Journal of Econometrics 216 (2020) 71–85

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Factor-adjusted regularized model selection

Jianqing Fan [a], Yuan Ke [b,*], Kaizheng Wang [a]

[a] Department of ORFE, Princeton University, USA
[b] Department of Statistics, University of Georgia, USA

ARTICLE INFO

Article history:
Available online 7 February 2020

JEL classification:
C52
C58

Keywords:
Model selection consistency
Correlated covariates
Factor model
Regularized *M*-estimator
Time series

ABSTRACT

This paper studies model selection consistency for high dimensional sparse regression when data exhibits both cross-sectional and serial dependency. Most commonly-used model selection methods fail to consistently recover the true model when the covariates are highly correlated. Motivated by econometric and financial studies, we consider the case where covariate dependence can be reduced through the factor model, and propose a consistency strategy named Factor-Adjusted Regularized Model Selection (FarmSelect). By learning the latent factors and idiosyncratic components and using both of them as predictors, FarmSelect transforms the problem from model selection with highly correlated covariates to that with weakly correlated ones via lifting. Model selection consistency, as well as optimal rates of convergence, are obtained under mild conditions. Numerical studies demonstrate the nice finite sample performance in terms of both model selection and out-of-sample prediction. Moreover, our method is flexible in the sense that it pays no price for weakly correlated and uncorrelated cases. Our method is applicable to a wide range of high dimensional sparse regression problems. An R-package *FarmSelect* is also provided for implementation.

1. Introduction

With the development of data collection and storage technologies, high dimensional time series characterize many contemporary research problems in economics, finance, genomics, statistics, machine learning and so on. Specifying an appropriate yet parsimonious model has become a key topic in high dimensional time series analysis. Parsimonious models are preferable due to their simplicity and interpretability. In classic econometric studies, extensive efforts have been made to identify the correct orders of time series models, see Akaike (1973), Schwarz (1978), Tsay and Tiao (1985), Choi (1992) and Tiao and Tsay (1989) among others. In addition, removing redundant coefficients can improve the prediction accuracy of time series. Professor George C. Tiao and his co-authors, among others, have contributed to this area by a series of pioneering works (Box and Tiao, 1976; Liu et al., 1992; Montgomery et al., 1998).

Over the past two decades, many model selection methods have been developed. A major part of them are based on the regularized *M*-estimation approach including the LASSO (Tibshirani, 1996), the SCAD (Fan and Li, 2001), the elastic net (Zou and Hastie, 2005), and the Dantzig selector (Candes and Tao, 2007), among others. These methods have attracted a large amount of theoretical and algorithmic studies. See Donoho and Elad (2003), Fan and Peng (2004), Efron et al. (2004), (Meinshausen and Bühlmann, 2006), Zhao and Yu (2006), Fan and Lv (2008), Zou and Li (2008), Bickel et al. (2009), Wainwright (2009), Zhang (2010), and references therein. However, most existing model selection schemes are not tailored for economic and financial applications as they assume covariates are cross-sectionally weakly correlated

* Correspondence to: 310 Herty Drive University of Georgia, Athens, GA 30602, USA.
E-mail addresses: jqfan@princeton.edu (J. Fan), yuan.ke@uga.edu (Y. Ke), kaizheng@princeton.edu (K. Wang).

https://doi.org/10.1016/j.jeconom.2020.01.006
0304-4076/

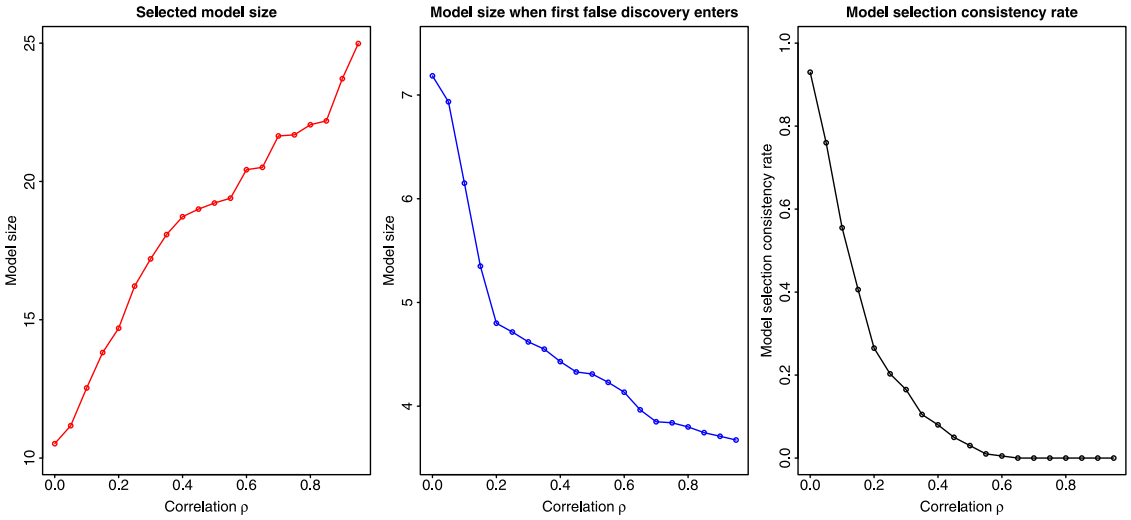

Fig. 1. LASSO model selection results with respect to the correlations.

and serially independent. These conditions are easily violated in economic and financial datasets. For example, economics studies (e.g. Stock and Watson, 2002; Bai and Ng, 2002) show that there exist strong co-movements among a large pool of macroeconomic variables. A stylized feature of the stock return data is cross-sectionally correlated among the stock returns. Furthermore, even if the weakly correlated assumption holds, one may still observe strong spurious correlations in a high dimensional sample.

To illustrate how cross-sectional correlations influence the model selection result, we consider a toy example of LASSO with an equally correlated design. Consider a sparse linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$. We choose sample size $n = 100$, dimensionality $p = 200$, $\boldsymbol{\beta}^* = (\beta_1, \ldots, \beta_{10}, \mathbf{0}^T_{(p-10)})^T$, and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}_n, 0.3\mathbf{I}_n)$. The nonzero coefficients $\beta_1, \ldots, \beta_{10}$ are drawn from i.i.d. Uniform [2, 5]. The covariates $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)^T$ are drawn from the normal distribution $N(\mathbf{0}_p, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a correlation matrix with all off-diagonal elements ρ for some $\rho \in [0, 1)$. Let ρ increase from 0 to 0.95 by a step size 0.05. For each given ρ, we simulate 200 replications and calculate the average model size selected by LASSO, the average model size when the first false discovery ($\mathbf{x}_j, j > 10$) enters the solution path and the model selection consistency rate. As is shown in Fig. 1, the correlation influences the model selection results in the following three aspects: (i) selected model size, (ii) early selection of false variables, (iii) model selection consistency rates. Therefore, when the covariates are highly correlated, there is little hope to exactly recover the active set from the solution path of LASSO. As to be shown later, the correlation has similar adverse impacts on other model selection methods (e.g. SCAD and elastic net).

To overcome the aforementioned problems caused by the cross-sectional correlation, this paper proposes a consistent strategy named Factor-Adjusted Regularized Model Selection (FarmSelect) for the case where covariates can be decorrelated via a few pervasive latent factors. More precisely, let x_{tj} be the tth ($t = 1, \ldots, n$) observation of the jth ($j = 1, \ldots, p$) covariate, and assume that $\mathbf{x}_t = (x_{t1}, \ldots, x_{tp})^T$ follows an approximate factor model

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

where $\mathbf{f}_t$ is a $K \times 1$ vector of latent factors, $\mathbf{B}$ is a $p \times K$ matrix of factor loadings, and $\mathbf{u}_t$ is a $p \times 1$ vector of idiosyncratic components that are uncorrelated with $\mathbf{f}_t$. The strategy of FarmSelect is to first learn the parameters in approximate factor model (1.1) for the covariates $\{\mathbf{x}_t\}_{t=1}^n$. Denote by $\widehat{\mathbf{f}}_t$ and $\widehat{\mathbf{B}}$ the obtained estimators of the factors and loadings respectively. Then by identifying the highly correlated low rank part by $\widehat{\mathbf{B}}\,\widehat{\mathbf{f}}_t$, we transform the problem from model selection with highly correlated covariates in $\mathbf{x}_t$ to that with weakly correlated or uncorrelated idiosyncratic components $\widehat{\mathbf{u}}_t := \mathbf{x}_t - \widehat{\mathbf{B}}\,\widehat{\mathbf{f}}_t$ and $\widehat{\mathbf{f}}_t$. This lifting step makes covariates weakly correlated. The second step amounts to solving a regularized profile likelihood problem. We study FarmSelect in detail by providing theoretical guarantees that FarmSelect can achieve model selection consistency as well as estimation consistency under mild conditions. Unlike traditional studies of model selection where the samples are assumed to be i.i.d., the serial dependency is allowed and thus our theories apply to time series data. Moreover, both theoretical and numerical studies show the flexibility of FarmSelect in the sense that it pays no price for weakly correlated cases. This property makes FarmSelect very powerful when the underlying correlations between active and inactive covariates are unknown.

FarmSelect is applicable to a wide range of high dimensional sparse regression related problems that include but are not limited to linear model, generalized linear model, Gaussian graphic model, robust linear model, and group LASSO. For the sparse linear regression, the proposed approach is equivalent to projecting the response variable and covariates onto the linear space orthogonal to the one spanned by the estimated factors. Existing algorithms that yield solution

paths of LASSO can be directly applied in the second step. To demonstrate the finite sample performance of FarmSelect, we study two simulated and one empirical example. The numerical results show FarmSelect can consistently select the true model even when the covariates are highly correlated while existing methods like LASSO, SCAD and elastic net fail to do so. An R-package FarmSelect (https://cran.r-project.org/web/packages/FarmSelect) is also provided to facilitate the implementation of our method.

Various methods have been studied to estimate the approximate factor model. Principal components analysis (PCA, Stock and Watson, 2002) is among one of the most popular ones. Data-driven estimation methods of the number of factors have been studied in extensive literature, such as Bai and Ng (2002), Luo et al. (2009), Hallin and Liška (2007), (Lam and Yao, 2012), and Ahn and Horenstein (2013) among others. Recently, a large amount of literature contributed to the asymptotic analysis of PCA under the ultra-high dimensional regime including Johnstone and Lu (2009), Fan et al. (2013), Shen et al. (2016) and Wang and Fan (2017), among others.

The rest of the paper is organized as follows. Section 2 overviews the problem setup including regularized M-estimators of sparse regression, the *irrepresentable condition*, and approximate factor models. Section 3 introduces the model selection methodology of FarmSelect and studies the sparse generalized linear model as a showcase example. Some issues related to the estimation of approximate factor models will be discussed in Section 3 as well. Section 4 presents the general theoretical results. Section 5 provides simulation studies and Section 6 studies the forecast of U.S. bond risk premia. Due to the limitation of space, all technical proofs are presented in a separate supplement file.

Here are some notations that will be used throughout the paper. $\mathbf{I}_n$ denotes the $n \times n$ identity matrix; $\mathbf{0}$ refers to the $n \times m$ zero matrix; $\mathbf{0}_n$ and $\mathbf{1}_n$ represent the all-zero and all-one vectors in $\mathbb{R}^n$, respectively. For a matrix $\mathbf{M}$, we denote its matrix entry-wise max norm as $\|\mathbf{M}\|_{\max} = \max_{i,j} |M_{ij}|$ and denote by $\|\mathbf{M}\|_F$ and $\|\mathbf{M}\|_p$ its Frobenius and induced p-norms, respectively. $\lambda_{\min}(\mathbf{M})$ denotes the minimum eigenvalue of $\mathbf{M}$ if it is symmetric. For $\mathbf{M} \in \mathbb{R}^{n\times m}$, $I \subseteq [n]$ and $J \subseteq [m]$, define $\mathbf{M}_{IJ} = (\mathbf{M}_{ij})_{i\in I, j\in J}$, $\mathbf{M}_{I\cdot} = (\mathbf{M}_{ij})_{i\in I, j\in[m]}$ and $\mathbf{M}_{\cdot J} = (\mathbf{M}_{ij})_{i\in[n], j\in J}$. For a vector $\mathbf{v} \in \mathbb{R}^p$ and $S \subseteq [p]$, define $\mathbf{v}_S = (\mathbf{v}_i)_{i\in S}$ to be its subvector. Let ∇ and ∇^2 be the gradient and Hessian operators. For $f : \mathbb{R}^p \to \mathbb{R}$ and $I, J \in [p]$, define $\nabla_I f(x) = (\nabla f(x))_I$ and $\nabla^2_{IJ} f(x) = (\nabla^2 f(x))_{IJ}$. $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ refers to the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

2. Problem setup

2.1. Regularized M-estimator

Let us begin with a family of high dimensional sparse regression problems in the following settings. From now on we suppose that $\{\mathbf{x}_t\}_{t=1}^n$ are $(p-1)$-dimensional random vectors of covariates with zero mean,[1] and $\{y_t\}_{t=1}^n$ are responses with each y_t sampled from some probability distribution $\mathbb{P}(z_t)$ parametrized by $z_t = \beta_0^* + \sum_{j=1}^{p-1} \beta_j^* \mathbf{x}_{tj} = (1, \mathbf{x}_t^T)\boldsymbol{\beta}^*$. Here $\boldsymbol{\beta}^* = (\beta_0^*, \ldots, \beta_{p-1}^*)^T \in \mathbb{R}^p$ is a sparse vector with $s \ll p$ non-zero elements. Let $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T \in \mathbb{R}^{n\times(p-1)}$ and $\mathbf{y} = (y_1, \cdots y_n)^T \in \mathbb{R}^n$ be the design matrix and response vector, respectively. Define $\mathbf{X}_1 = (\mathbf{1}_n, \mathbf{X}) \in \mathbb{R}^{n\times p}$, where the subscript 1 refers to the all-one column added to the original design matrix $\mathbf{X}$.

Let $L_n(\mathbf{y}, \mathbf{X}_1\boldsymbol{\beta})$ be some convex and differentiable loss function assigning a cost to any parameter $\boldsymbol{\beta} \in \mathbb{R}^p$. Suppose that $\boldsymbol{\beta}^*$ is the unique minimizer of the population risk $\mathrm{E}[L_n(\mathbf{y}, \mathbf{X}_1\boldsymbol{\beta})]$. Under the high-dimensional regime, it is natural to estimate $\boldsymbol{\beta}^*$ via a regularized M-estimator as follows:

$$\widetilde{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \{L_n(\mathbf{y}, \mathbf{X}_1\boldsymbol{\beta}) + \lambda R_n(\boldsymbol{\beta})\}, \tag{2.1}$$

where $R_n : \mathbb{R}^p \to \mathbb{R}_+$ is a norm that penalizes the use of a nonsparse vector $\boldsymbol{\beta}$ and $\lambda > 0$ is a tuning parameter.

A special case of this problem is the L_1 penalized likelihood estimation of generalized linear models. Suppose the conditional density function of Y given covariates $\mathbf{x}$ is a member of the exponential family, i.e.

$$f(y|\mathbf{x}, \boldsymbol{\beta}^*) \propto \exp[yz - b(z) + c(y)], \tag{2.2}$$

where $z = \beta_0^* + \sum_{j=1}^{p-1} \beta_j^* x_j = (1, \mathbf{x}^T)\boldsymbol{\beta}^*$, $b(\cdot)$ and $c(\cdot)$ are known functions, and $\boldsymbol{\beta}^*$ is an unknown coefficient vector of interest. It is commonly assumed that $b(\cdot)$ is strictly convex. Taking the loss function to be the negative log-likelihood function and the penalty function to be the L_1 norm, the regularized M-estimator of $\boldsymbol{\beta}^*$ admits the form

$$\widetilde{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{n} \sum_{t=1}^{n} [-y_t(1, \mathbf{x}_t^T)\boldsymbol{\beta} + b((1, \mathbf{x}_t^T)\boldsymbol{\beta})] + \lambda \|\boldsymbol{\beta}\|_1 \right\}. \tag{2.3}$$

[1] We use $(p-1)$ instead of p to denote the number of covariates so that there are p coefficients including the intercept. In addition, we center the covariates if they could have non-zero means. Whether this step is done or not does not affect the estimation of $\{\boldsymbol{\beta}_j^*\}_{j=1}^p$, but does affect the intercept β_0^*.

2.2. Irrepresentable condition

We expect a good estimator of (2.1) to achieve estimation as well as selection consistency. The former requires $\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\| \xrightarrow{\mathrm{P}} \mathbf{0}$ for some norm $\|\cdot\|$ as $n \to \infty$; while the latter requires $\mathrm{P}(\mathrm{supp}(\widetilde{\boldsymbol{\beta}}) = \mathrm{supp}(\boldsymbol{\beta}^*)) \to 1$ as $n \to \infty$. In general, the estimation consistency does not imply selection consistency and vice versa. To study the selection consistency, we consider a stronger condition named general sign consistency as follows.

Definition 2.1 (*Sign Consistency*)**.** *An estimate $\widetilde{\boldsymbol{\beta}}$ is sign consistent with respect to $\boldsymbol{\beta}^*$ if* $\lim_{n\to\infty} \mathrm{P}(\mathrm{sign}(\widetilde{\boldsymbol{\beta}}) = \mathrm{sign}(\boldsymbol{\beta}^*)) = 1$.

Zhao and Yu (2006) studied the LASSO estimator and showed there exists an *irrepresentable condition* which is sufficient and almost necessary for both sign and estimation consistencies for a sparse linear model. Without loss of generality, we assume $\mathrm{supp}(\boldsymbol{\beta}^*) = [s] = S$. Denote $(\mathbf{X}_1)_S$ and $(\mathbf{X}_1)_{S^c}$ as the submatrices of $\mathbf{X}_1$ defined by its first s columns and the rest $(p-s)$ columns, respectively. Then the *irrepresentable condition* requires some $\tau \in (0, 1]$, such that

$$\|(\mathbf{X}_1)_{S^c}^T(\mathbf{X}_1)_S[(\mathbf{X}_1)_S^T(\mathbf{X}_1)_S]^{-1}\|_\infty \le 1-\tau. \tag{2.4}$$

For general regularized M-estimator (2.1) to achieve both sign and estimation consistencies, Lee et al. (2015) proposed a generalized *irrepresentable condition*. When applied to the L_1 regularizer, it becomes

$$\|\nabla^2_{S^cS}L(\boldsymbol{\beta}^*)[\nabla^2_{SS}L(\boldsymbol{\beta}^*)]^{-1}\|_\infty \le 1-\tau, \tag{2.5}$$

for some $\tau \in (0, 1]$, where $L(\boldsymbol{\beta}) = L_n(\mathbf{y}, \mathbf{X}_1\boldsymbol{\beta})$. It is easy to check (2.5) is equivalent to (2.4) under the LASSO case. The generalized *irrepresentable condition* will easily get violated when there exist strong correlations between active and inactive variables. Even if it holds, the key parameter τ can be very close to zero, making it hard to select the correct model and obtain small estimation errors simultaneously.

2.3. Approximate factor model

To go beyond the assumption of weakly correlation, a natural extension is a conditional weak correlation. Suppose covariates are dependent through latent common factors. Given these common factors, the idiosyncratic components are weakly correlated. The factor model has been well studied in econometrics and statistics literature, we refer to Lawley and Maxwell (1971), Stock and Watson (2002), Bai and Ng (2002), Forni et al. (2005) and Fan et al. (2013), among others. For an overview, see Fan et al. (2019).

We assume that $\{\mathbf{x}_t\}_{t=1}^n \subseteq \mathbb{R}^{p-1}$ follows the approximate factor model

$$\mathbf{x}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t, \qquad t \in [n], \tag{2.6}$$

where $\{\mathbf{f}_t\}_{t=1}^n \subseteq \mathbb{R}^K$ are latent factors, $\mathbf{B} \in \mathbb{R}^{(p-1)\times K}$ is a loading matrix, and $\{\mathbf{u}_t\}_{t=1}^n \subseteq \mathbb{R}^{p-1}$ are idiosyncratic components. Note that $\mathbf{x}_t$ is the only observable quantity. Throughout the paper, K is assumed to be independent of n, which is frequently imposed in the literature of factor model (Fan et al., 2013). We assume that $\{\mathbf{f}_t, \mathbf{u}_t\}_{t=1}^n$ come from a time series $\{\mathbf{f}_t, \mathbf{u}_t\}_{t=-\infty}^{\infty}$. Denote $\mathbf{F} = (\mathbf{f}_1, \cdots, \mathbf{f}_n)^T \in \mathbb{R}^{n\times K}$ and $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)^T \in \mathbb{R}^{n\times(p-1)}$. Then (2.6) can be written in a more compact form:

$$\mathbf{X} = \mathbf{F}\mathbf{B}^T + \mathbf{U}. \tag{2.7}$$

We impose the following identifiability assumption (Fan et al., 2013). Here we only put the most basic assumption for factor model, and more can be found in Section 3.3 where estimation of factor model is discussed.

Assumption 2.1. Assume that $\mathrm{cov}(\mathbf{f}_t) = \mathbf{I}_K$, $\mathbf{B}^T\mathbf{B}$ is diagonal, and all the eigenvalues of $\mathbf{B}^T\mathbf{B}/p$ are bounded away from 0 and ∞ as $p \to \infty$.

3. Factor-adjusted regularized model selection

3.1. Methodology

To illustrate the main idea, we temporarily assume $\mathbf{f}_t$ and $\mathbf{u}_t$ to be observable. Define $\mathbf{B}_0 = (\mathbf{0}_K, \mathbf{B}^T)^T \in \mathbb{R}^{K\times p}$ and $\mathbf{U}_1 = (\mathbf{1}_n, \mathbf{U}) \in \mathbb{R}^{n\times p}$. By the approximate factor model (2.7), we have decompositions $\mathbf{X}_1 = \mathbf{F}\mathbf{B}_0^T + \mathbf{U}_1$ and

$$\mathbf{X}_1\boldsymbol{\beta} = \mathbf{F}\mathbf{B}_0^T\boldsymbol{\beta} + \mathbf{U}_1\boldsymbol{\beta} = \mathbf{F}\boldsymbol{\gamma} + \mathbf{U}_1\boldsymbol{\beta},$$

where $\boldsymbol{\gamma} = \mathbf{B}_0^T\boldsymbol{\beta} \in \mathbb{R}^K$. The regularized M-estimator (2.1) can be written as

$$\widetilde{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p,\ \boldsymbol{\gamma}=\mathbf{B}_0^T\boldsymbol{\beta}\in\mathbb{R}^K,} \{L_n(\mathbf{y}, \mathbf{F}\boldsymbol{\gamma} + \mathbf{U}_1\boldsymbol{\beta}) + \lambda R_n(\boldsymbol{\beta})\}.$$

Instead of using $\widetilde{\boldsymbol{\beta}}$ to estimate $\boldsymbol{\beta}^*$, we regard $\boldsymbol{\gamma}$ as nuisance parameters, drop the constraint $\boldsymbol{\gamma} = \mathbf{B}_0^T\boldsymbol{\beta}$, and consider a new estimator

$$\widehat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p,\ \boldsymbol{\gamma}\in\mathbb{R}^K}{\operatorname{argmin}} \{L_n(\mathbf{y}, \mathbf{F}\boldsymbol{\gamma} + \mathbf{U}_1\boldsymbol{\beta}) + \lambda R_n(\boldsymbol{\beta})\}, \tag{3.1}$$

namely $(\mathbf{u}_t^T, \mathbf{f}_t^T)^T$ are now regarded as new covariates. In other words, by lifting the covariate space from $\mathbb{R}^p$ to $\mathbb{R}^{p+K}$, the highly dependent covariates $\mathbf{x}_t$ are replaced by weakly dependent ones.

The theory for us to ignore the constraint $\boldsymbol{\gamma} = \mathbf{B}_0^T\boldsymbol{\beta}$ is given by the following lemma, whose proof is given by Appendix A in the supplement file.

Lemma 3.1. *Consider the generalized linear model (2.2), let* $L_n(\mathbf{y}, \mathbf{z}) = \frac{1}{n}\sum_{t=1}^n[-y_t z_t + b(z_t)]$, $\eta_t = y_t - b'((1, \mathbf{x}_t^T)\boldsymbol{\beta}^*)$ *and* $\mathbf{w}_t = (1, \mathbf{u}_t^T, \mathbf{f}_t^T)^T$. *If* $\mathrm{E}(\eta_t\mathbf{w}_t) = \mathbf{0}_{p+K}$, *then*

$$(\boldsymbol{\beta}^*, \mathbf{B}_0^T\boldsymbol{\beta}^*) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p,\ \boldsymbol{\gamma}\in\mathbb{R}^K}{\operatorname{argmin}} \mathrm{E}[L_n(\mathbf{y}, \mathbf{F}\boldsymbol{\gamma} + \mathbf{U}_1\boldsymbol{\beta})].$$

It is worth pointing out that the assumption $\mathrm{E}(\eta_t\mathbf{w}_t) = \mathbf{0}_{p+K}$ is very mild and natural. We just assume the residual η_t and augmented covariates $\mathbf{w}_t$ to be uncorrelated, which is almost as weak as the standard condition $\mathrm{E}(\eta_t|\mathbf{x}_t) = 0$ for the generalized linear model. For example, in the linear model $y_t = (1, \mathbf{x}_t^T)\boldsymbol{\beta}^* + \eta_t$, we strengthen the condition only from $\mathrm{E}(\eta_t\mathbf{x}_t) = 0$ to $\mathrm{E}(\eta_t\mathbf{f}_t) = 0$ and $\mathrm{E}(\eta_t\mathbf{u}_t) = 0$. In particular, the assumptions hold if η_t is independent of $\mathbf{u}_t$ and $\mathbf{f}_t$.

By construction, $(\mathbf{U}, \mathbf{F})$ has much weaker cross-sectional correlation than $\mathbf{X}$. Thus, the penalized profile likelihood (3.1) removes the effect of strong correlations caused by the latent factors. It can be implemented as follows:

Step 1: Initial estimation. Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be the design matrix. Fit the approximate factor model (2.7) and denote $\widehat{\mathbf{B}}$, $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{U}} = \mathbf{X} - \widehat{\mathbf{F}}\,\widehat{\mathbf{B}}^T$ the obtained estimators of $\mathbf{B}$, $\mathbf{F}$ and $\mathbf{U}$ respectively by using the principal component analysis (Bai, 2003; Fan et al., 2013, 2019). More specifically, the columns of $\widehat{\mathbf{F}}/\sqrt{n}$ are the eigenvectors of $\mathbf{X}\mathbf{X}^T$ corresponding to the top K eigenvalues, $\widehat{\mathbf{B}} = n^{-1}\mathbf{X}^T\widehat{\mathbf{F}}$. This is the same as $\widehat{\mathbf{B}} = (\sqrt{\lambda_1}\boldsymbol{\xi}_1, \ldots, \sqrt{\lambda_K}\boldsymbol{\xi}_K)$ and $\widehat{\mathbf{F}} = \mathbf{X}\widehat{\mathbf{B}}\,\mathrm{diag}(\lambda_1^{-1}\cdots, \lambda_K^{-1})$, where $\{\lambda_j\}_{j=1}^K$ and $\{\boldsymbol{\xi}_j\}_{j=1}^K$ are top K eigenvalues in descending order and their associated eigenvectors of the sample covariance matrix.

Step 2: Augmented M-estimation. Define $\widehat{\mathbf{W}} = (\mathbf{1}_n,\ \widehat{\mathbf{U}},\ \widehat{\mathbf{F}}) \in \mathbb{R}^{n\times(p+K)}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T \in \mathbb{R}^{p+K}$. Then $\widehat{\boldsymbol{\beta}}$ is obtained from the first p entries of the solution to the augmented problem

$$\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}\in\mathbb{R}^{p+K}}{\operatorname{argmin}} \left\{L_n(\mathbf{y}, \widehat{\mathbf{W}}\boldsymbol{\theta}) + \lambda R_n(\boldsymbol{\theta}_{[p]})\right\}. \tag{3.2}$$

We call the above two-step method as the factor-adjust regularized model selection (*FarmSelect*). If $\mathbf{u}_t$ is independent of $\mathbf{f}_t$ and the variables in the idiosyncratic component $\mathbf{u}_t$ are weakly correlated, then the columns in $\widehat{\mathbf{W}} = (\mathbf{1}_n, \widehat{\mathbf{U}}, \widehat{\mathbf{F}})$ are weakly correlated as long as $\mathbf{F}$ and $\mathbf{U}$ are well estimated. Hence, we successfully transform the problem from model selection with highly correlated covariates $\mathbf{X}$ in (2.1) to model selection with weakly correlated or uncorrelated ones by lifting the space to a higher dimension. The augmented problem (3.2) is a convex optimization problem which can be minimized via many existing convex optimization algorithms, such as coordinate descent (e.g. Friedman et al., 2010) and ADMM (Boyd et al., 2011).

3.2. Example: sparse linear model

Now we illustrate the FarmSelect procedure using sparse linear regression, where $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$. With aforementioned notation, we have

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}^* + \boldsymbol{\varepsilon} = \widehat{\mathbf{F}}\widehat{\mathbf{B}}_0^T\boldsymbol{\beta}^* + \widehat{\mathbf{U}}_1\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}. \tag{3.3}$$

The augmented M-estimator (3.2) for the sparse linear model is of the following form:

$$\widehat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p,\ \boldsymbol{\gamma}\in\mathbb{R}^K}{\operatorname{argmin}} \left\{\frac{1}{2n}\|\mathbf{y} - \widehat{\mathbf{F}}\boldsymbol{\gamma} - \widehat{\mathbf{U}}_1\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\right\}.$$

Solving the least-squares problem with respect to $\boldsymbol{\gamma}$, we have the penalized profile least-squares solution

$$\widehat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{\frac{1}{2n}\|(\mathbf{I}_n - \widehat{\mathbf{P}})(\mathbf{y} - \widehat{\mathbf{U}}_1\boldsymbol{\beta})\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\right\}, \tag{3.4}$$

where $\widehat{\mathbf{P}} = \widehat{\mathbf{F}}(\widehat{\mathbf{F}}^T\widehat{\mathbf{F}})^{-1}\widehat{\mathbf{F}}^T$ is the $n\times n$ projection matrix onto the column space of $\widehat{\mathbf{F}}$. As the decorrelation step does not depend on the choice of the regularizer $R(\cdot)$, FarmSelect can be applied to a wide range of penalized least squares problems such as SCAD, group LASSO, elastic net, fused LASSO, other folded concave penalties, and so on.

There is another way to understand this method. By left multiplying the projection matrix $(\mathbf{I}_n - \widehat{\mathbf{P}})$ to both sides of (3.3), we have

$$(\mathbf{I}_n - \widehat{\mathbf{P}})\mathbf{y} = (\mathbf{I}_n - \widehat{\mathbf{P}})\widehat{\mathbf{U}}_1\boldsymbol{\beta}^* + (\mathbf{I}_n - \widehat{\mathbf{P}})\boldsymbol{\varepsilon}, \tag{3.5}$$

where $(\mathbf{I}_n - \widehat{\mathbf{P}})\widehat{\mathbf{U}}_1$ can be treated as the decorrelated design matrix and $(\mathbf{I}_n - \widehat{\mathbf{P}})\mathbf{y}$ is the corresponding response variable. From (3.5) we see that the method in Kneip and Sarda (2011) coincides with FarmSelect in the linear case. However, the projection-based representation only makes sense in sparse linear regression. In contrast, our idea of profile likelihood directly generalizes to more general problems.

3.3. Estimating factor models

Principal component analysis (PCA) is frequently used to estimate latent factors for model (2.7). The estimated matrix of latent factors $\widehat{\mathbf{F}}$ is $\sqrt{n}$ times the eigenvectors corresponding to the K largest eigenvalues of the $n \times n$ matrix $\mathbf{X}\mathbf{X}^T$. Using the normalization $\mathbf{F}^T\mathbf{F}/n = \mathbf{I}_K$ yields $\widehat{\mathbf{B}} = \mathbf{X}^T\widehat{\mathbf{F}}/n$. Now we introduce the asymptotic properties of estimated factors and idiosyncratic components. We adopt the regularity assumptions in Fan et al. (2013), which are similar to the ones in Bai (2003) and other literature on high-dimensional factor analysis.

Assumption 3.1.

1. $\{\mathbf{f}_t, \mathbf{u}_t\}_{t=1}^{\infty}$ is strictly stationary. In addition, $\mathrm{E} f_{tk} = \mathrm{E}\, u_{tj} = \mathrm{E}(u_{tj} f_{tk}) = 0$ for all $i \in [n]$, $j \in [p-1]$ and $k \in [K]$;
2. There exist constants $c_1, c_2 > 0$ such that $\lambda_{\min}(\mathrm{cov}(\mathbf{u}_t)) > c_1$, $\|\mathrm{cov}(\mathbf{u}_t)\|_1 < c_2$ and $\min_{j,k\in[p-1]} \mathrm{var}(u_{tj}u_{tk}) > c_1$;
3. There exist $r_1, r_2 > 0$ and $b_1, b_2 > 0$ such that for any $s > 0$, $j \in [p-1]$ and $k \in [K]$, $\mathrm{P}(|u_{tj}| > s) \le \exp(-(s/b_1)^{r_1})$ and $\mathrm{P}(|f_{tk}| > s) \le \exp(-(s/b_2)^{r_2})$.

Assumption 3.2. Let $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_T^{\infty}$ denote the σ-algebras generated by $\{(\mathbf{f}_t, \mathbf{u}_t) : i \le 0\}$ and $\{(\mathbf{f}_t, \mathbf{u}_t) : i \ge T\}$ respectively. Assume the existence of $r_3, C > 0$ such that $3/r_1 + 3/(2r_2) + 1/r_3 > 1$ and for all $T \ge 1$,

$$\sup_{A\in\mathcal{F}_{-\infty}^0, B\in\mathcal{F}_T^{\infty}} |\mathrm{P}(A)\mathrm{P}(B) - \mathrm{P}(AB)| \le \exp(-CT^{r_3});$$

Assumption 3.3. There exists $M > 0$ such that for all $t, s \in [n]$, we have $\|\mathbf{B}\|_{\max} < M$, $\mathrm{E}\{p^{-1/2}[\mathbf{u}_t^T\mathbf{u}_s - \mathrm{E}(\mathbf{u}_t^T\mathbf{u}_s)]\}^4 < M$ and $\mathrm{E}\,\|p^{-1/2}\mathbf{B}^T\mathbf{u}_t\|_2^4 < M$.

We summarize useful properties of $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{U}}$ in Lemma 3.2, which directly follows from Lemmas 10–12 in Fan et al. (2013).

Lemma 3.2. *Let $\gamma^{-1} = 3/r_1 + 3/(2r_2) + 1/r_3 + 1$. Suppose that $\log p = o(n^{\gamma/6})$, $n = o(p^2)$, and Assumptions 2.1 and 3.1–3.3 hold. There exists a nonsingular matrix $\mathbf{H}_0 \in \mathbb{R}^{K\times K}$ such that*

1. $\|\widehat{\mathbf{F}}\mathbf{H}_0 - \mathbf{F}\|_{\max} = O_{\mathrm{P}}(\frac{1}{\sqrt{n}} + \frac{n^{1/4}}{\sqrt{p}})$;
2. $\max_{k\in[K]} n^{-1}\sum_{t=1}^n |(\widehat{\mathbf{F}}\mathbf{H}_0)_{jk} - f_{tk}|^2 = O_{\mathrm{P}}(\frac{1}{n} + \frac{1}{p})$
3. $\max_{j\in[p-1]} n^{-1}\sum_{t=1}^n |\widehat{u}_{ji} - u_{ji}|^2 = O_{\mathrm{P}}(\frac{\log p}{n} + \frac{1}{p})$;
4. $\|\widehat{\mathbf{U}} - \mathbf{U}\|_{\max} = o_{\mathrm{P}}(1)$.

A practical issue arises on how to choose the number of factors, i.e. K. As latent factors, loading and idiosyncratic components are all unobservable in the approximate factor model, the estimation of K is an intrinsic un-supervised learning problem. From the inference point of view, existing literature (Chamberlain and Rothschild, 1982; Stock and Watson, 2002; Bai and Ng, 2002, among others) usually assumes that there exists a non-negative integer K such that the first K population eigenvalues of $\mathbf{X}$ are diverging with p, while the rest $p - K$ eigenvalues are bounded. From the dimension reduction point of view, selecting K is to find a proper trade-off between goodness-of-fit and compactness of the model. In this paper, we follow a conditional sparsity perspective (Fan et al., 2013) regarding the role of K, where K is the smallest non-negative integer such that the idiosyncratic components $\mathbf{U} = \mathbf{X} - \mathbf{F}\mathbf{B}^T$ are weakly correlated. In this regard, for our purpose of model selection, a small overestimation of K does not seriously affect the adjusted model selection.

We adopt the modified ratio method, e.g. equation (10) in Chang et al. (2015), for the numerical studies in this paper due to its simplicity. Let $\lambda_k(\mathbf{X}\mathbf{X}^T)$ be the kth largest eigenvalue of $\mathbf{X}\mathbf{X}^T$, K_{max} be a prescribed upper bound and C_n be a constant that depends on n and p. The number of factors can be estimated by

$$\widehat{K} = \operatorname*{argmin}_{k\le K_{max}} \frac{\lambda_{k+1}(\mathbf{X}\mathbf{X}^T) + C_n}{\lambda_k(\mathbf{X}\mathbf{X}^T) + C_n} \tag{3.6}$$

for some given C_n. When $\mathbf{X}$ itself is weakly correlated, one can estimate K as 0.

Besides the modified ratio method, Bai and Ng (2002) studied the convergence and consistency estimation of K for high dimensional factor models. They proposed to estimate K by minimizing a family of information criteria. We refer to equation (9) in Bai and Ng (2002) for viable examples.

3.4. Factor-adjusted variable screening

Screening methods (e.g. Fan and Lv, 2008; Fan and Song, 2009; Wang and Leng, 2016) are computationally attractive and thus popular for ultra-high dimensional data analysis. However, the screening methods tend to include too many variables when there exist strong correlations among covariates (Fan and Lv, 2008; Wang and Leng, 2016). As an extension of FarmSelect, we propose the following conditional variable screening method to tackle this problem.

Step 1: *Initial estimation*. We fit the approximate factor model (2.7) to obtain $\widehat{\mathbf{B}}$, $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{U}}$.

Step 2: *Augmented marginal regression*. For $j \in [p-1]$, let $\widehat{\mathbf{U}}_j$ be the jth column of $\widehat{\mathbf{U}}$ and compute

$$(\widehat{\alpha}_j, \widehat{\beta}_j, \widehat{\boldsymbol{\gamma}}_j) = \underset{\alpha\in\mathbb{R}, \boldsymbol{\beta}\in\mathbb{R}, \boldsymbol{\gamma}\in\mathbb{R}^K}{\operatorname{argmin}} L_n(\mathbf{y}, \mathbf{1}_n\alpha + \widehat{\mathbf{U}}_j\beta + \widehat{\mathbf{F}}\boldsymbol{\gamma}). \tag{3.7}$$

Step 3 *Screening*. Return $\{j : |\widehat{\beta}_j| \geq \xi\}$ for some prescribed threshold ξ.

For sparse linear regression, our screening method reduces to the factor-profiled screening method proposed by Wang (2012).

4. Theoretical results

4.1. FarmSelect with approximate factor model

Now we establish theoretical guarantees of the FarmSelect estimator (3.2). Recall that $\boldsymbol{\beta}^*$ is equal to the first p entries of $\boldsymbol{\theta}^*$. Define $S = \operatorname{supp}(\boldsymbol{\theta}^*)$, $S_1 = \operatorname{supp}(\boldsymbol{\beta}^*)$ and $S_2 = [p+K]\backslash S$. When the covariates $\mathbf{X}$ admit the approximate factor model (2.7), the oracle procedure uses true augmented covariates $\mathbf{w}_t = (1, \mathbf{u}_t^T, \mathbf{f}_t^T)^T$ for $t \in [n]$ and solves

$$\min_{\boldsymbol{\theta}}\{L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}_{[p]}\|_1\},$$

where $\mathbf{W} = (\mathbf{w}_1^T, \ldots, \mathbf{w}_n^T)^T = (\mathbf{U}_1, \mathbf{F})$. However, $\mathbf{W}$ is not observable in practice. Hence we need to use its estimator $\widehat{\mathbf{W}}$ and solve

$$\min_{\boldsymbol{\theta}}\{L_n(\mathbf{y}, \widehat{\mathbf{W}}\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}_{[p]}\|_1\}.$$

Below the error induced by the factor estimation will be studied carefully. To deliver a clear discussion on the conditions and results, we focus on the FarmSelect estimator for the generalized linear model (2.3), and assume that the covariates are generated from the approximate factor model (2.7).

Assumption 4.1 (*Smoothness*)**.** $b(z) \in C^3(\mathbb{R})$. For some constants M_2 and M_3, we have $0 \leq b''(z) \leq M_2$ and $|b'''(z)| \leq M_3$, $\forall z$.

Assumption 4.2 (*Restricted strong convexity and irrepresentable condition*)**.** Let $\boldsymbol{\theta}^* = \begin{pmatrix} \boldsymbol{\beta}^* \\ \mathbf{B}_0^T\boldsymbol{\beta}^* \end{pmatrix}$. Assume the existence of $\kappa_2 > \kappa_\infty > 0$ and $\tau \in (0, 1)$ such that

$$\begin{aligned} &\|[\nabla^2_{SS} L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}^*)]^{-1}\|_\ell \leq \frac{1}{4\kappa_\ell}, \quad \text{for } \ell = 2 \text{ and } \infty, \\ &\|\nabla^2_{S_2S} L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}^*)[\nabla^2_{SS} L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}^*)]^{-1}\|_\infty \leq 1 - 2\tau. \end{aligned} \tag{4.1}$$

Assumption 4.3 (*Estimation of factor model*)**.** $\|\mathbf{W}\|_{\max} \leq \frac{M_0}{2}$ for some constant $M_0 > 0$. In addition, there exist $K \times K$ nonsingular matrix $\mathbf{H}_0$, and $\mathbf{H} = \begin{pmatrix} \mathbf{I}_p & \mathbf{0}_{p\times K} \\ \mathbf{0}_{K\times p} & \mathbf{H}_0 \end{pmatrix}$ such that for $\overline{\mathbf{W}} = \widehat{\mathbf{W}}\mathbf{H}$, we have $\|\overline{\mathbf{W}} - \mathbf{W}\|_{\max} \leq \frac{M_0}{2}$ and $\max_{j\in[p+K]} \left(\frac{1}{n}\sum_{t=1}^n |\overline{w}_{tj} - w_{tj}|^2\right)^{1/2} \leq \frac{2\kappa_\infty\tau}{3M_0M_2|S|}$.

Before presenting the main results, we make a few remarks on the assumptions.

1. Assumption 4.1 holds for a large family of generalized linear models. For example, linear model has $b(z) = \frac{1}{2}z^2$, $M_2 = 1$ and $M_3 = 0$; logistic model has $b(z) = \log(1 + e^z)$ and finite M_2, M_3.
2. The first inequality in (4.1) involves only a small matrix and holds easily, and the second inequality there is related to the generalized irrepresentable condition. Standard concentration inequalities (e.g. the Bernstein inequality for weakly dependent variables in Merlevède et al. (2011)) yield that Assumption 4.2 holds with high probability as long as $\mathrm{E}[\nabla^2 L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}^*)]$ satisfies similar conditions.

3. Here we show an example where the irrepresentable condition holds for the augmented covariates $\mathbf{W}$ but fails to hold for the original ones $\mathbf{X}$. Suppose the covariates $\{\mathbf{x}_i\}_{i=1}^n$ are generated from a single factor model $\mathbf{x}_i = \mathbf{b}f_i + \mathbf{u}_i$, where $\mathbf{b} = (1, 1, 2, \ldots, 2)^T \in \mathbb{R}^p$, $f_i \sim N(0, 1)$, $\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{I}_p)$ and it is independent of f_i. Consider a sparse linear model $y_i = \mathbf{x}_i^T\boldsymbol{\beta}^* + \varepsilon_i$, where $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. $N(0, \sigma^2)$ variables for some $\sigma > 0$ and they are independent of $\{\mathbf{f}_i, \mathbf{u}_i\}_{i=1}^n$; $\boldsymbol{\beta}^* = (1, 1, 0, \ldots, 0)^T \in \mathbb{R}^p$. Let $L_n(\mathbf{y}, \mathbf{z}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{z}\|_2^2$. Note that $\mathbf{w}_i = (\mathbf{u}_i^T, f_i)^T \in \mathbb{R}^{p+1}$, $\boldsymbol{\theta}^* = ((\boldsymbol{\beta}^*)^T, \mathbf{b}^T\boldsymbol{\beta}^*)^T \in \mathbb{R}^{p+1}$,

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta}^* + \varepsilon_i = f_i\mathbf{b}^T\boldsymbol{\beta}^* + \mathbf{u}_i^T\boldsymbol{\beta}^* + \varepsilon_i = \mathbf{w}_i^T\boldsymbol{\theta}^* + \varepsilon_i,$$

and $\nabla^2 L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}) = \frac{1}{n}\mathbf{W}^T\mathbf{W}$. Thanks to the independence, we have $\mathbf{w}_i \sim N(\mathbf{0}, \mathbf{I}_{p+1})$ and $\mathrm{E}\,\nabla^2 L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}) = \mathbf{I}_{p+1}$. Hence

$$\nabla^2_{S_2S}\,\mathrm{E}\,L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}^*)[\nabla^2_{SS}\,\mathrm{E}\,L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}^*)]^{-1} = \mathbf{0},$$

where $S = \{1, 2, p+1\}$ and $S_2 = \{3, \ldots, p\}$. With high probability, the irrepresentable condition also holds for the empirical quantity $L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}^*)$. On the other hand, we observe that $\mathbf{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \mathbf{b}\mathbf{b}^T + \mathbf{I}_p$. When $\mathbf{X}$ are used as covariates, we have $S = \{1, 2\}$ and $S_2 = \{3, \ldots, p\}$. From $\boldsymbol{\Sigma}_{SS} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ and $\boldsymbol{\Sigma}_{S^cS} = 2 \cdot \mathbf{1}_{(p-2)\times 2}$, we get $\boldsymbol{\Sigma}_{S^cS}\boldsymbol{\Sigma}_{SS}^{-1} = \frac{2}{3} \cdot \mathbf{1}_{(p-2)\times 2}$ and $\|\boldsymbol{\Sigma}_{S^cS}\boldsymbol{\Sigma}_{SS}^{-1}\|_\infty = \frac{4}{3} > 1$. Hence the irrepresentable condition is violated.

4. Under the conditions of Lemma 3.2, we have $\|\overline{\mathbf{W}} - \mathbf{W}\|_{\max} = o_{\mathrm{P}}(1)$ and $\max_{j\in[p+K]}\left(\frac{1}{n}\sum_{t=1}^n |\overline{w}_{tj} - w_{tj}|^2\right)^{1/2} = O_{\mathrm{P}}(\sqrt{\frac{\log p}{n}} + \frac{1}{\sqrt{p}})$, where $\overline{\mathbf{W}} = \widehat{\mathbf{W}}\mathbf{H}$, $\mathbf{H} = \begin{pmatrix} \mathbf{I}_p & \mathbf{0}_{p\times K} \\ \mathbf{0}_{K\times p} & \mathbf{H}_0 \end{pmatrix}$ and some proper $\mathbf{H}_0$. Hence $|S|^2(\frac{\log p}{n} + \frac{1}{p}) = O(1)$ can guarantee Assumption 4.3 to hold with high probability.

Theorem 4.1. *Suppose 4.1–4.3 hold. Define* $M = M_0^3 M_3 |S|^{3/2}$ *and*

$$\varepsilon = \max_{j\in[p+K]}\left|\frac{1}{n}\sum_{t=1}^n \overline{w}_{tj}[-y_t + b'((1, \mathbf{x}_t^T)\boldsymbol{\beta}^*)]\right|.$$

If $\frac{7\varepsilon}{\tau} < \lambda < \frac{\kappa_2\kappa_\infty\tau}{12M\sqrt{|S|}}$, *then we have* $\mathrm{supp}(\widehat{\boldsymbol{\beta}}) \subseteq \mathrm{supp}(\boldsymbol{\beta}^*)$ *and*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\infty \le \frac{6\lambda}{5\kappa_\infty}, \qquad \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le \frac{4\lambda\sqrt{|S|}}{\kappa_2}, \qquad \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \le \frac{6\lambda|S|}{5\kappa_\infty}.$$

In addition, if $\varepsilon < \frac{\kappa_2\kappa_\infty\tau^2}{12CM\sqrt{|S|}}$ *and* $\min\{|\boldsymbol{\beta}_j^*| : \boldsymbol{\beta}_j^* \ne 0,\ j \in [p]\} > \frac{6C\varepsilon}{5\kappa_\infty\tau}$ *hold for some* $C > 7$, *then by taking* $\lambda \in (\frac{7}{\tau}\varepsilon, \frac{C}{\tau}\varepsilon)$ *we can achieve the sign consistency* $\mathrm{sign}(\widehat{\boldsymbol{\beta}}) = \mathrm{sign}(\boldsymbol{\beta}^*)$.

By taking $\lambda \asymp \varepsilon$, one can achieve the sign consistency and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\infty/\varepsilon = O_{\mathrm{P}}(1)$, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2/\varepsilon = O_{\mathrm{P}}(\sqrt{|S|})$ and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1/\varepsilon = O_{\mathrm{P}}(|S|)$. Hence ε is a key quantity characterizing the error rate of our FarmSelect estimator, whose size is controlled using the following lemma.

Lemma 4.1. *Let* $\eta_t = y_t - b'((1, \mathbf{x}_t^T)\boldsymbol{\beta}^*)$ *and* $\mathbf{w}_t = (1, \mathbf{u}_t^T, \mathbf{f}_t^T)^T$. *Assume that* $\{\mathbf{w}_t, \eta_t\}_{-\infty}^\infty$ *is strictly stationary and satisfies the following conditions*

1. $\mathrm{E}(\eta_t\mathbf{w}_t) = \mathbf{0}$;
2. *There exist constants* $b, \gamma_1 > 0$ *such that* $\mathrm{P}(|\eta_t| > s) \le \exp(1 - (s/b)^{\gamma_1})$ *for all* $t \in \mathbb{Z}$ *and* $s \ge 0$;
3. *There exist constants* $c, \gamma_3 > 0$ *such that for all* $T \ge 1$,

$$\sup_{A\in\overline{\mathcal{F}}_{-\infty}^0, B\in\overline{\mathcal{F}}_T^\infty} |\mathrm{P}(A)\mathrm{P}(B) - \mathrm{P}(AB)| \le \exp(-cT^{\gamma_3}),$$

where $\overline{\mathcal{F}}_{-\infty}^0$ *and* $\overline{\mathcal{F}}_{-\infty}^0$ *denote the* σ*-algebras generated by* $\{(\mathbf{w}_t, \eta_t) : i \le 0\}$ *and* $\{(\mathbf{w}_t, \eta_t) : i \ge T\}$ *respectively;*

In addition, suppose that the assumptions in Lemma 3.2 hold. Then we have

$$\varepsilon = \max_{j\in[p+K]}\left|\frac{1}{n}\sum_{t=1}^n \overline{w}_{tj}\eta_t\right| = O_{\mathrm{P}}(\sqrt{\frac{\log p}{n}} + \frac{1}{\sqrt{p}}).$$

Recall that the assumption $\mathrm{E}(\eta_t\mathbf{w}_t) = \mathbf{0}$ has been used in Lemma 3.1 as a cornerstone of our FarmSelect methodology. The rest in the list are standard conditions similar to 3.1–3.3. All of them are mild and interpretable.

Lemma 4.1 asserts that $\varepsilon = O_{\mathrm{P}}(\sqrt{\frac{\log p}{n}} + \frac{1}{\sqrt{p}})$. The first term $\sqrt{\frac{\log p}{n}}$ corresponds to the optimal rate of convergence for high-dimensional M-estimator (e.g. Bickel et al., 2009). The second term $\frac{1}{\sqrt{p}}$ is the price we pay for factor estimation, which is negligible if $n = O(p \log p)$. In that high-dimensional regime, all the error bounds for $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\ell$ $(\ell = 1, 2, \infty)$ match the optimal ones in the literature.

4.2. Factor-adjusted variable screening

In this subsection, we study the factor-adjusted variable screening procedure described in Section 3.4. The lemma below considers the population version of the factor-adjusted screening procedure. It shows that as long as $\mathrm{E}(u_j y) \neq 0$ for all j's in the active set $\mathrm{supp}(\boldsymbol{\beta}^*)$, where u_j is the jth idiosyncratic component, the screening retains all the important variables. Furthermore, if $\mathrm{E}(u_j y) = 0$ for all j's outside the active set, the screening procedure exactly recovers the active set.

Lemma 4.2. *Let $\mathbf{f}$ be a K-dimensional random vector, u be a zero-mean random variable and is independent of $\mathbf{f}$, y be another random variable living in the same probability space, and $b \in C^2(\mathbb{R})$ such that $0 < b'' \leq M$. Assume that u, y and the coordinates of $\mathbf{f}$ all have finite second moments. Define*

$$(\alpha, \beta, \boldsymbol{\gamma}) = \operatorname*{argmin}_{\alpha\in\mathbb{R}, \beta\in\mathbb{R}, \boldsymbol{\gamma}\in\mathbb{R}^K} \mathrm{E}[b(\alpha + u\beta + \mathbf{f}^T\boldsymbol{\gamma}) - (\alpha + u\beta + \mathbf{f}^T\boldsymbol{\gamma})y].$$

We have the following.

1. $|\beta| \geq |\mathrm{E}(uy)|/(M \cdot \mathrm{E}\, u^2)$;
2. *If* $\mathrm{E}(uy) = 0$ *and* $\mathrm{P}(u = 0) = 0$, *then* $\beta = 0$.

Now we investigate the sure screening property of the factor-adjusted screening procedure. Recall that $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)^T \in \mathbb{R}^{n\times K}$ and $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)^T \in \mathbb{R}^{n\times p}$ are matrices of true factors and idiosyncratic components, respectively, and their estimated versions are $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{U}}$. We use $\mathbf{U}_j, \widehat{\mathbf{U}}_j \in \mathbb{R}^n$ to refer to the jth columns of $\mathbf{U}$ and $\widehat{\mathbf{U}}$. Define $L_{n,j}(\alpha, \beta, \boldsymbol{\gamma}) = L_n(\mathbf{y}, \mathbf{1}_n\alpha + \widehat{\mathbf{U}}_j\beta + \widehat{\mathbf{F}}\boldsymbol{\gamma})$ for $j \in [p]$, $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}$ and $\boldsymbol{\gamma} \in \mathbb{R}^K$. Let

$$(\alpha_j, \beta_j, \boldsymbol{\gamma}_j) = \operatorname*{argmin}_{\alpha\in\mathbb{R}, \beta\in\mathbb{R}, \boldsymbol{\gamma}\in\mathbb{R}^K} \mathrm{E}\, L_n(\mathbf{y}, \mathbf{1}_n\alpha + \mathbf{U}_j\beta + \mathbf{F}\boldsymbol{\gamma}) \tag{4.2}$$

be the population version of $(\widehat{\alpha}_j, \widehat{\beta}_j, \widehat{\boldsymbol{\gamma}}_j)$.

The following three assumptions are variants of those in Section 4.1, and hence they hold almost surely or with high probability in the cases we are interested in.

Assumption 4.4 (*Smoothness*)**.** $b \in C^3(\mathbb{R})$. For some constant M, we have $0 \leq b''(z) \leq M$ and $|b'''(z)| \leq M$, $\forall z$.

Assumption 4.5 (*Strong Convexity of Marginal Loss Functions*)**.** There exists some constant $\kappa > 0$ such that $\nabla^2 L_{n,j}(\alpha_j, \beta_j, \boldsymbol{\gamma}_j) \succeq \kappa\mathbf{I}$ for all $j \in [p]$.

Assumption 4.6 (*Estimation of Factor Model*)**.** There exist constants C, c and a nonsingular matrix $\mathbf{H}_0 \in \mathbb{R}^{K\times K}$ such that the followings happen. Let $\mathbf{H} = \begin{pmatrix} \mathbf{I}_p & \mathbf{0}_{p\times K} \\ \mathbf{0}_{K\times p} & \mathbf{H}_0 \end{pmatrix}$ and $\overline{\mathbf{W}} = \widehat{\mathbf{W}}\mathbf{H}$. We have $\|\mathbf{W}\|_{\max} \leq C$, $\|\overline{\mathbf{W}} - \mathbf{W}\|_{\max} \leq C$ and $\max_{j\in[p+K]}\left(\frac{1}{n}\sum_{t=1}^n |\overline{w}_{tj} - w_{tj}|^2\right)^{1/2} \leq c$.

Theorem 4.2. *Suppose that Assumptions 4.4–4.6 hold with high probability, and the constant c in Assumption 4.6 is small enough. For $j \in [p-1]$, define*

$$\boldsymbol{\varepsilon}_j = \frac{1}{n}\sum_{t=1}^n [b'(\alpha_j + \mathbf{f}_t^T\boldsymbol{\gamma}_j + u_{tj}\beta_j) - y_t](1, \mathbf{f}_t^T, u_{tj})^T.$$

If $\xi \leq \rho \min_{j\in\mathrm{supp}(\boldsymbol{\beta}^*)} |\mathrm{E}(u_j y)|/(M \cdot \mathrm{E}\, u_j^2)$ for some constant $\rho \in (0, 1)$ and

$$\max_{j\in\mathrm{supp}(\boldsymbol{\beta}^*)} \|\boldsymbol{\varepsilon}_j\|_2 = o_{\mathrm{P}}(\min_{j\in\mathrm{supp}(\boldsymbol{\beta}^*)} |\mathrm{E}(u_j y)|/\mathrm{E}\, u_j^2), \tag{4.3}$$

then we have

$$\mathbb{P}(\mathrm{supp}(\boldsymbol{\beta}^*) \subseteq \{j : |\widehat{\beta}_j| \geq \xi\}) = 1 - o(1).$$

Under the conditions in Lemma 4.1, we can prove that

$$\max_{j\in\mathrm{supp}(\boldsymbol{\beta}^*)} \|\boldsymbol{\varepsilon}_j\|_2 = O_{\mathrm{P}}\left(\sqrt{\frac{\log p}{n}} + \frac{1}{\sqrt{p}}\right).$$

Theorem 4.2 asserts that if $\min_{j\in\mathrm{supp}(\boldsymbol{\beta}^*)} |\mathrm{E}(u_j y)|/\mathrm{E}\, u_j^2$ grows faster than $\sqrt{\frac{\log p}{n}} + \frac{1}{\sqrt{p}}$, the factor-adjusted screening procedure enjoys the sure screening property (Fan and Lv, 2008; Fan and Song, 2009). The optimal choice of the screening threshold ξ can be discussed by following the analysis as in Fan and Song (2009). Here we do not pursue this result as it is not the main focus of the paper.

Table 1
Parameters calibrated from S&P 500 returns.

$\Sigma_{\mathbf{B}}$			$\widehat{\Phi}$			Σ_{η}			σ_u^2
0.5237	0	0	0.1897	−0.0375	−0.0223	0.9621	−0.0056	0.0182	
0	0.2884	0	0.0630	0.1553	0.0206	−0.0056	0.9715	−0.0078	0.246
0	0	0.2372	−0.0432	0.0102	0.4343	0 .0182	−0.0078	0.8094	

5. Simulation study

5.1. Example 1: Linear regression

We study a simulated example for high dimensional sparse linear regression with correlated covariates. The correlation structure is calibrated from S&P 500 monthly excess returns between 1980 and 2012. Throughout the numerical studies of this paper, the tuning parameter λ is selected by the 10-fold cross-validation. The model selection performance is measured by the model selection consistency rate and the sure screening rate. The former is the proportion of simulations that the selected model is identical to the true one and the latter is the proportion of simulations that the selected model contains all important variables.

Calibration and data generation process

We calculate the centered monthly excess returns for the stocks in S&P 500 index that have complete records from January 1980 to December 2012. The data, collected from CRSP,[2] contains the returns of 202 stocks with a time span of 396 months. Denote the centered monthly excess returns as $\mathbf{z}_t$, $t = 1, \ldots, 396$. The calibration and data generation procedure are outlined as follows.

(1) Fit $\mathbf{z}_t$ with a three factor model. We apply PCA on the sample covariance of $\{\mathbf{z}_t\}_{t=1}^{396}$ and denote λ_k and $\boldsymbol{\xi}_k$, $k = 1, 2, 3$, as the top three eigenvalues and corresponding eigenvectors. We estimate loadings $\tilde{\mathbf{B}} = (\sqrt{\lambda_1}\boldsymbol{\xi}_1, \sqrt{\lambda_2}\boldsymbol{\xi}_2, \sqrt{\lambda_3}\boldsymbol{\xi}_3)$ and $\tilde{\mathbf{f}}_t = (\lambda_1^{-1/2}\boldsymbol{\xi}_1^T\mathbf{z}_t, \lambda_2^{-1/2}\boldsymbol{\xi}_2^T\mathbf{z}_t, \lambda_3^{-1/2}\boldsymbol{\xi}_1^T\mathbf{z}_t)^T$.

(2) Calculate $\boldsymbol{\Sigma}_{\mathbf{B}}$ as the sample covariance of the rows of $\tilde{\mathbf{B}}$, which is diag$(\lambda_1, \lambda_2, \lambda_3)$. Generate loading matrix $\mathbf{B}$ whose rows are draws from i.i.d. $N(\mathbf{0},\ \boldsymbol{\Sigma}_{\mathbf{B}})$.

(3) Fit VAR(1) model $\tilde{\mathbf{f}}_t = \boldsymbol{\Phi}\tilde{\mathbf{f}}_{t-1} + \boldsymbol{\eta}_t$. Denote $\widehat{\boldsymbol{\Phi}}$ the estimate of $\boldsymbol{\Phi}$ and calculate $\boldsymbol{\Sigma}_{\eta} = \mathbf{I} - \widehat{\boldsymbol{\Phi}}\ \widehat{\boldsymbol{\Phi}}^T$. Generate $\mathbf{f}_t$ from the VAR(1) model $\mathbf{f}_t = \widehat{\boldsymbol{\Phi}}\mathbf{f}_{t-1} + \boldsymbol{\eta}_t$ with $\mathbf{f}_0 = \mathbf{0}$, where $\boldsymbol{\eta}_t$ is generated from i.i.d. $N(\mathbf{0},\ \boldsymbol{\Sigma}_{\eta})$.

(4) Calculate the residual $\tilde{\mathbf{u}}_t = \mathbf{z}_t - \tilde{\mathbf{B}}\tilde{\mathbf{f}}_t$ and $\boldsymbol{\Sigma}_{\mathbf{u}}$ the sample covariance matrix of $\tilde{\mathbf{u}}_t$. Denote σ_u^2 the average of the diagonal entries of $\boldsymbol{\Sigma}_{\mathbf{u}}$. Generate covariates $\mathbf{x}_t$ from a factor model $\mathbf{x}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t$ where the entries in $\mathbf{u}_t$ are drawn from i.i.d. $N(0,\ \sigma_u^2)$.

(5) Generate the response y_t from a sparse linear model $y_t = \mathbf{x}_t^T\boldsymbol{\beta}^* + \varepsilon_t$. The true coefficients are $\boldsymbol{\beta}^* = (\beta_1, \ldots, \beta_{10}, \mathbf{0}_{(p-10)}^T)^T$, and the nonzero coefficients $\beta_1, \ldots, \beta_{10}$ are drawn from i.i.d. Uniform(2,5). We draw ε_t from an AR(1) model $\varepsilon_t = 0.5\varepsilon_{t-1} + \gamma_t$ with $\gamma_t \sim N(0,\ 0.3)$.

The results of the calibrated parameters are presented in Table 1.

Impacts of irrepresentable condition

First, we show LASSO performs poorly in terms of model selection consistency rate when the *irrepresentable condition* is violated, while FarmSelect can consistently select the correct model. Let $n = 100$ and $p = 500$. Denote $\Gamma_\infty = \|\mathbf{X}_{S^c}^T\mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\|_\infty$. When $\Gamma_\infty < 1$ the *irrepresentable condition* holds and otherwise it is violated. We simulate 10,000 replications. For each replication, we calculate Γ_∞ and apply both LASSO and FarmSelect for model selection. Then we calculate the model selection consistency rate within each small interval around Γ_∞ (a nonparametric smoothing). The results are presented in Fig. 2. According to Fig. 2, both FarmSelect and LASSO have high model selection consistency rate when $\Gamma_\infty < 1$. This shows FarmSelect does not pay any price under the weak correlation scenario. As Γ_∞ grows beyond 1, the correct model selection rate of LASSO drops quickly. When the *irrepresentable condition* is strongly violated (e.g. $\Gamma_\infty > 1.5$), the correct model selection rate of LASSO is close to zero. On the contrary, FarmSelect has high selection consistency rates regardless of Γ_∞.

Impacts of sample size

Second, we examine the model selection consistency with a fixed dimensionality and an increasing sample size. We fix $p = 500$ and let n increase from 50 to 150. For each given sample size, we simulate 200 replications and calculate the model selection consistency rates and the sure screening rates for LASSO, SCAD, elastic net, and FarmSelect, respectively. For the elastic net, we set $\lambda_1 = \lambda_2 \equiv \lambda$. The results are presented in Figs. 3(a) and 3(b). Fig. 3(a) shows that model selection consistency rates of LASSO, SCAD, and elastic net do not enjoy fast convergence to 1 when the sample size increases,

[2] Center for Research in Security Prices Database, see http://www.crsp.com/ for more details.

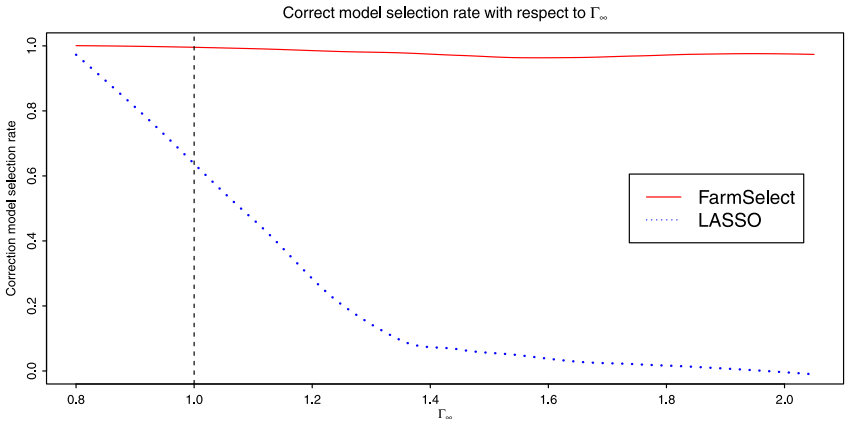

Fig. 2. Relationship between model selection consistency rate and *irrepresentable condition*. Among the 10,000 replications, more than 9,500 replications have $\Gamma_\infty > 1$ and more than 8,000 replications have $\Gamma_\infty > 1.5$.

while the one of FarmSelect equals to one as long as the sample size exceeds 100. Similar phenomena are observed from sure screening rates. To demonstrate the prediction performance, we report the mean estimation error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ for each method, which is a good indicator of the prediction error. The estimation errors are reported in Fig. 3(c). When the sample size is small, LASSO, SCAD, and elastic net have large estimation errors since they tend to select overfitted models.

Impacts of dimensionality

Third, we assess the model selection performance when the dimensionality p is growing beyond n and diverging. We fix $n = 100$ and let p grow from 200 to 1000. For each given p, we simulate 200 replications and calculate the model selection consistency rate of LASSO, SCAD, elastic net, and FarmSelect respectively. The model selection consistency rates are presented in Fig. 4(a). According to Fig. 4(a), the model selection consistency rate of FarmSelect stays close to 1 even as p increases, whereas the rates for the other three methods drop quickly. Again, we report the estimation errors in Fig. 4(b). As the dimensionality grows, FarmSelect has the least increase in estimation error.

5.2. Example 2: Logistic regression

We consider the following logistic regression model whose conditional probability function is:

$$\mathrm{P}(y_t = 1|\mathbf{X}_t) = \frac{\exp(\mathbf{X}_t^T\boldsymbol{\beta})}{1+\exp(\mathbf{X}_t^T\boldsymbol{\beta})}, \quad i = 1, \cdots, N. \tag{5.1}$$

We set sample size $n = 300$ and dimensionality $p = 300,\ 400,\ 500$. The true coefficients are set to be $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_{(1)}^T,\ \mathbf{0})^T$ with $\boldsymbol{\beta}_{(1)} = (6,\ 5,\ 4)^T$. Hence the true model size is 3.

The covariates $\mathbf{X}$ are generated from one of the following three models:

(1) Factor model $\mathbf{x}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t$ with $K = 3$. Factors are generated from a stationary VAR(1) model $\mathbf{f}_t = \boldsymbol{\Phi}\mathbf{f}_{t-1} + \boldsymbol{\eta}_t$ with $\mathbf{f}_0 = \mathbf{0}$. The (i, j)th entry of $\boldsymbol{\Phi}$ is set to be 0.5 when $i = j$ and $0.3^{|i-j|}$ when $i \neq j$. We draw $\mathbf{B}$, $\mathbf{u}_t$ and $\boldsymbol{\eta}_t$ from the i.i.d. standard Normal distribution.
(2) Equal correlated case. We draw $\mathbf{x}_t$ from i.i.d. $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ has diagonal elements 1 and off-diagonal elements 0.4.
(3) Uncorrelated case. We draw $\mathbf{x}_t$ from i.i.d. $N_p(\mathbf{0}, \mathbf{I})$

We compare the model selection performance of FarmSelect with LASSO and simulate 100 replications for each scenario. The model selection performance is measured by the selection consistency rate, sure screening rate and the average size of the selected model. The results are presented in Table 2. According to Table 2, FarmSelect pays no price for the uncorrelated case and outperforms LASSO for highly correlated cases.

6. Prediction of U.S. bond risk premia

In this section, we predict U.S. bond risk premia with a large panel of macroeconomic variables. The response variable is the monthly data of U.S. bond risk premia with a maturity of 2 to 5 years between January 1980 and December 2015 containing 432 data points. The bond risk premia are calculated as the one year return of an n year maturity bond excessing

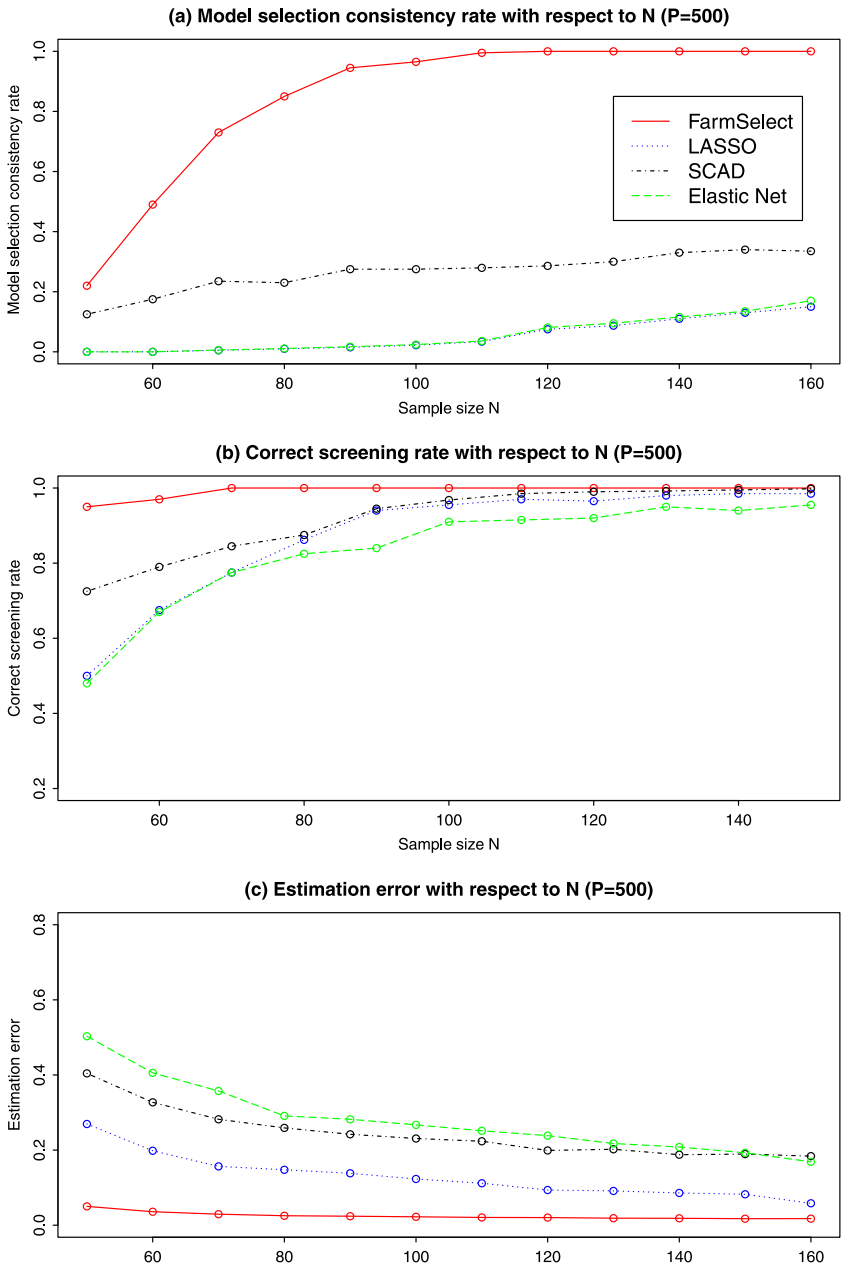

Fig. 3. From above to below: (a) Model selection consistency rates with fixed p and increasing n; (b) Sure screening rates with fixed p and increasing n; (c) Mean estimation errors $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ with fixed p and increasing n.

the risk-free rate. The covariates are 128 monthly U.S. macroeconomic variables in the FRED-MD database[3] (McCracken and Ng, 2016). The covariates in the FRED-MD dataset are strongly correlated and can be well explained by a few principal components. To see this, we apply principal component analysis to the covariates and draw the scree plot of the top 20 principal components in Fig. 5. The scree plot shows the first principal component solely explains more than 60% of the total variance. In addition, the first 5 principal components together explain more than 90% of the total variance.

We apply one month ahead rolling window prediction with a window size of 120 months. Within each window, we predict the U.S. bond risk premia by a high dimensional linear regression model of dimensionality 128. We compare the

[3] The FRED-MD is a monthly economic database updated by the Federal Reserve Bank of St. Louis which is publicly available at http://research.stlouisfed.org/econ/mccracken/sel/.

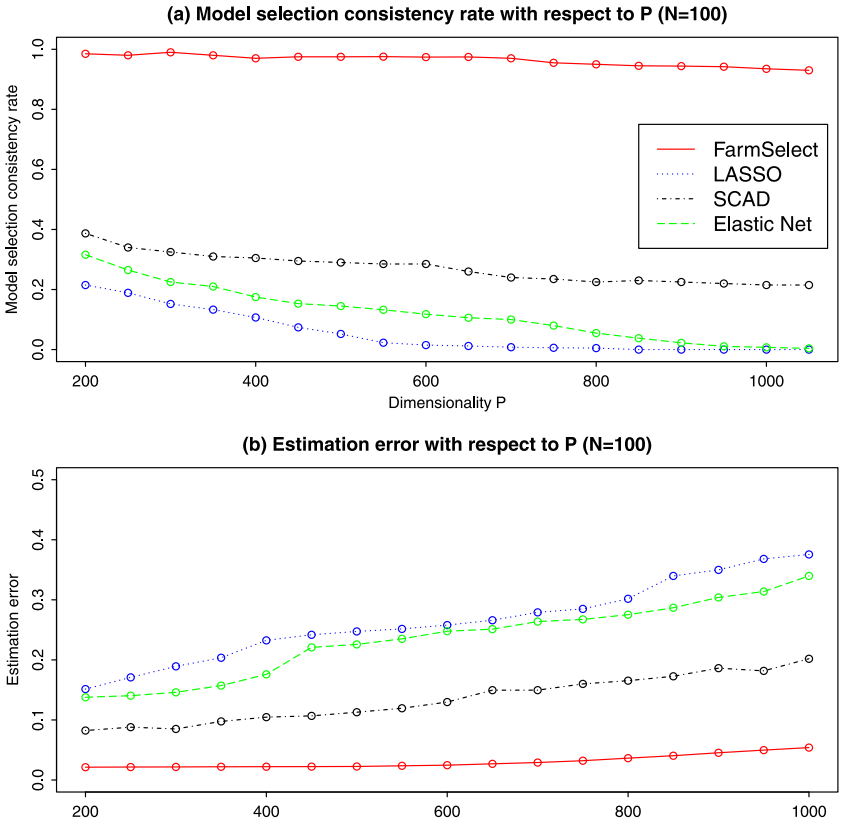

Fig. 4. From above to below: (a) Model selection consistency rates with fixed n and increasing p; (b) Mean estimation errors $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ with fixed n and increasing p.

Table 2
Model selection results of logistic regression ($n = 200$).

	Factor model with $K = 3$					
	FarmSelect			LASSO		
	Selection rate	Screening rate	Average model size	Selection rate	Screening rate	Average model size
$p = 300$	0.91	1.00	3.22	0.22	0.98	8.13
$p = 400$	0.90	0.99	3.14	0.17	0.97	7.66
$p = 500$	0.89	0.98	3.15	0.14	0.97	9.99
	Equal correlated case					
	FarmSelect			LASSO		
	Selection rate	Screening rate	Average model size	Selection rate	Screening rate	Average model size
$p = 300$	0.91	1.00	3.07	0.61	0.99	4.63
$p = 400$	0.91	1.00	3.06	0.54	0.99	4.67
$p = 500$	0.87	0.99	3.05	0.55	0.99	5.45
	Uncorrelated case					
	FarmSelect			LASSO		
	Selection rate	Screening rate	Average model size	Selection rate	Screening rate	Average model size
$p = 300$	1.00	1.00	3.00	0.88	1.00	4.05
$p = 400$	0.99	1.00	3.01	0.86	1.00	5.11
$p = 500$	0.99	1.00	3.02	0.85	1.00	3.57

proposed FarmSelect method with LASSO in terms of model selection and prediction. Besides, we include the principal component regression (PCR) in the competition of prediction. Instead of using the covariates as regressors directly, PCR regresses the dependent variable on the leading principal components of covariates. The FarmSelect is implemented by the *FarmSelect* R package with default settings. To be specific, the loss function is L_1, the number of factors is estimated by the modified eigen-ratio method and the regularized parameter is selected by multi-fold cross-validation. The LASSO method is implemented by the *glmnet* R package. The PCR method is implemented by the *pls* package in R.

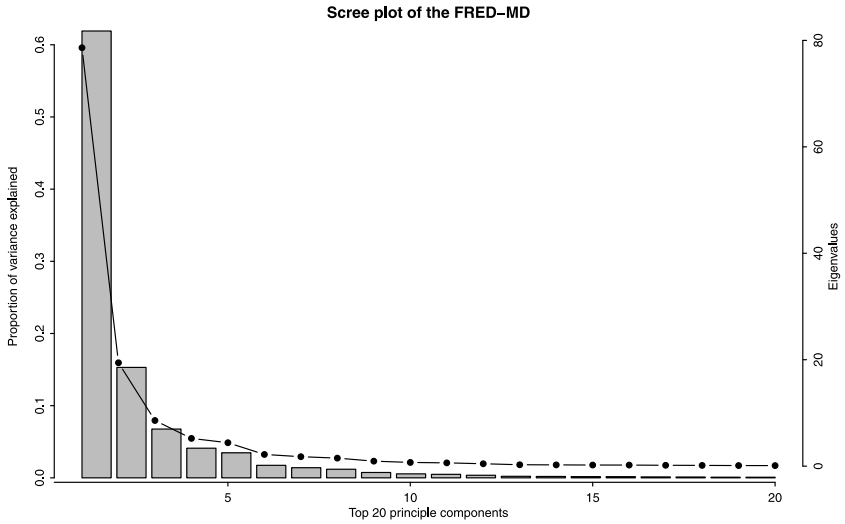

Fig. 5. Eigenvalues (dotted line) and proportion of variance explained (bar) by the top 20 principal components.

Table 3
Out of sample R^2 and average selected model size.

Maturity of Bond	Out of sample R^2			Average model size	
	FarmSelect	LASSO	PCR	FarmSelect	Lasso
2 Years	0.530	0.509	0.462	5.96	6.86
3 Years	0.526	0.523	0.483	5.71	7.09
4 Years	0.484	0.476	0.470	5.53	6.81
5 Years	0.481	0.475	0.477	5.90	6.84

Table 4
2 years Maturity: Top 5 variables with highest selection frequency.

	FarmSelect	
Rank	Name	Frequency
1	3-Month Commercial Paper Minus FEDFUNDS	133
2	Civilians Unemployed for 15–26 Weeks	84
3	Housing Starts, Midwest	71
4	Industrial Production: Durable Consumer Goods	70
5	Moody's Baa Corporate Bond Minus FEDFUNDS	65
	LASSO	
Rank	Name	Frequency
1	Total Reserves of Depository Institutions	209
2	Civilians Unemployed for 5–14 Weeks	185
3	Housing Starts, Midwest	110
4	3-Month Commercial Paper Minus FEDFUNDS	96
5	Civilians Unemployed for 15–26 Weeks	88

The prediction performance is evaluated by the out-of-sample R^2 which is defined as

$$R^2 = 1 - \frac{\sum_{t=121}^{432}(y_t - \widehat{y}_t)^2}{\sum_{t=121}^{432}(y_t - \bar{y}_t)^2},$$

where y_t is the response variable realized at time t, $\widehat{y}_t$ is the predicted y_t by one of the three methods above using the previous 120 months data, and $\bar{y}_t$ is the sample mean of the previous 120 months responses $(y_{t-120}, \ldots, y_{t-1})$, which represents a naive predictor. For FarmSelect and LASSO, we also report the average selected model size for prediction at time $t \in \{121, \ldots, 432\}$. The out-of-sample R^2 and average selected model size are reported in Table 3. The results in Table 3 show that FarmSelect selects parsimonious models and achieves the highest R^2's in all scenarios. On the contrary, LASSO may select redundant models as it ignores correlations among covariates. To see this, we rank the covariates according to the selected frequency. The top 5 selected covariates and their frequencies are listed in Table 4. According to Table 4, LASSO tends to select some highly correlated covariates simultaneously. For instance, LASSO includes both

Civilians Unemployed for 5–14 Weeks and Civilians Unemployed for 15–26 Weeks due to the strong correlation between them.

Acknowledgments

This research was supported in part by the National Science Foundation (NSF), USA grants DMS-1662139 and DMS-1712591, by the Office of Naval Research (ONR), USA grant N00014-19-1-2120, and by the National Institutes of Health (NIH), USA grant 2R01-GM072611-13.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2020.01.006.

References

Ahn, S.C., Horenstein, A.R., 2013. Eigenvalue ratio test for the number of factors. Econometrica 81, 1203–1227.
Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petroc, B.N., Caski, F. (Eds.), Second International Symposium in Information Theory. Akademiai Kiado, Budapest, pp. 276–281.
Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71, 135–171.
Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70, 191–221.
Bickel, P.J., Ritov, Y.A., Tsybakov, A.B., 2009. Simultaneous analysis of lasso and dantzig selector. Ann. Stat. 37, 1705–1732.
Box, G.E., Tiao, G.C., 1976. Comparison of forecast and actuality. J. R. Stat. Soc. Ser. C. Appl. Stat. 25 (3), 195–200.
Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends® Mach. Learn. 3, 1–122.
Candes, E., Tao, T., 2007. The Dantzig selector: Statistical estimation when p is much larger than n. Ann. Statist. 35, 2313–2351.
Chamberlain, G., Rothschild, M., 1982. Arbitrage, factor structure, and mean–variance analysis on large asset markets. Econometrica 51, 1305–1324.
Chang, J., Guo, B., Yao, Q., 2015. High dimensional stochastic regression with latent factors, endogeneity and nonlinearity. J. Econometrics 189, 297–312.
Choi, B.S., 1992. ARMA Model Identification. Springer-Verlag, New York.
Donoho, D.L., Elad, M., 2003. Optimally sparse representation in general (nonorthogonal) dictionaries via l_1 minimization. Proc. Natl. Acad. Sci. 100, 2197–2202.
Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Stat. 32, 407–499.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Stat. Assoc. 96, 1348–1360.
Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. 75, 603–680.
Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc. Ser. B Stat. Methodol. 70, 849–911.
Fan, J., Peng, H., 2004. On non-concave penalized likelihood with diverging number of parameters. Ann. Statist. 32, 928–961.
Fan, J., Song, R., 2009. Sure independence screening in generalized linear models with NP-dimensionality. Ann. Statist. 38, 3567–3604.
Fan, J., Wang, K., Zhong, Y., Zhu, Z., 2019. Robust high dimensional factor models with applications to statistical machine learning. Statist. Sci. (invited).
Forni, M., Hallin, M., Lippi, L., 2005. The generalized dynamic factor model: one-sided estimation and forecasting. J. Amer. Statist. Assoc. 100, 830–840.
Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1–22.
Hallin, M., Liška, R., 2007. Determining the number of factors in the general dynamic factor model. J. Amer. Statist. Assoc. 102, 603–617.
Johnstone, I.M., Lu, A.Y., 2009. On consistency and sparsity for principal components analysis in high dimensions. J. Amer. Statist. Assoc. 104, 682–693.
Kneip, A., Sarda, P., 2011. Factor models and variable selection in high-dimensional regression analysis. Ann. Statist. 39, 2410–2447.
Lam, C., Yao, Q., 2012. Factor modeling for high dimensional time-series: inference for the number of factors. Ann. Statist. 40, 694–726.
Lawley, D., Maxwell, A., 1971. Factor Analysis as a Statistical Method. Butterworths, London.
Lee, J.D., Sun, Y., Taylor, J.E., 2015. On model selection consistency of regularized M-estimators. Electron. J. Stat. 9, 608–642.
Liu, L.M., Hudak, G.B., Box, G.E., Muller, M.E., Tiao, G.C., 1992. Forecasting and Time Series Analysis using the SCA Statistical System, vol. 1(2). Scientific Computing Associates, DeKalb, IL.
Luo, R., Wang, H., Tsai, C.L., 2009. Contour projected dimension reduction. Ann. Statist. 37, 3743–3778.
McCracken, M., Ng, S., 2016. FRED-MD: A monthly database for macroeconomic research. J. Bus. Econom. Statist. 34, 574–589.
Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the lasso. Ann. Stat. 34, 1436–1462.
Merlevède, Florence, Peligrad, Magda, Rio, Emmanuel, 2011. A bernstein type inequality and moderate deviations for weakly dependent sequences. Probab. Theory Related Fields 151, 435–474.
Montgomery, A.L., Zarnowitz, V., Tsay, R.S., Tiao, G.C., 1998. Forecasting the US unemployment rate. J. Amer. Statist. Assoc. 93 (442), 478–493.
Schwarz, G.E., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.
Shen, D., Shen, H., Zhu, H., Marron, J.S., 2016. The statistics and mathematics of high dimension low sample size asymptotics. Statist. Sinica 26, 1747–1770.
Stock, J., Watson, M., 2002. Forecasting using principal components from a large number of predictors. J. Amer. Statist. Assoc. 97, 1167–1179.
Tiao, G.C., Tsay, R.S., 1989. Model specification in multivariate time series. J. R. Stat. Soc. Ser. B Stat. Methodol. 51, 157–213.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58, 267–288.
Tsay, R.S., Tiao, G.C., 1985. Use of canonical analysis in time series model identification. Biometrika 72, 299–315.
Wainwright, M.J., 2009. Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso) quadratic programming (Lasso). IEEE Trans. Inform. Theory 55, 2183–2202.
Wang, H., 2012. Factor profiled sure independence screening. Biometrika 99, 15–28.
Wang, W., Fan, J., 2017. Asymptotics of empirical eigen-structure for ultra-high dimensional spiked covariance model. Ann. Statist. 45, 1342–1374.
Wang, X., Leng, C., 2016. High dimensional ordinary least squares projection for screening variables. J. R. Stat. Soc. Ser. B Stat. Methodol. 78, 589–611.
Zhang, C.H., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. 38, 894–942.
Zhao, P., Yu, B., 2006. On model selection consistency of Lasso. J. Mach. Learn. Res. 7, 2541–2563.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67, 301–320.
Zou, H., Li, R., 2008. One-step sparse estimates in nonconcave penalized likelihood models. Ann. Stat. 36, 1509–1533.