Are Latent Factor Regression and Sparse Regression Adequate?

Jianqing Fan\textsuperscript{ab}, Zhipeng Lou\textsuperscript{b}, and Mengxin Yu\textsuperscript{b}\textsuperscript{c}

\textsuperscript{a}School of Data Science, Fudan University, Shanghai, China; \textsuperscript{b}Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ

ABSTRACT

We propose the Factor Augmented (sparse linear) Regression Model (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 Factor-Adjusted deBiased 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. Supplementary materials for this article are available online.

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, Liao, and Mincheva 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 $x$ follows from the model

$$ x = Bf + u, $$

(1.1)

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

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, Xue, and Yao 2017; Bing, Bunea, and Wegkamp 2019; Bunea, Strimas-Mackey, and Wegkamp 2020; Bing et al. 2021). It assumes that the latent factors drive both dependent and independent variables as follows:

$$ Y = f^\top y + \varepsilon, $$

$$ x = Bf + u. $$

(1.2)

Here $Y$ is the response variable and $\varepsilon \in \mathbb{R}$ is the random noise which is independent with the factor $f$. When the factors are unobserved, one usually learns the latent factors based on observed $x$ 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, Liao, and Mincheva 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 $f$, 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:

$$ Y = x^\top \beta + \varepsilon. $$

(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 $\beta \in \mathbb{R}^d$ 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 2008).
2. Second, it origins from the factor regression given in (1.2). In regression model (1.2) is a special case of (2.1) in which equivalent form, vectors quantifying the contribution of the latent factor \( f \) of the latent factor \( 1 \). First, it origins from Fan, Ke, and Wang (2020). In order comes from two perspectives. sparseregression.

In this article, we introduce the Factor Augmented (sparse linear) Regression Model (FARM) (1.4), which incorporates both the latent factor and the idiosyncratic component into the covariates,

\[
Y = f^\top \gamma^* + u^\top \beta^* + \varepsilon, \\
x = Bf + u, 
\] (1.4)

where \( \gamma^* \in \mathbb{R}^K \) and \( \beta^* \in \mathbb{R}^d \) are population parameter vectors quantifying the contribution of the latent factor \( f \) and the idiosyncratic component \( u \), respectively. Obviously, the factor regression model (1.2) is a special case of (2.1) in which \( \beta^* = 0 \). 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,

\[
Y = f^\top \phi^* + x^\top \beta^* + \varepsilon, \\
x = Bf + u, 
\] (1.5)

where \( \phi^* = \gamma^* - B^\top \beta^* \in \mathbb{R}^K \) quantifies the extra contribution of the latent factor \( f \) beyond the observed predictor \( x \). Therefore, FARM expands the space spanned by \( x \) into useful directions spanned by \( f \). It is clear that the sparse regression model (1.3) is also a special case of (1.4) with \( \phi^* = 0 \). 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. First, it origins from Fan, Ke, and Wang (2020). In order to get precise estimation of \( \beta^* \) based on highly correlated variables, they study the sparse regression estimation by substituting (1.1) into (1.3) and obtain

\[
Y = (Bf + u)^\top \beta^* + \varepsilon = f^\top (B^\top \beta^*) + u^\top \beta^* + \varepsilon. 
\] (1.6)

We observe from (1.6), when the sparse linear regression is adequate, for a given \( \beta^* \), the regression coefficient on \( f \) is fixed at \( \gamma^* = B^\top \beta^* \). 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 \( B^\top \beta^* \). 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 \( x \) into useful directions.

2. Second, it origins from the factor regression given in (1.2). In reality, the leading common factors \( f \) 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 \( u \) 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 \( u \) into the original factor regression.

In this article, 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 \( \ell_1 \)-penalty, we derive the \( \ell_2 \)-consistency results for parameter vectors \( \gamma^* \) and \( \beta^* \). 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 functions. For instance, quantile regression (Belloni and Chernozhukov 2011; Fan, Fan, and Barut 2014), support vector machine (Zhang et al. 2016; Peng, Wang, and Wu 2016), Huber regression (Fan, Li, and Wang 2017; Sun, Zhou, and Fan 2020), generalized linear model (Van de Geer 2008; Fan, Ke, and Wang 2020) and many other variants. In order to demonstrate the general applicability of our proposed methods, in our article, 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 \( \ell_1 \)-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 \( f \) given in our model is able to be treated as the unobserved confounding variables which affect both the covariate \( x \) 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 (Tibshirani 1996; Fan and Li 2001; Stock and Watson 2002; Zou 2006; Bai and Ng 2006; Zhang 2010; Fan, Xue, and Yao 2017; Fan, Ke, and Wang 2020; 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

\[
H_0 : Y = f^\top \gamma^* + \varepsilon \text{ versus } H_1 : Y = f^\top \gamma^* + u^\top \beta^* + \varepsilon. 
\] (1.7)

This amounts to testing \( H_0 : \beta^* = 0 \) under FARM model. To this end, we propose the Factor-Adjusted deBiased Test statistic (FabTest) \( \hat{\beta}_2 \), which serves as a de-sparsify version of the estimator \( \hat{\beta}_2 \), obtained under \( \ell_1 \)-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 \( \beta^* \) by following similar de-biasing procedure.
For validating the adequacy of the sparse linear regression model, we consider testing the hypotheses
\[ H_0 : Y = x^T \beta^* + \varepsilon \quad \text{versus} \quad H_1 : Y = f^T \varphi^* + x^T \beta^* + \varepsilon, \]
(1.8)
or \( \varphi^* = 0 \) 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 Asymptotic null distribution and the power of the test statistic of variables which cope well the curse of high dimensionality. For validating the adequacy of the factor regression, we consider testing the hypotheses
\[ H_0 : \gamma_1 = \ldots = \gamma_m = 0 \quad \text{versus} \quad H_1 : \gamma_1 \neq \ldots \neq \gamma_m \neq 0, \]
(1.9)
and investigate the corresponding statistical properties. Suppose \( \gamma_0 \in \mathbb{R}^m \) and \( \gamma_1 \in \mathbb{R}^m \), for our parameters of interest. Comparing with those closely related literature (Fan, Ke, and Wang 2020; Fan, Masini, and Medeiros 2021), 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, Zhou, and Fan (2020) to a more 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 \( \ell_1 \)-penalized adaptive Huber estimation, we establish statistical estimation results for our parameters of interest. Comparing with those closely related literature (Fan, Ke, and Wang 2020; Fan, Masini, and Medeiros 2021), 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, Zhou, and Fan (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 multmodal sparse linear regression model (Li and Li 2021), by testing whether the sparse linear regression for some given modes 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 \( \mathbf{y} = (y_1, \ldots, y_m)^T \in \mathbb{R}^m \), we denote its \( \ell_q \) norm as \( ||\mathbf{y}||_q = (\sum_{i=1}^{m} |y_i|^q)^{1/q}, 1 \leq q < \infty \), and write \( ||\mathbf{y}||_\infty = \max_i |y_i| \). For any integer \( m \), we denote \( [m] = \{1, \ldots, m\} \). The sub-Gaussian norm of a scalar random variable \( Z \) is defined as \( ||Z||_{\psi_2} = \inf \{ t > 0 : \mathbb{E} \exp(Z^2/t^2) \leq 2 \} \). For a random vector \( \mathbf{x} \in \mathbb{R}^m \), we use \( ||\mathbf{x}||_{\psi_2} = \sup_{||\mathbf{w}||_2 = 1} ||\mathbf{w}^T \mathbf{x}||_{\psi_2} \) to denote its sub-Gaussian norm. Let \( I(\cdot) \) denote the indicator function and let \( I_K \) denotes the identity matrix in \( \mathbb{R}^{K \times K} \). For a matrix \( \mathbf{A} = [A_{jk}] \), we define \( ||\mathbf{A}||_F = \sqrt{\sum_{j,k} A_{jk}^2} \), \( ||\mathbf{A}||_{max} = \max_j \max_k |A_{jk}| \) and \( ||\mathbf{A}||_\infty = \max_j \sum_k |A_{jk}| \) to be its Frobenius norm, element-wise max-norm and matrix \( \ell_\infty \)-norm, respectively. Moreover, we use \( \lambda_{\min}(\mathbf{A}) \) and \( \lambda_{\max}(\mathbf{A}) \) to denote the minimal and maximal eigenvalues of \( \mathbf{A} \), respectively. We use \( |\mathcal{A}| \) to denote the cardinality of set \( \mathcal{A} \). For two positive sequences \( \{a_n\}_{n \geq 1}, \{b_n\}_{n \geq 1} \), we write \( a_n = O(b_n) \) if there exists a positive constant \( C \) such that \( a_n \leq C \cdot b_n \) and we write \( a_n = o(b_n) \) if \( a_n/b_n \to 0 \). In addition, \( a_n = O_p(b_n) \) and \( a_n = o_p(b_n) \) 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 article 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. 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 \( n \) (iid) random samples \( \{(x_t, Y_t)\}_{t=1}^n \) from \( (x, Y) \), which satisfy that
\[
x_t = B f_t + u_t \quad \text{and} \quad Y_t = f_t^T \gamma^* + u_t^T \beta^* + \varepsilon_t, \quad t = 1, \ldots, n, \]
(2.1)
where \( f_1, \ldots, f_n \in \mathbb{R}^K, u_1, \ldots, u_n \in \mathbb{R}^d \) and \( \varepsilon_1, \ldots, \varepsilon_n \in \mathbb{R} \) are iid realizations of \( f, u \) and \( \varepsilon \), respectively. To ease the presentation, we rewrite (2.1) in a more compact matrix form as follows,
\[
X = FB^T + U, \quad Y = FY^* + U^* + \varepsilon, \]
(2.2)
where $X = (x_1, \ldots, x_n)^\top$, $F = (f_1, \ldots, f_n)^\top$, $U = (u_1, \ldots, u_n)^\top$, $Y = (Y_1, \ldots, Y_n)^\top$ and $E = (e_1, \ldots, e_n)^\top$. Throughout the whole article, we assume only get access to observations $\{(x_t, Y_t)\}_{t=1}^n$. Both the latent factors $F$ and the idiosyncratic components $U$ are unobserved and need to be estimated from the observed predictors $X$. Thus, in the following, we shall first illustrate how to estimate $F$ and $U$ and then proceed with the regularized estimation for model (2.2).

### 2.1. Factor Estimation

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

$$
cov(f) = I_K \text{ and } B^\top B \text{ is diagonal.}
$$

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

$$
(\hat{F}, \hat{B}) = \arg\min_{F \in \mathbb{R}^{n \times K}, B \in \mathbb{R}^{d \times K}} ||X - FB^\top||_F^2
$$

subject to $\frac{1}{n} F^\top F = I_K$ and $B^\top B$ is diagonal.

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

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

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

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

**Assumption 2.3.** Let $\Sigma = \text{cov}(u)$. There exists a constant $\Upsilon > 0$ such that $||B||_{\max} \leq \Upsilon$ and $\mathbb{E}|u^\top u - \text{tr}(\Sigma)|^4 \leq \Upsilon d^2$.

**Assumption 2.4.** There exist a positive constant $\kappa < 1$ such that $\kappa \leq \lambda_{\min}(\Sigma), ||\Sigma||_1 \leq 1/\kappa$ and $\min_{1 \leq k \leq d} \text{var}(u_k u_k^\top) \geq \kappa$.

**Remark 1.** Assumptions 2.1–2.4 are standard assumptions in the studies of large dimensional factor model. We refer to Bai (2003), Fan, Liao, and Mincheva (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 $H = n^{-1}V^{-1}\hat{F}F^\top B$, where $V \in \mathbb{R}^{K \times K}$ is a diagonal matrix consisting of the first $K$ largest eigenvalues of the matrix $n^{-1}XX^\top$. Then, under Assumptions 2.1–2.4, we have

1. $||\hat{F} - FH^\top||_F^2 = O_p(n/d + 1/n)$.
2. For any $I \subset \{1, 2, \ldots, d\}$, we have $\max_{\ell \in I} \sum_{i=1}^n |\hat{u}_{\ell} - u_{\ell}|^2 = O_p(\log |I| + n/d)$.
3. $||H^\top H - I_K||_F^2 = O_p(1/n + 1/d)$.
4. $\max_{\ell \in \{1, 2, \ldots, d\}} ||\hat{b}_\ell - Hb_\ell||_2^2 = O_p((\log d)/n)$.

**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, Guo, and Zheng (2022)). Our theories always work as long as we replace $K$ by any consistent estimator $\hat{K}$, that is, we only require $\mathbb{P}(\hat{K} = K) \to 1$, as $n \to \infty$.

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

$$
\hat{K} = \arg\max_{K \leq K} \frac{\lambda_k(XX^\top)}{\lambda_{k+1}(XX^\top)},
$$

where $1 \leq K \leq n$ 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 $\beta^\star$ is sparse. Then the regularized estimator for the unknown parameter vectors $\beta^\star$ and $\gamma^\star$ of our factor augmented linear model is defined as follows:

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

(2.4)

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

We let $\hat{Y} = (I_n - \hat{P})X$ and $\hat{U} = (I_n - \hat{P})X$. Hence, $\hat{F}^\top \hat{U} = 0$ and it is straightforward to verify that the solution of (2.4) is equivalent to

$$
\hat{\beta}_\lambda = \arg\min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|\hat{Y} - \hat{U}\beta\|_2^2 + \lambda \|\beta\|_1 \right\},
$$

$$
\hat{\gamma} = (\hat{F}^\top \hat{F})^{-1}\hat{F}^\top Y = \frac{1}{n} \hat{F}^\top Y.
$$
For any subset \( S \) of \([1, \ldots, d]\), we define the convex cone
\[ C(S, 3) = \{ \delta \in \mathbb{R}^d : ||\delta_S||_1 \leq 3||\delta_S||_1 \} \]
For simplicity of notation, we write
\[ \mathcal{V}_{n,d} = \frac{n}{d} + \sqrt{\frac{\log d}{n}} + \sqrt{n \log d}. \tag{2.5} \]
To investigate the consistency property of \((\hat{\beta}, \hat{y})\), we impose the following moment condition on the random noise \( \epsilon \).

**Assumption 2.5.** There exists a positive constant \( c_1 < \infty \) such that \( ||\epsilon||_{\psi_2} \leq c_1 \).

**Theorem 2.2.** Recall \( \varphi^* = \gamma^* - B^T \beta^* \in \mathbb{R}^K \). Under Assumptions 2.1–2.5, we have
\[ ||\hat{y} - H\gamma^*||_2 = O_P \left( \frac{1}{\sqrt{n}} + \left( \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{d}} \right) ||\varphi^*||_2 \right) \]
\[ + ||\beta^*||_1 \left( \frac{\log |S_*|}{n} + \frac{1}{\sqrt{d}} \right) \right) \right), \]
where \( S_* = \{ j \in [d] : \beta^*_j \neq 0 \} \) and \( |S_*| \) is its cardinality. Furthermore, if \( |S_*| \left( \frac{\log d}{n} + \frac{1}{\sqrt{n}} \right) = o(1) \), then, by taking \( \lambda = (I_0/n)||\hat{U}^T(\hat{Y} - \hat{U}\beta^*)||_\infty \) for some constant \( I_0 \geq 2 \), we have
\[ \hat{\beta}_\lambda - \beta^* \in C(S_*, 3) \]
\[ ||\hat{\beta}_\lambda - \beta^*||_2 = O_P \left( \frac{|S_*| \log \log d}{n} + \mathcal{V}_{n,d}||\varphi^*||_2 \sqrt{|S_*|} \right) \tag{2.6} \]

**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 \( x \) is a sub-Gaussian random vector with bounded sub-Gaussian norm \( ||x||_{\psi_2} \). 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 \( f \) and \( u \) are Gaussian random vectors and the underlying \( x \) satisfies the factor model (1.1). Then \( x \) is also a Gaussian random vector with \( \text{cov}(x) = BB^T + \Sigma \). Under the pervasiveness condition (Assumption 2.2) and Assumption 2.4, it is straightforward to verify that \( ||x||_{\psi_2} = \sqrt{3/2} \lambda_{\max}(BB^T + \Sigma) \propto d \), which violates the assumption on bounded sub-Gaussian norm. In contrast, our model can circumvent such issue because we decompose the covariate \( x \) into \((f, u)\), and we only need impose sub-Gaussian assumption on \((f, u)\). 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, Ke, and Wang (2020) with weaker assumptions. First, we did not impose the irrepresentable condition on the design matrix \( U \), only the lower bound on \( \Sigma = \text{cov}(u) \) is required. In addition, although Fan, Ke, and Wang (2020) also decompose the covariate \( x \) into \((f, u)\) in order to get precise estimator for \( \beta^* \), they mainly focus on the linear model \( Y = x^T \beta^* + \epsilon \) which corresponds to the special case with \( \varphi^* = 0 \) in our results given in Theorem 2.2.

**Remark 5.** Our study is different from the related work by Fan, Masini, and Medeiros (2021), 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. Second, 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, Zhou, and Fan 2020).

Let \( \rho_{\omega}() \) denote the Huber function,
\[ \rho_{\omega}(z) = \begin{cases} \frac{z^2}{2}, & \text{if } |z| \leq \omega, \\ \omega(z^2 - \omega_1^2), & \text{if } |z| > \omega, \end{cases} \]
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 \((\beta^*, \gamma^*)\) is given by
\[ (\hat{\beta}_h, \hat{y}_h) = \arg \min_{\beta \in \mathbb{R}^d, \gamma \in \mathbb{R}^K} \left\{ \frac{1}{n} \sum_{t=1}^n \rho_{\omega}(y_t - \hat{u}_t^T \beta - \hat{f}_t^T \gamma) + \lambda ||\beta||_1 \right\}, \tag{2.7} \]
where \( \lambda > 0 \) is a tuning parameter and \( \omega \) depends on sample size, dimension, and noise level. For simplicity of notation, we write \( \hat{\varphi}_h = (\hat{\beta}_h^T, \hat{y}_h^T) \in \mathbb{R}^{d+K} \) and \( \hat{\varphi} = (\beta^T, \gamma^T) \in \mathbb{R}^{d+K} \), where \( \hat{\beta} = \hat{\beta}^* + n^{-1}\hat{F}\varphi^* \). The following theorem establishes the statistical consistency of \( \hat{\varphi}_h \).

**Proposition 2.3.** Assume that \( \mathbb{E}|\epsilon|^{1+\theta} < \infty \) for some constant \( \theta > 0 \). Let
\[ \omega \asymp \left( \frac{n}{\log d} \right)^{-\frac{1}{1+\theta}} \quad \text{and} \quad \lambda \asymp \left( \frac{\log d}{n} \right)^{-\frac{\theta}{1+\theta}}. \]
Furthermore, we assume that \((|S_*| + K)(\log d)^{1/2} = o(n),\)
\[ \frac{\log n}{n + \sqrt{\log d}} ||\varphi^*||_2 = o(\omega) \quad \text{and} \quad \mathcal{V}_{n,d}||\varphi^*||_2 = O(\omega \log d). \tag{2.8} \]
Then, under Assumptions 2.1–2.4, we have
\[ ||\hat{\varphi}_h - \hat{\varphi}||_1 = O_P \left( \left( |S_*| + K \right) \left( \frac{\log d}{n} \right)^{-\frac{\theta}{1+\theta}} \right) \right). \]

We establish the \( \ell_1 \)-statistical rate for the parameters in model (1.4) [also (1.5)] by only assuming the existence of \((1 + \theta)\)th moment of the noise distribution. Specifically, when \( \theta \geq 1 \), the results reduce to the same rate as the sub-Gaussian assumption of \( \epsilon \). Our result serves as an extension of Sun, Zhou, and Fan (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 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 heavy-tailed noises constant. As our learned covariates contain statistical errors, thus, we first introduce a de-biased version of $\hat{\beta}_\lambda$ given in (2.4).

3.1. Bias Correction

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

$$\hat{\beta}_{\lambda} = \hat{\beta}_\lambda + \frac{1}{n} \hat{\Theta} \hat{U}^\top (Y - \hat{U} \hat{\beta}_\lambda).$$

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

$$\hat{\beta}_{\lambda} - \beta^* = \frac{1}{n} \hat{\Theta} \hat{U}^\top \varepsilon + \frac{1}{n} \hat{\Theta} \hat{U}^\top F \phi^* + (I_d - \hat{\Theta} \Sigma) (\hat{\beta}_{\lambda} - \beta^*),$$

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

One observes that constructing the de-biased estimator $\hat{\beta}_{\lambda}$ given above requires an estimator $\hat{\Theta}$. 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, Liu, and Luo (2011), Javanmard and Montanari (2014), and Avella-Medina et al. (2018). In our work, we do not restrict $\hat{\Theta}$ to be any specific one, but require to satisfy the following general conditions.

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

$$||I_d - \hat{\Theta} \Sigma||_{\max} = O_p(\Lambda_{\max}) \text{ and } ||\hat{\Theta} - \Theta||_{\infty} = O_p(\Delta_{\infty}).$$

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

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

$$\Lambda_{\max} = O\left(\sqrt{\frac{\log d}{n} + \frac{1}{d}}\right) \text{ and } \Delta_{\infty} = O\left(\max_{j|d} |S_j| \sqrt{\frac{\log d}{n} + \frac{1}{d}}\right),$$

by using node-wise regression (Zhang and Zhang 2014; van de Geer et al. 2014), where $|S_j| = \sum_{k=1}^d I(\Theta_{jk} \neq 0)$ quantifies the sparsity of $j$th column of the precision matrix $\Theta$ for each $1 \leq j \leq d$. In Appendix E.1, we will provide a detailed analysis on estimating $\Sigma^{-1}$ 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 $||\hat{\beta}_j - \beta^*||_{\infty}$ in the high dimensional setting. To this end, we apply the Gaussian approximation result given in Chernozhukov, Chetverikov, and Kato (2013, 2017), Chernozhukov, Chetverikov, and Koike (2020) for high dimensional random vectors. More specifically, we let $Z = (Z_1, \ldots, Z_d)^\top \in \mathbb{R}^d$ be a zero-mean Gaussian random vector with the same covariance matrix as that of $n^{-1/2} \Theta U^\top \varepsilon$, that is,

$$\text{cov}(Z) = \text{cov}\left(\frac{1}{\sqrt{n}} \Theta U^\top \varepsilon\right) = \sigma^2 \Theta.$$ (3.4)

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

Theorem 3.1. Recall $\varphi^* = \psi^* - B^\top \beta^* \in \mathbb{R}^K$. We assume that $(\log d)^3/n \to 0$,

$$(\Lambda_{\max} |\varphi^*|_{\infty} + \Delta_{\infty}) \log d \to 0 \text{ and } \left(\frac{\log n}{\log d} + \frac{1}{\log d}\right) ||\Theta||_{\infty} \log d \to 0, \quad \text{(3.5)}$$

with $\Lambda_{\max}$. Then under Assumption 3.1, we have

$$\sup_{x > 0} \left|\mathbb{P}\left(\sqrt{n}||\hat{\beta}_j - \beta^*||_{\infty} \leq x\right) - \mathbb{P}\left(||Z||_{\infty} \leq x\right)\right| \to 0.$$ 

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

$$\psi_{\infty, \alpha} = \mathbb{P}\left(\sqrt{n}||\hat{\beta}_j - \beta^*||_{\infty} > c_{1-\alpha}\right).$$ (3.6)

3.3. Gaussian Multiplier Bootstrap

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

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

$$\hat{L} = \frac{1}{\sqrt{n}} ||\hat{\Theta} \hat{U}^\top \xi||_{\infty}, \text{ where } \xi = (\xi_1, \xi_2, \ldots, \xi_n)^\top.$$
2. Repeat the first step independently for \( B \) times and obtain \( \tilde{L}_1, \ldots, \tilde{L}_B \). Estimate the critical value \( c_{1-\alpha} \) via 1 – \( \alpha \) quantile of the empirical distribution of the bootstrap statistics:

\[
\hat{c}_{1-\alpha} = \inf \{ t \geq 0 : H_B(t) \geq 1 - \alpha \},
\]

where \( H_B(t) = \frac{1}{B} \sum_{b=1}^B \mathbb{I} \{ \tilde{L}_b \leq t \} \).

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

**Assumption 3.2.** There exists a \( \Delta_{\max} > 0 \) such that \( \| \Theta - \Theta \|_{\max} = O_{\text{p}}(\Delta_{\max}) \).

**Assumption 3.3.** There exists a \( 0 < \Delta_{\sigma} \leq 1 \) such that \( |\hat{\sigma}/\sigma - 1| = O_{\text{p}}(\Delta_{\sigma}) \).

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

Let \( P^* (\cdot) = P(\cdot | X, Y) \) 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

\[
\Lambda_{\max} \| \Theta \|_{\infty} + \Delta_{\max} + \Delta_{\sigma} = o \left( \frac{1}{\log d} \right).
\]

Then, under conditions of Theorem 3.1, we have

\[
sup_{x > 0} \left| P \left( \sqrt{n} \| \hat{\beta}_j - \beta^* \|_\infty / \hat{\sigma} \leq x \right) - P^* \left( \hat{L} \leq x \right) \right| \rightarrow 0.
\]

**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, Bühlmann, and Zhang 2017) simultaneous confidence intervals for \( \beta^* \). For each \( 1 \leq j \leq d \), a (1 - \( \alpha \))-confidence interval for \( \beta^*_j \) is given by

\[
CI_{\alpha}(\beta^*_j) = \left\{ \hat{\beta}_{j,\lambda} - \hat{\sigma} z_{1-\alpha/2} / \sqrt{n}, \hat{\beta}_{j,\lambda} - \hat{\sigma} z_{1-\alpha/2} / \sqrt{n} \right\},
\]

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

\[
H_{0,G} : \beta^*_j = \beta^*_j \text{ for all } j \in G \quad \text{versus} \quad H_{1,G} : \beta^*_j \neq \beta^*_j \text{ for some } j \in G.
\]

In particular, when \( \beta^*_j = 0 \) for all \( j \in G \), this reduces to testing the significance of a group of parameters. We obtain that the asymptotic distribution of \( \max_{j \in G} |Z_j| \) 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

\[
H_0 : Y = x^T \beta^* + \varepsilon \quad \text{versus} \quad H_1 : Y = f^T \varphi^* + x^T \beta^* + \varepsilon,
\]

which is equivalent to test whether \( \varphi^* = y^* - B^T \beta^* = 0 \). Since \( B \) is an unknown dense matrix, simultaneously testing this linear equation will suffer from the curse of dimensionality.

On the other hand, for any set \( S \subset \{d\} \) with \( S_* \subset S \), we have \( B_S^T \beta_S^* = B^T \beta^* \). Hence, it suffices to compare the following two linear models in reduced dimension:

\[
H_0 : Y = x_S^T \beta_S^* + \varepsilon \quad \text{versus} \quad H_1 : Y = f^T \varphi^* + x_S^T \beta_S^* + \varepsilon.
\]

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, Li, and Wu 2014; Barut, Fan, and Verhasselt 2016; Chu, Liu, and Reimherr 2016; Wang and Leng 2016).

We propose an ANOVA-type test for (4.1) with two stages. In the first stage, the dataset is split into two datasets \( (Y^{(1)},X^{(1)}) \) and \((Y^{(2)},X^{(2)})\), with sample sizes \( m \) and \( n - m \), respectively. We use \((Y^{(1)},X^{(1)})\) to screen variables. Let \( \tilde{S}_1 \) denote the set of variables selected. In the second stage, we leverage the selected \( \tilde{S}_1 \) and remaining data \((Y^{(2)},X^{(2)})\) to perform hypothesis testing based on the ANOVA-type test statistic for low-dimensional model (4.2) with \( S \) replaced by \( \tilde{S}_1 \). 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 \( \tilde{S}_1 \).

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

\[
P \left( |\tilde{S}_1| \leq s_n \text{ and } S_* \subset \tilde{S}_1 \right) \rightarrow 1, \text{ as } n \rightarrow \infty.
\]
A simple procedure that satisfies the above assumption is the following factor-adjusted marginal screening based on the data \((Y^{(1)}, X^{(1)})\).

1. Estimation. Compute the latent factor estimator \(\hat{F}^{(1)}\), idiosyncratic component \(\hat{U}^{(1)}\) based on \(X^{(1)}\), and \(\hat{Y}^{(1)} = (I_m - F^{(1)T}F^{(1)})F^{(1)T}Y^{(1)}\), the residual after factor regression.

2. Marginal regression. Compute the least square estimate \(\hat{\beta}_{\ell,M} = \hat{U}^{(1)T}\hat{Y}^{(1)}/(\hat{U}^{(1)T}\hat{U}^{(1)})\) for each \(1 \leq \ell \leq d\).

3. Screening. Let \(\hat{S}_\phi := \{\ell \in [d] : |\hat{\beta}_{\ell,M}| > \phi\}\) for some prescribed \(\phi > 0\).

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

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

\[
\mathcal{O}\left(\frac{||\phi||^*}{d} + \frac{||\beta^*||}{\sqrt{d/m}}\right) \leq \phi \leq \mathcal{O}\left(\min_{\ell \in [d]} |\beta^*|_{\ell,M}\right).
\]

where \(\beta^*_{\ell,M} = \Sigma^T_{\ell} \beta^* / \Sigma_{\ell}\). Here \(\Sigma_{\ell}\) denotes the \(\ell\)th column of \(\Sigma\). Then, under the Assumptions 2.1–2.5, we have

\[
\mathbb{P}\left(S \subset \hat{S}_\phi\right) \to 1, \text{ as } m \to \infty.
\]

Furthermore, we assume that \(\min_{\ell \in S_\phi} |\beta^*_{\ell,M}| \geq c_4 m^{-k}\) for some positive constant \(k < 1/2\). Then for any \(\phi = c_4 m^{-k}\) with \(c_4 \leq c_5/(1 + \delta)\), we have

\[
\mathbb{P}\left(|\hat{S}_\phi| \leq \frac{c_6^2 m^{2k}}{\lambda_{\min}(\Sigma)(1 - \delta)^2}\right) \to 1 \text{ as } m \to \infty.
\]

**Remark 10.** From the conclusion of Proposition 4.1, we obtain sure screening property by using our first dataset with sample size \(m = n^a\) for some \(a < 1\) as long as the signal satisfies \(\min_{\ell \in S_\phi} |\beta^*_{\ell,M}| \geq c_4 m^{-k}\). Thus, the size of the remaining dataset for constructing the test statistic in our second step is \(n - n^a \approx n\). 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, Ke, and Wang (2020) proposed a similar sure screening estimator which is a special case of our Proposition 4.1 with \(\phi^* = \gamma^* - B^T \hat{\beta}^* = 0\). Moreover, we also provide an upper bound for the number of selected variables whereas Fan, Ke, and Wang (2020) 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 \(S\) replaced by \(\hat{S}_1\), which is given by

\[
Q^{(2)}(\alpha) = \left\|I_{n-m} - P_{\hat{S}_1}^{(2)}\right\|_2^2 - \left\|I_{n-m} - P_{\hat{F}^{(2)}_{\phi} - P_{U_{\hat{S}_1}^{(2)}}}\right\|_2^2.
\]

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

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

\[
s_n \left(\log d + \frac{1}{d}\right) \to 0 \text{ and } \Delta_n \to 0.
\]

We obtain

\[
\sup_{x > 0}\left|\mathbb{P}\left(Q_n^{(2)} \leq x \hat{d}^2 \mid H_0\right) - \mathbb{P}(X_K^2 \leq x)\right| \to 0, \text{ as } n \to \infty.
\]

**Theorem 4.2** yields a level \(\alpha\) test for (4.1), with critical region

\[
\left\{Q_n^{(2)} > \hat{d}^2 \chi_{K,1-\alpha}^2\right\},
\]

where \(\chi_{K,1-\alpha}^2\) is the \((1 - \alpha)\) quantile of \(\chi_K^2\) distribution.

**Remark 12.** Under stronger conditions such as irrepresentable condition (Zhou and Yu 2006) or RIP condition (Candes and Tao 2007), the \(\hat{S}\) achieved by certain explicit regularization (Zhou and Yu 2006; Fan and Li 2011; Shi et al. 2019; Fan, Ke, and Wang 2020) or implicit regularization accompanied with early stopping and signal truncation (Zhao, Yang, and He 2019; Fan, Yang, and Yu 2021) enjoys variable selection consistency \(\mathbb{P}(\hat{S} = S) \to 1\). In this scenario, we take the test statistic as

\[
Q_n = \left\|(P_{\hat{F}} + P_{\hat{U}_\phi} - P_{X_{\hat{S}}}) Y^{(2)}\right\|_2^2
\]

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

\[
\sup_{x > 0}\left|\mathbb{P}\left(Q_n \leq x \hat{d}^2 \mid H_0\right) - \mathbb{P}(X_K^2 \leq x)\right| \to 0,
\]

by following similar proof idea with Theorem 4.2.

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

**Theorem 4.3.** Define

\[
D(\alpha, \theta) = \left\{\varphi \in \mathbb{R}^K : \frac{n||\varphi||^2}{1 + K \sqrt{n}||B||_{\max}/\lambda_{\min}(\Sigma)} \geq \sigma^2(2 + \delta)(\chi_{K,1-\alpha}^2 + \chi_{K,1-\theta}^2)\right\},
\]

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 \(\theta\) is a threshold such that for any \(\varphi^* \in D(\alpha, \theta)\), the power of the test is larger than \(1 - \theta\). To be more specific, we assume that

\[
||\varphi^*||_2(\sqrt{n/d} + 1/\sqrt{n}) \to 0.
\]

Then, under the conditions of Theorem 4.2, we have

\[
\inf_{\varphi^* \in D(\alpha, \theta)} \mathbb{P}(\psi_a = 1 \mid H_1) \geq 1 - \theta,
\]

where \(\psi_a = \frac{Q_n^{(2)} - \hat{d}^2 \chi_{K,1-\alpha}^2}{\chi_{K,1-\alpha}^2}\).

**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 Section F.4, supplementary materials for more details.
5. Numerical Studies

5.1. Accuracy of Estimation

For data generation, we let number of factors $K = 2$, dimension of covariate $d = 1000$, $\gamma^* = (0.5, 0.5)$, the first $s = 3$ entries of $\beta^*$ be 0.5 and remaining $d - s$ entries be 0. Throughout this section, we generate every entry of $F, U$ from the standard Gaussian distribution and let every entry of $B$ be generated from the uniform distribution $\text{Unif}(-1, 1)$. We choose the noise distribution of $e$ 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. Using Lasso directly on measurements $(X, Y)$ leads to much worse results due to the inadequacy of the model. In addition, as shown in Fan, Ke, and Wang (2020), 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 reduced sample sizes help reduce the computation cost on the regularized adaptive Huber estimation using cross-validation to choose the parameter $\omega$. 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 $X = FB^1 + U$ using the following two models with entries of $B$ generated from $\text{Unif}(-1, 1)$.

1. We generate every row of $F \in \mathbb{R}^{n \times K}, U \in \mathbb{R}^{n \times d}$ from $N(0, I_K)$ and $N(0, I_d)$, respectively.
2. We let the $r$th row $f_r \in \mathbb{R}^K$ of $F \in \mathbb{R}^{n \times K}$ follow $f_r = \Phi f_{r-1} + \xi$, where $\Phi \in \mathbb{R}^{K \times K}$ with $\Phi_{ij} = 0.5^{\left| i - j \right| + 1}, i, j \in [K]$. In addition, $\{\xi_i\}_i \geq 1$ are drawn independently from $N(0, I_K)$. We generate every row of $U$ from $N(0, \Sigma)$ where $\Sigma_{ij} = 0.6^{|i - j|}, i, j \in [d]$.

The response vector follows $Y = F\gamma^* + UB^1 + \varepsilon$ in (4.1) with every entry of $\varepsilon \in \mathbb{R}^n$ being generated independently from either from $N(0, 0.5^2)$ or uniform distribution $\text{Unif}(-\sqrt{3}/2, \sqrt{3}/2)$. We set $Y^* = (0.5, 0.5)$ and $B^1 = (w, w, 0, \ldots, 0)$, where $w \geq 0$. When $w = 0$, the null hypothesis $Y = F\gamma^* + \varepsilon$ holds and the simulation results correspond to the size of the test. Otherwise, they correspond to the power of the test.

For $n = 200, K = 2, d = [200, 500]$ and all $w \in \{0.05, 0.10, 0.15, 0.20\}$, we generate the data from each model and compute the testing results based on procedures in Section 3 and 2000 simulations with $\alpha = 0.05$. For every replication, we conduct bootstrap 2000 times to compute the critical value $c_{1 - \alpha}$ given in Section 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.

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 section 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, Jiang, and Lan 2019) to select $\hat{S}_1$ and apply the refitted cross-validation (Fan,
Guo, and Hao (2012) to estimate $\sigma^2$. 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 $\epsilon$ are iid from $N(0, 0.5^2)$ or Unif($-\sqrt{3}/2, \sqrt{3}/2$). The covariate $X \in \mathbb{R}^{n \times d}$ follows the factor model $X = FB^T + U$. We generate $F$, $U$ and $B$ in the same way as those in Section 5.2. In addition, the response variable follows $Y = F\varphi^* + X\beta^* + \epsilon$ in (4.1) with $\beta^* = (0.8, 0.8, 0.8, 0.8, 0.0, \ldots, 0)$ and $\varphi^* = v^*1_{K \times 1}$ for several different values of $v \geq 0$. 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 $\hat{S}$ and covariates when $\hat{S}$ is not a consistent estimator of $S_{\nu}$. 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 ($\chi^2_{K}$) via the quantile-quantile plot in Figure 7 in Section B.2, supplementary materials. Figure 7 reveals that the test statistic constructed without using sample splitting has heavier right tail than that of the $\chi^2_{K}$ 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 misspecification. 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 Section B.1 for more details. For high-dimensional inference, as illustrated in Sections 5.2 and 5.3, when the null hypotheses hold, the size of the test is well-controlled. On the other hand, when the null hypotheses 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 nonstationary 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 examining 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 forecasting results of FARM with Lasso (sparse linear regression), PCR (latent factor regression), Ridge (Ridge regression), El-Net (ElasticNet) used in Coulombe, Marcellino, and Stevanovi´c (2021), Bianchi, Büchner, and Tamoni (2021) and FarmSelect (Factor adjusted Lasso) used in Fan, Ke, and Wang (2020). 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 $\{(x_{t-90}, y_{t-90}), \ldots, (x_{t-1}, y_{t-1})\}$ to
by using FARM as the alternative model. As computing the boot-

latent factor regression and sparse linear regression respectively
used model for making predictions via machine learning.

conclude that the dense model may not explain this dataset.

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).

Table 3. Out-of-sample $R^2$ for predicting “HOUSTNE” and “GS5” data using different models in different time periods.

| 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      |

In Table 4, we present the $p$-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.

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.

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

NOTE: The LA_Factor and SP_Linear denote latent factor regression and sparse linear regression model, respectively.

Table 4 presents the $p$-values 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 $\hat{\beta}$ and worse prediction outcomes when the covariates have a factor structure (highly-correlated). Fan, Ke, and Wang (2020) 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 $\beta^*$ 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.

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.

6. Conclusion and Discussion

In this article, 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 use 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 discuss our connection with the sparse and dense model proposed by Giannone, Lenza, and Primiceri (2021). They consider the model

$$Y = x_1^\top \beta^* + z_1^\top \phi^* + \epsilon,$$

where $z_t$ is a low-dimensional vector whose regression parameter $\phi^*$ is considered to be dense and $x_t$ is a high-dimensional covariate whose regression coefficient $\beta^*$ 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 \( z_t = f_t \) and \( x_t \) 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 \( \beta^* \) elementwisely or
groupwisely via both theories and numerical studies. Last but
not least, it is mentioned in their article, 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 colin-
earity 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 decom-
compose covariates into the factor and idiosyncratic component
and run factor adjusted regression via FARM. This showcases
the necessity 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 \( x_t \) to \( Y \). Due
to the space limit, we put them in the Section A, supplementary
materials.

**Supplementary Materials**

The supplementary material contains additional simulation results, tech-
nical proofs, and relevant codes for implementing the methodology of this
paper.

**Funding**

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

**ORCID**

Mengxin Yu https://orcid.org/0000-0002-6818-4083

**References**

Ahn, S. C., and Horenstein, A. R. (2013), "Eigenvalue Ratio Test for the
Number of Factors," *Econometrica*, 81, 1203–1227. [1079]
Avella-Medina, M., Battey, H. S., Fan, J., and Li, Q. (2018), "Robust
Estimation of High-Dimensional Covariance and Precision Matrices,
*Biometrika*, 105, 271–284. [1081]
Bai, J. (2003), "Inferential Theory for Factor Models of Large Dimensions,"
*Econometrica*, 71, 135–171. [1076,1079]
Bai, J., and Li, K. (2012), "Statistical Analysis of Factor Models of High
Dimension," *The Annals of Statistics*, 40, 436–465. [1076]
Bai, J., and Ng, S. (2002), "Determining the Number of Factors in Approx-
imate Factor Models," *Econometrica*, 70, 191–221. [1079]
——— (2006), "Confidence Intervals for Diffusion Index Forecasts and
Inference for Factor-Augmented Regressions," *Econometrica*, 74, 1133–
1150. [1076,1077]
——— (2008), "Forecasting Economic Time Series Using Targeted Predic-
tors," *Journal of Econometrics*, 146, 304–317. [1076]

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006), "Prediction by
Supervised Principal Components," *Journal of the American Statistical
Association*, 101, 119–137. [1076]
Barut, E., Fan, J., and Verhasselt, A. (2016), "Conditional Sure Independence
Screening," *Journal of the American Statistical Association*, 111, 1266–
1277. [1082]
Belloni, A., and Chernozhukov, V. (2011), "\( \ell_1 \)-Penalized Quantile Regres-
sion in High-Dimensional Sparse Models," *The Annals of Statistics*, 39,
82–130. [1077]
Bianchi, D., Büchner, M., and Tamoni, A. (2021), "Bond Risk Premiums with
Machine Learning," *The Review of Financial Studies*, 34, 1046–1089.
[1085]
Bing, X., Bunea, F., and Wegkamp, M. (2019), "Inference in Latent Factor
Regression with Clusterable Features," arXiv:1905.12696. [1076]
Bing, X., Bunea, F., Strimas-Mackey, S., and Wegkamp, M. (2021), "Pre-
diction under Latent Factor Regression: Adaptive pcr, Interpolating
Predictors and Beyond," *Journal of Machine Learning Research*, 22, 1–50.
[1076,1077]
Bunea, F., Strimas-Mackey, S., and Wegkamp, M. (2020), "Interpolating
Predictors in High-Dimensional Factor Regression," arXiv:2002.02525.
[1076]
Cai, T., Liu, W., and Luo, X. (2011), "A Constrained \( \ell_1 \) Minimization
Approach to Sparse Precision Matrix Estimation," *Journal of the American
Statistical Association*, 106, 594–607. [1081]
Candes, E., and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation
When \( p \) is much Larger than \( n^2 \)." *The Annals of Statistics*, 35, 2313–2351.
[1077,1083]
Chernozhukov, V., Chetverikov, D., and Kato, K. (2013), "Gaussian Approx-
imations and Multiplier Bootstrap for Maxima of Sums of High-
Dimensional Random Vectors," *The Annals of Statistics*, 41, 2786–2819.
[1081]
——— (2017), "Central Limit Theorems and Bootstrap in High Dimen-
sions," *Annals of Probability*, 45, 2309–2352. [1081]
Chernozhukov, V., Chetverikov, D., and Koike, Y. (2020), "Nearly Optimal
Central Limit Theorem and Bootstrap Approximations in High Dimen-
sions," arXiv preprint arXiv:2012.09513. [1081]
Chu, W., Li, R., and Reimherr, M. (2016), "Feature Screening for Time-
Varying Coefficient Models with Ultrahigh Dimensional Longitudinal
Data," *The Annals of Applied Statistics*, 10, 596–617. [1082]
Coulombe, P. G., Leroux, M., Stevanovic, D., and Surprenant, S. (2021),
"Macroeconomic Data Transformations Matter," *International Journal of
Forecasting*, 37, 1338–1354. [1085]
Coulombe, P. G., Marcellino, M., and Stevanović, D. (2021), "Can Machine
Learning Catch the Covid-19 Recession?" *National Institute Economic
Review*, 256, 71–109. [1085]
Dezerez, R., Buhlmann, P., and Zhang, C.-H. (2017), "High-Dimensional
Simultaneous Inference with the Bootstrap," *Test*, 26, 685–719. [1082]
Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle
Regression," *The Annals of Statistics*, 32, 407–499. [1076]
Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized
Likelihood and its Oracle Properties," *Journal of the American Statistical
Association*, 96, 1348–1360. [1076,1077]
Fan, J., and Liao, Y. (2020), "Learning Latent Factors from Diversified
Projections and its Applications to Over-Estimated and Weak Factors,
*Journal of the American Statistical Association*, 117, 909–924. [1076]
Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh
Dimensional Feature Space," *Journal of the Royal Statistical Society*, Series
B, 70, 849–911. [1078,1082,1084]
——— (2011), "Nonconcave Penalized Likelihood with NP-
Dimensionality," *IEEE Transactions on Information Theory*, 57,
5467–5484. [1083]
Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized
Linear Models with NP-Dimensionality," *The Annals of Statistics*, 38,
3567–3604. [1082]
Fan, J., Guo, S., and Hao, N. (2012), "Variance Estimation Using Refitted
Cross-Validation in Ultrahigh Dimensional Regression," *Journal of the
Royal Statistical Society*, Series B, 74, 37–65. [1082,1085]
Fan, J., Liao, Y., and Mincheva, M. (2013), "Large Covariance Estimation
by Thresholding Principal Orthogonal Complements," *Journal of the
