilde{\boldsymbol{\Sigma}} \boldsymbol{v}}{\|\boldsymbol{v}\|_2^2} \geqslant \frac{\boldsymbol{v}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{v}}{\|\boldsymbol{v}\|_2^2} - \frac{\|\widetilde{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}\|_{\max}\|\boldsymbol{v}\|_1^2}{\|\boldsymbol{v}\|_2^2} \geqslant \frac{\boldsymbol{v}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{v}}{\|\boldsymbol{v}\|_2^2} - 16|\mathcal{S}|\|\widetilde{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}\|_{\max}, \tag{C.2}$$

where $\widehat{\boldsymbol{\Sigma}} = n^{-1}\boldsymbol{U}^\top\boldsymbol{U}$. Combining this with Lemma C.6 and (C.1), we obtain

$$\mathbb{P}(\mathcal{E}_1) := \mathbb{P}\left(16|\mathcal{S}|\|\widetilde{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}\|_{\max} \leqslant \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{8}\right) \to 1, \text{ as } n \to \infty.$$

For $\phi > 0$, define the sparse set $\mathbb{K}(\phi) = \{\boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v}\|_0 \leqslant \phi, \|\boldsymbol{v}\|_2 \leqslant 1\}$. Taking $\phi_\diamond = 128|\mathcal{S}|$. It follows from (C.1) that $\phi_\diamond = o(n/\log d)$. Then, by Lemma 15 in Loh and Wainwright [2012] and Assumption 2.1, it follows that

$$\mathbb{P}(\mathcal{E}_2) := \mathbb{P}\left(\sup_{\boldsymbol{v} \in \mathbb{K}(2\phi_\diamond)} |\boldsymbol{v}^\top(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{v}| \leqslant \frac{\lambda_{\max}(\boldsymbol{\Sigma})}{54}\right) \to 1, \text{ as } n \to \infty.$$

Under the event $\mathcal{E}_2$, by Lemma 13 in Loh and Wainwright [2012],

$$\boldsymbol{v}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{v} \geqslant \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{2}\|\boldsymbol{v}\|_2^2 - \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{\phi_\diamond}\|\boldsymbol{v}\|_1^2 \geqslant \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{2}\|\boldsymbol{v}\|_2^2 - \frac{16|\mathcal{S}|\lambda_{\min}(\boldsymbol{\Sigma})}{\phi_\diamond}\|\boldsymbol{v}\|_2^2.$$

Combining this with (C.2), under $\mathcal{E}_1 \cap \mathcal{E}_2$, we obtain

$$\frac{\boldsymbol{v}^\top \widetilde{\boldsymbol{\Sigma}} \boldsymbol{v}}{\|\boldsymbol{v}\|_2^2} \geqslant \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{2} - 16|\mathcal{S}|\|\widetilde{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}\|_{\max} - \frac{16|\mathcal{S}|\lambda_{\min}(\boldsymbol{\Sigma})}{\phi_\diamond} \geqslant \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{4}.$$

□

Lemma C.3. *Under Assumptions 2.1–2.4, for any vector $\boldsymbol{\phi} \in \mathbb{R}^K$ with $\|\boldsymbol{\phi}\|_2 = 1$, we have*

$$\|\widehat{\boldsymbol{U}}^\top \boldsymbol{F}\boldsymbol{\phi}\|_\infty = O_{\mathbb{P}}(\mathcal{V}_{n,d}). \tag{C.3}$$

Proof of Lemma C.3. Define $\mathcal{E}_\lambda = \{\lambda_{\min}(\boldsymbol{H}^\top \boldsymbol{H}) \geqslant 1/2\}$. It follows from Lemma 2.1 that $\mathbb{P}(\mathcal{E}_\lambda) \to 1$. Since $\widehat{\boldsymbol{F}}^\top \widehat{\boldsymbol{U}} = \boldsymbol{O}$, under $\mathcal{E}_\lambda$, we have

$$\begin{aligned}\|\widehat{\boldsymbol{U}}^\top \boldsymbol{F}\boldsymbol{\phi}\|_\infty = \|\widehat{\boldsymbol{U}}^\top (\boldsymbol{F} - \widehat{\boldsymbol{F}}\boldsymbol{H}^{-\top})\boldsymbol{\phi}\|_\infty &\leqslant \|(\widehat{\boldsymbol{U}} - \boldsymbol{U})^\top (\boldsymbol{F}\boldsymbol{H}^\top - \widehat{\boldsymbol{F}})\boldsymbol{H}^{-\top}\boldsymbol{\phi}\|_\infty \\ &\quad + \|\boldsymbol{U}^\top (\boldsymbol{F}\boldsymbol{H}^\top - \widehat{\boldsymbol{F}})\boldsymbol{H}^{-\top}\boldsymbol{\phi}\|_\infty =: \Delta^\diamond + \Delta^\circ.\end{aligned}$$

We first bound $\Delta^\diamond$. By Lemma 2.1 and the Cauchy-Schwarz inequality, it follows that

$$\begin{aligned}\Delta^\diamond &\leqslant \left(\max_{j\in[d]} \sum_{t=1}^n |\widehat{u}_{tj} - u_{tj}|^2\right)^{1/2} \|\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H}^\top\|_{\mathbb{F}} \|\boldsymbol{H}^{-\top}\boldsymbol{\phi}\|_2 \\ &= O_{\mathbb{P}}\left\{\sqrt{\left(\log d + \frac{n}{d}\right)\left(\frac{n}{d} + \frac{1}{n}\right)}\right\} = O_{\mathbb{P}}(\mathcal{V}_{n,d}).\end{aligned}$$

We now bound Δ°. Recall that $\boldsymbol{H} = n^{-1}\boldsymbol{V}^{-1}\widehat{\boldsymbol{F}}^\top \boldsymbol{F}\boldsymbol{B}^\top \boldsymbol{B}$ and $\widehat{\boldsymbol{F}}\boldsymbol{V} = n^{-1}\boldsymbol{X}\boldsymbol{X}^\top \widehat{\boldsymbol{F}}$. Hence

$$\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H}^\top = \frac{1}{n}\boldsymbol{F}\boldsymbol{B}^\top \boldsymbol{U}^\top \widehat{\boldsymbol{F}}\boldsymbol{V}^{-1} + \frac{1}{n}\boldsymbol{U}\boldsymbol{B}\boldsymbol{F}^\top \widehat{\boldsymbol{F}}\boldsymbol{V}^{-1} + \frac{1}{n}\boldsymbol{U}\boldsymbol{U}^\top \widehat{\boldsymbol{F}}\boldsymbol{V}^{-1}.$$

By the triangle inequality, it follows that

$$\begin{aligned}\Delta^\circ \leqslant \frac{1}{n}\Big(&\|\boldsymbol{U}^\top \boldsymbol{F}\boldsymbol{B}^\top \boldsymbol{U}^\top \widehat{\boldsymbol{F}}\boldsymbol{V}^{-1}\boldsymbol{H}^{-\top}\boldsymbol{\phi}\|_\infty + \|\boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{B}\boldsymbol{F}^\top \widehat{\boldsymbol{F}}\boldsymbol{V}^{-1}\boldsymbol{H}^{-\top}\boldsymbol{\phi}\|_\infty \\ &+ \|\boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{U}^\top \widehat{\boldsymbol{F}}\boldsymbol{V}^{-1}\boldsymbol{H}^{-\top}\boldsymbol{\phi}\|_\infty\Big) =: \Delta_1^\circ + \Delta_2^\circ + \Delta_3^\circ.\end{aligned}$$

By Assumptions 2.1 and 2.3, we have $\mathbb{E}\|\boldsymbol{B}^\top \boldsymbol{U}^\top\|_{\mathbb{F}}^2 = n\mathrm{tr}(\boldsymbol{B}^\top \boldsymbol{\Sigma}\boldsymbol{B})$ and

$$\mathbb{E}\|\boldsymbol{B}^\top \boldsymbol{U}^\top \boldsymbol{F}\|_{\mathbb{F}}^2 = n\mathbb{E}\|\boldsymbol{B}^\top \boldsymbol{u}\|_2^2 \|\boldsymbol{f}\|_2^2 \leqslant n\sqrt{\mathbb{E}\|\boldsymbol{B}^\top \boldsymbol{u}\|_2^4}\sqrt{\mathbb{E}\|\boldsymbol{f}\|_2^4} = O(nd).$$

Consequently, by Lemma 2.1, it follows that

$$\|\boldsymbol{B}^\top \boldsymbol{U}^\top \widehat{\boldsymbol{F}}\|_{\mathbb{F}} \leqslant \|\boldsymbol{B}^\top \boldsymbol{U}^\top (\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H}^\top)\|_{\mathbb{F}} + \|\boldsymbol{B}^\top \boldsymbol{U}^\top \boldsymbol{F}\boldsymbol{H}^\top\|_{\mathbb{F}}$$

$$\leqslant \|\boldsymbol{B}^\top \boldsymbol{U}^\top\|_{\mathbb{F}} \|\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H}^\top\|_{\mathbb{F}} + \|\boldsymbol{H}\|_2 \|\boldsymbol{B}^\top \boldsymbol{U}^\top \boldsymbol{F}\|_{\mathbb{F}}$$
$$= O_{\mathbb{P}}\left\{\sqrt{nd\left(\frac{n}{d} + \frac{1}{n}\right)} + \sqrt{nd}\right\} = O_{\mathbb{P}}\left(n + \sqrt{nd}\right).$$

Combining this with $\max_{j\in[d]} \|\boldsymbol{e}_j^\top \boldsymbol{U}^\top \boldsymbol{F}\|_2^2 = \max_{j\in[d]} \|\sum_{t=1}^n u_{tj} \boldsymbol{f}_t\|_2^2 = O_{\mathbb{P}}(n \log d)$, we obtain

$$\begin{aligned}
\Delta_1^\circ &= \max_{j\in[d]} \frac{1}{n} |\boldsymbol{e}_j^\top \boldsymbol{U}^\top \boldsymbol{F}\boldsymbol{B}^\top \boldsymbol{U}^\top \widehat{\boldsymbol{F}} \boldsymbol{V}^{-1} \boldsymbol{H}^{-\top} \boldsymbol{\phi}| \\
&\leqslant \max_{j\in[d]} \frac{1}{n} \|\boldsymbol{e}_j^\top \boldsymbol{U}^\top \boldsymbol{F}\|_2 \|\boldsymbol{B}^\top \boldsymbol{U}^\top \widehat{\boldsymbol{F}}\|_{\mathbb{F}} \|\boldsymbol{V}^{-1}\|_2 \|\boldsymbol{H}^{-\top}\|_2 \\
&= O_{\mathbb{P}}\left\{\sqrt{(\log d)/d} + \sqrt{n \log d}/d\right\}.
\end{aligned}$$

Write $\boldsymbol{B} = (\widetilde{\boldsymbol{b}}_1, \ldots, \widetilde{\boldsymbol{b}}_K) \in \mathbb{R}^{d\times K}$. By Assumptions 2.3 and 2.4,

$$\max_{k\in[K]} \max_{j\in[d]} \mathbb{E}\left(u_j \boldsymbol{u}^\top \widetilde{\boldsymbol{b}}_k\right) = \max_{k\in[K]} \max_{j\in[d]} \boldsymbol{\Sigma}_j^\top \widetilde{\boldsymbol{b}}_k \leqslant \max_{k\in[K]} \max_{j\in[d]} \|\boldsymbol{\Sigma}_j\|_1 \|\widetilde{\boldsymbol{b}}_k\|_\infty \leqslant \frac{\Upsilon}{\kappa},$$

where $\boldsymbol{\Sigma}_j \in \mathbb{R}^d$ denotes the j-th column of the covariance matrix $\boldsymbol{\Sigma}$. Moreover, by Assumption 2.1, for each $k \in [K]$, $\{u_{tj} \boldsymbol{u}_t^\top \widetilde{\boldsymbol{b}}_k\}_{t=1}^n$ is a sequence of i.i.d. sub-exponential random variables with

$$\max_{k\in[K]} \max_{j\in[d]} \|u_{tj} \boldsymbol{u}_t^\top \widetilde{\boldsymbol{b}}_k\|_{\psi_1} \leqslant \max_{j\in[d]} \|u_{tj}\|_{\psi_2} \max_{k\in[K]} \|\boldsymbol{u}_t^\top \widetilde{\boldsymbol{b}}_k\|_{\psi_2} \leqslant c_0^2 \max_{k\in[K]} \|\widetilde{\boldsymbol{b}}_k\|_2 \leqslant c_0^2 \Upsilon \sqrt{d}.$$

Hence, it follows that

$$\max_{j\in[d]} \left\|\sum_{t=1}^n u_{tj} \boldsymbol{u}_t^\top \boldsymbol{B}\right\|_2 \leqslant n \max_{j\in[d]} \left\|\mathbb{E}(u_j \boldsymbol{u}^\top \boldsymbol{B})\right\|_2 + \max_{j\in[d]} \left\|\sum_{t=1}^n \mathbb{E}_0(u_{tj} \boldsymbol{u}_t^\top \boldsymbol{B})\right\|_2 = O_{\mathbb{P}}\left(n + \sqrt{nd \log d}\right),$$

where $\mathbb{E}_0(\cdot) = \cdot - \mathbb{E}(\cdot)$. Consequently, we obtain

$$\begin{aligned}
\Delta_2^\circ &= \max_{j\in[d]} \frac{1}{n} \left|\sum_{t=1}^n u_{tj} \boldsymbol{u}_t^\top \boldsymbol{B} \sum_{s=1}^n \boldsymbol{f}_s \widehat{\boldsymbol{f}}_s^\top \boldsymbol{V}^{-1} \boldsymbol{H}^{-\top} \boldsymbol{\phi}\right| \\
&\leqslant \frac{1}{n} \|\boldsymbol{V}^{-1}\|_2 \|\boldsymbol{H}^{-\top}\|_2 \left\|\sum_{s=1}^n \widehat{\boldsymbol{f}}_s \boldsymbol{f}_s^\top\right\|_2 \max_{j\in[d]} \left\|\sum_{t=1}^n u_{tj} \boldsymbol{u}_t^\top \boldsymbol{B}\right\|_2 \\
&= O_{\mathbb{P}}\left\{n/d + \sqrt{n(\log d)/d}\right\}.
\end{aligned}$$

By the triangle inequality,

$$\Delta_3^\circ \leqslant \frac{1}{n}\left(\max_{j\in[d]} \|\boldsymbol{e}_j \boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{F}\boldsymbol{H}^\top\|_2 + \max_{j\in[d]} \|\boldsymbol{e}_j \boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{U}^\top (\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H}^\top)\|_2\right) \|\boldsymbol{V}^{-1}\|_2 \|\boldsymbol{H}^{-\top}\|_2$$

$$=: \Delta_{31}^{\circ} + \Delta_{32}^{\circ}$$

By Assumptions 2.3 and Lemma 2.1, it follows that

$$\sum_{t=1}^{n} \mathbb{E} \left\| \sum_{s=1}^{n} \mathbb{E}_0(\boldsymbol{u}_t^\top \boldsymbol{u}_s) \boldsymbol{f}_s \right\|_2^2 \lesssim n\mathbb{E}\|\mathbb{E}_0(\boldsymbol{u}^\top \boldsymbol{u})\boldsymbol{f}\|_2^2 + n^2 \mathbb{E}\|\boldsymbol{u}_1^\top \boldsymbol{u}_2 \boldsymbol{f}_2\|_2^2 \lesssim n^2 d.$$

Combining this with the fact that $\max_{j\in[d]} \|\boldsymbol{e}_j^\top \boldsymbol{U}^\top \boldsymbol{F}\|_2^2 = O_{\mathbb{P}}(n \log d)$, we obtain

$$\begin{aligned}
\Delta_{31}^{\circ} &\leqslant \max_{j\in[d]} \frac{\mathrm{tr}(\boldsymbol{\Sigma})}{n} \left\| \sum_{t=1}^{n} u_{tj} \boldsymbol{f}_t \right\|_2 \|\boldsymbol{H}\|_2 \|\boldsymbol{V}^{-1}\|_2 \|\boldsymbol{H}^{-\top}\|_2 \\
&\quad + \max_{j\in[d]} \frac{1}{n} \left\| \sum_{t=1}^{n} u_{tj} \sum_{s=1}^{n} \mathbb{E}_0(\boldsymbol{u}_t^\top \boldsymbol{u}_s) \boldsymbol{f}_s \right\|_2 \|\boldsymbol{V}^{-1}\|_2 \|\boldsymbol{H}^{-\top}\|_2 \\
&= O_{\mathbb{P}}\left\{ \sqrt{(\log d)/n} + \sqrt{n/d} \right\}.
\end{aligned}$$

Similarly, by Lemma 2.1, we have $\Delta_{32}^{\circ} = O_{\mathbb{P}}(n/d + 1/n + \sqrt{n/d})$. Putting all these pieces together, we obtain (C.3). □

Lemma C.4. *Under Assumptions 2.1–2.5, we have*

$$\|(\widehat{\boldsymbol{U}} - \boldsymbol{U})^\top \mathcal{E}\|_\infty = O_{\mathbb{P}}\left(\sqrt{\log d + \frac{n}{d}} \right).$$

Proof of Lemma C.4. Recall that $\widehat{\boldsymbol{U}} = \boldsymbol{X} - \widehat{\boldsymbol{F}}\widehat{\boldsymbol{B}}^\top$. By the triangle inequality,

$$\begin{aligned}
\|(\widehat{\boldsymbol{U}} - \boldsymbol{U})^\top \mathcal{E}\|_\infty &\leqslant \|(\widehat{\boldsymbol{B}} - \boldsymbol{B}\boldsymbol{H}^\top)\widehat{\boldsymbol{F}}^\top \mathcal{E}\|_\infty + \|\boldsymbol{B}\boldsymbol{H}^\top(\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H}^\top)^\top \mathcal{E}\|_\infty \\
&\quad + \|\boldsymbol{B}(\boldsymbol{H}^\top \boldsymbol{H} - \boldsymbol{I}_K)\boldsymbol{F}^\top \mathcal{E}\|_\infty =: \Delta_1 + \Delta_2 + \Delta_3.
\end{aligned}$$

By Lemma 2.1 and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\Delta_1 &\leqslant \max_{j\in[d]} \|\widehat{\boldsymbol{b}}_j - \boldsymbol{H}\boldsymbol{b}_j\|_2 \|\widehat{\boldsymbol{F}}^\top \mathcal{E}\|_2 = O_{\mathbb{P}}\left(\sqrt{\log d} \right) \\
\Delta_2 &\leqslant \max_{j\in[d]} \|\boldsymbol{b}_j\|_2 \|\boldsymbol{H}\|_2 \|(\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H}^\top)^\top \mathcal{E}\|_2 = O_{\mathbb{P}}\left(1/\sqrt{n} + \sqrt{n/d} \right) \\
\Delta_3 &\leqslant \max_{j\in[d]} \|\boldsymbol{b}_j\|_2 \|\boldsymbol{H}^\top \boldsymbol{H} - \boldsymbol{I}_K\|_{\mathbb{F}} \|\boldsymbol{F}^\top \mathcal{E}\|_2 = O_{\mathbb{P}}\left(1 + \sqrt{n/d} \right).
\end{aligned}$$

□

Lemma C.5. *Under Assumptions 2.1–2.5, we have*

$$\|\widehat{\boldsymbol{U}}^\top(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{U}}\boldsymbol{\beta}^\star)\|_\infty = O_{\mathbb{P}}\left(\sqrt{n \log d} + \mathcal{V}_{n,d} \|\boldsymbol{\varphi}^\star\|_2 \right).$$

Proof of Lemma C.5. By Lemmas C.3 and C.4, we have $\|\widehat{\boldsymbol{U}}^\top \boldsymbol{F}\boldsymbol{\varphi}^\star\|_\infty = O_{\mathbb{P}}(\mathcal{V}_{n,d}\|\boldsymbol{\varphi}^\star\|_2)$ and

$$\|\widehat{\boldsymbol{U}}^\top \mathcal{E}\|_\infty \leqslant \|(\widehat{\boldsymbol{U}} - \boldsymbol{U})^\top \mathcal{E}\|_\infty + \|\boldsymbol{U}^\top \mathcal{E}\|_\infty = O_{\mathbb{P}}\left(\sqrt{n \log d}\right).$$

Consequently, as $\widehat{\boldsymbol{U}}^\top(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{U}}\boldsymbol{\beta}^\star) = \widehat{\boldsymbol{U}}^\top \mathcal{E} + \widehat{\boldsymbol{U}}^\top \boldsymbol{F}\boldsymbol{\varphi}^\star$, it follows that

$$\|\widehat{\boldsymbol{U}}^\top(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{U}}\boldsymbol{\beta}^\star)\|_\infty \leqslant \|\widehat{\boldsymbol{U}}^\top \mathcal{E}\|_\infty + \|\widehat{\boldsymbol{U}}^\top \boldsymbol{F}\boldsymbol{\varphi}^\star\|_\infty = O_{\mathbb{P}}\left(\sqrt{n \log d} + \mathcal{V}_{n,d}\|\boldsymbol{\varphi}^\star\|_2\right).$$

□

Lemma C.6. *Under Assumptions 2.1–2.4, we have*

$$\|\widehat{\boldsymbol{U}}^\top \widehat{\boldsymbol{U}} - \boldsymbol{U}^\top \boldsymbol{U}\|_{\max} = O_{\mathbb{P}}\left(\frac{n}{d} + \log d\right).$$

Proof of Lemma C.6. By the triangle inequality, we have

$$\|\widehat{\boldsymbol{U}}^\top \widehat{\boldsymbol{U}} - \boldsymbol{U}^\top \boldsymbol{U}\|_{\max} \leqslant 2\|\widehat{\boldsymbol{U}}^\top(\widehat{\boldsymbol{U}} - \boldsymbol{U})\|_{\max} + \|(\widehat{\boldsymbol{U}} - \boldsymbol{U})^\top(\widehat{\boldsymbol{U}} - \boldsymbol{U})\|_{\max}.$$

Recall that $\boldsymbol{X} = \boldsymbol{F}\boldsymbol{B}^\top + \boldsymbol{U} = \widehat{\boldsymbol{F}}\widehat{\boldsymbol{B}}^\top + \widehat{\boldsymbol{U}}$ and $\widehat{\boldsymbol{F}}^\top \widehat{\boldsymbol{U}} = \boldsymbol{O}$. Hence, it follows from Lemma C.3 that

$$\|\widehat{\boldsymbol{U}}^\top(\widehat{\boldsymbol{U}} - \boldsymbol{U})\|_{\max} = \|\widehat{\boldsymbol{U}}^\top \boldsymbol{F}\boldsymbol{B}^\top\|_{\max} = \max_{j\in[d]} \|\widehat{\boldsymbol{U}}\boldsymbol{F}\boldsymbol{b}_j\|_\infty = O_{\mathbb{P}}(\mathcal{V}_{n,d}) = O_{\mathbb{P}}\left(\frac{n}{d} + \log d\right).$$

Similarly, by Lemma 2.1,

$$\|(\widehat{\boldsymbol{U}} - \boldsymbol{U})^\top(\widehat{\boldsymbol{U}} - \boldsymbol{U})\|_{\max} \leqslant \max_{j\in[d]} \sum_{t=1}^{n} |\widehat{u}_{tj} - u_{tj}|^2 = O_{\mathbb{P}}\left(\frac{n}{d} + \log d\right).$$

□

D Proof of Results in Section 2

D.1 Proof of Theorem 2.2

Proof of Theorem 2.2. Recall that $\widehat{\boldsymbol{F}}^\top \widehat{\boldsymbol{U}} = \boldsymbol{O}$ and $\widehat{\boldsymbol{\gamma}} = n^{-1}\widehat{\boldsymbol{F}}^\top \boldsymbol{Y}$. Hence

$$\widehat{\boldsymbol{\gamma}} - \boldsymbol{H}\boldsymbol{\gamma}^\star = \frac{1}{n}\widehat{\boldsymbol{F}}^\top \mathcal{E} + \frac{1}{n}\widehat{\boldsymbol{F}}^\top(\boldsymbol{F} - \widehat{\boldsymbol{F}}\boldsymbol{H})\boldsymbol{\varphi}^\star + (\widehat{\boldsymbol{B}}^\top - \boldsymbol{H}\boldsymbol{B}^\top)\boldsymbol{\beta}^\star.$$

By Proposition 2.1, we have

$$\|\widehat{\boldsymbol{F}}^\top(\boldsymbol{F} - \widehat{\boldsymbol{F}}\boldsymbol{H})\boldsymbol{\varphi}^\star\|_2 \leqslant \|\widehat{\boldsymbol{F}}\|_{\mathbb{F}}\|\boldsymbol{F} - \widehat{\boldsymbol{F}}\boldsymbol{H}\|_{\mathbb{F}}\|\boldsymbol{\varphi}^\star\|_2 = O_{\mathbb{P}}\left\{\left(\sqrt{n} + \frac{n}{\sqrt{d}}\right)\|\boldsymbol{\varphi}^\star\|_2\right\},$$

$$\|(\widehat{\boldsymbol{B}} - \boldsymbol{B}\boldsymbol{H}^\top)^\top \boldsymbol{\beta}^\star\|_2 \leqslant \max_{j \in \mathcal{S}_\star} \|\widehat{\boldsymbol{b}}_j - \boldsymbol{H}\boldsymbol{b}_j\|_2 \|\boldsymbol{\beta}^\star\|_1 = O_{\mathbb{P}}\left(\|\boldsymbol{\beta}^\star\|_1 \sqrt{\frac{\log|\mathcal{S}_\star|}{n} + \frac{1}{d}}\right).$$

Combined with the fact that $\|\widehat{\boldsymbol{F}}^\top \mathcal{E}\|_2 = O_{\mathbb{P}}(\sqrt{n})$, we obtain the bound for $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{H}\boldsymbol{\gamma}^\star\|_2$ by the triangle inequality. We now bound $\|\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star\|_2$. Applying Lemma C.1 with Lemmas C.5 and C.2, we obtain (2.6).

□

D.2 Proof of Proposition 2.3

Proof. We introduce several key ingredients for proving Proposition 2.3 in the following several Lemmas. First, the following Lemma D.1 provides upper bounds for several key terms. Here we let $\widehat{\boldsymbol{\nu}}_t = (\widehat{\boldsymbol{u}}_t^\top, \widehat{\boldsymbol{f}}_t^\top)^\top \in \mathbb{R}^{d+K}$.

Lemma D.1. *Under Assumptions 2.1–2.4, we have*

$$\max_{t\in[n]} \|\widehat{\boldsymbol{\nu}}_t\|_\infty = O_{\mathbb{P}}\left(\sqrt{\log d}\right) \quad \text{and} \quad \max_{t\in[n]} |\boldsymbol{e}_t^\top (\boldsymbol{I}_n - \widehat{\boldsymbol{P}})\boldsymbol{F}\boldsymbol{\varphi}^\star| = O_{\mathbb{P}}\left(\frac{\log n}{n + \sqrt{d}} \|\boldsymbol{\varphi}^\star\|_2\right).$$

Furthermore, under the assumption, for any positive constant $\tau < \infty$, we have

$$\max_{j\in[d]} \sum_{t=1}^n \mathbb{I}\{|\varepsilon_t| > \tau\omega\} |\widehat{\nu}_{tj}|^2 = O_{\mathbb{P}}(\log d).$$

Second, we provide an upper bound for the first order derivative of our loss function $\rho_\omega(\cdot)$ evaluated at the global optimizer. In the following, we let $\psi_\omega(\cdot)$ be the first-order derivative function of $\rho_\omega(\cdot)$ and let $\mathcal{S}_\diamond = \mathcal{S}_\star \cup \{d+1, \ldots, d+K\}$. Recall that $\widehat{\boldsymbol{\phi}}_h = (\widehat{\boldsymbol{\beta}}_h^\top, \widehat{\boldsymbol{\gamma}}_h^\top)^\top \in \mathbb{R}^{d+K}$ and $\widetilde{\boldsymbol{\phi}} = (\boldsymbol{\beta}^{\star\top}, \widetilde{\boldsymbol{\gamma}}^\top)^\top \in \mathbb{R}^{d+K}$, where $\widetilde{\boldsymbol{\gamma}} = \widehat{\boldsymbol{B}}^\top \boldsymbol{\beta}^\star + n^{-1}\widehat{\boldsymbol{F}}^\top \boldsymbol{F}\boldsymbol{\varphi}^\star$. Hence we have $\boldsymbol{Y} - \widehat{\boldsymbol{F}}\widetilde{\boldsymbol{\gamma}} - \widehat{\boldsymbol{U}}\boldsymbol{\beta}^\star = (\boldsymbol{I}_n - \widehat{\boldsymbol{P}})\boldsymbol{F}\boldsymbol{\varphi}^\star + \mathcal{E}$.

Lemma D.2. *Assume that $n = O(d \log d)$ and $\mathbb{E}|\varepsilon|^{1+\vartheta} < \infty$ for some constant $\vartheta > 0$. Then, under Assumptions 2.1–2.4, we have*

$$\left\| \sum_{t=1}^n \psi_\omega(y_t - \widehat{\boldsymbol{\nu}}_t^\top \widetilde{\boldsymbol{\phi}}) \widehat{\boldsymbol{\nu}}_t \right\|_\infty = O_{\mathbb{P}}\left(\mathcal{V}_{n,d} \|\boldsymbol{\varphi}^\star\|_2 + \sqrt{n\omega^{1-(\vartheta\wedge 1)} \log d} + \omega \log d\right).$$

Third, we prove the local strong convexity of our loss function $\rho_\omega(\cdot)$ in the following Lemma D.3.

Lemma D.3. *Let $\widehat{\boldsymbol{\Gamma}}_n = n^{-1}\sum_{t=1}^n \mathbb{I}\{|\varepsilon_t| \leqslant \omega/3\} \widehat{\boldsymbol{\nu}}_t \widehat{\boldsymbol{\nu}}_t^\top$. Assume that $|\mathcal{S}_\diamond| \log d = o(n)$ and $\mathbb{E}|\varepsilon|^{1+\vartheta} < \infty$ for some constant $\vartheta > 0$. Then, under Assumptions 2.1–2.4, there exists a constant $\varrho_0 > 0$ such that*

$$\min_{0 \neq \boldsymbol{v} \in \mathcal{C}(\mathcal{S}_\diamond, 3)} \frac{\boldsymbol{v}^\top \widehat{\boldsymbol{\Gamma}}_n \boldsymbol{v}}{\|\boldsymbol{v}\|_2^2} \geqslant \varrho_0. \tag{D.1}$$

For some constant $\kappa_0 > 0$, we define

$$\mathcal{E}_\Delta = \left\{ \max_{t\in[n]} |\boldsymbol{e}_t^\top (\boldsymbol{I}_n - \widehat{\boldsymbol{P}}) \boldsymbol{F} \boldsymbol{\varphi}^\star| \leqslant \frac{\omega}{3} \right\} \text{ and } \mathcal{E}_\nu = \left\{ \max_{t\in[n]} \|\widehat{\boldsymbol{\nu}}_t\|_\infty \leqslant \frac{\omega \varrho_0}{36(1+\kappa_0)\lambda |\mathcal{S}_\diamond|} =: M_\nu \right\}.$$

Next, we prove that $\mathbb{P}(\mathcal{E}_\Delta) \to 1$ and $\mathbb{P}(\mathcal{E}_\nu) \to 1$. By Lemma D.1 and (2.8), we have

$$\max_{t\in[n]} |\boldsymbol{e}_t^\top (\boldsymbol{I}_n - \widehat{\boldsymbol{P}}) \boldsymbol{F} \boldsymbol{\varphi}^\star| = O_{\mathbb{P}} \left(\frac{\log n}{n + \sqrt{d}} \|\boldsymbol{\varphi}^\star\|_2 \right) = o_{\mathbb{P}}(\omega).$$

Therefore $\mathbb{P}(\mathcal{E}_\Delta) \to 1$. As $\max_{t\in[n]} \|\widehat{\boldsymbol{\nu}}_t\|_\infty = O_{\mathbb{P}}(\sqrt{\log d})$ and $|\mathcal{S}_\diamond|(\log d)^{3/2} = o(n)$, then

$$\frac{\max_{t\in[n]} \|\widehat{\boldsymbol{\nu}}_t\|_\infty}{M_\nu} = O_{\mathbb{P}} \left(\frac{\lambda |\mathcal{S}_\diamond| \sqrt{\log d}}{\omega} \right) = o_{\mathbb{P}}(1).$$

Hence $\mathbb{P}(\mathcal{E}_\nu) \to 1$.

By the choice of ω, we have

$$\frac{2}{n} \left\| \sum_{t=1}^n \psi_\omega \left(y_t - \widehat{\boldsymbol{\nu}}_t^\top \widetilde{\boldsymbol{\phi}} \right) \widehat{\boldsymbol{\nu}}_t \right\|_\infty = O_{\mathbb{P}} \left(\mathcal{V}_{n,d} \|\boldsymbol{\varphi}^\star\|_2 + \sqrt{n \omega^{1-(\vartheta \wedge 1)} \log d} + \omega \log d \right) = O_{\mathbb{P}}(\omega \log d).$$

Combining all conclusions given above, we conclude the proof of Proposition 2.3 by the following Lemma D.4.

Lemma D.4. *Assume that $\mathbb{E}|\varepsilon|^{1+\vartheta} < \infty$ for some constant $\vartheta > 0$, and $\max_{t\in[n]} \|\widehat{\boldsymbol{\nu}}_t\|_\infty \leqslant M_\nu$ for some $M_\nu \geqslant 0$. Let*

$$\lambda \geqslant \frac{2}{n} \left\| \sum_{t=1}^n \psi_\omega \left(y_t - \widehat{\boldsymbol{\nu}}_t^\top \widetilde{\boldsymbol{\phi}} \right) \widehat{\boldsymbol{\nu}}_t \right\|_\infty \text{ and } \omega \geqslant 3 \max_{t\in[n]} |\boldsymbol{e}_t^\top (\boldsymbol{I}_n - \widehat{\boldsymbol{P}}) \boldsymbol{F} \boldsymbol{\varphi}^\star|. \tag{D.2}$$

Furthermore, we assume that there exists a constant $\varrho_0 > 36\lambda M_\nu |\mathcal{S}_\diamond|/\omega$ such that

$$\min_{0 \neq \boldsymbol{v} \in \mathcal{C}(\mathcal{S}_\diamond, 3)} \frac{\boldsymbol{v}^\top \widehat{\boldsymbol{\Gamma}}_n \boldsymbol{v}}{\|\boldsymbol{v}\|_2^2} \geqslant \varrho_0. \tag{D.3}$$

Then we have $\|\widehat{\boldsymbol{\phi}}_h - \widetilde{\boldsymbol{\phi}}\|_1 \leqslant 12\lambda |\mathcal{S}_\diamond|/\varrho_0$.

□

D.3 Proof of Lemma D.1

Proof. By Lemma 2.1,

$$\max_{t\in[n]}\|\widehat{\boldsymbol{f}}_t\|_\infty \leqslant \max_{t\in[n]}\|\boldsymbol{H}\boldsymbol{f}_t\|_\infty + \max_{t\in[n]}\|\widehat{\boldsymbol{f}}_t-\boldsymbol{H}\boldsymbol{f}_t\|_\infty = O_{\mathbb{P}}\left(\sqrt{\log n}+\frac{\log n}{n+\sqrt{d}}\right)=O_{\mathbb{P}}\left(\sqrt{\log n}\right).$$

Decompose $\boldsymbol{u}_t-\widehat{\boldsymbol{u}}_t=(\widehat{\boldsymbol{B}}-\boldsymbol{B}\boldsymbol{H}^{-1})(\widehat{\boldsymbol{f}}_t-\boldsymbol{H}\boldsymbol{f}_t)+(\widehat{\boldsymbol{B}}-\boldsymbol{B}\boldsymbol{H}^{-1})\boldsymbol{H}\boldsymbol{f}_t+\boldsymbol{B}\boldsymbol{H}^{-1}(\widehat{\boldsymbol{f}}_t-\boldsymbol{H}\boldsymbol{f}_t)$. Then, by Lemma 2.1,

$$\begin{aligned}\max_{t\in[n]}\|\boldsymbol{u}_t-\widehat{\boldsymbol{u}}_t\|_\infty &\leqslant K\|\widehat{\boldsymbol{B}}-\boldsymbol{B}\boldsymbol{H}^{-1}\|_{\max}\|\widehat{\boldsymbol{F}}-\boldsymbol{F}\boldsymbol{H}^\top\|_{\max}\\ &\quad+\sqrt{K}\|\widehat{\boldsymbol{B}}-\boldsymbol{B}\boldsymbol{H}^{-1}\|_{\max}\|\boldsymbol{H}\|_2\max_{t\in[n]}\|\boldsymbol{f}_t\|_2\\ &\quad+K\|\boldsymbol{H}^{-1}\|_2\|\boldsymbol{B}\|_{\max}\|\widehat{\boldsymbol{F}}-\boldsymbol{F}\boldsymbol{H}^\top\|_{\max}\\ &=O_{\mathbb{P}}\left(\sqrt{(\log d)(\log n)/n}\right)=O_{\mathbb{P}}\left(\sqrt{\log d}\right).\end{aligned}$$

Consequently, by Assumption 2.1, we have

$$\max_{t\in[n]}\|\widehat{\boldsymbol{u}}_t\|_\infty \leqslant \max_{t\in[n]}\|\boldsymbol{u}_t\|_\infty+\max_{t\in[n]}\|\widehat{\boldsymbol{u}}_t-\boldsymbol{u}_t\|_\infty=O_{\mathbb{P}}\left(\sqrt{\log d}\right).$$

We now bound $\max_{t\in[n]}|\boldsymbol{e}_t^\top(\boldsymbol{I}_n-\widehat{\boldsymbol{P}})\boldsymbol{F}\boldsymbol{\varphi}^\star|$. Observe that $(\boldsymbol{I}_n-\widehat{\boldsymbol{P}})\widehat{\boldsymbol{F}}=\boldsymbol{O}$. Hence

$$\begin{aligned}\max_{t\in[n]}|\boldsymbol{e}_t^\top(\boldsymbol{I}_n-\widehat{\boldsymbol{P}})\boldsymbol{F}\boldsymbol{\varphi}^\star| &\leqslant \max_{t\in[n]}|\boldsymbol{e}_t^\top(\boldsymbol{I}_n-\widehat{\boldsymbol{P}})(\boldsymbol{F}\boldsymbol{H}^\top-\widehat{\boldsymbol{F}})\boldsymbol{H}^{-\top}\boldsymbol{\varphi}^\star|\\ &\leqslant \max_{t\in[n]}|(\widehat{\boldsymbol{f}}_t-\boldsymbol{H}\boldsymbol{f}_t)^\top\boldsymbol{H}^{-\top}\boldsymbol{\varphi}^\star|+\max_{t\in[n]}\frac{1}{n}|\widehat{\boldsymbol{f}}_t^\top\widehat{\boldsymbol{F}}^\top(\boldsymbol{F}\boldsymbol{H}^\top-\widehat{\boldsymbol{F}})\boldsymbol{H}^{-\top}\boldsymbol{\varphi}^\star|\\ &\leqslant \max_{t\in[n]}\|\widehat{\boldsymbol{f}}_t-\boldsymbol{H}\boldsymbol{f}_t\|_2\|\boldsymbol{H}^{-\top}\|_2\|\boldsymbol{\varphi}^\star\|_2+\max_{t\in[n]}\frac{1}{n}\|\widehat{\boldsymbol{f}}_t\|_2\|\widehat{\boldsymbol{F}}\|_{\mathbb{F}}\|\boldsymbol{F}\boldsymbol{H}^\top-\widehat{\boldsymbol{F}}\|_{\mathbb{F}}\|\boldsymbol{H}^{-\top}\|_2\|\boldsymbol{\varphi}^\star\|_2\\ &=O_{\mathbb{P}}\left(\frac{\log n}{n+\sqrt{d}}\|\boldsymbol{\varphi}^\star\|_2+\frac{1}{n}\sqrt{\log n}\sqrt{n}\sqrt{\frac{1}{n}+\frac{n}{d}}\|\boldsymbol{\varphi}^\star\|_2\right)=O_{\mathbb{P}}\left(\frac{\log n}{n+\sqrt{d}}\|\boldsymbol{\varphi}^\star\|_2\right).\end{aligned}$$

Recall that $n^{-1}\widehat{\boldsymbol{F}}^\top\widehat{\boldsymbol{F}}=\boldsymbol{I}_K$. Hence

$$\mathbb{E}\left(\max_{k\in[K]}\sum_{t=1}^n\mathbb{I}\{|\varepsilon_t|>\tau\omega\}|\widehat{f}_{tk}|^2\right)\leqslant\mathbb{E}\left(\sum_{t=1}^n\mathbb{I}\{|\varepsilon_t|>\tau\omega\}\|\widehat{\boldsymbol{f}}_t\|_2^2\right)\leqslant\frac{nK\mathbb{E}|\varepsilon|^{1+(\vartheta\wedge1)}}{(\tau\omega)^{1+(\vartheta\wedge1)}}.$$

Then it suffices to bound $\max_{j\in[d]}\sum_{t=1}^n\sum_{t=1}^n\mathbb{I}\{|\varepsilon_t|>\tau\omega\}u_{tj}^2$ as $\max_{j\in[d]}\sum_{t=1}^n(\widehat{u}_{tj}-u_{tj})^2=O_{\mathbb{P}}(\log d+n/d)$. By Assumption 2.1, we have $\mathbb{E}\exp(u_{tj}^2/c_0^2)\leqslant 2$ uniformly for $j\in[d]$. By Jensen's inequality and

$$\mathbb{E}\left(\max_{j\in[d]}\frac{1}{c_0^2}\sum_{t=1}^n\mathbb{I}\{|\varepsilon_t|>\tau\omega\}u_{tj}^2\right)\leqslant\log\left\{\sum_{j=1}^d\mathbb{E}\exp\left(\frac{1}{c_0^2}\sum_{t=1}^n\mathbb{I}\{|\varepsilon_t|>\tau\omega\}u_{tj}^2\right)\right\}$$

$$\leqslant \log\left[d\max_{j\in[d]}\left\{\mathbb{E}\exp\left(\frac{u_{tj}^2}{c_0^2}\right)\mathbb{P}(|\varepsilon|>\tau\omega)+\mathbb{P}(|\varepsilon|\leqslant\tau\omega)\right\}^n\right]$$
$$\leqslant \log d+\frac{n\mathbb{E}|\varepsilon|^{1+(\vartheta\wedge 1)}}{(\tau\omega)^{1+(\vartheta\wedge 1)}}.$$

□

D.4 Proof of Lemma D.2

Proof. Recall that $\boldsymbol{Y}-\widehat{\boldsymbol{U}}\boldsymbol{\beta}^\star-\widehat{\boldsymbol{F}}\widetilde{\boldsymbol{\gamma}}=\mathcal{E}+(\boldsymbol{I}_n-\widehat{\boldsymbol{P}})\boldsymbol{F}\boldsymbol{\varphi}^\star$. Hence, by Taylor's formula,

$$\begin{aligned}\sum_{t=1}^n\psi_\omega(y_t-\widehat{\boldsymbol{\nu}}_t^\top\widetilde{\boldsymbol{\phi}})\widehat{\boldsymbol{\nu}}_t&=\sum_{t=1}^n\psi_\omega(\varepsilon_t)(\widehat{\boldsymbol{\nu}}_t-\widetilde{\boldsymbol{\nu}}_t)+\sum_{t=1}^n\psi_\omega(\varepsilon_t)\widetilde{\boldsymbol{\nu}}_t\\&\quad+\sum_{t=1}^n\boldsymbol{e}_t^\top(\boldsymbol{I}_n-\widehat{\boldsymbol{P}})\boldsymbol{F}\boldsymbol{\varphi}^\star\widehat{\boldsymbol{\nu}}_t\int_0^1\mathbb{I}\left\{|\varepsilon_t+s\boldsymbol{e}_t^\top(\boldsymbol{I}_n-\widehat{\boldsymbol{P}})\boldsymbol{F}\boldsymbol{\varphi}^\star|\leqslant\omega\right\}ds\\&=:\boldsymbol{\Delta}_1+\boldsymbol{\Delta}_2+\boldsymbol{\Delta}_3,\end{aligned}$$

where $\widetilde{\boldsymbol{\nu}}_t=(\boldsymbol{u}_t^\top,\boldsymbol{f}_t^\top\boldsymbol{H}^\top)^\top\in\mathbb{R}^{d+K}$. Note that $\mathbb{E}|\psi_\omega(\varepsilon)|^2\leqslant\mathbb{E}|\varepsilon|^{1+(\vartheta\wedge1)}\omega^{1-(\vartheta\wedge1)}$. Hence, by Lemma 2.1 and the Cauchy-Schwarz inequality,

$$\|\boldsymbol{\Delta}_1\|_\infty^2\leqslant\sum_{t=1}^n\{\psi_\omega(\varepsilon_t)\}^2\max_{1\leqslant j\leqslant d+K}\sum_{t=1}^n(\widehat{\nu}_{tj}-\nu_{tj})^2=O_{\mathbb{P}}\left\{n\omega^{1-(\vartheta\wedge1)}\left(\log d+\frac{n}{d}\right)\right\}.$$

By Lemma C.6 in Sun et al. [2020] and Lemma 2.1, we have

$$\|\boldsymbol{\Delta}_2\|_\infty=O_{\mathbb{P}}\left(\sqrt{n\omega^{1-(\vartheta\wedge1)}\log d}+\omega\log d\right).$$

Decompose

$$\begin{aligned}\boldsymbol{\Delta}_3&=\sum_{t=1}^n\boldsymbol{e}_t^\top(\boldsymbol{I}_n-\widehat{\boldsymbol{P}})\boldsymbol{F}\boldsymbol{\varphi}^\star\widehat{\boldsymbol{\nu}}_t-\sum_{t=1}^n\boldsymbol{e}_t^\top(\boldsymbol{I}_n-\widehat{\boldsymbol{P}})\boldsymbol{F}\boldsymbol{\varphi}^\star\widehat{\boldsymbol{\nu}}_t\int_0^1\mathbb{I}\left\{|\varepsilon_t+s\boldsymbol{e}_t^\top(\boldsymbol{I}_n-\widehat{\boldsymbol{P}})\boldsymbol{F}\boldsymbol{\varphi}^\star|>\omega\right\}ds\\&=:\boldsymbol{\Delta}_{31}+\boldsymbol{\Delta}_{32}.\end{aligned}$$

Observe that $\sum_{t=1}^n\boldsymbol{e}_t^\top(\boldsymbol{I}_n-\widehat{\boldsymbol{P}})\boldsymbol{F}\boldsymbol{\varphi}^\star\widehat{\boldsymbol{f}}_t=\widehat{\boldsymbol{F}}^\top(\boldsymbol{I}_n-\widehat{\boldsymbol{P}})\boldsymbol{F}\boldsymbol{\varphi}^\star=0$. Hence, by Lemma C.3,

$$\|\boldsymbol{\Delta}_{31}\|_\infty=\|\widehat{\boldsymbol{U}}^\top(\boldsymbol{I}_n-\widehat{\boldsymbol{P}})\boldsymbol{F}\boldsymbol{\varphi}^\star\|_\infty=O_{\mathbb{P}}(\mathcal{V}_{n,d}\|\boldsymbol{\varphi}^\star\|_2).$$

Since

$$\sum_{t=1}^n|\boldsymbol{e}_t^\top(\boldsymbol{I}_n-\widehat{\boldsymbol{P}})\boldsymbol{F}\boldsymbol{\varphi}^\star|^2\leqslant\|\boldsymbol{F}\boldsymbol{H}^\top-\widehat{\boldsymbol{F}}\|_{\mathbb{F}}^2\|\boldsymbol{H}^{-\top}\|_2^2\|\boldsymbol{\varphi}^\star\|_2^2=O_{\mathbb{P}}\left\{\left(\frac{1}{n}+\frac{n}{d}\right)\|\boldsymbol{\varphi}^\star\|_2^2\right\}.$$

Combined with the fact that $\mathbb{P}(|\varepsilon_t| > 2\omega/3) \lesssim (\log d)/n$, we obtain

$$\boldsymbol{\Delta}_{32}^{\diamond} := \sum_{t=1}^{n} \mathbb{I}\left\{|\varepsilon_t| > \frac{2\omega}{3}\right\} |\boldsymbol{e}_t^\top (\boldsymbol{I}_n - \widehat{\boldsymbol{P}}) \boldsymbol{F} \boldsymbol{\varphi}^\star|^2 = o_{\mathbb{P}}\left(\frac{\mathcal{V}_{n,d}^2}{\log d}\right).$$

Consequently, by the Cauchy-Schwarz inequality and Lemma D.1,

$$\|\boldsymbol{\Delta}_{32}\|_\infty \leqslant \left(\boldsymbol{\Delta}_{32}^{\diamond} \max_{k\in[K]} \sum_{t=1}^{n} \mathbb{I}\left\{|\varepsilon_t| > \frac{2\omega}{3}\right\} \widehat{\nu}_{tj}^2\right)^{1/2} = o_{\mathbb{P}}(\mathcal{V}_{n,d} \|\boldsymbol{\varphi}^\star\|_2).$$

□

D.5 Proof of Lemma D.3

Proof. For simplicity of notation, write

$$\widetilde{\boldsymbol{\Gamma}}_n = \frac{1}{n}\sum_{t=1}^{n} \mathbb{I}\left\{|\varepsilon_t| \leqslant \frac{\omega}{3}\right\} \widetilde{\boldsymbol{\nu}}_t \widetilde{\boldsymbol{\nu}}_t^\top \text{ and } \boldsymbol{\Gamma}_n = \frac{1}{n}\sum_{t=1}^{n} \mathbb{I}\left\{|\varepsilon_t| \leqslant \frac{\omega}{3}\right\} \boldsymbol{\nu}_t \boldsymbol{\nu}_t^\top,$$

where $\boldsymbol{\nu}_t = (\boldsymbol{u}_t^\top, \boldsymbol{f}_t^\top)^\top \in \mathbb{R}^{d+K}$. For any $\boldsymbol{v} \in \mathcal{C}(\mathcal{S}_\diamond, 3)$, we write $\boldsymbol{v} = (\boldsymbol{v}_{[d]}^\top, \boldsymbol{v}_{[-d]}^\top)^\top \in \mathbb{R}^{d+K}$, where $\boldsymbol{v}_{[d]} \in \mathbb{R}^d$ and $\boldsymbol{v}_{[-d]} \in \mathbb{R}^K$. We first bound $\sup_{0\neq\boldsymbol{v}\in\mathcal{C}(\mathcal{S}_\diamond,3)} |\boldsymbol{v}^\top(\widehat{\boldsymbol{\Gamma}}_n - \widetilde{\boldsymbol{\Gamma}}_n)\boldsymbol{v}|$. Decompose

$$\begin{aligned}
\boldsymbol{v}^\top(\widehat{\boldsymbol{\Gamma}}_n - \widetilde{\boldsymbol{\Gamma}}_n)\boldsymbol{v} &= \frac{1}{n}\sum_{t=1}^{n} \mathbb{I}\left\{|\varepsilon_t| \leqslant \frac{\omega}{3}\right\} \left\{(\widehat{\boldsymbol{u}}_t^\top \boldsymbol{v}_{[d]})^2 - (\boldsymbol{u}_t^\top \boldsymbol{v}_{[d]})^2\right\} \\
&\quad + \frac{2}{n}\sum_{t=1}^{n} \mathbb{I}\left\{|\varepsilon_t| \leqslant \frac{\omega}{3}\right\} \left(\boldsymbol{v}_{[d]}^\top \widehat{\boldsymbol{u}}_t \widehat{\boldsymbol{f}}_t^\top \boldsymbol{v}_{[-d]} - \boldsymbol{v}_{[d]}^\top \boldsymbol{u}_t \boldsymbol{f}_t^\top \boldsymbol{H}^\top \boldsymbol{v}_{[-d]}\right) \\
&\quad + \frac{1}{n}\sum_{t=1}^{n} \mathbb{I}\left\{|\varepsilon_t| \leqslant \frac{\omega}{3}\right\} \left\{(\widehat{\boldsymbol{f}}_t^\top \boldsymbol{v}_{[-d]})^2 - (\boldsymbol{f}_t^\top \boldsymbol{H}^\top \boldsymbol{v}_{[-d]})^2\right\} \\
&=: \frac{1}{n}\{\boldsymbol{\Delta}_1(\boldsymbol{v}) + 2\boldsymbol{\Delta}_2(\boldsymbol{v}) + \boldsymbol{\Delta}_3(\boldsymbol{v})\}.
\end{aligned}$$

Let $\boldsymbol{D} = \mathrm{diag}\{D_{11}, \ldots, D_{nn}\} \in \mathbb{R}^{n\times n}$ denote an identity matrix, where $D_{tt} = \mathbb{I}\{|\varepsilon_t| \leqslant \omega/3\}$ for each $t \in [n]$. Then, for any $\boldsymbol{v} \in \mathcal{C}(\mathcal{S}_\diamond, 3)$,

$$\begin{aligned}
\boldsymbol{\Delta}_1(\boldsymbol{v}) &= \boldsymbol{v}_{[d]}^\top \left(\widehat{\boldsymbol{U}}^\top \boldsymbol{D} \widehat{\boldsymbol{U}} - \boldsymbol{U}^\top \boldsymbol{D} \boldsymbol{U}\right) \boldsymbol{v}_{[d]} \\
&= \boldsymbol{v}_{[d]}^\top \left(\widehat{\boldsymbol{U}}^\top \widehat{\boldsymbol{U}} - \boldsymbol{U}^\top \boldsymbol{U}\right) \boldsymbol{v}_{[d]} + \boldsymbol{v}_{[d]}^\top \left(\widehat{\boldsymbol{U}} - \boldsymbol{U}\right)^\top \bar{\boldsymbol{D}} \boldsymbol{U} \boldsymbol{v}_{[d]} + \boldsymbol{v}_{[d]}^\top \widehat{\boldsymbol{U}}^\top \bar{\boldsymbol{D}} \left(\widehat{\boldsymbol{U}} - \boldsymbol{U}\right) \boldsymbol{v}_{[d]} \\
&=: \boldsymbol{\Delta}_{11}(\boldsymbol{v}) + \boldsymbol{\Delta}_{12}(\boldsymbol{v}) + \boldsymbol{\Delta}_{13}(\boldsymbol{v}).
\end{aligned}$$

where $\bar{\boldsymbol{D}} = \mathrm{diag}\{1 - D_{11}, \ldots, 1 - D_{nn}\} \in \mathbb{R}^{n\times n}$. By Lemma C.6, we have

$$\sup_{0\neq \boldsymbol{v}\in\mathcal{C}(\mathcal{S}_\diamond,3)} \frac{|\boldsymbol{\Delta}_{11}(\boldsymbol{v})|}{\|\boldsymbol{v}\|_2^2} \leqslant \sup_{0\neq \boldsymbol{v}\in\mathcal{C}(\mathcal{S}_\diamond,3)} \frac{\|\widehat{\boldsymbol{U}}^\top\widehat{\boldsymbol{U}} - \boldsymbol{U}^\top\boldsymbol{U}\|_{\max}\|\boldsymbol{v}_{[d]}\|_1^2}{\|\boldsymbol{v}\|_2^2}$$
$$\leqslant 32|\mathcal{S}_\diamond|\|\widehat{\boldsymbol{U}}^\top\widehat{\boldsymbol{U}} - \boldsymbol{U}^\top\boldsymbol{U}\|_{\max} = O_{\mathbb{P}}(|\mathcal{S}_\diamond|\log d).$$

By the Cauchy-Schwarz inequality and Lemma D.1,

$$\sup_{0\neq \boldsymbol{v}\in\mathcal{C}(\mathcal{S}_\diamond,3)} \frac{|\boldsymbol{\Delta}_{12}(\boldsymbol{v})|}{\|\boldsymbol{v}\|_2^2} \leqslant 32|\mathcal{S}_\diamond| \max_{1\leqslant j,\ell\leqslant d}\left|\sum_{t=1}^n \mathbb{I}\left\{|\varepsilon_t| > \frac{\omega}{3}\right\}\widehat{u}_{tj}(\widehat{u}_{t\ell} - u_{t\ell})\right|$$
$$\leqslant 32|\mathcal{S}_\diamond| \max_{1\leqslant j,\ell\leqslant d}\left[\sum_{t=1}^n \mathbb{I}\left\{|\varepsilon_t| > \frac{\omega}{3}\right\}u_{tj}^2 \sum_{s=1}^n \mathbb{I}\left\{|\varepsilon_t| > \frac{\omega}{3}\right\}(\widehat{u}_{t\ell} - u_{t\ell})^2\right]^{1/2}$$
$$= O_{\mathbb{P}}\left\{|\mathcal{S}_\diamond|(\log d)^{1/2}\left(\log d + \frac{n}{d}\right)^{1/2}\right\} = O_{\mathbb{P}}(|\mathcal{S}_\diamond|\log d).$$

Similarly, we have $\sup_{0\neq \boldsymbol{v}\in\mathcal{C}(\mathcal{S}_\diamond,3)}\{|\boldsymbol{\Delta}_{13}(\boldsymbol{v})|/\|\boldsymbol{v}\|_2^2\} = O_{\mathbb{P}}(|\mathcal{S}_\diamond|\log d)$. We now upper bound $|\boldsymbol{\Delta}_2(\boldsymbol{v})|$. Decompose

$$\begin{aligned}\boldsymbol{\Delta}_2(\boldsymbol{v}) &= \boldsymbol{v}_{[d]}^\top\left(\widehat{\boldsymbol{U}}^\top\boldsymbol{D}\widehat{\boldsymbol{F}} - \boldsymbol{U}^\top\boldsymbol{D}\boldsymbol{F}\boldsymbol{H}^\top\right)\boldsymbol{v}_{[-d]}\\ &= \boldsymbol{v}_{[d]}^\top\boldsymbol{U}^\top\boldsymbol{F}\boldsymbol{H}^\top\boldsymbol{v}_{[-d]} + \boldsymbol{v}_{[d]}^\top\left(\widehat{\boldsymbol{U}} - \boldsymbol{U}\right)\bar{\boldsymbol{D}}\boldsymbol{F}\boldsymbol{H}^\top\boldsymbol{v}_{[-d]} + \boldsymbol{v}_{[d]}^\top\widehat{\boldsymbol{U}}^\top\bar{\boldsymbol{D}}\left(\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H}^\top\right)\boldsymbol{v}_{[-d]}\\ &=: \boldsymbol{\Delta}_{21}(\boldsymbol{v}) + \boldsymbol{\Delta}_{22}(\boldsymbol{v}) + \boldsymbol{\Delta}_{23}(\boldsymbol{v}).\end{aligned}$$

By Assumption 2.1 and the Cauchy-Schwarz inequality, it follows that for any $0 \neq \boldsymbol{v} \in \mathcal{C}(\mathcal{S}_\diamond, 3)$,

$$\begin{aligned}\frac{|\boldsymbol{\Delta}_{21}(\boldsymbol{v})|}{\|\boldsymbol{v}\|_2^2} &\leqslant \frac{\|\boldsymbol{v}_{[d]}\|_1\|\boldsymbol{U}^\top\boldsymbol{F}\boldsymbol{H}^\top\boldsymbol{v}_{[-d]}\|_\infty}{\|\boldsymbol{v}\|_2^2}\\ &\leqslant \max_{j\in[d]}\left\|\sum_{t=1}^n u_{tj}\boldsymbol{f}_t\right\|_2\|\boldsymbol{H}^\top\|_2\frac{(4\sqrt{|\mathcal{S}_\star|}\|\boldsymbol{v}_{\mathcal{S}_\star}\|_2 + 3\sqrt{K}\|\boldsymbol{v}_{[-d]}\|_2)\|\boldsymbol{v}_{[-d]}\|_2}{\|\boldsymbol{v}\|_2^2}\\ &\leqslant 7\max_{j\in[d]}\left\|\sum_{t=1}^n u_{tj}\boldsymbol{f}_t\right\|_2\|\boldsymbol{H}^\top\|_2\sqrt{|\mathcal{S}_\diamond|} = O_{\mathbb{P}}\left(\sqrt{n|\mathcal{S}_\diamond|\log d}\right).\end{aligned}$$

Similarly, we have

$$\sup_{0\neq \boldsymbol{v}\in\mathcal{C}(\mathcal{S}_\diamond,3)} \frac{|\boldsymbol{\Delta}_{22}(\boldsymbol{v})|}{\|\boldsymbol{v}\|_2^2} \leqslant 7\max_{j\in[d]}\left\|\sum_{t=1}^n \mathbb{I}\left\{|\varepsilon_t| > \frac{\omega}{3}\right\}(\widehat{u}_{tj} - u_{tj})\boldsymbol{f}_t\right\|_2\|\boldsymbol{H}^\top\|_2\sqrt{|\mathcal{S}_\diamond|}$$
$$= O_{\mathbb{P}}\left\{\sqrt{|\mathcal{S}_\diamond|(\log d)\left(\log d + \frac{n}{d}\right)}\right\} = o_{\mathbb{P}}\left(\sqrt{n|\mathcal{S}_\diamond|\log d}\right),$$

and $\sup_{0\neq \boldsymbol{v}\in\mathcal{C}(\mathcal{S}_\diamond,3)}\{|\boldsymbol{\Delta}_{23}(\boldsymbol{v})|/\|\boldsymbol{v}\|_2^2\} = o_{\mathbb{P}}(\sqrt{n|\mathcal{S}_\diamond|\log d})$. Therefore $\sup_{0\neq \boldsymbol{v}\in\mathcal{C}(\mathcal{S}_\diamond,3)}\{|\boldsymbol{\Delta}_{2}(\boldsymbol{v})|/\|\boldsymbol{v}\|_2^2\} = O_{\mathbb{P}}(\sqrt{n|\mathcal{S}_\diamond|\log d})$. Decompose

$$\begin{aligned}
\boldsymbol{\Delta}_3(\boldsymbol{v}) &= \boldsymbol{v}_{[-d]}^\top \left(\widehat{\boldsymbol{F}}^\top \boldsymbol{D}\widehat{\boldsymbol{F}} - \boldsymbol{H}\boldsymbol{F}^\top \boldsymbol{D}\boldsymbol{F}\boldsymbol{H}^\top\right)\boldsymbol{v}_{[-d]} \\
&= \boldsymbol{v}_{[-d]}^\top \left(n\boldsymbol{I}_K - \boldsymbol{H}\boldsymbol{F}^\top\boldsymbol{F}\boldsymbol{H}^\top\right)\boldsymbol{v}_{[-d]} + \boldsymbol{v}_{[-d]}^\top\left(\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H}^\top\right)^\top \bar{\boldsymbol{D}}\boldsymbol{F}\boldsymbol{H}^\top\boldsymbol{v}_{[-d]} \\
&+ \boldsymbol{v}_{[-d]}^\top\widehat{\boldsymbol{F}}^\top\bar{\boldsymbol{D}}\left(\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H}^\top\right)\boldsymbol{v}_{[-d]} \\
&=: \boldsymbol{\Delta}_{31}(\boldsymbol{v}) + \boldsymbol{\Delta}_{32}(\boldsymbol{v}) + \boldsymbol{\Delta}_{33}(\boldsymbol{v}).
\end{aligned}$$

By Lemma 2.1, we have

$$\begin{aligned}
\sup_{0\neq \boldsymbol{v}\in\mathcal{C}(\mathcal{S}_\diamond,3)}\frac{|\boldsymbol{\Delta}_{31}(\boldsymbol{v})|}{\|\boldsymbol{v}\|_2^2} &\leqslant \sup_{0\neq \boldsymbol{v}\in\mathcal{C}(\mathcal{S}_\diamond,3)}\frac{\|n\boldsymbol{I}_K - \boldsymbol{H}\boldsymbol{F}^\top\boldsymbol{F}\boldsymbol{H}^\top\|_2\|\boldsymbol{v}_{[-d]}\|_2^2}{\|\boldsymbol{v}\|_2^2} \\
&\leqslant n\|\boldsymbol{I}_K - \boldsymbol{H}\boldsymbol{H}^\top\|_2 + \|\boldsymbol{H}\|_2\|n\boldsymbol{I}_K - \boldsymbol{F}^\top\boldsymbol{F}\|_2\|\boldsymbol{H}^\top\|_2 \\
&= O_{\mathbb{P}}\left(\sqrt{n} + n/\sqrt{d}\right), \\
\sup_{0\neq \boldsymbol{v}\in\mathcal{C}(\mathcal{S}_\diamond,3)}\frac{|\boldsymbol{\Delta}_{32}(\boldsymbol{v})|}{\|\boldsymbol{v}\|_2^2} &\leqslant \left\|\sum_{t=1}^n \mathbb{I}\left\{|\varepsilon_t| > \frac{\omega}{3}\right\}(\widehat{\boldsymbol{f}}_t - \boldsymbol{H}\boldsymbol{f}_t)\boldsymbol{f}_t^\top\right\|_2 \|\boldsymbol{H}^\top\|_2 \\
&\leqslant \left(\sum_{t=1}^n \mathbb{I}\left\{|\varepsilon_t| > \frac{\omega}{3}\right\}\|\widehat{\boldsymbol{f}}_t - \boldsymbol{H}\boldsymbol{f}_t\|_2^2 \sum_{t=1}^n \mathbb{I}\left\{|\varepsilon_t| > \frac{\omega}{3}\right\}\|\boldsymbol{f}_t\|_2^2\right)^{1/2}\|\boldsymbol{H}^\top\|_2 \\
&= o_{\mathbb{P}}\left(\sqrt{n} + n/\sqrt{d}\right),
\end{aligned}$$

and $\sup_{0\neq \boldsymbol{v}\in\mathcal{C}(\mathcal{S}_\diamond,3)}\{|\boldsymbol{\Delta}_{33}(\boldsymbol{v})|/\|\boldsymbol{v}\|_2^2\} = o_{\mathbb{P}}(\sqrt{n} + n/\sqrt{d})$. Putting all these pieces together, we obtain

$$\sup_{0\neq \boldsymbol{v}\in\mathcal{C}(\mathcal{S}_\diamond,3)}\frac{|\boldsymbol{v}^\top(\widehat{\boldsymbol{\Gamma}}_n - \widetilde{\boldsymbol{\Gamma}}_n)\boldsymbol{v}|}{\|\boldsymbol{v}\|_2^2} = O_{\mathbb{P}}\left(\frac{|\mathcal{S}_\diamond|\log d}{n} + \sqrt{\frac{|\mathcal{S}_\diamond|\log d}{n}} + \frac{1}{\sqrt{n}+\sqrt{d}}\right) = o_{\mathbb{P}}(1).$$

For any $\boldsymbol{v}\in\mathcal{C}(\mathcal{S}_\diamond,3)$, write $\widetilde{\boldsymbol{v}} = (\boldsymbol{v}_{[d]}^\top, \boldsymbol{v}_{[-d]}^\top\boldsymbol{H})^\top \in \mathbb{R}^{d+K}$. Define $\mathcal{E}_H = \{\lambda_{\min}(\boldsymbol{H}^\top\boldsymbol{H}) \geqslant 1/4\}$. It follows form Lemma 2.1 that $\mathbb{P}(\mathcal{E}_H)\to 1$. Under $\mathcal{E}_H$, we have

$$\|\widetilde{\boldsymbol{v}}_{\mathcal{S}_\diamond^c}\|_1 = \|\boldsymbol{v}_{\mathcal{S}_\diamond^c}\|_1 \leqslant 3\|\boldsymbol{v}_{\mathcal{S}_\star}\|_1 + 3\|\boldsymbol{v}_{[-d]}\|_1 \leqslant 3\|\boldsymbol{v}_{\mathcal{S}_\star}\|_1 + 6\sqrt{K}\|\boldsymbol{H}^\top\boldsymbol{v}_{[-d]}\|_1 \leqslant 6\sqrt{K}\|\widetilde{\boldsymbol{v}}_{\mathcal{S}_\diamond}\|_1.$$

Combined with the fact that

$$\|\widetilde{\boldsymbol{v}}\|_2^2 = \|\boldsymbol{v}_{[d]}\|_2^2 + \|\boldsymbol{H}^\top\boldsymbol{v}_{[-d]}\|_2^2 \geqslant \|\boldsymbol{v}_{[d]}\|_2^2 + \frac{1}{4}\|\boldsymbol{v}_{[-d]}\|_2^2 \geqslant \frac{1}{4}\|\boldsymbol{v}\|_2^2,$$

we obtain

$$\inf_{0\neq \boldsymbol{v}\in\mathcal{C}(\mathcal{S}_\diamond,3)}\frac{\boldsymbol{v}^\top\widetilde{\boldsymbol{\Gamma}}_n\boldsymbol{v}}{\|\boldsymbol{v}\|_2^2} \geqslant \inf_{0\neq \widetilde{\boldsymbol{v}}\in\mathcal{C}(\mathcal{S}_\diamond,6\sqrt{K})}\frac{\widetilde{\boldsymbol{v}}^\top\boldsymbol{\Gamma}_n\widetilde{\boldsymbol{v}}}{\|\widetilde{\boldsymbol{v}}\|_2^2} \geqslant \inf_{0\neq \widetilde{\boldsymbol{v}}\in\mathcal{C}(\mathcal{S}_\diamond,6\sqrt{K})}\frac{\widetilde{\boldsymbol{v}}^\top\boldsymbol{\Gamma}_n\widetilde{\boldsymbol{v}}}{4\|\widetilde{\boldsymbol{v}}\|_2^2}.$$

Observe that $\mathbb{I}\{|\varepsilon_t| \leqslant \omega/3\}\boldsymbol{\nu}_t \in \mathbb{R}^{d+K}$, $t = 1, \ldots, n$, are i.i.d. sub-Gaussian random vectors. Hence, similar to Lemma C.2, $\mathbb{P}(|\varepsilon| \leqslant \omega/3) \geqslant 1/2$ and $\lambda_{\min}\{\mathrm{Cov}(\boldsymbol{\nu}_t)\} \geqslant (\lambda_{\min}(\boldsymbol{\Sigma}) \wedge 1)$, we obtain

$$\mathbb{P}\left\{\inf_{0\neq\widetilde{\boldsymbol{v}}\in\mathcal{C}(\mathcal{S}_\diamond,6\sqrt{K})} \frac{\widetilde{\boldsymbol{v}}^\top \boldsymbol{\Gamma}_n \widetilde{\boldsymbol{v}}}{\|\widetilde{\boldsymbol{v}}\|_2^2} \geqslant \frac{1}{8}(\lambda_{\min}(\boldsymbol{\Sigma}) \wedge 1)\right\} \to 1.$$

Putting all these pieces together, we obtain (D.1). □

D.6 Proof of Lemma D.4

Proof. By the definition of $(\widehat{\boldsymbol{\beta}}_h, \widehat{\boldsymbol{\gamma}}_h)$, we have

$$\frac{1}{n}\sum_{t=1}^n \rho_\omega\left(y_t - \widehat{\boldsymbol{\nu}}_t^\top \widehat{\boldsymbol{\phi}}_h\right) + \lambda\|\widehat{\boldsymbol{\beta}}_h\|_1 \leqslant \frac{1}{n}\sum_{t=1}^n \rho_\omega\left(y_t - \widehat{\boldsymbol{\nu}}_t^\top \widetilde{\boldsymbol{\phi}}\right) + \lambda\|\boldsymbol{\beta}^\star\|_1. \tag{D.4}$$

Let $\boldsymbol{\delta}_\phi = \widehat{\boldsymbol{\phi}}_h - \widetilde{\boldsymbol{\phi}}$. Since $\rho_\omega(\cdot)$ is convex, it follows from (D.2) that

$$\lambda\|\boldsymbol{\beta}^\star\|_1 - \lambda\|\widehat{\boldsymbol{\beta}}_h\|_1 \geqslant \frac{1}{n}\sum_{t=1}^n \psi_\omega\left(y_t - \widehat{\boldsymbol{\nu}}_t^\top \widetilde{\boldsymbol{\phi}}\right)\widehat{\boldsymbol{\nu}}_t^\top \boldsymbol{\delta}_\phi \geqslant -\frac{\lambda}{2}\|\boldsymbol{\delta}_\phi\|_1.$$

Observe that $\|\boldsymbol{\delta}_{\phi,\mathcal{S}_\diamond^c}\|_1 = \|\widehat{\boldsymbol{\beta}}_{h,\mathcal{S}_\star^c}\|_1$ and $\|\boldsymbol{\beta}^\star\|_1 = \|\widetilde{\boldsymbol{\phi}}_{\mathcal{S}_\star}\|_1$. Hence

$$\begin{aligned}
0 &\leqslant \lambda\|\widetilde{\boldsymbol{\phi}}_{\mathcal{S}_\star}\|_1 - \lambda\|\widehat{\boldsymbol{\phi}}_{h,\mathcal{S}_\star}\|_1 - \lambda\|\boldsymbol{\delta}_{\phi,\mathcal{S}_\diamond^c}\|_1 + \frac{\lambda}{2}\|\boldsymbol{\delta}_{\phi,\mathcal{S}_\diamond}\|_1 + \frac{\lambda}{2}\|\boldsymbol{\delta}_{\phi,\mathcal{S}_\diamond^c}\|_1 \\
&\leqslant \lambda\|\boldsymbol{\delta}_{\phi,\mathcal{S}_\diamond}\|_1 - \frac{\lambda}{2}\|\boldsymbol{\delta}_{\phi,\mathcal{S}_\diamond^c}\|_1 + \frac{\lambda}{2}\|\boldsymbol{\delta}_{\phi,\mathcal{S}_\diamond}\|_1 \\
&= \frac{3\lambda}{2}\|\boldsymbol{\delta}_{\phi,\mathcal{S}_\diamond}\|_1 - \frac{\lambda}{2}\|\boldsymbol{\delta}_{\phi,\mathcal{S}_\diamond^c}\|_1.
\end{aligned}$$

Therefore $\boldsymbol{\delta}_\phi \in \mathcal{C}(\mathcal{S}_\diamond, 3)$ and it follows from (D.3) that

$$\boldsymbol{\delta}_\phi^\top \widehat{\boldsymbol{\Gamma}}_n \boldsymbol{\delta}_\phi \geqslant \varrho_0\|\boldsymbol{\delta}_\phi\|_2^2. \tag{D.5}$$

Since $\max_{t\in[n]}\|\widehat{\boldsymbol{\nu}}_t\|_\infty \leqslant M_\nu$ and $\max_{t\in[n]}|\boldsymbol{e}_t^\top(\boldsymbol{I}_n - \widehat{\boldsymbol{P}})\boldsymbol{F}\boldsymbol{\varphi}^\star| \leqslant \omega/3$, for each $t \in [n]$, we have

$$\begin{aligned}
&\mathbb{I}\left\{|\varepsilon_t + \boldsymbol{e}_t^\top(\boldsymbol{I}_n - \widehat{\boldsymbol{P}})\boldsymbol{F}\boldsymbol{\varphi}^\star + s\widehat{\boldsymbol{\nu}}_t^\top\boldsymbol{\delta}_\phi| \leqslant \omega\right\} \\
&\geqslant \mathbb{I}\left\{|\varepsilon_t| + \max_{t\in[n]}|\boldsymbol{e}_t^\top(\boldsymbol{I}_n - \widehat{\boldsymbol{P}})\boldsymbol{F}\boldsymbol{\varphi}^\star| + |s|\max_{t\in[n]}\|\widehat{\boldsymbol{\nu}}_t\|_\infty\|\boldsymbol{\delta}_\phi\|_1 \leqslant \omega\right\} \\
&\geqslant \mathbb{I}\left\{|\varepsilon_t| \leqslant \frac{\omega}{3}\right\}\mathbb{I}\left\{|s| \leqslant \frac{\omega}{3M_\nu\|\boldsymbol{\delta}_\phi\|_1}\right\}.
\end{aligned}$$

Combined with (D.4) and (D.5), we obtain

$$\begin{aligned}
\frac{3\lambda}{2}\|\boldsymbol{\delta}_{\phi,\mathcal{S}_\diamond}\|_1 &\geqslant \frac{1}{n}\sum_{t=1}^{n}(\widehat{\boldsymbol{\nu}}_t^\top\boldsymbol{\delta}_\phi)^2\int_0^1(1-s)\mathbb{I}\left\{|\varepsilon_t+\boldsymbol{e}_t^\top(\boldsymbol{I}_n-\widehat{\boldsymbol{P}})\boldsymbol{F}\boldsymbol{\varphi}^\star+s\widehat{\boldsymbol{\nu}}_t^\top\boldsymbol{\delta}_\phi|\leqslant\omega\right\}ds\\
&\geqslant \frac{1}{n}\sum_{t=1}^{n}\mathbb{I}\left\{|\varepsilon_t|\leqslant\frac{\omega}{3}\right\}(\widehat{\boldsymbol{\nu}}_t^\top\boldsymbol{\delta}_{\boldsymbol{\phi}})^2\int_0^{1\wedge\frac{\omega}{3M_\nu\|\boldsymbol{\delta}_\phi\|_1}}(1-s)ds\\
&\geqslant \frac{1}{2}\left(1\wedge\frac{\omega}{3M_\nu\|\boldsymbol{\delta}_\phi\|_1}\right)\frac{1}{n}\sum_{t=1}^{n}\mathbb{I}\left\{|\varepsilon_t|\leqslant\frac{\omega}{3}\right\}(\widehat{\boldsymbol{\nu}}_t^\top\boldsymbol{\delta}_\phi)^2\\
&\geqslant \frac{\varrho_0}{2}\left(1\wedge\frac{\omega}{3M_\nu\|\boldsymbol{\delta}_\phi\|_1}\right)\|\boldsymbol{\delta}_\phi\|_2^2\\
&\geqslant \frac{\varrho_0}{2}\left(1\wedge\frac{\omega}{12M_\nu\|\boldsymbol{\delta}_{\phi,\mathcal{S}_\diamond}\|_1}\right)\frac{\|\boldsymbol{\delta}_{\phi,\mathcal{S}_\diamond}\|_1^2}{|\mathcal{S}_\diamond|}.
\end{aligned}$$

If $12M_\nu\|\boldsymbol{\delta}_{\phi,\mathcal{S}_\diamond}\|_1\leqslant\omega$, then $\|\boldsymbol{\delta}_{\phi,\mathcal{S}_\diamond}\|_1\leqslant 3\lambda|\mathcal{S}_\diamond|/\varrho_0$. Consequently, since $\omega\varrho_0>36\lambda M_\nu|\mathcal{S}_\diamond|$, it follows that

$$\|\boldsymbol{\delta}_{\phi,\mathcal{S}_\diamond}\|_1\leqslant\min\left\{\frac{\omega}{12M_\nu},\frac{3\lambda|\mathcal{S}_\diamond|}{\varrho_0}\right\}=\frac{3\lambda|\mathcal{S}_\diamond|}{\varrho_0}.$$

Otherwise, if $12M_\nu\|\boldsymbol{\delta}_{\phi,\mathcal{S}_\diamond}\|_1>\omega$, then $36\lambda M_\nu|\mathcal{S}_\diamond|\geqslant\omega\varrho_0$, which contradicts to the assumption that $\varrho_0>36\lambda M_\nu|\mathcal{S}_\diamond|/\omega$. □

E Proof of Results in Section 3

E.1 Example with node-wise regression

Example E.1. In this example, we use node-wise regression in van de Geer et al. [2014] to estimate $\widehat{\boldsymbol{\Theta}}$. More specifically, for each $j\in[d]$, we first get

$$\widehat{\boldsymbol{\omega}}_j=\underset{\boldsymbol{\omega}\in\mathbb{R}^{d-1}}{\arg\min}\left\{\frac{1}{2n}\sum_{t=1}^{n}|\widehat{u}_{tj}-\boldsymbol{\omega}^\top\widehat{\boldsymbol{u}}_{t,-j}|^2+\lambda_j\|\boldsymbol{\omega}\|_1\right\}$$

with components of $\widehat{\boldsymbol{\omega}}_j=\{\widehat{\omega}_{j\ell};\ell=1,2,\ldots,d,\ell\neq j\}$, where $\lambda_j>0$ is a tuning parameter. We next obtain

$$\widehat{\nu}_j^2=\frac{1}{n}\sum_{t=1}^{n}|\widehat{u}_{tj}-\widehat{\boldsymbol{\omega}}_j^\top\widehat{\boldsymbol{u}}_{t,-j}|^2+\lambda_j\|\widehat{\boldsymbol{\omega}}_j\|_1.$$

Then the estimator for $\boldsymbol{\Theta}$ is given by $\widehat{\boldsymbol{\Theta}}=(\widehat{\Theta}_{j\ell})\in\mathbb{R}^{d\times d}$, where $\widehat{\Theta}_{jj}=1/\widehat{\nu}_j^2$ and $\widehat{\Theta}_{j\ell}=-\widehat{\omega}_{j\ell}/\widehat{\nu}_j^2$ for $j\neq\ell$.

We shall first impose a sparsity assumption on the columns of the precision matrix $\boldsymbol{\Theta}$. To be more specific, for each $k \in [K]$, we let $\mathcal{S}_j = \{\ell \neq j : \Theta_{j\ell} \neq 0\}$ denote the support of the j-th column of $\boldsymbol{\Theta}$. We then define the population parameter

$$\boldsymbol{\omega}_j^\star = \operatorname*{arg\,min}_{\boldsymbol{\omega} \in \mathbb{R}^{d-1}} \mathbb{E}|u_j - \boldsymbol{\omega}^\top \boldsymbol{u}_{-j}|^2 \text{ and } \nu_j^2 = \mathbb{E}|u_j - \boldsymbol{\omega}_j^{\star\top} \boldsymbol{u}_{-j}|^2.$$

In addition, for each $j \in [d]$, let $\widetilde{\boldsymbol{\omega}}_j^\star$ denote the d-dimensional vector with components $\widetilde{\omega}_{jj}^\star = 1$ and $\widetilde{\omega}_{j\ell}^\star = -\omega_{j\ell}^\star$ for $\ell \neq j$.

The following proposition provides theoretical guarantees for the estimator $\widehat{\boldsymbol{\Theta}}$ constructed above.

Proposition E.1. *Let Assumptions 2.1–2.4 hold. Assume that*

$$\max_{j\in[d]} |\mathcal{S}_j| \left(\frac{\log d}{n} + \frac{1}{d}\right) \to 0 \text{ and } \max_{j\in[d]} \frac{\mathcal{V}_{n,d}}{n} \sqrt{|\mathcal{S}_j|} \|\boldsymbol{B}^\top \widetilde{\boldsymbol{\omega}}_j^\star\|_2 \to 0. \tag{E.1}$$

Then, with suitably chosen $\lambda_j \asymp n^{-1} \|\widehat{\boldsymbol{U}}_{-j}^\top \widehat{\boldsymbol{U}} \widetilde{\boldsymbol{\omega}}_j^\star\|_\infty$ uniformly for $j \in [d]$, we have

$$\begin{aligned}
\|\boldsymbol{I}_d - \widehat{\boldsymbol{\Theta}}\widetilde{\boldsymbol{\Sigma}}\|_{\max} &= O_{\mathbb{P}}\left(\sqrt{\frac{\log d}{n}} + \frac{1}{\sqrt{d}} + \max_{j\in[d]} \frac{\mathcal{V}_{n,d}}{n} \|\boldsymbol{B}^\top \widetilde{\boldsymbol{\omega}}_j^\star\|_2\right), \\
\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{\max} &= O_{\mathbb{P}}\left(\max_{j\in[d]} \sqrt{|\mathcal{S}_j| \left(\frac{\log d}{n} + \frac{1}{d}\right)} + \max_{j\in[d]} \frac{\mathcal{V}_{n,d}}{n} \sqrt{|\mathcal{S}_j|} \|\boldsymbol{B}^\top \widetilde{\boldsymbol{\omega}}_j^\star\|_2\right), \\
\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{\infty} &= O_{\mathbb{P}}\left(\max_{j\in[d]} |\mathcal{S}_j| \sqrt{\frac{\log d}{n} + \frac{1}{d}} + \max_{j\in[d]} \frac{\mathcal{V}_{n,d}}{n} |\mathcal{S}_j| \|\boldsymbol{B}^\top \widetilde{\boldsymbol{\omega}}_j^\star\|_2\right).
\end{aligned} \tag{E.2}$$

E.2 Proof of Proposition E.1

Lemma E.2. *Under Assumptions 2.1–2.4, uniformly for $j \in [d]$, we have*

$$\|\widehat{\boldsymbol{U}}_{-j}^\top \widehat{\boldsymbol{U}} \widetilde{\boldsymbol{\omega}}_j^\star\|_\infty = O_{\mathbb{P}}\left(\sqrt{n \log d} + \frac{n}{\sqrt{d}} + \mathcal{V}_{n,d} \|\boldsymbol{B}^\top \widetilde{\boldsymbol{\omega}}_j^\star\|_2\right).$$

Proof of Lemma E.2. By the triangle inequality, for each $j \in [d]$,

$$\|\widehat{\boldsymbol{U}}_{-j}^\top \widehat{\boldsymbol{U}} \widetilde{\boldsymbol{\omega}}_j^\star\|_\infty \leqslant \|\boldsymbol{U}_{-j}^\top \boldsymbol{\chi}_j\|_\infty + \|\widehat{\boldsymbol{U}}_{-j}^\top (\widehat{\boldsymbol{U}} - \boldsymbol{U}) \widetilde{\boldsymbol{\omega}}_j^\star\|_\infty + \|(\widehat{\boldsymbol{U}} - \boldsymbol{U})^\top \boldsymbol{\chi}_j\|_\infty$$

where $\boldsymbol{\chi}_j = \boldsymbol{U} \widetilde{\boldsymbol{\omega}}_j^\star$. By the definition of $\widetilde{\boldsymbol{\omega}}_j^\star$ and Assumption 2.4, it follows that

$$\|\widetilde{\boldsymbol{\omega}}_j^\star\|_2^2 \leqslant \frac{\widetilde{\boldsymbol{\omega}}_j^{\star\top} \boldsymbol{\Sigma} \widetilde{\boldsymbol{\omega}}_j^\star}{\lambda_{\min}(\boldsymbol{\Sigma})} = \frac{1}{\Theta_{jj} \lambda_{\min}(\boldsymbol{\Sigma})} \leqslant \frac{\lambda_{\max}(\boldsymbol{\Sigma})}{\lambda_{\min}(\boldsymbol{\Sigma})} \leqslant \frac{1}{\kappa^2}.$$

Hence $\{\boldsymbol{u}_t^\top \widetilde{\boldsymbol{\omega}}_j^\star\}_{t=1}^n$ are i.i.d. zero-mean sub-Gaussian random variables with $\|\boldsymbol{u}_t^\top \widetilde{\boldsymbol{\omega}}_j^\star\|_{\psi_2} \leqslant c_0/\kappa$ and for any $\ell \neq j$, $\{u_{t\ell}\boldsymbol{u}_t^\top \widetilde{\boldsymbol{\omega}}_j^\star\}_{t=1}^n$ are i.i.d. zero-mean sub-exponential random variables with $\max_{\ell\neq j}\|u_{t\ell}\boldsymbol{u}_t^\top \widetilde{\boldsymbol{\omega}}_j^\star\|_{\psi_1} \leqslant c_0^2/\kappa$. Consequently, by Lemmas 2.1 and C.3, we obtain $\|\widehat{\boldsymbol{U}}_{-j}^\top \boldsymbol{\chi}_j\|_\infty^2 = O_{\mathbb{P}}(n\log d)$, $\|\widehat{\boldsymbol{U}}_{-j}^\top(\widehat{\boldsymbol{U}} - \boldsymbol{U})\widetilde{\boldsymbol{\omega}}_j^\star\|_\infty \leqslant \|\widehat{\boldsymbol{U}}^\top \boldsymbol{F}\boldsymbol{B}^\top \widetilde{\boldsymbol{\omega}}_j^\star\|_\infty = O_{\mathbb{P}}(\mathcal{V}_{n,d}\|\boldsymbol{B}^\top \widetilde{\boldsymbol{\omega}}_j^\star\|_2)$ and

$$\|(\widehat{\boldsymbol{U}} - \boldsymbol{U})^\top \boldsymbol{\chi}_j\|_\infty^2 \leqslant \|\boldsymbol{\chi}_j\|_2^2 \max_{j\in[d]} \sum_{t=1}^n |\widehat{u}_{tj} - u_{tj}|^2 = O_{\mathbb{P}}\left(n\log d + \frac{n^2}{d}\right).$$

□

Proof of Proposition E.1. Following the proof of Lemma C.2, under (E.1) and Assumptions 2.1–2.4, there exists a positive constant $\widetilde{\kappa}$ such that

$$\mathbb{P}\left(\min_{j\in[d]} \inf_{\boldsymbol{h}\in\mathbb{R}^{d-1}:\|\boldsymbol{h}_{\mathcal{S}_j^c}\|_1 \leqslant 3\|\boldsymbol{h}_{\mathcal{S}_j}\|_1} \frac{\boldsymbol{h}^\top \widehat{\boldsymbol{U}}_{-j}^\top \widehat{\boldsymbol{U}}_{-j}\boldsymbol{h}}{n\|\boldsymbol{h}\|_2^2} \geqslant \widetilde{\kappa}\right) \to 1, \quad \text{as } n \to \infty.$$

Then, by Lemma C.1, we have $\|\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^\star\|_1 = O_{\mathbb{P}}(\lambda_j|\mathcal{S}_j|)$ and $\|\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^\star\|_2 = O_{\mathbb{P}}(\lambda_j\sqrt{|\mathcal{S}_j|})$ uniformly for all $j \in [d]$. Similar to the proof of Theorem 2.4 in van de Geer et al. [2014], we obtain $|\widehat{\nu}_j^2 - \nu_j^2| \xrightarrow{\mathbb{P}} 0$ uniformly for $j \in [d]$. Putting all these pieces together, we obtain (E.2). □

E.3 Proof of Theorem 3.1

Proof of Theorem 3.1. We first derive a Gaussian approximation result for $n^{-1/2}\boldsymbol{\Theta}\boldsymbol{U}^\top\mathcal{E}$. To this end, we shall apply Theorem 2.1 in Chernozhukov et al. [2017] and verify the conditions therein. By Assumption 2.4, we have $\lambda_{\max}(\boldsymbol{\Sigma}) \leqslant \|\boldsymbol{\Sigma}\|_1 \leqslant 1/\kappa$. Hence $\min_{j\in[d]}\Theta_{jj} \geqslant \lambda_{\min}(\boldsymbol{\Theta}) \geqslant \kappa > 0$. Combined with Assumption 2.1, for each $j \in [d]$, $\{\boldsymbol{\Theta}_j^\top \boldsymbol{u}_t\varepsilon_t\}_{t=1}^n$ are i.i.d. zero-mean sub-exponential random variables with $\mathrm{Cov}(\boldsymbol{\Theta}_j^\top \boldsymbol{u}_t\varepsilon_t) = \sigma^2\Theta_{jj} \geqslant \sigma^2\kappa$ and

$$\max_{j\in[d]} \|\boldsymbol{\Theta}_j^\top \boldsymbol{u}_t\varepsilon_t\|_{\psi_1} \leqslant c_1 \max_{j\in[d]} \|\boldsymbol{\Theta}_j^\top \boldsymbol{u}_t\|_{\psi_2} \leqslant \frac{c_0c_1}{\kappa}.$$

Then, by Theorem 2.2 given in Chernozhukov et al. [2020],

$$\rho^\natural := \sup_{x>0} \left|\mathbb{P}\left(n^{-1/2}\|\boldsymbol{\Theta}\boldsymbol{U}^\top\mathcal{E}\|_\infty \leqslant x\right) - \mathbb{P}(\|\boldsymbol{Z}\|_\infty \leqslant x)\right| \to 0.$$

Next, we will quantify the difference between our test statistics and $n^{-1}\boldsymbol{\Theta}\boldsymbol{U}^\top\mathcal{E}$. We first introduce Lemma E.3 below.

Lemma E.3. *Assume that $\mathcal{V}_{n,d}\|\boldsymbol{\varphi}^\star\|_2 \lesssim \sqrt{n\log d}$. Then, under Assumption 3.1 and conditions of Theorem 2.2, we have*

$$\|n(\widetilde{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star) - \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{U}}^\top\mathcal{E}\|_\infty = O_{\mathbb{P}}\left(\Lambda_{\max}|\mathcal{S}_\star|\sqrt{n\log d} + \mathcal{V}_{n,d}\|\boldsymbol{\Theta}\|_\infty\|\boldsymbol{\varphi}^\star\|_2\right),$$

$$\|\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{U}}^\top\mathcal{E} - \boldsymbol{\Theta}\boldsymbol{U}^\top\mathcal{E}\|_\infty = O_{\mathbb{P}}\left(\|\boldsymbol{\Theta}\|_\infty\sqrt{\log d + \frac{n}{d}} + \Delta_\infty\sqrt{n\log d}\right).$$

By Lemma E.3, we have $\sqrt{n}\|\widetilde{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star - n^{-1}\boldsymbol{\Theta}\boldsymbol{U}^\top\mathcal{E}\|_\infty = O_{\mathbb{P}}(\Delta_{n,d})$, where

$$\Delta_{n,d} = \Lambda_{\max}|\mathcal{S}_\star|\sqrt{\log d} + \frac{\mathcal{V}_{n,d}\|\boldsymbol{\Theta}\|_\infty\|\boldsymbol{\varphi}^\star\|_2}{\sqrt{n}} + \|\boldsymbol{\Theta}\|_\infty\sqrt{\frac{\log d}{n} + \frac{1}{d}} + \Delta_\infty\sqrt{\log d}.$$

It follows from (3.5) that $\Delta_{n,d}\sqrt{\log d} \to 0$. Taking $\widetilde{\Delta}_{n,d} = \sqrt{\Delta_{n,d}}/(\log d)^{1/4}$, then we have

$$\widetilde{\Delta}_{n,d}\sqrt{\log d} \to 0 \ \text{ and } \ \mathbb{P}\left(\sqrt{n}\|\widetilde{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star - n^{-1}\boldsymbol{\Theta}\boldsymbol{U}^\top\mathcal{E}\|_\infty > \widetilde{\Delta}_{n,d}\right) \to 0.$$

Consequently, by Lemma in Chernozhukov et al. [2017] and Lemma E.3, we obtain

$$\begin{aligned}
&\sup_{x>0}\left|\mathbb{P}\left(\sqrt{n}\|\widetilde{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|_\infty \leqslant x\right) - \mathbb{P}(n^{-1/2}\|\boldsymbol{\Theta}\boldsymbol{U}^\top\mathcal{E}\|_\infty \leqslant x)\right| \\
&\leqslant \mathbb{P}\left(\sqrt{n}\|\widetilde{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star - n^{-1}\boldsymbol{\Theta}\boldsymbol{U}^\top\mathcal{E}\|_\infty > \widetilde{\Delta}_{n,d}\right) + 2\rho^\natural + \sup_{x>0}\mathbb{P}(x < \|\boldsymbol{Z}\|_\infty \leqslant x + \widetilde{\Delta}_{n,d}) \\
&\leqslant \mathbb{P}\left(\sqrt{n}\|\widetilde{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star - n^{-1}\boldsymbol{\Theta}\boldsymbol{U}^\top\mathcal{E}\|_\infty > \widetilde{\Delta}_{n,d}\right) + 2\rho^\natural + C\widetilde{\Delta}_{n,d}\sqrt{\log d} \to 0.
\end{aligned}$$

□

E.4 Proof of Theorem 3.2

Proof of Theorem 3.2. By Theorem 3.1, it suffices to prove that

$$\rho^\star := \sup_{x>0}\left|\mathbb{P}(\|\boldsymbol{Z}\|_\infty \leqslant x) - \mathbb{P}^\star\left(\widehat{L} \leqslant x\right)\right| \xrightarrow{\mathbb{P}} 0.$$

Note that $n^{-1/2}\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{U}}^\top\boldsymbol{\xi}$ is a zero-mean Gaussian random vector with covariance matrix

$$\mathrm{Cov}^\star\left(\frac{1}{\sqrt{n}}\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{U}}^\top\boldsymbol{\xi}\right) = \widehat{\sigma}^2\widehat{\boldsymbol{\Theta}}\widetilde{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Theta}}^\top.$$

By (3.7) and Assumptions 3.1–3.2, we have

$$\begin{aligned}
\|\widehat{\sigma}^2\widehat{\boldsymbol{\Theta}}\widetilde{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Theta}}^\top - \sigma^2\boldsymbol{\Theta}\|_{\max} &\leqslant \widehat{\sigma}^2\|\widehat{\boldsymbol{\Theta}}\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{I}_d\|_{\max}\|\widehat{\boldsymbol{\Theta}}\|_\infty + \widehat{\sigma}^2\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{\max} + |\widehat{\sigma}^2 - \sigma^2|\|\boldsymbol{\Theta}\|_{\max} \\
&= O_{\mathbb{P}}(\Lambda_{\max}\|\boldsymbol{\Theta}\|_\infty + \Delta_{\max} + \Delta_\sigma) = O_{\mathbb{P}}\left(\frac{1}{\log d}\right).
\end{aligned}$$

Then it follows from Lemma 2.1 in Chernozhukov et al. [2020] that $\rho^\star \xrightarrow{\mathbb{P}} 0$. □

E.5 Proof of Lemma E.3

Proof of Lemma E.3. By (3.3) and the triangle inequality,

$$\|n(\widetilde{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star) - \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{U}}^\top \mathcal{E}\|_\infty \leqslant \|\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{U}}^\top \boldsymbol{F}\boldsymbol{\varphi}^\star\|_\infty + n\|(\boldsymbol{I}_d - \widehat{\boldsymbol{\Theta}}\widetilde{\boldsymbol{\Sigma}})(\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star)\|_\infty.$$

Since $\mathcal{V}_{n,d}\|\boldsymbol{\varphi}^\star\|_2 \lesssim \sqrt{n\log d}$, it follows from Theorem 2.2 that $\|\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star\|_1 = O_{\mathbb{P}}(|\mathcal{S}_\star|\sqrt{(\log d)/n})$. Combining this with Assumption 3.1, we obtain

$$\|(\boldsymbol{I} - \widehat{\boldsymbol{\Theta}}\widetilde{\boldsymbol{\Sigma}})(\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star)\|_\infty \leqslant \|\boldsymbol{I}_d - \widehat{\boldsymbol{\Theta}}\widetilde{\boldsymbol{\Sigma}}\|_{\max}\|\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star\|_1 = O_{\mathbb{P}}\left(\Lambda_{\max}|\mathcal{S}_\star|\sqrt{\frac{\log d}{n}}\right).$$

By Assumption 3.1, we have $\|\widehat{\boldsymbol{\Theta}}\|_\infty \leqslant \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_\infty + \|\boldsymbol{\Theta}\|_\infty = O_{\mathbb{P}}(\|\boldsymbol{\Theta}\|_\infty)$. Then, it follows from Lemma C.3 that

$$\|\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{U}}^\top \boldsymbol{F}\boldsymbol{\varphi}^\star\|_\infty \leqslant \|\widehat{\boldsymbol{\Theta}}\|_\infty\|\widehat{\boldsymbol{U}}^\top \boldsymbol{F}\boldsymbol{\varphi}^\star\|_\infty = O_{\mathbb{P}}(\mathcal{V}_{n,d}\|\boldsymbol{\Theta}\|_\infty\|\boldsymbol{\varphi}^\star\|_2).$$

By Lemma C.4 and Assumption 2.1, we have $\|\boldsymbol{U}^\top\mathcal{E}\|_\infty^2 = O_{\mathbb{P}}(n\log d)$ and

$$\begin{aligned}\|\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{U}}^\top\mathcal{E} - \boldsymbol{\Theta}\boldsymbol{U}^\top\mathcal{E}\|_\infty &\leqslant \|\widehat{\boldsymbol{\Theta}}\|_\infty\|(\widehat{\boldsymbol{U}} - \boldsymbol{U})^\top\mathcal{E}\|_\infty + \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_\infty\|\boldsymbol{U}^\top\mathcal{E}\|_\infty \\ &= O_{\mathbb{P}}\left(\|\boldsymbol{\Theta}\|_\infty\sqrt{\log d + \frac{n}{d}} + \Delta_\infty\sqrt{n\log d}\right).\end{aligned}$$

□

F Proof of Results in Section 4

F.1 Proof of Proposition 4.1

Proof. For each $\ell \in [d]$, let $\beta^\star_{\ell,M} \in \mathbb{R}$ denote the corresponding population version of the marginal least square estimator $\widehat{\beta}_{\ell,M}$, that is,

$$\beta^\star_{\ell,M} = \arg\min_{\beta\in\mathbb{R}} \mathbb{E}(Y - u_\ell\beta)^2 = \frac{\boldsymbol{\Sigma}_\ell^\top\boldsymbol{\beta}^\star}{\Sigma_{\ell\ell}}.$$

For each $\ell \in [d]$, we have

$$\begin{aligned}\widehat{\beta}_{\ell,M} - \beta^\star_{\ell,M} &= \frac{1}{\widehat{\boldsymbol{U}}_\ell^\top\widehat{\boldsymbol{U}}_\ell}\left\{\widehat{\boldsymbol{U}}_\ell^\top(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{U}}\boldsymbol{\beta}^\star) + \left(\widehat{\boldsymbol{U}}_\ell^\top\widehat{\boldsymbol{U}} - m\boldsymbol{e}_\ell^\top\boldsymbol{\Sigma}\right)\boldsymbol{\beta}^\star + \left(\widehat{\boldsymbol{U}}_\ell^\top\widehat{\boldsymbol{U}}_\ell - m\Sigma_{\ell\ell}\right)\beta^\star_{\ell,M}\right\} \\ &=: \delta_{\ell,1} + \delta_{\ell,2} + \delta_{\ell,3}.\end{aligned}$$

Let $\boldsymbol{U}_\ell = (u_{1\ell}, \ldots, u_{m\ell})^\top \in \mathbb{R}^m$ denote the ℓ-th column of the design matrix $\boldsymbol{U}$. By Assumption 2.1, $u_{1\ell}, \ldots, u_{m\ell}$ are i.i.d. sub-Gaussian random variables with $\|u_{1\ell}\|_{\psi_2} \leqslant c_0$. Hence, for any $x > 0$, by Bernstein's inequality,

$$\mathbb{P}\left(\max_{\ell\in[d]} |\boldsymbol{U}_\ell^\top \boldsymbol{U}_\ell - m\Sigma_{\ell\ell}| > x\right) \leqslant 2d \exp\left\{-c \min\left(\frac{x^2}{mc_0^4}, \frac{x}{c_0^2}\right)\right\}.$$

Here $c > 0$ is an absolute constant. Combined with Lemma 2.1, we obtain that

$$\max_{\ell\in[d]} |\widehat{\boldsymbol{U}}_\ell^\top \widehat{\boldsymbol{U}}_\ell - m\Sigma_{\ell\ell}| \leqslant \max_{\ell\in[d]} |\widehat{\boldsymbol{U}}_\ell^\top \widehat{\boldsymbol{U}}_\ell - \boldsymbol{U}_\ell^\top \boldsymbol{U}_\ell| + \max_{\ell\in[d]} |\boldsymbol{U}_\ell^\top \boldsymbol{U}_\ell - m\Sigma_{\ell\ell}| = O_{\mathbb{P}}\left(\sqrt{m \log d}\right).$$

Since $\log d = o(m)$, we have

$$\mathbb{P}(\mathcal{E}_U) := \mathbb{P}\left(\min_{\ell\in[d]} \widehat{\boldsymbol{U}}_\ell^\top \widehat{\boldsymbol{U}}_\ell \geqslant \frac{m}{2} \min_{\ell\in[d]} \Sigma_{\ell\ell}\right) \to 1.$$

Under $\mathcal{E}_U$, it follows from Lemma C.5 that

$$\max_{\ell\in[d]} |\delta_{\ell,1}| \leqslant \frac{\|\widehat{\boldsymbol{U}}^\top(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{U}}\boldsymbol{\beta}^\star)\|_\infty}{\min_{\ell\in[d]} \widehat{\boldsymbol{U}}_\ell^\top \widehat{\boldsymbol{U}}_\ell} \leqslant \frac{2\|\widehat{\boldsymbol{U}}^\top(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{U}}\boldsymbol{\beta}^\star)\|_\infty}{m \min_{\ell\in[d]} \Sigma_{\ell\ell}} = O_{\mathbb{P}}\left(\frac{\mathcal{V}_{m,d}}{m}\|\boldsymbol{\varphi}^\star\|_2 + \sqrt{\frac{\log d}{m}}\right).$$

Similarly, by Lemma C.6, we have

$$\begin{aligned}\max_{\ell\in[d]} |\delta_{\ell,2}| &\leqslant \frac{2\|\widehat{\boldsymbol{U}}^\top \widehat{\boldsymbol{U}} - \boldsymbol{U}^\top \boldsymbol{U}\|_{\max}\|\boldsymbol{\beta}^\star\|_1 + 2\|\boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{\beta}^\star - m\boldsymbol{\Sigma}\boldsymbol{\beta}^\star\|_\infty}{m \min_{\ell\in[d]} \Sigma_{\ell\ell}} \\ &= O_{\mathbb{P}}\left(\|\boldsymbol{\beta}^\star\|_1 \frac{\log d}{m} + \|\boldsymbol{\beta}^\star\|_2 \sqrt{\frac{\log d}{m}}\right),\end{aligned}$$

and

$$\max_{\ell\in[d]} |\delta_{\ell,3}| \leqslant \frac{\max_{\ell\in[d]} |\widehat{\boldsymbol{U}}_\ell^\top \widehat{\boldsymbol{U}}_\ell - m\Sigma_{\ell\ell}||\beta^\star_{\ell,M}|}{m \min_{\ell\in[d]} \Sigma_{\ell\ell}} = O_{\mathbb{P}}\left(\|\boldsymbol{\beta}^\star\|_\infty \sqrt{\frac{\log d}{m}}\right).$$

Putting all these pieces together, it follows from (4.3) that $\max_{\ell\in[d]} |\widehat{\beta}_{\ell,M} - \beta^\star_{\ell,M}| = o(\phi)$, which implies that

$$\mathbb{P}\left(\mathcal{E}_\phi\right) := \mathbb{P}\left(\max_{\ell\in\mathcal{S}_\star} |\widehat{\beta}_{\ell,M} - \beta^\star_{\ell,M}| \leqslant \bar{c}\phi\right) \to 1, \quad \text{as} \quad m \to \infty.$$

Under $\mathcal{E}_\phi$, by (4.3),

$$\min_{\ell\in\mathcal{S}_\star} |\widehat{\beta}_{\ell,M}| \geqslant \min_{\ell\in\mathcal{S}_\star} |\beta^\star_{\ell,M}| - \max_{\ell\in\mathcal{S}_\star} |\widehat{\beta}_{\ell,M} - \beta^\star_{\ell,M}| \geqslant (1 + \bar{c})\phi - \bar{c}\phi = \phi.$$

Consequently, we obtain

$$\mathbb{P}\left(\mathcal{S}_\star \subset \widehat{\mathcal{S}}_\phi\right) \geqslant \mathbb{P}\left(\min_{\ell\in\mathcal{S}_\star} |\widehat{\beta}_{\ell,M}| > \phi\right) \geqslant \mathbb{P}(\mathcal{E}_\phi) \to 1.$$

Under $\mathcal{E}_\phi$, elementary calculations show that

$$\begin{aligned}
|\widehat{\mathcal{S}}_\phi| &= \sum_{\ell=1}^{d} \mathbb{I}\left\{|\widehat{\beta}_{\ell,M}| > \phi\right\} \leqslant \sum_{\ell=1}^{d} \mathbb{I}\left\{\max_{\ell\in[d]} |\widehat{\beta}_{\ell,M} - \beta^\star_{\ell,M}| + |\beta^\star_{\ell,M}| > \phi\right\} \\
&\leqslant \sum_{\ell=1}^{d} \mathbb{I}\left\{|\beta^\star_{\ell,M}| > (1-\bar{c})\phi\right\} \leqslant \sum_{\ell=1}^{d} \frac{|\beta^\star_{\ell,M}|^2}{(1-\bar{c})^2\phi^2} \\
&\leqslant \frac{\|\boldsymbol{\Sigma}\boldsymbol{\beta}^\star\|_2^2}{\lambda^2_{\min}(\boldsymbol{\Sigma})(1-\bar{c})^2\phi^2} = \frac{c_\diamond^2 m^{2\kappa}\|\boldsymbol{\Sigma}\boldsymbol{\beta}^\star\|_2^2}{\lambda^2_{\min}(\boldsymbol{\Sigma})(1-\bar{c})^2}.
\end{aligned}$$

□

F.2 Proof of Theorem 4.2

Proof. Before proceeding to the proof of the theorem, we first introduce the following Lemma.

Lemma F.1. *Let $\widetilde{\mathcal{S}}$ be an estimator for $\mathcal{S}_\star$ such that $\mathbb{P}(|\widetilde{\mathcal{S}}| \leqslant s_n) \to 1$ and $\mathbb{P}(\mathcal{S}_\star \subset \widetilde{\mathcal{S}}) \to 1$. Moreover, we assume that $\widetilde{\mathcal{S}}$ is independent of $(\boldsymbol{X}, \boldsymbol{Y})$ and*

$$s_n\left(\frac{\log d}{n} + \frac{1}{d}\right) \to 0. \tag{F.1}$$

Then, under Assumptions 2.1–2.5, we have that as $n \to \infty$,

$$\widetilde{\rho}_K = \sup_{x>0}\left|\mathbb{P}\left(\frac{\mathcal{E}^\top \boldsymbol{P}_{\widetilde{\mathcal{S}}}\mathcal{E}}{\sigma^2} \leqslant x\right) - \mathbb{P}(\chi^2_K \leqslant x)\right| \to 0, \quad \text{where } \boldsymbol{P}_{\widetilde{\mathcal{S}}} = \boldsymbol{P}_{\widehat{\boldsymbol{F}}} + \boldsymbol{P}_{\widehat{\boldsymbol{U}}_{\widetilde{\mathcal{S}}}} - \boldsymbol{P}_{\boldsymbol{X}_{\widetilde{\mathcal{S}}}}.$$

By Lemma F.1, we have

$$\rho_K = \sup_{x>0}\left|\mathbb{P}\left(\frac{Q_n^{(2)}}{\sigma^2} \leqslant x\right) - \mathbb{P}(\chi^2_K \leqslant x)\right| \to 0 \tag{F.2}$$

Hence, it suffices to prove that

$$\rho_\sigma = \sup_{x>0}\left|\mathbb{P}\left(\frac{Q_n^{(2)}}{\widehat{\sigma}^2} \leqslant x\right) - \mathbb{P}\left(\frac{Q_n^{(2)}}{\sigma^2} \leqslant x\right)\right| \to 0.$$

After conducting some elementary calculations, we obtain

$$\rho_\sigma \leqslant \mathbb{P}\left(\frac{Q_n^{(2)}}{\sigma^2}\left|\frac{\sigma^2}{\widehat{\sigma}^2} - 1\right| > \sqrt{\Delta_\sigma}\right) + \sup_{x>0}\mathbb{P}\left(x < \chi^2_K \leqslant x + \sqrt{\Delta_\sigma}\right) + 2\rho_K =: \delta_1 + \delta_2 + 2\rho_K.$$

As $\Delta_\sigma \to 0$, we have $\delta_2 \to 0$. Moreover, by Assumption 3.3, we have $|\sigma^2/\widehat{\sigma}^2 - 1| \leqslant |\sigma/\widehat{\sigma} - 1| = O_{\mathbb{P}}(\Delta_\sigma)$. Combined with (F.2), we obtain $\delta_1 \to 0$. Thus, we claim our conclusion of Theorem 4.2. □

F.3 Proof of Lemma F.1

Proof. As $\widetilde{\mathcal{S}}$ is independent of $(\boldsymbol{X}, \boldsymbol{Y})$, it follows that

$$\widetilde{\rho}_K \leqslant \mathbb{P}(|\widetilde{\mathcal{S}}| > s_n) + \sup_{\mathcal{S}\subset[d]:|\mathcal{S}|\leqslant s_n} \sup_{x>0} \left| \mathbb{P}\left(\frac{\mathcal{E}^\top \boldsymbol{P}_{\mathcal{S}} \mathcal{E}}{\sigma^2} \leqslant x \right) - \mathbb{P}(\chi^2_K \leqslant x) \right|.$$

Since $\mathbb{P}(|\widetilde{\mathcal{S}}| > s_n) \to 0$ as $n \to \infty$, it suffices to prove that

$$\rho^\diamond := \sup_{\mathcal{S}\subset[d]:|\mathcal{S}|\leqslant s_n} \sup_{x>0} \left| \mathbb{P}\left(\frac{\mathcal{E}^\top \boldsymbol{P}_{\mathcal{S}} \mathcal{E}}{\sigma^2} \leqslant x \right) - \mathbb{P}(\chi^2_K \leqslant x) \right| \to 0.$$

For any set $\mathcal{S} \subset [d]$ such that $|\mathcal{S}| \leqslant s_n$ and $\mathcal{S}_\star \subset \mathcal{S}$, we define

$$\mathcal{A}_{\mathcal{S}} = \left\{ \lambda_{\min}(\widehat{\boldsymbol{U}}_{\mathcal{S}}^\top \widehat{\boldsymbol{U}}_{\mathcal{S}}) \geqslant \frac{n\lambda_0}{2} \right\}.$$

Then by Theorem 1.1 given in Bentkus [2005], it follows that

$$\sup_{x>0} \left| \mathbb{P}\left(\frac{\mathcal{E}^\top \boldsymbol{P}_{\mathcal{S}} \mathcal{E}}{\sigma^2} \leqslant x \right) - \mathbb{P}(\chi^2_K \leqslant x) \right| \leqslant \mathbb{P}(\mathcal{A}_{\mathcal{S}}^c) + C\mathbb{E}|\varepsilon_t|^3 K^{5/4} \mathbb{E}\left(\max_{t\in[n]} \sqrt{\boldsymbol{P}_{\mathcal{S},tt}} \right) \mathbb{I}\{\mathcal{A}_{\mathcal{S}}\}, \tag{F.3}$$

where $C > 0$ is an absolute constant. Now we proceed to prove (F.3). Recall that $\boldsymbol{P}_{\mathcal{S}} = \boldsymbol{P}_{\widehat{\boldsymbol{F}}} + \boldsymbol{P}_{\widehat{\boldsymbol{U}}_{\mathcal{S}}} - \boldsymbol{P}_{\boldsymbol{X}_{\mathcal{S}}}$ is a projection matrix onto the linear space generated by the columns of $\widetilde{\boldsymbol{F}} = (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_{\mathcal{S}}})\widehat{\boldsymbol{F}} \in \mathbb{R}^{n\times K}$. Hence, under $\mathcal{A}_{\mathcal{S}}$, we can write

$$\boldsymbol{P}_{\mathcal{S}} = \widetilde{\boldsymbol{F}}(\widetilde{\boldsymbol{F}}^\top \widetilde{\boldsymbol{F}})^{-1} \widetilde{\boldsymbol{F}}^\top = \widetilde{\boldsymbol{F}}(\widetilde{\boldsymbol{F}}^\top \widetilde{\boldsymbol{F}})^{-1/2} (\widetilde{\boldsymbol{F}}^\top \widetilde{\boldsymbol{F}})^{-1/2} \widetilde{\boldsymbol{F}}^\top =: \boldsymbol{W}\boldsymbol{W}^\top,$$

where $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_n)^\top \in \mathbb{R}^{n\times K}$. Therefore,

$$\mathcal{E}^\top \boldsymbol{P}_{\mathcal{S}} \mathcal{E} = \mathcal{E}^\top \boldsymbol{W}\boldsymbol{W}^\top \mathcal{E} = \left\| \sum_{t=1}^n \boldsymbol{w}_t \varepsilon_t \right\|_2^2,$$

which further implies that for any $x > 0$,

$$\mathbb{P}\left(\mathcal{E}^\top \boldsymbol{P}_{\mathcal{S}} \mathcal{E} \leqslant x \right) = \mathbb{P}\left(\sum_{t=1}^n \boldsymbol{w}_t \varepsilon_t \in \mathcal{B}_K(\sqrt{x}) \right),$$

where $\mathcal{B}_K(r)$ denotes the K-dimensional ball centered at the origin with radius $r > 0$. Observe that

$$\mathrm{Cov}\left(\sum_{t=1}^n \boldsymbol{w}_t \varepsilon_t \right) = \sum_{t=1}^n \boldsymbol{w}_t \boldsymbol{w}_t^\top = \boldsymbol{W}^\top \boldsymbol{W} = (\widetilde{\boldsymbol{F}}^\top \widetilde{\boldsymbol{F}})^{-1/2} \widetilde{\boldsymbol{F}}^\top \widetilde{\boldsymbol{F}} (\widetilde{\boldsymbol{F}}^\top \widetilde{\boldsymbol{F}})^{-1/2} = \boldsymbol{I}_K.$$

To apply Theorem 1.1 in Bentkus [2005], let $\mathcal{Z} \in \mathbb{R}^K$ be a zero-mean Gaussian random vector with covariance matrix $\mathrm{Cov}(\mathcal{Z}) = \boldsymbol{I}_K$, then $\|\mathcal{Z}\|^2 \sim \chi^2_K$ and

$$\begin{aligned}
&\sup_{x>0} \left| \mathbb{P}\left(\mathcal{E}^\top \boldsymbol{P}_{\mathcal{S}} \mathcal{E} \leqslant x | \boldsymbol{X} \right) - \mathbb{P}(\chi^2_K \leqslant x) \right| \\
&= \sup_{r>0} \left| \mathbb{P}\left(\sum_{t=1}^n \boldsymbol{w}_t \varepsilon_t \in \mathcal{B}_K(r) | \boldsymbol{X} \right) - \mathbb{P}(\mathcal{Z} \in \mathcal{B}_K(r)) \right| \\
&\leqslant cK^{1/4} \sum_{t=1}^n \mathbb{E}\|\boldsymbol{w}_t \varepsilon_t\|_2^3 = cK^{1/4} \sum_{t=1}^n \|\boldsymbol{w}_t\|_2^3 \mathbb{E}|\varepsilon_t|^3 \\
&\leqslant cK^{1/4} \mathbb{E}|\varepsilon|^3 \max_{t\in[n]} \|\boldsymbol{w}_t\|_2 \sum_{t=1}^n \|\boldsymbol{w}_t\|_2^2.
\end{aligned}$$

Observe that the diagonal elements of the matrix $\boldsymbol{P}_{\mathcal{S}}$ are $\|\boldsymbol{w}_1\|_2^2, \ldots, \|\boldsymbol{w}_n\|_2^2$. Hence

$$\sum_{t=1}^n \|\boldsymbol{w}_t\|_2^2 = \mathrm{tr}(\boldsymbol{P}_{\mathcal{S}}) = K \quad \text{and} \quad \max_{t\in[n]} \|\boldsymbol{w}_t\|_2 = \max_{t\in[n]} \sqrt{\boldsymbol{P}_{\mathcal{S},tt}}.$$

Consequently, we obtain

$$\sup_{x>0} \left| \mathbb{P}\left(\mathcal{E}^\top \boldsymbol{P}_{\mathcal{S}} \mathcal{E} \leqslant x | \boldsymbol{X} \right) - \mathbb{P}(\chi^2_K \leqslant x) \right| \leqslant cK^{5/4} \mathbb{E}|\varepsilon|^3 \max_{t\in[n]} \sqrt{\boldsymbol{P}_{\mathcal{S},tt}},$$

and

$$\sup_{x>0} \left| \mathbb{P}\left(\mathcal{E}^\top \boldsymbol{P}_{\mathcal{S}} \mathcal{E} \leqslant x \right) - \mathbb{P}(\chi^2_K \leqslant x) \right| \leqslant cK^{5/4} \mathbb{E}|\varepsilon|^3 \mathbb{E}\left(\max_{t\in[n]} \sqrt{\boldsymbol{P}_{\mathcal{S},tt}} \mathbb{I}\{\mathcal{A}_{\mathcal{S}}\} \right) + \mathbb{P}(\mathcal{A}_{\mathcal{S}}^c). \tag{F.4}$$

Recall that both $\boldsymbol{P}_{\widehat{\boldsymbol{F}}} + \boldsymbol{P}_{\widehat{\boldsymbol{U}}_{\mathcal{S}}} - \boldsymbol{P}_{\boldsymbol{X}_{\mathcal{S}}}$ and $\boldsymbol{P}_{\boldsymbol{X}_{\mathcal{S}}}$ are orthogonal projection matrices. Hence

$$\max_{t\in[n]} \boldsymbol{P}_{\mathcal{S},tt} = \max_{t\in[n]} \left(\boldsymbol{P}_{\widehat{\boldsymbol{F}}} + \boldsymbol{P}_{\widehat{\boldsymbol{U}}_{\mathcal{S}}} - \boldsymbol{P}_{\boldsymbol{X}_{\mathcal{S}}} \right)_{tt} \leqslant \max_{t\in[n]} \boldsymbol{P}_{\widehat{\boldsymbol{F}},tt} + \max_{t\in[n]} \boldsymbol{P}_{\widehat{\boldsymbol{U}}_{\mathcal{S}},tt}.$$

By Lemma 2.1, it follows that

$$\max_{t\in[n]} \boldsymbol{P}_{\widehat{\boldsymbol{F}},tt} = \max_{t\in[n]} \frac{1}{n} \|\widehat{\boldsymbol{f}}_t\|_2^2 \lesssim \max_{t\in[n]} \frac{1}{n} \|\widehat{\boldsymbol{f}}_t - \boldsymbol{H}\boldsymbol{f}_t\|_2^2 + \max_{t\in[n]} \frac{1}{n} \|\boldsymbol{H}\|_2^2 \|\boldsymbol{f}_t\|_2^2 = O_{\mathbb{P}}\left(\frac{\log n}{n} \right). \tag{F.5}$$

Now we bound $\max_{t\in[n]} \boldsymbol{P}_{\widehat{\boldsymbol{U}}_{\mathcal{S}},tt}$. Under event $\mathcal{A}_{\mathcal{S}}$, we have

$$\max_{t\in[n]} \boldsymbol{P}_{\widehat{\boldsymbol{U}}_{\mathcal{S}},tt} \leqslant \frac{1}{\lambda_{\min}(\widehat{\boldsymbol{U}}_{\mathcal{S}}^\top \widehat{\boldsymbol{U}}_{\mathcal{S}})} \max_{t\in[n]} \sum_{\ell\in\mathcal{S}} |\widehat{u}_{t\ell}|^2 \leqslant \frac{2}{n\lambda_0} \max_{t\in[n]} \sum_{\ell\in\mathcal{S}} |\widehat{u}_{t\ell}|^2.$$

Recall that $\widehat{\boldsymbol{U}} = (\boldsymbol{I}_n - \widehat{\boldsymbol{P}})\boldsymbol{X} = (\boldsymbol{I}_n - \widehat{\boldsymbol{P}})\boldsymbol{F}\boldsymbol{B}^\top + (\boldsymbol{I}_n - \widehat{\boldsymbol{P}})\boldsymbol{U}$ and $\widehat{\boldsymbol{F}}^\top\widehat{\boldsymbol{U}} = \boldsymbol{O}$. Hence, under the event $\mathcal{E}_\lambda$, we have

$$\widehat{\boldsymbol{U}} = (\boldsymbol{I}_n - \widehat{\boldsymbol{P}})(\boldsymbol{F}\boldsymbol{H}^\top - \widehat{\boldsymbol{F}})\boldsymbol{H}^{-\top}\boldsymbol{B}^\top - \widehat{\boldsymbol{P}}(\boldsymbol{U} - \widehat{\boldsymbol{U}}) + \boldsymbol{U}.$$

Consequently, it follows that

$$\begin{aligned}\max_{t\in[n]}\sum_{\ell\in\mathcal{S}}|\widehat{u}_{t\ell}|^2 &\lesssim \max_{t\in[n]}\sum_{\ell\in\mathcal{S}}|\boldsymbol{e}_t^\top(\boldsymbol{I}_n - \widehat{\boldsymbol{P}})(\boldsymbol{F}\boldsymbol{H}^\top - \widehat{\boldsymbol{F}})\boldsymbol{H}^{-\top}\boldsymbol{b}_\ell|^2 \\ &\quad + \max_{t\in[n]}\sum_{\ell\in\mathcal{S}}|\boldsymbol{e}_t^\top\widehat{\boldsymbol{P}}(\boldsymbol{U}_{\cdot\ell} - \widehat{\boldsymbol{U}}_{\cdot\ell})|^2 + \max_{t\in[n]}\sum_{\ell\in\mathcal{S}}|u_{t\ell}|^2 \\ &=: \Delta_1^\natural + \Delta_2^\natural + \Delta_3^\natural.\end{aligned}$$

By Lemma 2.1, we have

$$\begin{aligned}\Delta_1^\natural &\leqslant \|\boldsymbol{F}\boldsymbol{H}^\top - \widehat{\boldsymbol{F}}\|_{\mathbb{F}}^2\|\boldsymbol{H}^{-\top}\|_2^2\sum_{\ell\in\mathcal{S}}\|\boldsymbol{b}_\ell\|_2^2 = O_{\mathbb{P}}\left(\frac{|\mathcal{S}|}{n} + \frac{|\mathcal{S}|n}{d}\right), \\ \Delta_2^\natural &\leqslant \frac{|\mathcal{S}|}{n}\max_{t\in[n]}\|\widehat{\boldsymbol{f}}_t\|_2^2\max_{\ell\in[d]}\sum_{s=1}^n|\widehat{u}_{s\ell} - u_{s\ell}|^2 = O_{\mathbb{P}}\left\{|\mathcal{S}|\log n\left(\frac{\log d}{n} + \frac{1}{d}\right)\right\},\end{aligned}$$

and $\Delta_3^\natural = O_{\mathbb{P}}(|\mathcal{S}| + \log n)$. Therefore, under event $\mathcal{A}_{\mathcal{S}}$, we have

$$\max_{t\in[n]}\boldsymbol{P}_{\widehat{\boldsymbol{U}}_{\mathcal{S}},tt} \leqslant \frac{2}{n\lambda_0}\max_{t\in[n]}\sum_{\ell\in\mathcal{S}}|\widehat{u}_{t\ell}|^2 = O_{\mathbb{P}}\left(\frac{s_n}{n\wedge d} + \frac{\log n}{n}\right).$$

Combined with (F.5), we obtain

$$\max_{t\in[n]}\boldsymbol{P}_{\mathcal{S},tt} = O_{\mathbb{P}}(\varpi_n), \quad \text{where } \varpi_n = \frac{s_n}{n\wedge d} + \frac{\log n}{n}.$$

Note that $\varpi_n \to 0$ by (F.1). As $\boldsymbol{P}_{\mathcal{S}}$ is an orthogonal matrix, we have $\max_{t\in[n]}\boldsymbol{P}_{\mathcal{S},tt} \leqslant 1$ and consequently

$$\sup_{\mathcal{S}\subset[d]:|\mathcal{S}|\leqslant s_n}\mathbb{E}\left(\max_{t\in[n]}\boldsymbol{P}_{\mathcal{S},tt}\right)\mathbb{I}\{\mathcal{A}_{\mathcal{S}}\} \leqslant \sqrt{\varpi_n} + \sup_{\mathcal{S}\subset[d]:|\mathcal{S}|\leqslant s_n}\mathbb{P}\left(\max_{t\in[n]}\boldsymbol{P}_{\mathcal{S},tt} > \sqrt{\varpi_n}\right) \to 0.$$

In view of (F.3), we obtain $\rho^\diamond \to 0$ once we prove that

$$\sup_{\mathcal{S}\subset[d]:|\mathcal{S}|\leqslant s_n}\mathbb{P}(\mathcal{A}_{\mathcal{S}}^c) = \sup_{\mathcal{S}\subset[d]:|\mathcal{S}|\leqslant s_n}\mathbb{P}\left\{\lambda_{\min}(\widehat{\boldsymbol{U}}_{\mathcal{S}}^\top\widehat{\boldsymbol{U}}_{\mathcal{S}}) \leqslant \frac{n\lambda_0}{2}\right\} \to 0. \tag{F.6}$$

Next, we will prove (F.6). Recall that $\widetilde{\boldsymbol{\Sigma}} = n^{-1}\widehat{\boldsymbol{U}}^\top\widehat{\boldsymbol{U}}$ and $\widehat{\boldsymbol{\Sigma}} = n^{-1}\boldsymbol{U}^\top\boldsymbol{U}$. By Weyl's theorem on eigenvalues, it follows that

$$\sup_{\mathcal{S}\subset[d]:|\mathcal{S}|\leqslant s_n}\mathbb{P}\left\{\lambda_{\min}(\widehat{\boldsymbol{U}}_{\mathcal{S}}^\top\widehat{\boldsymbol{U}}_{\mathcal{S}}) \leqslant \frac{n\lambda_0}{2}\right\} \leqslant \sup_{\mathcal{S}\subset[d]:|\mathcal{S}|\leqslant s_n}\mathbb{P}\left(\|\widetilde{\boldsymbol{\Sigma}}_{\mathcal{S}\mathcal{S}} - \widehat{\boldsymbol{\Sigma}}_{\mathcal{S}\mathcal{S}}\| > \frac{\lambda_0}{4}\right)$$

$$+ \sup_{\mathcal{S}\subset[d]:|\mathcal{S}|\leqslant s_n} \mathbb{P}\left(\|\widetilde{\boldsymbol{\Sigma}}_{\mathcal{S}\mathcal{S}} - \widehat{\boldsymbol{\Sigma}}_{\mathcal{S}\mathcal{S}}\| > \frac{\lambda_0}{4}\right) =: \pi_1 + \pi_2.$$

We first bound π_1. By Lemma C.6 and (F.1),

$$\sup_{\mathcal{S}\subset[d]:|\mathcal{S}|\leqslant s_n} \|\widetilde{\boldsymbol{\Sigma}}_{\mathcal{S}\mathcal{S}} - \widehat{\boldsymbol{\Sigma}}_{\mathcal{S}\mathcal{S}}\| \leqslant \sup_{\mathcal{S}\subset[d]:|\mathcal{S}|\leqslant s_n} |\mathcal{S}|\|\widetilde{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}\|_{\max} = O_{\mathbb{P}}\left(\frac{s_n \log d}{n} + \frac{s_n}{d}\right) = o_{\mathbb{P}}(1). \tag{F.7}$$

Hence $\pi_1 \to 0$ as $n \to \infty$. Now we bound π_2. For any subset $\mathcal{S} \subset [d]$ such that $|\mathcal{S}| \leqslant s_n$, by Lemma 5.4 in Vershynin [2012] and Assumption 2.1,

$$\begin{aligned}\mathbb{P}\left(\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}\mathcal{S}} - \boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}\| > \frac{\lambda_0}{4}\right) &\leqslant 9^{|\mathcal{S}|} \sup_{\boldsymbol{v}\in\mathbb{S}^{d-1}:\boldsymbol{v}_{\mathcal{S}^c}=0} \mathbb{P}\left(\left|\frac{1}{n}\sum_{t=1}^n (\boldsymbol{u}_t^\top \boldsymbol{v})^2 - \boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v}\right| > \frac{\lambda_0}{8}\right) \\ &\leqslant 2\exp\left\{|\mathcal{S}|\log 9 - C\min\left(n\lambda_0^2, n\lambda_0\right)\right\}. \end{aligned} \tag{F.8}$$

Consequently, by (F.1), we obtain

$$\begin{aligned}\pi_2 &\leqslant \sup_{\mathcal{S}\subset[d]:|\mathcal{S}|\leqslant s_n} 2\exp\left(|\mathcal{S}|\log 9 - Cn\min\{\lambda_0^2, \lambda_0\}\right) \\ &\leqslant 2\exp\left(s_n \log 9 - Cn\min\{\lambda_0^2, \lambda_0\}\right) \to 0.\end{aligned}$$

□

Lemma F.2. *Assume that*

$$\|\boldsymbol{\varphi}^\star\|_2 \left(\sqrt{n/d} + 1/\sqrt{n}\right) \to 0. \tag{F.9}$$

Define

$$\mathcal{H}(\alpha, \theta) = \left\{\boldsymbol{\varphi} \in \mathbb{R}^K : \frac{n\|\boldsymbol{\varphi}\|_2^2}{1 + Ks_n\Upsilon^2/\lambda_{\min}(\boldsymbol{\Sigma})} \geqslant \sigma^2(2+\delta)(\chi^2_{K,1-\alpha} + \chi^2_{K,1-\theta})\right\}$$

for some $\delta > 0$. Then, under the conditions of Lemma F.1, we have

$$\inf_{\boldsymbol{\varphi}^\star\in\mathcal{H}(\alpha,\theta)} \mathbb{P}\left(\frac{\|\boldsymbol{P}_{\widetilde{\mathcal{S}}}\boldsymbol{Y}\|_2^2}{\sigma^2} > \chi^2_{K,1-\alpha}\right) \geqslant 1 - \theta. \tag{F.10}$$

Proof of Lemma F.2. Recall that $\boldsymbol{Y} = \boldsymbol{F}\boldsymbol{\varphi}^\star + \boldsymbol{X}\boldsymbol{\beta}^\star + \mathcal{E} = \boldsymbol{Y} = \boldsymbol{F}\boldsymbol{\varphi}^\star + \boldsymbol{X}_{\widetilde{\mathcal{S}}}\boldsymbol{\beta}^\star_{\widetilde{\mathcal{S}}} + \mathcal{E}$. Hence $\boldsymbol{P}_{\widetilde{\mathcal{S}}}\boldsymbol{Y} = \boldsymbol{P}_{\widetilde{\mathcal{S}}}\mathcal{E} + \boldsymbol{P}_{\widetilde{\mathcal{S}}}\boldsymbol{F}\boldsymbol{\varphi}^\star$. Without loss of generality, we assume that $\sigma = 1$. By the Cauchy-Schwarz inequality,

$$\|\boldsymbol{P}_{\widetilde{\mathcal{S}}}\boldsymbol{Y}\|_2^2 \geqslant \frac{1}{2}\|\boldsymbol{P}_{\widetilde{\mathcal{S}}}\boldsymbol{F}\boldsymbol{\varphi}^\star\|_2^2 - \|\boldsymbol{P}_{\widetilde{\mathcal{S}}}\mathcal{E}\|_2^2.$$

Therefore, for any $\boldsymbol{\varphi}^\star \in \mathcal{H}(\alpha, \theta)$, we have

$$\mathbb{P}\left(\|\boldsymbol{P}_{\widetilde{\mathcal{S}}}\boldsymbol{Y}\|_2^2 > \chi^2_{K,1-\alpha}\right) \geqslant \mathbb{P}\left(\frac{1}{2}\|\boldsymbol{P}_{\widetilde{\mathcal{S}}}\boldsymbol{F}\boldsymbol{\varphi}^\star\|_2^2 > \chi^2_{K,1-\alpha} + \|\boldsymbol{P}_{\widetilde{\mathcal{S}}}\mathcal{E}\|_2^2\right).$$

We first bound $\|\boldsymbol{P}_{\widetilde{\mathcal{S}}}\boldsymbol{F}\boldsymbol{\varphi}^\star\|_2^2$. By Lemma 2.1 and the Cauchy-Schwarz inequality, for $\epsilon > 0$,

$$\|\boldsymbol{P}_{\widetilde{\mathcal{S}}}\boldsymbol{F}\boldsymbol{\varphi}^\star\|_2^2 \geqslant (1-\epsilon)\|\boldsymbol{P}_{\widetilde{\mathcal{S}}}\widehat{\boldsymbol{F}}\boldsymbol{H}\boldsymbol{\varphi}^\star\|_2^2 - \left(\frac{1}{\epsilon} - 1\right)\|\boldsymbol{P}_{\widetilde{\mathcal{S}}}(\boldsymbol{F} - \widehat{\boldsymbol{F}}\boldsymbol{H})\boldsymbol{\varphi}^\star\|_2^2.$$

As $\boldsymbol{P}_{\widetilde{\mathcal{S}}}$ is a projection matrix, it follows from Lemma 2.1 and (F.9) that

$$\|\boldsymbol{P}_{\widetilde{\mathcal{S}}}(\boldsymbol{F} - \widehat{\boldsymbol{F}}\boldsymbol{H})\boldsymbol{\varphi}^\star\|_2 \leqslant \|\boldsymbol{F} - \widehat{\boldsymbol{F}}\boldsymbol{H}\|_{\mathbb{F}}\|\boldsymbol{\varphi}^\star\|_2 = O_{\mathbb{P}}\left(\|\boldsymbol{\varphi}^\star\|_2\sqrt{n/d} + \|\boldsymbol{\varphi}^\star\|_2\sqrt{1/n}\right) = o_{\mathbb{P}}(1).$$

Recall that $\boldsymbol{P}_{\widetilde{\mathcal{S}}} = \boldsymbol{P}_{\widehat{\boldsymbol{F}}} + \boldsymbol{P}_{\widehat{U}_{\widetilde{\mathcal{S}}}} - \boldsymbol{P}_{\boldsymbol{X}_{\widetilde{\mathcal{S}}}}$ and $\widetilde{\boldsymbol{\Sigma}} = n^{-1}\widehat{\boldsymbol{U}}^\top\widehat{\boldsymbol{U}}$. Hence

$$\begin{aligned}
\|\boldsymbol{P}_{\widetilde{\mathcal{S}}}\widehat{\boldsymbol{F}}\boldsymbol{H}\boldsymbol{\varphi}^\star\|_2^2 &= n(\boldsymbol{H}\boldsymbol{\varphi}^\star)^\top\left\{\boldsymbol{I}_K - \widehat{\boldsymbol{B}}_{\widetilde{\mathcal{S}}}^\top\left(\widehat{\boldsymbol{B}}_{\widetilde{\mathcal{S}}}\widehat{\boldsymbol{B}}_{\widetilde{\mathcal{S}}}^\top + \widetilde{\boldsymbol{\Sigma}}_{\widetilde{\mathcal{S}}\widetilde{\mathcal{S}}}\right)^{-1}\widehat{\boldsymbol{B}}_{\widetilde{\mathcal{S}}}\right\}(\boldsymbol{H}\boldsymbol{\varphi}^\star) \\
&= n(\boldsymbol{H}\boldsymbol{\varphi}^\star)^\top\left(\boldsymbol{I}_K + \widehat{\boldsymbol{B}}_{\widetilde{\mathcal{S}}}^\top\widetilde{\boldsymbol{\Sigma}}_{\widetilde{\mathcal{S}}\widetilde{\mathcal{S}}}^{-1}\widehat{\boldsymbol{B}}_{\widetilde{\mathcal{S}}}\right)^{-1}(\boldsymbol{H}\boldsymbol{\varphi}^\star) \\
&= n\boldsymbol{\varphi}^{\star\top}\left\{(\boldsymbol{H}^\top\boldsymbol{H})^{-1} + \boldsymbol{H}^{-1}\widehat{\boldsymbol{B}}_{\widetilde{\mathcal{S}}}^\top\widetilde{\boldsymbol{\Sigma}}_{\widetilde{\mathcal{S}}\widetilde{\mathcal{S}}}^{-1}\widehat{\boldsymbol{B}}_{\widetilde{\mathcal{S}}}\boldsymbol{H}^{-\top}\right\}^{-1}\boldsymbol{\varphi}^\star =: n\boldsymbol{\varphi}^{\star\top}\widehat{\mathbb{W}}^{-1}\boldsymbol{\varphi}^\star.
\end{aligned}$$

Let $\mathbb{W} = \boldsymbol{I}_K + \boldsymbol{B}_{\widetilde{\mathcal{S}}}^\top\boldsymbol{\Sigma}_{\widetilde{\mathcal{S}}\widetilde{\mathcal{S}}}^{-1}\boldsymbol{B}_{\widetilde{\mathcal{S}}}$. We first upper bound $\|\widehat{\mathbb{W}} - \mathbb{W}\|_2$. For any $\boldsymbol{h} \in \mathbb{R}^{|\widetilde{\mathcal{S}}|}$ with $\|\boldsymbol{h}\|_2 = 1$, we decompose $\boldsymbol{h}^\top(\widehat{\mathbb{W}} - \mathbb{W})\boldsymbol{h} = \Psi_1 + \Psi_2 + \Psi_3 + \Psi_4$, where

$$\begin{aligned}
\Psi_1 &= \boldsymbol{h}^\top(\boldsymbol{H}^\top\boldsymbol{H})^{-1}\left(\boldsymbol{I}_K - \boldsymbol{H}^\top\boldsymbol{H}\right)\boldsymbol{h}, \\
\Psi_2 &= \boldsymbol{h}^\top\boldsymbol{H}^{-1}\left(\widehat{\boldsymbol{B}}_{\widetilde{\mathcal{S}}} - \boldsymbol{B}_{\widetilde{\mathcal{S}}}\boldsymbol{H}^\top\right)^\top\widetilde{\boldsymbol{\Sigma}}_{\widetilde{\mathcal{S}}\widetilde{\mathcal{S}}}^{-1}\widehat{\boldsymbol{B}}_{\widetilde{\mathcal{S}}}\boldsymbol{H}^{-\top}\boldsymbol{h}, \\
\Psi_3 &= \boldsymbol{h}^\top\boldsymbol{B}_{\widetilde{\mathcal{S}}}^\top\widetilde{\boldsymbol{\Sigma}}_{\widetilde{\mathcal{S}}\widetilde{\mathcal{S}}}^{-1}\left(\boldsymbol{\Sigma}_{\widetilde{\mathcal{S}}\widetilde{\mathcal{S}}} - \widetilde{\boldsymbol{\Sigma}}_{\widetilde{\mathcal{S}}\widetilde{\mathcal{S}}}\right)\boldsymbol{\Sigma}_{\widetilde{\mathcal{S}}\widetilde{\mathcal{S}}}^{-1}\widehat{\boldsymbol{B}}_{\widetilde{\mathcal{S}}}\boldsymbol{H}^{-\top}\boldsymbol{h}, \\
\Psi_4 &= \boldsymbol{h}^\top\boldsymbol{B}_{\widetilde{\mathcal{S}}}^\top\boldsymbol{\Sigma}_{\widetilde{\mathcal{S}}\widetilde{\mathcal{S}}}^{-1}\left(\widehat{\boldsymbol{B}}_{\widetilde{\mathcal{S}}} - \boldsymbol{B}_{\widetilde{\mathcal{S}}}\boldsymbol{H}^\top\right)\boldsymbol{H}^{-\top}\boldsymbol{h}.
\end{aligned}$$

By Lemma 2.1, we have $\|\boldsymbol{H}^\top\boldsymbol{H} - \boldsymbol{I}_K\|_{\mathbb{F}} = o_{\mathbb{P}}(1)$ and $\mathbb{P}\{\lambda_{\min}(\boldsymbol{H}^\top\boldsymbol{H}) \geqslant 1/2\} \to 1$. Therefore

$$\Psi_1 \leqslant \frac{\|\boldsymbol{H}^\top\boldsymbol{H} - \boldsymbol{I}_K\|_{\mathbb{F}}}{\lambda_{\min}(\boldsymbol{H}^\top\boldsymbol{H})} = o_{\mathbb{P}}(1)$$

and $\mathbb{P}(\|\boldsymbol{H}^{-\top}\boldsymbol{h}\|_2^2 \leqslant 2) \to 1$. Since $\mathbb{P}(|\widetilde{\mathcal{S}}| \leqslant s_n) \to 1$, it follows that

$$\left\|\left(\widehat{\boldsymbol{B}}_{\widetilde{\mathcal{S}}} - \boldsymbol{B}_{\widetilde{\mathcal{S}}}\boldsymbol{H}^\top\right)\boldsymbol{H}^{-\top}\boldsymbol{h}\right\|^2 \leqslant |\widetilde{\mathcal{S}}|\max_{j\in[d]}\|\widehat{\boldsymbol{b}}_j - \boldsymbol{H}\boldsymbol{b}_j\|^2\|\boldsymbol{H}^{-\top}\boldsymbol{h}\|^2 = O_{\mathbb{P}}\left\{s_n\left(\frac{\log d}{n} + \frac{1}{d}\right)\right\}.$$

Combining this with (F.1) and the fact that $\|\boldsymbol{B}_{\widetilde{\mathcal{S}}}\boldsymbol{h}\|^2 \leqslant K|\widetilde{\mathcal{S}}|\mathcal{K}^2$, we obtain

$$\Psi_2 \leqslant \frac{\|\widehat{\boldsymbol{B}}_{\widetilde{\mathcal{S}}}\boldsymbol{H}^{-\top}\boldsymbol{h}\|}{\lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}_{\widetilde{\mathcal{S}}\widetilde{\mathcal{S}}})}\left\|\left(\widehat{\boldsymbol{B}}_{\widetilde{\mathcal{S}}} - \boldsymbol{B}_{\widetilde{\mathcal{S}}}\boldsymbol{H}^{\top}\right)\boldsymbol{H}^{-\top}\boldsymbol{h}\right\| = o_{\mathbb{P}}\left(s_n^{1/2}\right).$$

Similarly, by (F.7) and (F.8), we have $\|\widetilde{\boldsymbol{\Sigma}}_{\widetilde{\mathcal{S}}\widetilde{\mathcal{S}}} - \boldsymbol{\Sigma}_{\widetilde{\mathcal{S}}\widetilde{\mathcal{S}}}\|_2 = o_{\mathbb{P}}(1)$,

$$\Psi_3 \leqslant \frac{\|\boldsymbol{B}_{\widetilde{\mathcal{S}}}\boldsymbol{h}\|\|\widetilde{\boldsymbol{\Sigma}}_{\widetilde{\mathcal{S}}\widetilde{\mathcal{S}}} - \boldsymbol{\Sigma}_{\widetilde{\mathcal{S}}\widetilde{\mathcal{S}}}\|_2\|\widehat{\boldsymbol{B}}_{\widetilde{\mathcal{S}}}\boldsymbol{H}^{-\top}\boldsymbol{h}\|}{\lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}_{\widetilde{\mathcal{S}}\widetilde{\mathcal{S}}})\lambda_{\min}(\boldsymbol{\Sigma}_{\widetilde{\mathcal{S}}\widetilde{\mathcal{S}}})} = o_{\mathbb{P}}\left(s_n^{1/2}\right).$$

and

$$\Psi_4 \leqslant \frac{\|\boldsymbol{B}_{\widetilde{\mathcal{S}}}\boldsymbol{h}\|}{\lambda_{\min}(\boldsymbol{\Sigma}_{\widetilde{\mathcal{S}}\widetilde{\mathcal{S}}})}\left\|\left(\widehat{\boldsymbol{B}}_{\widetilde{\mathcal{S}}} - \boldsymbol{B}_{\widetilde{\mathcal{S}}}\boldsymbol{H}^{\top}\right)\boldsymbol{H}^{-\top}\boldsymbol{h}\right\| = o_{\mathbb{P}}\left(s_n^{1/2}\right).$$

Consequently, we obtain $\|\widehat{\mathbb{W}} - \mathbb{W}\|_2 = o_{\mathbb{P}}(\sqrt{s_n})$. Combining this with the fact that

$$\mathbb{P}\left\{\lambda_{\max}(\mathbb{W}) \leqslant 1 + \frac{Ks_n\|\boldsymbol{B}\|_{\max}^2}{\lambda_{\min}(\boldsymbol{\Sigma})}\right\} \to 1,$$

we obtain

$$\mathbb{P}\left\{\|\widehat{\mathbb{W}} - \mathbb{W}\|_2 \leqslant \epsilon^{\diamond}\left(1 + \frac{Ks_n\|\boldsymbol{B}\|_{\max}^2}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)\right\} \to 1.$$

Note that

$$\|\boldsymbol{P}_{\widetilde{\mathcal{S}}}\widehat{\boldsymbol{F}}\boldsymbol{H}\boldsymbol{\varphi}^{\star}\|^2 \geqslant \frac{n\|\boldsymbol{\varphi}^{\star}\|^2}{\lambda_{\max}(\widehat{\mathbb{W}})} \geqslant \frac{n\|\boldsymbol{\varphi}^{\star}\|^2}{\lambda_{\max}(\mathbb{W}) + \|\widehat{\mathbb{W}} - \mathbb{W}\|_2}.$$

Hence

$$\mathbb{P}\left\{\|\boldsymbol{P}_{\widetilde{\mathcal{S}}}\widehat{\boldsymbol{F}}\boldsymbol{H}\boldsymbol{\varphi}^{\star}\|^2 \geqslant \frac{n\|\boldsymbol{\varphi}^{\star}\|^2}{(1+\epsilon^{\diamond})(1 + Ks_n\|\boldsymbol{B}\|_{\max}^2/\lambda_{\min}(\boldsymbol{\Sigma}))}\right\} \to 1.$$

Putting all these pieces together, we obtain (F.10). □

F.4 Application to multi-modal sparse regression model

Dataset with multiple types are now frequently collected for some mutual experimental subjects. This data structure is called as multimodal data and is becoming more and more popular in many fields. Factor analysis is commonly used in integrative analysis of multimodal data, and is particularly useful to overcome the curse of high dimensionality and high correlations. Recently Li and Li [2021] study sparse linear multi-modal regression model with factor structures. However, the hypothesis testing

problem on whether sparse regression for a given modal (a certain group of modal) is adequate hasn't been investigated.

Thus, in this subsection, we extend our results given in the last subsection to the multi-modal sparse regression model. Our observations are $(\boldsymbol{Y}, \boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_M)$ in which we assume covariate $\boldsymbol{X}_i$ is generated from i-th group (modal). Moreover, for every $i \in [M]$, $\boldsymbol{X}_i$ is assumed to have its own factor structure $\boldsymbol{X}_i = \boldsymbol{F}_i \boldsymbol{B}_i^\top + \boldsymbol{U}_i$. We next consider the hypothesis test as follows:

$$H_0 : \boldsymbol{Y} = \sum_{i=1}^{L} \boldsymbol{X}_i \boldsymbol{\beta}_i^\star + \sum_{i=L+1}^{M} \boldsymbol{X}_i \boldsymbol{\beta}_i^\star + \mathcal{E} \text{ versus } H_1 : \boldsymbol{Y} = \sum_{i=1}^{L} (\boldsymbol{F}_i \boldsymbol{\gamma}_i^\star + \boldsymbol{U}_i \boldsymbol{\beta}_i^\star) + \sum_{i=L+1}^{M} \boldsymbol{X}_i \boldsymbol{\beta}_i^\star + \mathcal{E}.$$

Here we aim at simultaneously testing whether the sparse regression is adequate for the first L modals. L is assumed to be any fixed number in $[M]$. The hypothesis testing problem in the previous section is a special case to this with $L = M = 1$.

In order to proceed the hypothesis testing procedure, similar to section 4, we separate our dataset into two parts with size m and $n - m$ respectively. We let $\boldsymbol{X}_i^{(j)}$ and $\boldsymbol{Y}^{(j)}, j \in \{1, 2\}$ denote the i-th modal and our response respectively in the j-th part of our splitted data. Next, we decompose every $\boldsymbol{X}_i^{(1)}$ into $(\widehat{\boldsymbol{F}}_i^{(1)}, \widehat{\boldsymbol{U}}_i^{(1)})$ and then use sure screening to select $(\widehat{S}_1, \widehat{S}_2, \widehat{S}_3, \cdots, \widehat{S}_M)$ by using $(\boldsymbol{Y}^{(1)}, \widehat{\boldsymbol{U}}_1^{(1)}, \widehat{\boldsymbol{U}}_2^{(1)}, \cdots, \widehat{\boldsymbol{U}}_L^{(1)}, \widehat{\boldsymbol{U}}_{L+1}^{(1)}, \cdots, \widehat{\boldsymbol{U}}_M^{(1)})$. Here we also make an assumption that

Assumption F.1. Here we assume

$$\mathbb{P}\left(\text{for all } i \in [M],\ S_i^\star \subset \widehat{S}_i \text{ and } \sum_{i=1}^{M} |S_i| \leqslant s_n^M \right) \to 1. \tag{F.11}$$

Next, we also impose an assumption on the covariance of our factors and idiosyncratic components .

Assumption F.2. We assume

$$\mathrm{Cov}(\boldsymbol{f}_{1,t}, \boldsymbol{f}_{2,t}, \cdots, \boldsymbol{f}_{M,t}) = I_K,\ \mathrm{Cov}(\boldsymbol{u}_{1,t}, \boldsymbol{u}_{2,t}, \cdots, \boldsymbol{u}_{M,t}) = \Sigma_u,$$

with $0 < \lambda_{\min}(\Sigma_u), \lambda_{\max}(\Sigma_u) \leqslant C$ with C being a absolute constant. Moreover, $(\boldsymbol{f}_{1,t}, \ldots, \boldsymbol{f}_{M,t}), t \in [n]$ is assumed to be uncorrelated with $(\boldsymbol{u}_{1,t}, \ldots, \boldsymbol{u}_{M,t})$, where $\boldsymbol{f}_{i,t}$ and $\boldsymbol{u}_{i,t}$ are the t-th row of $\boldsymbol{F}_i$ and $\boldsymbol{U}_i$ respectively.

Next, we use the second part of the data to construct our test statistic. For simplicity we denote

$$\boldsymbol{Z}_1 = (\boldsymbol{X}_{\widehat{S}_1}^{(2)}, \cdots, \boldsymbol{X}_{\widehat{S}_L}^{(2)}, \widehat{\boldsymbol{F}}_{L+1}^{(2)}, \widehat{\boldsymbol{U}}_{\widehat{S}_{L+1}}^{(2)}, \cdots, \widehat{\boldsymbol{F}}_M^{(2)}, \widehat{\boldsymbol{U}}_{\widehat{S}_M}^{(2)})$$

and

$$\boldsymbol{Z}_2 = (\widehat{\boldsymbol{F}}_1^{(2)}, \widehat{\boldsymbol{U}}_{\widehat{S}_1}^{(2)}, \cdots, \widehat{\boldsymbol{F}}_L^{(2)}, \widehat{\boldsymbol{U}}_{\widehat{S}_L}^{(2)}, \widehat{\boldsymbol{F}}_{L+1}^{(2)}, \widehat{\boldsymbol{U}}_{\widehat{S}_{L+1}}^{(2)}, \cdots, \widehat{\boldsymbol{F}}_M^{(2)}, \widehat{\boldsymbol{U}}_{\widehat{S}_M}^{(2)}).$$

The test statistic is given by

$$Q_{n,M}^{(2)} = \left\| (\boldsymbol{I}_{n-m} - \boldsymbol{P}_{\boldsymbol{Z}_1}) \boldsymbol{Y}^{(2)} \right\|_2^2 - \left\| (\boldsymbol{I}_{n-m} - \boldsymbol{P}_{\boldsymbol{Z}_2}) \boldsymbol{Y}^{(2)} \right\|_2^2 = \left\| (\boldsymbol{P}_{\boldsymbol{Z}_2} - \boldsymbol{P}_{\boldsymbol{Z}_1}) \boldsymbol{Y}^{(2)} \right\|_2^2.$$

We finally summarize the asymptotic behaviors of our test statistic in the following Corollary F.3.

Corollary F.3. *Let Assumptions 2.1–2.5 and Assumption F.1 hold with*

$$s_n^M\left(\frac{\log d}{n}+\frac{1}{d}\right)\to 0 \ \text{ and } \ \Delta_\sigma\to 0,$$

where Δ_σ is given in Assumption 3.3. Then we have

$$\sup_{x>0}\left|\mathbb{P}\left(Q_{n,M}^{(2)}/\widehat{\sigma}^2\leqslant x|H_0\right)-\mathbb{P}\left(\chi^2_{K^\star}\leqslant x\right)\right|\to 0, \ \text{ as } \ n\to\infty,$$

in which $K^\star=\sum_{i=1}^L K_i$. In addition, we define $\boldsymbol{\varphi}=(\boldsymbol{\varphi}_1,\boldsymbol{\varphi}_2,\cdots,\boldsymbol{\varphi}_L)$ and

$$\mathcal{D}^\star(\alpha,\theta)=\left\{\boldsymbol{\varphi}\in\mathbb{R}^K:\frac{n\|\boldsymbol{\varphi}\|^2}{1+K^\star s_n^M\Upsilon^2/\lambda_{\min}(\Sigma_u)}\geqslant\sigma^2(2+\delta)(\chi^2_{K^\star,1-\alpha}+\chi^2_{K^\star,1-\theta})\right\},$$

where $\delta>0$ is some constant. Then when $s_n^M=o(n^{1/6})$, we have

$$\inf_{\boldsymbol{\varphi}^\star\in\mathcal{D}^\star(\alpha,\theta)}\mathbb{P}(\psi_\alpha^\star=1|H_1)\geqslant 1-\theta, \tag{F.12}$$

in which

$$\psi_\alpha^\star=\mathbb{I}\left\{Q_{n,M}^{(2)}/\widehat{\sigma}^2\geqslant\chi^2_{K^\star,1-\alpha}\right\}.$$

F.5 Proof of Corollary F.3

Proof. Here we assume $(\widetilde{S}_1,\cdots,\widetilde{S}_M)$ are independent with $(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_M,\boldsymbol{Y})$.

First, $\boldsymbol{P}_{\boldsymbol{Z}_2}-\boldsymbol{P}_{\boldsymbol{Z}_1}$ is a projection matrix onto the space of

$$(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{Z}_1})\boldsymbol{Z}_2.$$

Recall that

$$\boldsymbol{Z}_2=(\widehat{\boldsymbol{F}}_1^{(2)},\widehat{\boldsymbol{U}}_{\widehat{S}_1}^{(2)},\cdots,\widehat{\boldsymbol{F}}_L^{(2)},\widehat{\boldsymbol{U}}_{\widehat{S}_L}^{(2)},\widehat{\boldsymbol{F}}_{L+1}^{(2)},\widehat{\boldsymbol{U}}_{\widehat{S}_{L+1}}^{(2)},\cdots,\widehat{\boldsymbol{F}}_M^{(2)},\widehat{\boldsymbol{U}}_{\widehat{S}_M}^{(2)}).$$

Note that for each $1\leqslant l\leqslant L$,

$$(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{Z}_1})(\widehat{\boldsymbol{F}}_l^{(2)},\widehat{\boldsymbol{U}}_{\widehat{S}_l}^{(2)})=(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{Z}_1})(\widehat{\boldsymbol{F}}_l^{(2)},\boldsymbol{X}_{\widehat{S}_l}^{(2)}-\widehat{\boldsymbol{F}}_l^{(2)}\widehat{\boldsymbol{B}}_{\widehat{S}_l}^{(2)}).$$

For any vector $\boldsymbol{\phi}$, if we there exists an $\widetilde{\boldsymbol{\phi}}$ such that $\boldsymbol{Z}_1\boldsymbol{\phi}=\boldsymbol{Z}_2\widetilde{\boldsymbol{\phi}}$, we obtain $col(\boldsymbol{Z}_1)\subset col(\boldsymbol{Z}_2)$. Thus, we have $\boldsymbol{P}_{S_M^*}:=\boldsymbol{P}_{\boldsymbol{Z}_2}-\boldsymbol{P}_{\boldsymbol{Z}_1}$ is a projection matrix onto the column space spanned by $\widetilde{\boldsymbol{F}}^\star:=(\boldsymbol{I}_n-\boldsymbol{P}_{\boldsymbol{Z}_1})(\widehat{\boldsymbol{F}}_1,\cdots,\widehat{\boldsymbol{F}}_L)\in\mathbb{R}^{n\times K^\star}$, in which $K^\star=\sum_{i=1}^L K_i$. $\widetilde{\boldsymbol{F}}^\star:=(\boldsymbol{I}_n-\boldsymbol{P}_{\boldsymbol{Z}_1})(\widehat{\boldsymbol{F}}_1,\cdots,\widehat{\boldsymbol{F}}_L)\in\mathbb{R}^{n\times K^\star}$.

Similar with the proof of Lemma F.1, we are able to write $\boldsymbol{P}_{\mathcal{S}_M^\star} = \boldsymbol{W}^\star \boldsymbol{W}^{*\top}$, and we also obtain

$$\sup_{x>0} \left| \mathbb{P}(\mathcal{E}^\top \boldsymbol{P}_{\mathcal{S}_M^\star} \mathcal{E} \leqslant x | \boldsymbol{X}) - \mathbb{P}(\chi_K^{*2} \leqslant x) \right| \leqslant cK^{*1/4} \mathbb{E}|\varepsilon_t|^3 \max_{t\in[n]} \|\boldsymbol{w}_i^\star\| \sum_{t=1}^n \|\boldsymbol{w}_t^\star\|^2$$

with $\max_{t\in[n]} \|\boldsymbol{w}_i^\star\| = \max_{t\in[n]} \sqrt{\boldsymbol{P}_{S_M^\star,tt}}$.

Recall we have $\boldsymbol{P}_{S_M^*} := \boldsymbol{P}_{\boldsymbol{Z}_2} - \boldsymbol{P}_{\boldsymbol{Z}_1}$ and

$$\boldsymbol{Z}_2 = (\widehat{\boldsymbol{F}}_1^{(2)}, \widehat{\boldsymbol{U}}_{\widehat{S}_1}^{(2)}, \cdots, \widehat{\boldsymbol{F}}_L^{(2)}, \widehat{\boldsymbol{U}}_{\widehat{S}_L}^{(2)}, \widehat{\boldsymbol{F}}_{L+1}^{(2)}, \widehat{\boldsymbol{U}}_{\widehat{S}_{L+1}}^{(2)}, \cdots, \widehat{\boldsymbol{F}}_M^{(2)}, \widehat{\boldsymbol{U}}_{\widehat{S}_M}^{(2)}).$$

We first denote the event

$$\mathcal{A}_{Z_2} = \left\{ \lambda_{\min}(\boldsymbol{Z}_2^\top \boldsymbol{Z}_2) \geqslant \frac{n\lambda_0}{2} \right\},$$

in which λ_0 is a absolute constant. We know from (F.4), in order to prove Corollary F.3, we only need to prove **(i).** $\mathbb{E}[\max_{t\in[n]} \sqrt{\boldsymbol{P}_{S_M^\star,tt}} \mathbb{I}_{\mathcal{A}_{Z_2}}] \to 0$ and **(ii).** $\mathbb{P}(\mathcal{A}_{Z_2}^c) \to 0$.

Next, we prove the term **(i)**.

Since $\boldsymbol{P}_{\boldsymbol{Z}_1}$ is a projection matrix with $\boldsymbol{P}_{\boldsymbol{Z}_1,tt} > 0$ for any $t \in [n]$, we have

$$\max_{t\in[n]} \sqrt{\boldsymbol{P}_{S_M^\star,tt}} \leqslant \max_{t\in[n]} \sqrt{\boldsymbol{P}_{\boldsymbol{Z}_2,tt}}.$$

The next step is prove that $\mathbb{E}[\max_{t\in[n]} \sqrt{\boldsymbol{P}_{\boldsymbol{Z}_2,tt}} \mathbb{I}_{\mathcal{A}_{Z_2}}] \to 0$. We first prove that

$$\begin{aligned}
\max_{t\in[n]} \boldsymbol{P}_{\boldsymbol{Z}_2,tt} &\leqslant \frac{1}{\lambda_{\min}(\boldsymbol{Z}_2^\top \boldsymbol{Z}_2)} \max_{t\in[n]} \left(\sum_{i=1}^M |\boldsymbol{f}_t^{(i)}|^2 + \sum_{i=1}^M \sum_{\ell\in\widetilde{\mathcal{S}}_i} |\widehat{u}_{t\ell}|^2 \right) \\
&\leqslant \frac{1}{\lambda_{\min}(\boldsymbol{Z}_2^\top \boldsymbol{Z}_2)} \max_{t\in[n]} \sum_{i=1}^M |\boldsymbol{f}_t^{(i)}|^2 + \max_{t\in[n]} \sum_{i=1}^M \sum_{\ell\in\widetilde{\mathcal{S}}_i} |\widehat{u}_{t\ell}|^2
\end{aligned}$$

Leveraging similar techniques given in section F.3, we have

$$\mathbb{E}\left[\max_{t\in[n]} \sqrt{\boldsymbol{P}_{S_M^\star,tt}} \mathbb{I}_{\mathcal{A}_{Z_2}} \right] \leqslant \mathbb{E}\left[\max_{t\in[n]} \sqrt{\boldsymbol{P}_{\boldsymbol{Z}_2,tt}} \mathbb{I}_{\mathcal{A}_{Z_2}} \right] \to 0$$

Next, using certain concentration inequalities given in section F.3, we prove that there exists an positive constant λ_0 such that

$$\mathbb{P}(\mathcal{A}_{Z_2}^c) := \mathbb{P}\left(\frac{\lambda_{\min}(\boldsymbol{Z}_2^\top \boldsymbol{Z}_2)}{2} \leqslant \frac{n\lambda_0}{2} \right) \to 0.$$

Thus, we claim our conclusion for the first part of Corollary F.3.

We now prove (F.12). The proof details are the almost the same with the proof of Lemma F.2, so we only describe the outline here. We let

$$\mathcal{Z}_2 = (\widehat{\boldsymbol{F}}_1^{(2)}, \ldots, \widehat{\boldsymbol{F}}_M^{(2)}, \widehat{\boldsymbol{U}}_{\hat{\mathcal{S}}_1}^{(2)}, \ldots, \widehat{\boldsymbol{U}}_{\hat{\mathcal{S}}_M}^{(2)}).$$

Note that $\boldsymbol{P}_{\boldsymbol{Z}_2} = \boldsymbol{P}_{\mathcal{Z}_2}$. Hence it is equivalent to bound $\max_{t\in[n]} \boldsymbol{P}_{\mathcal{Z}_2,tt}$. Let

$$\mathcal{Z}_1 = (\boldsymbol{X}_{\hat{\mathcal{S}}_1}, \ldots, \boldsymbol{X}_{\hat{\mathcal{S}}_L}, \widehat{\boldsymbol{U}}_{\hat{\mathcal{S}}_{L+1}}, \ldots, \widehat{\boldsymbol{U}}_{\hat{\mathcal{S}}_M}, \widehat{\boldsymbol{F}}_{L+1}, \ldots, \widehat{\boldsymbol{F}}_M).$$

Then $\boldsymbol{P}_{\boldsymbol{Z}_2} - \boldsymbol{P}_{\boldsymbol{Z}_1} = \boldsymbol{P}_{\mathcal{Z}_2} - \boldsymbol{P}_{\mathcal{Z}_1}$, which is a projection matrix onto the column space of $(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{Z}_1})(\widehat{\boldsymbol{F}}_1, \ldots, \widehat{\boldsymbol{F}}_L) =: (\boldsymbol{I} - \boldsymbol{P}_{\mathcal{Z}_1})\widehat{\mathbb{F}}$. Here we denote $\widehat{\mathbb{F}} := (\widehat{\boldsymbol{F}}_1, \ldots, \widehat{\boldsymbol{F}}_L)$.

Under the alternative hypothesis, $\boldsymbol{Y} = \sum_{m=1}^{L}(\boldsymbol{F}_m\boldsymbol{\varphi}_m^\star + \boldsymbol{X}_m\boldsymbol{\beta}_m^\star) + \sum_{m=L+1}^{M}(\boldsymbol{F}_m\boldsymbol{\varphi}_m^\star + \boldsymbol{X}_m\boldsymbol{\beta}_m^\star) + \mathcal{E}$. We have,

$$\begin{aligned}\|(\boldsymbol{P}_{\mathcal{Z}_2} - \boldsymbol{P}_{\mathcal{Z}_1}) \sum_{m=L+1}^{M} (\boldsymbol{F}_m\boldsymbol{\varphi}_m^\star + \boldsymbol{X}_m\boldsymbol{\beta}_m^\star)\|^2 &= \|(\boldsymbol{P}_{\mathcal{Z}_2} - \boldsymbol{P}_{\mathcal{Z}_1}) \sum_{m=L+1}^{M} \boldsymbol{F}_m\boldsymbol{\varphi}_m^\star\| \\ &= o(1) + \|(\boldsymbol{P}_{\mathcal{Z}_2} - \boldsymbol{P}_{\mathcal{Z}_1}) \sum_{m=L+1}^{M} \widehat{\boldsymbol{F}}_m\boldsymbol{H}_m\boldsymbol{\varphi}_m^\star\| = o(1)\end{aligned}$$

and

$$\|(\boldsymbol{P}_{\mathcal{Z}_2} - \boldsymbol{P}_{\mathcal{Z}_1})\mathcal{E}\|^2 \sim \chi^2_{K^\star},$$

where $K^\star = \sum_{i=1}^{L} K_i$. According to the definition of $\mathcal{Z}_1$, we obtain

$$\begin{aligned}\mathcal{Z}_1 &= (\boldsymbol{X}_{\hat{\mathcal{S}}_1}, \ldots, \boldsymbol{X}_{\hat{\mathcal{S}}_L}, \widehat{\boldsymbol{U}}_{\hat{\mathcal{S}}_{L+1}}, \ldots, \widehat{\boldsymbol{U}}_{\hat{\mathcal{S}}_M}, \widehat{\boldsymbol{F}}_{L+1}, \ldots, \widehat{\boldsymbol{F}}_M) \\ &= (\widehat{\boldsymbol{U}}_{\hat{\mathcal{S}}_1}, \ldots, \widehat{\boldsymbol{U}}_{\hat{\mathcal{S}}_L}, \widehat{\boldsymbol{U}}_{\hat{\mathcal{S}}_{L+1}}, \ldots, \widehat{\boldsymbol{U}}_{\hat{\mathcal{S}}_M}, \widehat{\boldsymbol{F}}_{L+1}, \ldots, \widehat{\boldsymbol{F}}_M) \\ &\quad + (\widehat{\boldsymbol{F}}_1, \ldots, \widehat{\boldsymbol{F}}_L) \begin{pmatrix} \widehat{\boldsymbol{B}}_{1,\hat{\mathcal{S}}_1}^\top & \cdots & \cdots \\ & 0 & \\ & & \boldsymbol{O} \\ & & \widehat{\boldsymbol{B}}_{L,\hat{\mathcal{S}}_L}^\top \end{pmatrix} =: \widehat{\mathbb{U}} + \widehat{\mathbb{F}}\widehat{\mathbb{B}}^\top\end{aligned}$$

In addition, we let $\mathbb{F} := (\boldsymbol{F}_1, \ldots, \boldsymbol{F}_L)$ and $\mathbb{H} = (\boldsymbol{H}_1, \ldots, \boldsymbol{H}_L)$. We obtain

$$\begin{aligned}\|(\boldsymbol{P}_{\mathcal{Z}_2} - \boldsymbol{P}_{\mathcal{Z}_1}) \sum_{m=1}^{L} (\boldsymbol{F}_m\boldsymbol{\varphi}_m^\star + \boldsymbol{X}_m\boldsymbol{\beta}_m^\star)\|^2 &= \|(\boldsymbol{P}_{\mathcal{Z}_2} - \boldsymbol{P}_{\mathcal{Z}_1}) \sum_{m=1}^{L} \boldsymbol{F}_m\boldsymbol{\varphi}_m^\star\|^2 = \|(\boldsymbol{P}_{\mathcal{Z}_2} - \boldsymbol{P}_{\mathcal{Z}_1})\mathbb{F}\boldsymbol{\varphi}^\star\|^2 \\ &= o(1) + \|(\boldsymbol{P}_{\mathcal{Z}_2} - \boldsymbol{P}_{\mathcal{Z}_1})\widehat{\mathbb{F}}\mathbb{H}\boldsymbol{\varphi}^\star\|^2 = o(1) + (\mathbb{H}\boldsymbol{\varphi}^\star)^\top\widehat{\mathbb{F}}^\top(\boldsymbol{P}_{\mathcal{Z}_2} - \boldsymbol{P}_{\mathcal{Z}_1})\widehat{\mathbb{F}}(\mathbb{H}\boldsymbol{\varphi}^\star) \\ &= o(1) + (\mathbb{H}\boldsymbol{\varphi}^\star)^\top\widehat{\mathbb{F}}^\top(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{Z}_1})\widehat{\mathbb{F}}(\mathbb{H}\boldsymbol{\varphi}^\star) \\ &\asymp (\mathbb{H}\boldsymbol{\varphi}^\star)^\top\widehat{\mathbb{F}}^\top(\boldsymbol{I} - \widehat{\mathbb{F}}\widehat{\mathbb{B}}^\top(\widehat{\mathbb{B}}\widehat{\mathbb{F}}^\top\widehat{\mathbb{F}}\widehat{\mathbb{B}}^\top + \widehat{\mathbb{U}}^\top\widehat{\mathbb{U}})^{-1}\widehat{\mathbb{B}}\widehat{\mathbb{F}}^\top)\widehat{\mathbb{F}}(\mathbb{H}\boldsymbol{\varphi}^\star)\end{aligned}$$

$$\asymp (\mathbb{H}\boldsymbol{\varphi}^\star)^\top \widehat{\mathbb{F}}^\top (\boldsymbol{I} + \widehat{\mathbb{F}}\widehat{\mathbb{B}}^\top (\widehat{\mathbb{U}}^\top \widehat{\mathbb{U}})^{-1} \widehat{\mathbb{B}}\widehat{\mathbb{F}}^\top)^{-1} \widehat{\mathbb{F}} (\mathbb{H}\boldsymbol{\varphi}^\star)$$
$$\geqslant \frac{\|\widehat{\mathbb{F}}(\mathbb{H}\boldsymbol{\varphi}^\star)\|^2}{1 + \lambda_{\max}(\widehat{\mathbb{F}}\widehat{\mathbb{B}}^\top (\widehat{\mathbb{U}}^\top \widehat{\mathbb{U}})^{-1} \widehat{\mathbb{B}}\widehat{\mathbb{F}}^\top)} \geqslant \frac{\lambda_{\min}(\widehat{\mathbb{F}}^\top \widehat{\mathbb{F}}) \|\boldsymbol{\varphi}^\star\|^2}{1 + \frac{\lambda_{\max}(\widehat{\mathbb{B}}^\top \widehat{\mathbb{B}}) \lambda_{\max}(\widehat{\mathbb{F}}\widehat{\mathbb{F}}^\top)}{\lambda_{\min}(\widehat{\mathbb{U}}^\top \widehat{\mathbb{U}})}}.$$

The remaining proof steps follow the same procedure of Lemma F.2. □

References

S. Basu and G. Michailidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.

V. Bentkus. A lyapunov-type bound in R^d. *Theory of Probability & Its Applications*, 49(2):311–323, 2005.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

V. Chernozhukov, D. Chetverikov, and K. Kato. Central limit theorems and bootstrap in high dimensions. *Ann. Probab.*, 45(4):2309–2352, 2017.

V. Chernozhukov, D. Chetverikov, and Y. Koike. Nearly optimal central limit theorem and bootstrap approximations in high dimensions. *arXiv preprint arXiv:2012.09513*, 2020.

Y. Deshpande, A. Javanmard, and M. Mehrabi. Online debiasing for adaptively collected high-dimensional data with applications to time series analysis. *Journal of the American Statistical Association*, pages 1–14, 2021.

J. Fan, Y. Ke, and K. Wang. Factor-adjusted regularized model selection. *J. Econometrics*, 216(1): 71–85, 2020.

D. Giannone, M. Lenza, and G. E. Primiceri. Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89(5):2409–2437, 2021.

F. Han, H. Lu, and H. Liu. A direct estimation of high dimensional stationary vector autoregressions. *Journal of Machine Learning Research*, 2015.

A. Kneip and P. Sarda. Factor models and variable selection in high-dimensional regression analysis. *The Annals of Statistics*, 39(5):2410–2447, 2011.

Q. Li and L. Li. Integrative factor regression and its inference for multimodal data analysis. *J. Amer. Statist. Assoc.*, 113(521):1–15, 2021.

J. Lin and G. Michailidis. System identification of high-dimensional linear dynamical systems with serially correlated output noise components. *IEEE Transactions on Signal Processing*, 68:5573–5587, 2020.

P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.*, 40(3):1637–1664, 2012.

R. P. Masini, M. C. Medeiros, and E. F. Mendes. Regularized estimation of high-dimensional vector autoregressions with weakly dependent innovations. *Journal of Time Series Analysis*, 43(4):532–557, 2022.

Q. Sun, W.-X. Zhou, and J. Fan. Adaptive Huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 2014.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.

K. C. Wong, Z. Li, and A. Tewari. Lasso guarantees for β-mixing heavy-tailed time series. *The Annals of Statistics*, 48(2):1124–1142, 2020.

W.-B. Wu and Y. N. Wu. Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics*, 10(1):352–379, 2016.

Y. Yan, Y. Chen, and J. Fan. Inference for heteroskedastic pca with missing data. *arXiv preprint arXiv:2107.12365*, 2021.

P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.