High-dimensional Macroeconomic Forecasting and Variable Selection via Penalized Regression

Yoshimasa Uematsu[1]

University of Southern California

Shinya Tanaka[2]

Aoyama Gakuin University

May 30, 2018

Abstract

This paper studies high-dimensional forecasting and variable selection via folded-concave penalized regressions. The penalized regression approach leads to sparse estimates of the regression coefficients, and allows the case where dimensionality of the model is much larger than the sample size. The first half of the paper discusses theoretical aspects of the penalized regression under a time series setting. Specifically, we show the oracle inequality with ultrahigh-dimensional time-dependent regressors. The latter half of the paper shows the validity of the penalized regression in two motivating empirical applications. First, we consider forecasting of quarterly U.S. GDP with high-dimensional monthly dataset using unrestricted MIDAS framework with penalization. Second, we examine how well the penalized regression screens a hidden portfolio from a NYSE stock price large dataset. Both applications show that the penalized regression provides remarkable results in terms of forecasting performance and variable selection.

Keywords: *Ultrahigh-dimensional time series, Penalized regression, Oracle inequality, Macroeconomic forecasting, Mixed data sampling (MIDAS), Portfolio selection.*

JEL classification: C13, C32, C52, C53, C55

1 Introduction

Recent advancements in macroeconomic data collection have led to an increased focus on high-dimensional time series analysis. A more efficient and precise analysis could be realized

[1] Corresponding address: Yoshimasa Uematsu, Data Sciences and Operations Department, University of Southern California, Los Angeles, CA 90089, USA. E-mail: yuematsu@marshall.usc.edu.

[2] Department of Economics, Aoyama Gakuin University, 4-4-25 Shibuya, Shibuya-ku, Tokyo 150-8366, Japan. E-mail: shinya.tanaka1882@aoyamagakuin.jp.

if we elicit information appropriately from a large number of explanatory variables. However, the performance varies depending on the dimensionality and which estimation method is considered. Without appropriate dimension reduction, performance may be poor owing to accumulated estimation losses from redundant or unimportant variables. After seminal papers on diffusion index (DI) forecasting, such as Stock and Watson (2002), a factor model is now a common tool for forecasting with large datasets. Specifically, Stock and Watson (2012) showed that factor-based forecasts have a good performance in comparison with existing forecasting methods, including autoregressive forecast, pretest methods, Bayesian model averaging, empirical Bayes, and bagging. They concluded that the DI successfully works to reduce the dimension of the regression and it seems difficult to outperform it without introducing any drastic changes to a forecast model. Recent extensions of factor models which could be applicable to forecasting include Bai and Liao (2016), Fan, Ke and Liao (2016a), Fan, Liao and Wang (2016b), Hansen and Liao (2016), and Fan, Xue and Yao (2017).

In addition to such factor approaches, *sparse modeling* is another direction to dimension reduction and has rapidly progressed in statistics and econometrics. A sparse model is assumed to contain only a few relevant covariates while many irrelevant ones with zero coefficients. One of the advantages is to allow for ultrahigh dimensionality of covariates, where the number of regressors diverges sub-exponentially. The model can be estimated by a *penalized regression*, such as the Lasso by Tibshirani (1996) and Frank and Friedman (1993) as a special case, the smoothly clipped absolute deviation (SCAD) penalized regression by Fan and Li (2001), and regression with the minimax concave penalty (MCP) by Zhang (2010). The unknown sparsity can be recovered by the penalized regression to pursue both prediction efficiency and variable selection consistency. In these days, the sparse modeling is also of great concern to macroeconometricians because it can handle large macroeconomic dataset effectively and expected to cast as an alternative of the factor model; in particular, see Bai and Ng (2008), De Mol, Giannone, and Reichlin (2008), Li and Chen (2014), Marsilli (2014), and Nicholson, Matteson, and Bien (2015). It is worth mentioning here that they mainly focused on the ℓ_1-penalty (Lasso) though it may lack model selection consistency while the SCAD and MCP can have. Moreover, to the best of our knowledge, it is hard to find a statistical theory for these SCAD-type penalties in a time series context. With such motivations, the paper sheds light on the validity of the penalized regression through comprehensive theoretical and empirical investigations that are suitable for macroeconomic time series.

Time series analysis and macroeconomic forecasting in high-dimensional settings have been well developed in late years. Basu and Michailidis (2015) establishes the oracle inequality for the lasso with weak dependence under Gaussian assumption. High dimensional Vector autoregressions (VAR) have been explored by several authors, including Song and Bickel (2011), Callot and Kock (2014), and Kock and Callot (2015). A Bayesian VAR in high dimension is investigated by Banbura, Giannone and Reichlin (2010), Koop (2013) and Gefang (2014). Kock (2016) considers adaptive lasso in not only stationary but also non-stationary autoregressions. Davis, Zang, and Zheng (2016) proposes a two-stage approach to estimating sparse VARs based on the partial coherence and t-test statistic together with the BIC. For panel data models in high dimensions, Kock (2013) considers estimation of models with random and fixed effects and establishes the oracle property. Kock and Tang (2018) achieves the oracle inequalities and investigates uniformly valid inference for panel data models with fixed effects. Regarding panel VAR models, Schnucker (2017) proposes a new Lasso-type estimator and conducts forecasting as an empirical example. Koop and Korobilis (2018) explores Bayesian panel VAR with time-varying parameters and stochastic volatility. In terms of variable selection in predictive regressions, Ng (2013) investigates finite sample properties of forecast values based on the BIC, AIC and Lasso, and finds the Lasso has rather stable compared to the others. Kallestrup-Lamb, Kock, and Kristensen (2016) uses the (adaptive) lasso in high-dimensional linear and logistic regressions to analyze the retirement decision of workers. Another direction for forecasting can be found in Kock and Teräsvirta (2014) and Kock and Teräsvirta (2016); they consider forecasting during an economic crisis based on neural network models and novel three automated modeling techniques. Li, Zbonakova, and Härdle (2017) combines propagation-separation approach and SCAD-penalized regression to perform variable selection and capture parameter instability simultaneously. Smeekes and Wijler (2018) finds through both simulation and empirical studies that the lasso-type estimators can outperform factor approaches when there exists a non-sphericity in the idiosyncratic component or cointegrating relation among variables. Kim and Swanson (2018) examines similar analyses and also concludes machine learning and shrinkage methods are useful for forecasting.

In the first half of this paper, we provide theoretical properties of the penalized regression estimator with the SCAD-type penalties as well as ℓ_1-penalty under suitable conditions for macroeconometrics, from the perspective of prediction efficiency. In light of recent development of the literature, this has not been fully addressed in the literature. We first derive a non-asymptotic upper bound for the prediction loss called the *oracle inequality* that can be enjoyed by the Lasso and SCAD-type penalties. This ensures that the forecast value is reliable with implying optimality in the asymptotic sense. Next, we also discuss inferential aspects known as the *oracle property* for the SCAD-type penalization and its limitation. This property gives correct selection of the subset of predictors and estimation of the non-zero coefficients as efficiently as would be possible if we knew which variables were irrelevant. Note that this property will basically be endowed to the SCAD-type penalties rather than the Lasso. For theoretical details on the oracle property for time series models, see Medeiros and Mendes (2016).

In the second half, we check the empirical validity of the penalized regression in macroeconometrics by two applications. First, in order to observe the validity of oracle inequality, we consider to forecast quarterly U.S. real GDP with a large number of monthly predictors using MIDAS (MIxed DAta Sampling) regression framework originally proposed by Ghysels, Sinko, and Valkanov (2007). Since the total number of parameters is much larger than that of observations, this situation should be treated as an ultra-high dimensional problem. In contrast to the original MIDAS model of Ghysels, Sinko, and Valkanov (2007), the penalized regression enables us to forecast the quarterly GDP using a large number of monthly predictors without imposing a distributed lag structure on the regression coefficients. We find that the forecasting performance of the penalized regression is better than that of the factor-based MIDAS (F-MIDAS) regression by Marcellino and Schumacher (2010) and is competitive with a nowcasting model based on the state-space representation in real-time forecasting. Second, to observe accuracy of model selection, we investigate how well the penalized regression can screen a hidden fund manager's portfolio from a large-dimensional NYSE stock price data set. We construct artificial portfolios, and then confirm that the SCAD-type penalized regression effectively detects the relevant stocks better than the Lasso. These two convincing empirical applications motivate us to apply the penalized regression to macroeconomic time series broadly.

The remainder of this paper is organized as follows. Section 2 specifies a ultrahigh-dimensional time series regression model and the estimation scheme. A statistical validity of the method is explored in Section 3 by deriving the oracle inequality. A limitation of the inferential aspects is also discussed. Section 4 illustrates how we can apply the penalized regression to macroeconomic time series by two empirical analyses, such as forecasting and portfolio screening. Section 5 concludes. The proofs and miscellaneous results are collected in Appendix.

2 Regression Model

The regression model to be considered is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}, \tag{1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_T)^\top$ is a response vector, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)^\top$ is a covariate matrix with $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tp})^\top$, $\boldsymbol{u} = (u_1, \ldots, u_T)^\top$ is an error vector, and $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0A}^\top, \boldsymbol{\beta}_{0B}^\top)^\top$ is a p-dimensional unknown sparse parameter vector with $\boldsymbol{\beta}_{0A} = (\beta_{0,1}, \ldots, \beta_{0,s})^\top$ the s-dimensional subvector of nonzero elements and $\boldsymbol{\beta}_{0B} = \boldsymbol{0}$. We also denote jth column vector of $\boldsymbol{X}$ by $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{Tj})^\top$. Further, we write $\boldsymbol{X} = (\boldsymbol{X}_A, \boldsymbol{X}_B)$ corresponding to the decomposition of the parameter vector. Throughout the paper, we assume that for each j, $\{x_{tj}u_t\}_t$ is a martingale difference sequence with respect to an appropriate filtration.

The objective of the paper is how we construct an efficient h-step ahead forecast of y_{T+h} and how we select variables consistently when dimension p is much larger than T. More specifically, we consider an *ultrahigh-dimensional* case, meaning that p diverges sub-exponentially (non-polynomially). In such cases, $\boldsymbol{X}$ may contain many irrelevant columns, so that the sparsity assumption on $\boldsymbol{\beta}_0$ is appropriate. At the same time, the degree of sparsity s may also diverge, but $s < T$ must be satisfied. The estimation procedure should select a relevant model as well as consistently estimate the parameter vector. The estimator $\hat{\boldsymbol{\beta}}$ is defined as a minimizer of the objective function

$$Q_T(\boldsymbol{\beta}) \equiv (2T)^{-1}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \|p_\lambda(\boldsymbol{\beta})\|_1 \tag{2}$$

over $\boldsymbol{\beta} \in \mathbb{R}^p$, where $p_\lambda(\boldsymbol{\beta}) \equiv (p_\lambda(|\beta_1|), \ldots, p_\lambda(|\beta_p|))^\top$ and $p_\lambda(v)$, for $v \geq 0$, is a penalty function indexed by a regularization parameter $\lambda (= \lambda_T) > 0$. The penalty function p_λ takes forms such as the ℓ_1-penalty (Lasso) by Tibshirani (1996), SCAD penalty by Fan and Li (2001), and MCP by Zhang (2010). These penalties belong to a family of so-called *folded-concave penalties* because of their functional forms. The statistical properties have been developed for models with a deterministic covariate and i.i.d. Gaussian errors in the literature on high-dimensional statistics. We thoroughly investigate these properties, while relaxing the assumptions sufficiently to include many time series models.

We introduce the three penalties to be used. Let v denote a positive variable. The ℓ_1-penalty is given by $p_\lambda(v) = \lambda v$, and we then obtain $p'_\lambda(v) = \lambda$ and $p''_\lambda(v) = 0$. The SCAD penalty is defined by

$$p_\lambda(v) = \lambda v 1\{v \le \lambda\} + \frac{a\lambda v - 0.5(v^2 + \lambda^2)}{a-1} 1\{\lambda < v \le a\lambda\} + \frac{\lambda^2(a^2-1)}{2(a-1)} 1\{v > a\lambda\}.$$

Its derivative is

$$p'_\lambda(v) = \lambda \left\{ 1(v \le \lambda) + \frac{(a\lambda - v)_+}{(a-1)\lambda} 1(v > \lambda) \right\},$$

for some $a > 2$. Then we have $p''_\lambda(v) = -(a-1)^{-1} 1\{v \in (\lambda, a\lambda)\}$. The MCP is defined by

$$p_\lambda(v) = \left(\lambda v - \frac{v^2}{2a} \right) 1\{v \le a\lambda\} + \frac{1}{2} a\lambda^2 1\{v > a\lambda\}.$$

Its derivative is $p'_\lambda(v) = a^{-1}(a\lambda - v)_+$ for some $a \ge 1$. Thus, we have $p''_\lambda(v) = -a^{-1} 1\{v < a\lambda\}$.

3 Theoretical Result

In this section, we establish an important theoretical result, the *oracle inequality* for time series models. The oracle inequality gives an optimal non-asymptotic error bound for estimation and prediction in the sense that the error bound is of the same order of magnitude up to a logarithmic factor as those we would have if we a priori knew the relevant variables; see Buhlman and van de Geer (2011). This result strongly supports the use of penalized regressions in terms of forecasting accuracy, even in ultrahigh-dimensional settings. The existing researches have shown the oracle inequality under i.i.d. Gaussian errors and deterministic covariates, but in the paper we extend the result to apply to a certain class of time series models.

Assumption 1. Penalty function $p_\lambda(\cdot)$ is characterized as follows:

(a) $p_\lambda(v)$ is concave in $v \in [0, \infty)$ with $p_\lambda(0) = 0$;

(b) $p_\lambda(v)$ is nondecreasingin $v \in [0, \infty)$;

(c) $p_\lambda(v)$ has a continuous derivative $p'_\lambda(v)$ for $v \in (0, \infty)$ with $p'_\lambda(0+) = \lambda$.

Assumption 1 determines a family of folded-concave penalties that bridges ℓ_0- and ℓ_1-penalties. The SCAD and MCP are included in this family. The ℓ_1-penalty also satisfies this as the boundary of this class.

We define the gradient vector and Hessian matrix of $(2T)^{-1}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ as $G_T(\boldsymbol{\beta}) \equiv -\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/T$ and $\boldsymbol{H}_T \equiv \boldsymbol{X}^\top\boldsymbol{X}/T$, respectively. Denoting $\boldsymbol{G}_{0T} \equiv G_T(\boldsymbol{\beta}_0)$, we may write

$$\boldsymbol{G}_{0T} = -\frac{1}{T}\begin{pmatrix} \boldsymbol{X}_A^\top \boldsymbol{u} \\ \boldsymbol{X}_B^\top \boldsymbol{u} \end{pmatrix} \equiv \begin{pmatrix} \boldsymbol{G}_{0AT} \\ \boldsymbol{G}_{0BT} \end{pmatrix},$$

$$\boldsymbol{H}_T = \frac{1}{T}\begin{pmatrix} \boldsymbol{X}_A^\top \boldsymbol{X}_A & \boldsymbol{X}_A^\top \boldsymbol{X}_B \\ \boldsymbol{X}_B^\top \boldsymbol{X}_A & \boldsymbol{X}_B^\top \boldsymbol{X}_B \end{pmatrix} \equiv \begin{pmatrix} \boldsymbol{H}_{AAT} & \boldsymbol{H}_{ABT} \\ \boldsymbol{H}_{BAT} & \boldsymbol{H}_{BBT} \end{pmatrix}.$$

3.1 Oracle Inequality

We derive optimal non-asymptotic error bounds for estimation and prediction called the oracle inequality. In the literature, Ch. 6 of Buhlman and van de Geer (2011) presented a complete guide for the inequality using the ℓ_1-penalty with fixed predictors and i.i.d. normal errors. We extend the result in two ways. First, the inequality holds for the general model (1). Second, we prove that the upper bounds for the errors under ℓ_1- and the other folded-concave penalizations characterized by Assumption 1 are the same up to a constant factor. This indicates that the forecasting error bounds decay with the same asymptotic rate, irrespective of the folded-concave penalty used. We first derive the bounds under two high-level assumptions in Section 3.1.1. We next consider the conditions under which the two high-level assumptions are actually verified in a reasonable time series setting in Section 3.1.2.

3.1.1 Generl result from the literature

First we review a general result known in the existing literature. We start with the following high-level assumptions:

Assumption 2. There are a positive sequence $\lambda = \lambda_{pT}$ and a positive constant c_1 such that $\mathcal{E}_1^c$, the complement of the event $\mathcal{E}_1 = \{\|\boldsymbol{G}_{0T}\|_\infty \leq \lambda/2\}$, satisfies $P(\mathcal{E}_1^c) = O(p^{-c_1})$.

Assumption 3. For $\mathbb{V} = \{\boldsymbol{v} \in \mathbb{R}^p : \|\boldsymbol{v}_B\|_1 \leq 3\|\boldsymbol{v}_A\|_1\}$, there are positive constants c_2 and γ such that $\mathcal{E}_2^c$, the complement of the event $\mathcal{E}_2 = \left\{\min_{\boldsymbol{v}\in\mathbb{V}} T^{-1}\|\boldsymbol{X}\boldsymbol{v}\|_2^2/\|\boldsymbol{v}\|_2^2 \geq \gamma\right\}$, satisfies

$P(\mathcal{E}_2^c) = O(p^{-c_2})$.

These two assumptions fully control the randomness of the regression model, irrespective of the dependence structure and tail behaviors. Assumption 2 requires that the gradient vector $\boldsymbol{G}_{0T}$ to behave less fluctuate and converge to zero with an appropriate rate determined by λ. Assumption 3 is a stochastic version of the *restricted strong convexity* studied by Negahban, Ravikumar, Wainwright, and Yu (2012). This postulates that the minimum eigenvalue of sub-matrices of Hessian matrix $\boldsymbol{H}_T$ should not be too small and bounded by γ from below.

Under the assumptions listed above, we can derive the following theorem:

Theorem 1 (Oracle inequality)**.** *Let Assumptions 1–3 hold. Then, for any minimizer* $\hat{\boldsymbol{\beta}}$ *of* $Q_T(\boldsymbol{\beta})$*, the following inequalities hold simultaneously with probability at least* $1 - O(p^{-c_1}) - O(p^{-c_2})$*:*

(a) (Estimation error in ℓ_2*-norm)* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \leq 12 s^{1/2}\lambda/\gamma$*;*

(b) (Estimation error in ℓ_1*-norm)* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq 48 s\lambda/\gamma$*;*

(c) (Prediction loss) $T^{-1/2}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2 \leq 144 s^{1/2}\lambda/\gamma^{1/2}$*.*

Theorem 1 provides *non-asymptotic* error bounds for any combination of p, T, and s. Note that the results have richer information than the conventional asymptotic results. First, the asymptotic results are implied by the non-asymptotic results. In fact, if γ is assumed to be fixed, the error bounds (a) and (b) converge to zero as long as λ goes to zero relatively faster than $s^{1/2}$ or s and it implies consistency of the estimator. More specifically, in a simple setting with i.i.d. Gaussian u_t and nonrandom $\boldsymbol{X}_t$, it is known that λ should be given by $O((\log p/T)^{1/2})$, leading to the explicit convergence rates $O((s\log p/T)^{1/2})$ for (a) and (c), and $O(s(\log p/T)^{1/2})$ for (b). Second, the non-asymptotic bounds reveal how correlation between the covariates affects the estimation and prediction accuracy. If γ becomes small, we can observe the upper bound tends to loose. Result (c) exhibits an optimal bound for the prediction loss in the ℓ_2-norm in the sense of Buhlman and van de Geer (2011). This result justifies using any penalty function specified by Assumption 1 when the aim is forecasting due to the following two reasons. First, the bound in (c) means

$$T^{-1}\sum_{t=1}^{T}\left(\boldsymbol{x}_t^\top\hat{\boldsymbol{\beta}} - \boldsymbol{x}_t^\top\boldsymbol{\beta}\right)^2 = T^{-1}\sum_{t=1}^{T}\left(\hat{y}_t - E[y_t|\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t]\right)^2 \to 0$$

with high probability, under an appropriate choice of λ. This fact implies the forecast values converges to its conditional expectations in mean square. Second, the bound (c) gives the best possible bound of the forecast value under ultrahigh dimensionality. Consider a simple case such that $\boldsymbol{X}$ is deterministic, $\boldsymbol{u}$ is i.i.d. with a unit variance, and $s = p < T$. Then, the squared risk of the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$ becomes

$$T^{-1}\mathrm{E}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0)\|_2^2 = T^{-1}\mathrm{E}[\boldsymbol{u}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{u}] = T^{-1}\mathrm{tr}\boldsymbol{I} = s/T.$$

Likewise, consider the case $p \geq T > s$. If we knew the true model A, we could choose the correct s variables from $\boldsymbol{X}$, leading to the risk s/T. On the other hand, since A is unknown in practice, we must pay for an extra cost for not knowing A. However, we can choose λ as $O((\log p/T)^{1/2})$ in the simplest case and it means the extra cost is only $\log p$ compared to conventional $(p < T)$ case.

3.1.2 When does Theorem 1 hold?

Theorem 1 has established the non-asymptotic error bounds for the penalized regression estimators and prediction error under general but high-level assumptions. In applications, Assumptions 2 and 3 should be verified for each model we attempt to employ. Here we consider a specific time series model. We first introduce two classes of random variables.

Definition 1. A random variable $X \in \mathbb{R}$ is said to be *sub-Gaussian* with variance proxy α^2 if $\mathrm{E}[X] = 0$ and its moment generating function satisfies $\mathrm{E}[\exp(sX)] \leq \exp(\alpha^2 s^2/2)$ for all $s \in \mathbb{R}$. In this case, we write $X \sim \mathrm{subG}(\alpha^2)$.

Definition 2. A random variable $Y \in \mathbb{R}$ is said to be *sub-exponential* with parameter γ if $\mathrm{E}[Y] = 0$ and its moment generating function satisfies $\mathrm{E}[\exp(sY)] \leq \exp(\gamma^2 s^2/2)$ for all $|s| \leq b^{-1}$. In this case, we write $Y \sim \mathrm{subE}(\gamma, b)$. Furthermore, we denote $\mathrm{subE}(\gamma)$ when $b = \gamma$.

Definition 1 forms a family of random variables whose tails decay at least as fast as the Gaussian tail. More specifically, $X \sim \mathrm{subG}(\alpha^2)$ has the tail probability $P(|X| > x) \leq 2\exp\{-x^2/(2\alpha^2)\}$. Note that α^2 is not the variance of $X \sim \mathrm{subG}(\alpha^2)$. In fact, it is known that $\mathrm{E}[X^2] \leq 4\alpha^2$, so that the variance can be larger than α^2. Similarly, for $Y \sim \mathrm{subE}(\gamma, b)$, we have the tail probability $P(|Y| > y) \leq 2\exp\{-y^2/(2\gamma^2)\}$ for all $y \in [0, \gamma^2/b]$. An important fact is that the product of two independent subG random variables becomes subE;

see Lemma 1 in Appendix for the details. Based on this property, the proofs of propositions below will go through by using Bernstein's inequality for subE random variables.

In order to specify the processes of $\boldsymbol{X}$ and $\boldsymbol{u}$, we assume the same structure as Ahn and Horenstein (2013); the covariate $\boldsymbol{X}$ and the error $\boldsymbol{u}$ are respectively given by

$$\boldsymbol{X} = \boldsymbol{R}^{1/2}\boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}, \qquad \boldsymbol{u} = \boldsymbol{S}^{1/2}\boldsymbol{e}\tau, \tag{3}$$

where $\boldsymbol{R}^{1/2} = (r_{st}) \in \mathbb{R}^{T\times T}$, $\boldsymbol{S}^{1/2} = (s_{st}) \in \mathbb{R}^{T\times T}$, $\boldsymbol{\Sigma}^{1/2} = (\sigma_{ij}) \in \mathbb{R}^{p\times p}$, and $\tau > 0$ are deterministic and $\boldsymbol{Z} = (z_{ti}) \in \mathbb{R}^{T\times p}$ and $\boldsymbol{e} = (e_t) \in \mathbb{R}^{T}$ are composed of random variables. More specifically, they are characterized by the following assumption:

Assumption 4. The following conditions hold:

(a) $z_{ti}, e_t \sim$ i.i.d. subG(α^2) for some positive constant α;

(b) $\boldsymbol{R}$, $\boldsymbol{S}$, and $\boldsymbol{\Sigma}$ are symmetric and positive definite non-random matrices; the minimum and maximum eigenvalues of which are all bounded from below and above by positive constants $c_{\min}$ and $c_{\max}$, respectively;

(c) $\boldsymbol{R}^{1/2}$ and $\boldsymbol{S}^{1/2}$ are lower triangular matrices such that $r_{tt} = s_{tt} = 1$ and

$$\max_{t\in\{1,\ldots,T\}} \|\boldsymbol{r}_{\cdot t}\|_1 + \max_{s\in\{1,\ldots,T\}} \|\boldsymbol{r}_{s\cdot}\|_2 + \max_{s\in\{1,\ldots,T\}} \|\boldsymbol{s}_{s\cdot}\|_2 = O(1).$$

Moreover, $\boldsymbol{\Sigma}^{1/2}$ is a positive definite matrix satisfying $\max_{j\in\{1,\ldots,p\}} \|\boldsymbol{\sigma}_{\cdot j}\|_2 = O(1)$.

Matrices in condition (c) are defined based on the Cholesky decomposition and Spectral decomposition under condition (b). Model (3) with Assumption 4 covers a wide range of time series processes with cross-sectional dependences. A simple example of $\boldsymbol{R}^{1/2}$ and $\boldsymbol{\Sigma}^{1/2}$ is given by setting $r_{t,t-1} = \theta_r$ and $\sigma_{i,i-1} = \varphi_\sigma$ for some constants θ_r and φ_σ satisfying $|\theta_r| < 1$ and $|\varphi_\sigma| < \infty$ with other entries all zero. Obviously, this formulation satisfies condition (c) with reducing model (3) to an MA(1) process. Other weak stationary processes with cross-sectional dependences can be expressed in a similar manner.

Proposition 1. *Let Assumption 4 hold with* $\lambda = c_0(\log p/T)^{1/2}$*, with choosing*

$$c_0 = 8\tau e\alpha^2 \max_{j\in\{1,\ldots,p\}} \|\boldsymbol{\sigma}_{\cdot j}\|_2 \max_{t\in\{1,\ldots,T\}} \|\boldsymbol{r}_{\cdot t}\|_1 \max_{s\in\{1,\ldots,T\}} \|\boldsymbol{s}_{s\cdot}\|_2 (2+2\nu)^{1/2}$$

for $\nu > 0$ *arbitrary but fixed constant. Then, Assumption 2 is satisfied with* $P(\mathcal{E}_1^c) = 2p^{-\nu}$.

Proposition 2. *Let Assumption 4 hold and assume* $s(\log p/T)^{1/2} = o(1)$*, with choosing* $\gamma = c_{\min}^2 - 16\gamma_0 s(\log p/T)^{1/2}$*, where*

$$\gamma_0 = 32\alpha^2 \max_{j\in\{1,\ldots,p\}} \|\boldsymbol{\sigma}_{\cdot j}\|_2 \max_{t\in\{1,\ldots,T\}} \|\boldsymbol{r}_{\cdot t}\|_1 \max_{s\in\{1,\ldots,T\}} \|\boldsymbol{r}_{s\cdot}\|_2 (4+2\nu)^{1/2}$$

for $\nu > 0$ *arbitrary but fixed constant. Then, Assumption 3 is satisfied with* $P(\mathcal{E}_2^c) \leq 4p^{-\nu}$*.*

Combining Propositions 1 and 2 leads to the non-asymptotic bounds in the time series setting specified by Assumption 4. Note that $c_0 + \gamma_0 = O(1)$ by Assumption 4(c). Thus, the rate of convergence is completely determined by λ and s.

Corollary 1. *Let Assumptions 1 and 4 hold with the constants being the same as in Propositions 1 and 2. Then, for any minimizer* $\hat{\boldsymbol{\beta}}$ *of* $Q_T(\boldsymbol{\beta})$*, results (a)–(c) of Theorem 1 hold with probability at least* $1 - 6p^{-\nu}$*. In addition, if we suppose* $\log p = O(T^{\delta})$ *and* $s = O(T^{\delta_0})$ *for constants* $\delta, \delta_0 \in (0,1)$ *such that* $\delta + 2\delta^0 < 1$*, results (a)–(c) imply consistency.*

The latter statement in Corollary 1 is easily verified since $s\lambda = o(1)$ and condition $s(\log p/T)^{1/2} = o(1)$ in Proposition 2 are equivalent to condition $\delta + 2\delta^0 < 1$ in the corollary, the region of which is included in the region implied by $s^{1/2}\lambda = o(1)$. Under the condition, it turns out that the rate of convergence is the same as the conventional rate, $O\left((s\log p/T)^{1/2}\right)$, obtained with i.i.d. normal errors and fixed covariates. From the obtained results so far, we can anticipate that the rate of convergence will change when mitigating the summability conditions in Assumption 4(c). If some norms diverge (slowly) in Assumption 4(c), the resulting rate will become slower. Another factor is a fatter tail behavior than the assumed subG random variables. For more information on effects of heavy-tailedness, see Section C in Appendix.

3.2 Discussion: Inference Based on Penalized Regressions

We have observed that the penalized regressions can achieve the oracle inequality under reasonable assumptions which indicate efficient prediction. At the same time, we want to know inferential aspects; that is, we are interested in properties of model selection and asymptotic distribution of the Lasso and SCAD-type penalized regressions. It is well known from the literature that the Lasso has limitation in capacity of selecting the underlying true submodel while the SCAD-type penalized regression have possibility to do so. In fact, the SCAD-type penalization has a chance to enjoy the oracle property:

Property 1 (Oracle property)**.** *Under some assumptions, there exists a local minimizer* $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_A^\top, \hat{\boldsymbol{\beta}}_B^\top)^\top$ *of* $Q_T(\boldsymbol{\beta})$ *such that*

(a) (Sparsity) $\hat{\boldsymbol{\beta}}_B = \mathbf{0}$ *with probability approaching one;*

(b) (Rate of convergence) $\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\|_2 = O_p((s/T)^{1/2})$*;*

(c) (Asymptotic normality) $T^{1/2}\boldsymbol{b}^\top \boldsymbol{I}_{0AA}^{-1/2} \boldsymbol{J}_{0AA}^\top (\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}) \to_d N(0,1)$ *for any* $\boldsymbol{b} \in \mathbb{R}^s$ *satisfying* $\|\boldsymbol{b}\|_2^2 = 1$*, where* $\boldsymbol{I}_{0AA} = T\mathrm{E}[\boldsymbol{G}_{0AT}\boldsymbol{G}_{0AT}^\top]$ *and* $\boldsymbol{J}_{0AA} = \mathrm{E}[\boldsymbol{H}_{AAT}]$.

Results (a) and (b) means that the SCAD-type penalized regression has the variable selection consistency. Result (c) guarantees that the non-zero coefficients can be efficiently estimated as if A were known in advance; see Fan and Lv (2011) for a complete guide in an i.i.d. setting.

The point is that this property holds only in a very ideal case. Unfortunately, some of the required assumptions are hardly satisfied in many high-dimensional macroeconometric applications. For example, a condition suffer from collinearity caused by strong correlations among many macroeconomic variables. Actually, under strong collinearity, the model selection consistency tends to be violated; see Section D in Appendix. In addition, a more serious problem arises due to the so-called *beta-min* condition:

$$\min_{j \in A} |\beta_{0,j}| \gg \lambda = O\left((\log p/T)^{1/2}\right).$$

This is necessary for the oracle property to distinguish the nonzero coefficient of relevant variables from zero, with ruling out variables that have small magnitude of coefficients. However, in a series of articles by Leeb and Pötscher, they have claimed that under the condition inference lacks uniform validity over sequences of models that include even minor deviations from conditions implying perfect model selection; see Section 6 of Chernozhukov, Hansen and Spindler (2015). As a result, asymptotic normal approximations in (c) may become very poor; see e.g., Leeb and Potscher (2008) and Potscher and Leeb (2009). To avoid this problem, a number of methods have been proposed recently, such as the double selection by Belloni, Chernozhukov and Hansen (2014a), partialling out by Belloni, Chernozhukov and Wang (2014b), and de-biasing by Javanmard and Montanari (2014) and Zhang and Zhang (2014). For a comprehensive list of references, see Chernozhukov, Hansen and Spindler (2015).

4 Empirical Examples

In this section, we provide two empirical examples that illustrate how well the penalized regression works in macroeconometric analyses. The first example is motivated by the oracle inequality that justifies to employ the penalized regression for forecasting. We focus on macroeconomic forecasting with mixed-frequency data. The second example considers variable selection the SCAD-type penalized estimator is expected to have by investigating hidden portfolio screening.

4.1 Forecasting Quarterly U.S. GDP with a Large Number of Predictors

4.1.1 Penalized MIDAS regression

In this section, we illustrate how to apply the penalized regression model to macroeconomic time series using the MIDAS (MIxed DAta Sampling) regression framework. The MIDAS regression model was originally proposed by Ghysels, Sinko, and Valkanov (2007) and it is now one of the standard tools for macroeconomic forecasting with mixed-frequency data, as well as the now-casting model based on the state-space representation: see e.g., Giannone, Reichlin, and Small (2008) and Banbura and Modugno (2014). The original (or basic) MIDAS regression model has an advantage of describing a forecasting model with a large number of lags of the predictors in a simple and parsimonious way by employing a distributed lag structure with a few hyperparameters. However, the distributed lag structure seems quite restrictive, and more importantly, the original MIDAS is not suitable when the number of predictors is very large. In fact, the original MIDAS can just reduce the dimensionality originated from lags, but not the dimensionality of predictors itself. For example, consider the original MIDAS regression model including a constant term with N macroeconomic time series, K ($K \ll N$) hyperparameters, and ℓ lags. In this case, the total number of parameters to be estimated is $NK+1$, which is reduced from $NK(\ell+1)+1$. Thus, the original MIDAS only successfully reduces the total number of parameters regarding the lags, but it invokes a serious efficiency loss or even it makes the model inestimable once N is getting very large. On the other hand, the penalized regression scheme enables us to directly estimate the high-dimensional MIDAS regression without imposing any distributed lag structures, since the coefficients of irrelevant predictors are automatically estimated to be zero.

Here we propose the *penalized MIDAS regression* model that applies the penalized regression framework to the mixed-frequency data. This can reduce the dimensionality caused from both N and ℓ. Let $\{y_t, \boldsymbol{x}_{t/m}^{(m)}\}$ be the MIDAS process in line with Andreou, Ghysels, and Kourtellos (2010), where the scalar y_t is the low-frequency variable observed at $t = 1, \ldots, T$, and the $(N+1)$-dimensional vector $\boldsymbol{x}_{t/m}^{(m)} = \left(1, x_{1,t/m}^{(m)}, \ldots, x_{N,t/m}^{(m)}\right)^\top$ is a set of higher-frequency variables observed m times between t and $t-1$. For example, $m = 3$ if we forecast a quarterly variable with monthly predictors. We consider the h-quarter-ahead mixed-frequency forecasting regression model with ℓ lags,

$$y_t = \boldsymbol{x}_{t-h}^\top \boldsymbol{\beta}_0 + u_t, \qquad t = 1, \ldots, T, \tag{4}$$

where $\boldsymbol{x}_{t-h} = \left(1, \boldsymbol{x}_{1,t-h,\ell}^{(m)\top}, \ldots, \boldsymbol{x}_{N,t-h,\ell}^{(m)\top}\right)^\top$ with

$$\boldsymbol{x}_{k,t-h,\ell}^{(m)\top} = \left(x_{k,t-h}^{(m)}, x_{k,t-h-1/m}^{(m)}, \ldots, x_{k,t-h-\ell/m}^{(m)}\right)$$

for $k = 1, 2, \ldots, N$, $\boldsymbol{\beta}_0 = (\beta_{0,0}, \beta_{0,1}, \ldots, \beta_{0,N(\ell+1)})^\top$ is a *sparse* parameter vector and u_t is an error term. Here the case $h < 1$ $(h = 0, 1/m, 2/m, \ldots, (m-1)/m)$ corresponds to nowcast; we forecast a low-frequency variable with the "latest" high-frequency variables released between $t-1$ and t. For instance, if we consider a quarterly/monthly $(m = 3)$ case, $h = 0$ $(1/3)$ means that we forecast a quarterly variable in 2015Q2 with monthly data in June (May) 2015 or later. Note that model (4) has the same structure as (1) with $p := N(\ell+1)+1$ but it differs from the original MIDAS in the sense that our model does not employ the distributed lag structure on $\boldsymbol{x}_{t-h}$, while the original MIDAS uses $\boldsymbol{x}_{t-h}(\theta) = \left(1, x_{1,t-h}^{(m)}(\theta_1), \ldots, x_{N,t-h}^{(m)}(\theta_N)\right)^\top$ instead of $\boldsymbol{x}_{t-h}$ such that $x_{k,t}^{(m)}(\theta_k) = \sum_{j=1}^{\ell} w_{j,k}(\theta_k) L^{j/m} x_{k,t/m}^{(m)}$ for $k = 1, 2, \ldots, N$, where $w_{j,k}(\theta_k) \in (0,1)$ is a weight term that satisfies $\sum_{j=1}^{\ell} w_{j,k}(\theta_k) = 1$.

Note that the penalized MIDAS regression model (4) stands on the sparsity assumption on forecasting regression coefficients. Although the assumption seems restrictive, it is rather plausible in a macroeconomic forecasting point of view because it is natural to consider that there is a small set of key predictors that contain rich information with non-zero coefficients while there are lots of redundant predictors with zero coefficients when we forecast macroeconomic time series. Bai and Ng (2008) shared the same spirit and they call the key predictors as *targeted predictors*. Here we may also consider the broader situation where the redundant predictors would have very low but non-zero forecasting power. Since the penalized regression makes their coefficient estimates zero even in this case, we can interpret the sparsity assumption leads to an approximation of a true forecasting model. Note further that Marsilli (2014) proposed a similar method to our penalized MIDAS regression, but his model completely differs from ours; he employed the original MIDAS parsimonious parameterization of distributed lags.

4.1.2 Data

U.S. quarterly real GDP growth is taken from the FRED database. The sample period is from 1959Q4 to 2017Q3. We retrieve 117 U.S. monthly macroeconomic time series ($N = 117$) from the FRED–MD database and the series are appropriately detrended according to a guideline given in McCracken and Ng (2015). Note that the FRED–MD database originally contains a total of 128 series, but we remove 11 series because the CBOE S&P 100 Volatility Index (VXOCLSx), Consumer sentiment index (UMCSENTx), Trade weighted U.S. dollar index of major currencies (TWEXMMTH), New orders for nondefense capital goods (ANDENOx), New orders for consumer goods (ACOGNO), and New private housing permits (PERMIT, PERMITNE, PERMITMW, PERMITS, PERMITW) have no observations from 1959. Reserves of depository institutions nonborrowed (NONBORRES) is not used as well since our preliminary inspection found that it contained extreme changes in February 2008, which would contaminate our analysis. The sample period of the detrended monthly series is from April 1959 (1959:3) to September 2017 (2017:9).

4.1.3 Forecasting strategy

We evaluate the out-of-sample forecasting performance by mean squared forecast errors (MSFE) in the evaluation period from 2000Q1 to 2017Q3. The parameter estimates are obtained from each estimation period; the initial period is 1959Q4–1999Q4 and the next one extends the end point to 2000Q1 with the starting point 1959Q4 being fixed. For example, the initial forecast error in 2000Q1 is calculated using the estimates from the initial estimation period 1959Q4–1999Q4, and the second forecast error in 2000Q2 uses the estimates from the second estimation period 1959Q4–2000Q1. We suppose that the forecast regression consists of eight lags ($\ell = 8$), so that the total number of parameters for the forecasting regression to be estimated is $N(\ell + 1) + 1 = 1054$, including a constant term. The penalized MIDAS regression is expected to be robust to the choice of ℓ, as long as we choose ℓ to be moderately large, because the penalized regression conducts dimension reduction as well as parameter estimation. To investigate the forecasting performance of the penalized MIDAS regression model with a variety of horizons, we examine cases where $h = 0, 1/3, 2/3, 1, 4/3, 5/3, 2$ in the same manner as Clements and Galvao (2008) and Marcellino and Schumacher (2010). The cases $h = 0, 1/3$, and $2/3$ correspond to nowcasting in the sense that we forecast contemporaneous or very short-forecast-horizon quarterly GDP growth using monthly series before the official announcement of the GDP, while the case $h = 2$ is a forecast with a relatively long horizon. The sample size of the estimation period T gradually increases and varies depending on h; for example, T ranges from 161 to 231 if $h = 0$, and from 159 to 229 if $h = 2$. Finally, we need to predetermine the values of tuning parameter a and regularization parameter λ. Following the guidelines by Breheny and Huang (2011) with our preliminary inspection of the entire samples, we set $a = 12$ for the SCAD and MCP, although the performance could be improved by a more careful choice. The value of λ is selected by 5-fold cross-validation. The validity of the CV selection was confirmed by Chetverikov, Liao, Chernozhukov (2017) theoretically and Uematsu and Tanaka (2016) in simulation. All the computations are conducted using R 3.4.4 and we use `ncvreg` package of Breheny and Huang (2011) for the penalized regression and `midasr` for constructing MIDAS regressors.

4.1.4 Forecast methods

We consider the following three evaluation periods: (*i*) Overall (2000Q1–2017Q3), (*ii*) 1st subsample (2000Q1–2007Q4), and (*iii*) 2nd subsample (2008Q1–2017Q3) to avoid the period the unprecedented turmoil of the U.S. economy stemming from the subprime mortgage crisis and the ensuing collapse of Lehman Brothers in 2008 that would introduce parameter instability and distort the forecast evaluation. As a result, the forecast performance is evaluated in complete data from a total of 70 (overall), 32 (1st subsample) and 38 (2nd subsample) squared forecast errors, respectively. Due to limitation in the number of observations, we consider only the single economic shock. It would be better to consider other shocks, such as the Black Monday in 1987Q3 and the collapse dot-com bubble around 2001, if we had enough samples. The competitors are a naive AR(4) forecast, the factor MIDAS proposed by Marcellino and Schumacher (2010), and the OLS post-Lasso proposed by Belloni and Chernozhukov (2013).

The factor MIDAS is expected to be one of the strong competitors since the factor-based forecast performs well in forecasting real variables; see e.g., Stock and Watson (2002), Stock and Watson (2012), De Mol, Giannone, and Reichlin (2008). The factor MIDAS considered here is based on the basic MIDAS structure with the exponential Almon lag of two hyperparameters. This is implemented by `midasr` package in R. The number of factors is assumed to be seven ($r = 7$) based on the information criterion IC_{p2} by Bai and Ng (2002). Although we can consider the *unrestricted* factor MIDAS as in Marcellino and Schumacher (2010), which is free from the distributed lag structure, we do not because of its intractability caused by high dimensionality.

The OLS post-Lasso is expected to perform at least as well as the Lasso and to have better performance in some cases. Roughly speaking, the OLS post-Lasso implements the Lasso as a first screening for effective predictors and then run the OLS only with the selected predictors. More precisely, it consists of the following five steps:

1. Let $I_1, \ldots, I_5$ be 5 subsamples for the 5-fold CV and $\{\lambda_1, \ldots, \lambda_M\}$ be a set of potential regularization parameters. We run the Lasso with each $\lambda \in \{\lambda_1, \ldots, \lambda_M\}$ in kth subsample I_k for $k = 1, \ldots, 5$, and get the active sets $\widehat{A}_1(\lambda), \ldots, \widehat{A}_5(\lambda)$.

2. Run the OLS of y_t on $\{\boldsymbol{x}_j\}$ such that $j \in \widehat{A}_k(\lambda)$ in each subsample I_k and get $\widehat{\boldsymbol{\beta}}_1^{OLS}(\lambda), \ldots, \widehat{\boldsymbol{\beta}}_5^{OLS}(\lambda)$.

3. Determine λ^* as to satisfy

$$\lambda^* = \underset{\lambda \in \{\lambda_1, \ldots, \lambda_M\}}{\operatorname{argmin}} \sum_{k=1}^{5} |I_k|^{-1} \sum_{t \in I_k} \left(y_t - \sum_{j \in \widehat{A}_k(\lambda)} x_{tj} \hat{\beta}_{k,j}^{OLS}(\lambda) \right)^2 ,$$

   where and $\hat{\beta}_{k,j}^{OLS}(\lambda)$ is a jth element of the OLS estimator $\widehat{\boldsymbol{\beta}}_k^{OLS}(\lambda)$.

4. Estimate model (4) by the Lasso with λ^* in the whole sample $\cup_{k=1}^5 I_k = \{1, 2, \ldots, T\}$ and obtain the active set $\widehat{A}(\lambda^*) = \{j : \hat{\beta}_j(\lambda^*) \neq 0\}$, where $\hat{\beta}_j(\lambda^*)$ denotes jth element of the Lasso estimate $\widehat{\boldsymbol{\beta}}$ with λ^* for $j = 1, 2, \ldots, N(\ell + 1) + 1$.

5. Run the OLS only with active covariates $\{\boldsymbol{x}_j\}$ such that $j \in \widehat{A}(\lambda^*)$, and we obtain the OLS post-Lasso estimator.

It should be stressed that determining λ^* in the 3rd step, the CV procedure is different from that in the (conventional) Lasso because we need to select λ^* so as to minimize the sum of squared residuals (SSR) that are obtained by the OLS with screened covariates in each k subsample. In this example we set $M = 100$ and potential regularization parameters

$\lambda_1, \dots, \lambda_M$ are obtained by the same rule as `ncvreg`. Here we can also define the OLS post-MCP and OLS post-SCAD in the same manner. We employ the OLS post-MCP and the OLS post-SCAD as well as the OLS post-Lasso and hereafter we call them as the *OLS post-selection estimators.*

4.1.5 Forecast performance

Tables 1–3 report the mean squared forecast errors (MSFE) of the penalized MIDAS regression with the SCAD, MCP, Lasso, OLS post-selection estimators, factor MIDAS, and naive AR(4) in the overall sample, 1st subsample, and 2nd subsample, respectively. The median forecast errors are also shown in parentheses as a robust measure for contamination by outliers. In the tables, the lower table "with y_{t-1}" provides the results when the lagged-dependent variable y_{t-1} is included as a regressor in (4) in addition to $\boldsymbol{x}_{t-h}$, whereas the upper table "without y_{t-1}" gives the results without the lagged-dependent variable. All the values are standardized by the naive AR(4) forecast. First we see the results on the overall sample. In the nowcasting ($0 \leq h < 1$) cases, Table 1 shows that every method is much better than the naive AR(4) forecast (i.e., the values are less than one) with a few exceptions, but that the MCP, SCAD, and Lasso outperform the factor MIDAS and OLS post-selection estimators in the overall sample as a whole, in terms of both the mean and median squared forecast errors. Inclusion of the lagged dependent variable does not essentially affect the forecasting performance of the MCP, SCAD, Lasso, and OLS post-selection estimators, but improves the MSFEs of the factor MIDAS. Here we should mention that the MSFE of the factor MIDAS when $h = 1/3$ is much worse than the other methods. It is mainly due to outliers of forecast values around the subprime mortgage crisis. The factor MIDAS is not necessarily much worse than the other estimators for the other h's although the forecasts of the Factor MIDAS tend to be volatile. We find the post-selection estimators work well in terms of MSFE in some cases ($h = 2/3$ and $h = 1$), however, they do not necessarily work better than the MCP, SCAD, and Lasso both in mean and median measures. Let us see the forecast performance when $h \geq 1$. The table shows that all the forecasts have similar forecast performances; they perform well when $h = 1$, however, when $h > 1$, they are all beaten by the AR(4) forecast. The results are not surprising because previous studies, such as Clements and Galvao (2008) and Marcellino and Schumacher (2010), reported the same tendency. Here we also find the OLS post-SCAD behaves like the OLS post-Lasso. It is because the OLS post-SCAD has almost the same active set $\widehat{A}(\lambda^*)$ in the screening step. Next, we see the results for the subsamples. Tables 2 and 3 show the forecasting performances for the first and second subsamples, respectively. As a whole, both of the results are the same as the whole sample case; the forecasts with the MCP, SCAD, and Lasso are superior to those with the Factor MIDAS, OLS-post-selection estimators, and naive AR in the nowcasting cases, although they are not necessarily reliable when we consider the forecast with a long horizon.

In consequence, our analysis shows that the forecasts with mixed frequency by the MCP, SCAD and Lasso have good forecast performances in nowcasting though it does not seem to be a primary tool for forecasting with relatively long horizons. At the same time, we also find that the forecast performances of the post-inference estimators are not so convincing contrary to our expectation. The reason may come from strong time dependences in covariates. Section E in Appendix examine performances of the MCP, SCAD, Lasso, and OLS post-Lasso with the simulated data, where the covariates are moderately cross-sectionally dependent but time-independent. We find from the simulation that the OLS post-Lasso performs better than the other penalization methods, which supports our conjecture; see Section E for details. Furthermore, we have seen that the penalized regression estimators perform well in nowcasting with a complete data, so far. However, when we actually conduct real-time forecasting of quarterly GDP with monthly data, a complete dataset may not be available because of possible publication lags. Thus, it happens to face an incomplete, so-called *jagged (ragged)-edge* dataset that contains missing values in some latest months. Section F in Appendix investigates the forecast performance with such jagged-edge data and finds favorable results comparing with the state-space maximum-likelihood method; see Section F for details.

4.2 Screening Effective Portfolio from a Large Number of Potential Securities

Recently, several studies on portfolio selection have focused on the penalized regression as it can select stocks in a portfolio among a large number of potential stocks. Brodie, Daubechies, Giannone, De Mol and Loris (2009) found out that the penalized regression is useful in selecting optimal portfolio in terms of the out-of-sample performance measured by the Sharpe ratio; Fan, Zhang, and Yu (2012) introduced gross-exposure constraints to admit short sales in the estimation of an optimal portfolio; Carrasco and Noumon (2012) focused on estimating a precision matrix of returns. They noticed that the penalized regression was quite useful to stabilize the estimation of the covariance matrix and provided better finite sample performances than traditional methods.

To the best of our knowledge, the existing literature focused mainly on yieldability. However, it seems interesting to examine the consistent estimation of hidden weights of the portfolio; that is, we are curious in screening how fund managers construct their portfolio from a large number of potential securities. Unlike the other high-dimensional estimation methods, such as the factor and the Ridge regression, the SCAD-type penalized regression may enable us to reveal their portfolio from a large dataset of stock prices. It is because the SCAD-type penalized regression is expected to have a chance to select variables consistency as stated in Property 1. In this section, we examine how well the penalized regression usefully works in this direction using a large NYSE stock price dataset.

4.2.1 Portfolio construction

Suppose a fund manager faces p potential stocks, where x_{it} is the rate of return of the ith $(i = 1, 2, \ldots, p)$ stock at time t. Let $\boldsymbol{x}_t = [x_{1t}, x_{2t}, \ldots, x_{pt}]^\top$ be the p-dimensional rates of the return vector at t and $\boldsymbol{\omega}_0$ be the p-dimensional weight vector of the portfolio that satisfies $\|\boldsymbol{\omega}_0\|_0 = s \ (\ll p)$, $\boldsymbol{\iota}'\boldsymbol{\omega}_0 = 1$ and $\|\boldsymbol{\omega}_0\|_1 = \zeta_w$, where $\zeta_w \in [1, \infty)$ and $\boldsymbol{\iota}$ is a p-dimensional vector with all the elements being one. That is, the portfolio is constructed by s stocks from p potential stocks. We assume the fund manager constructs her portfolio as

$$y_t = \boldsymbol{x}_t^\top \boldsymbol{\omega}_0 + u_t, \qquad t = 1, \ldots, T, \tag{5}$$

where u_t is a miscellaneous component that includes all assets in the portfolio other than stocks, such as T-bills and corporate bonds. Further we assume that $\boldsymbol{x}_t$ and u_t are independent of each other and $u_t \sim i.i.d.N(0, \sigma_u^2)$, where $\sigma_u^2 = T^{-1}\boldsymbol{\omega}_{0A}^\top \boldsymbol{X}_A^\top \boldsymbol{X}_A \boldsymbol{\omega}_{0A}/\mathrm{SNR}$, $\boldsymbol{\omega}_{0A}$ is a nonzero s-dimensional subvector of $\boldsymbol{\omega}_0$, $\boldsymbol{X}_A$ is $T \times s$ submatrix of $\boldsymbol{X}$ that corresponds to $\boldsymbol{\omega}_{0A}$, and $\mathrm{SNR} = V(\boldsymbol{x}_t^\top \boldsymbol{\omega}_0)/V(u_t)$. Although we might consider the case in which $\boldsymbol{x}_t$ and u_t are dependent by extending the results of Fan and Liao (2014), this is beyond the scope of our research, and we regard $\boldsymbol{x}_t$ and u_t as independent here.

The portfolio allows short sales if $\zeta_w > 1$ with ζ_w determining a constraint on the short sales as shown in Fan, Zhang, and Yu (2012): Let $w_0^+ = (\zeta_w + 1)/2$ and $w_0^- = (\zeta_w - 1)/2$. Then w_0^+ and w_0^- correspond to the total proportions of long and short sales, respectively, since $w_0^+ + w_0^- = \zeta_w = \|\boldsymbol{\omega}_0\|_1$ and $w_0^+ - w_0^- = 1$, and w_0^- becomes larger as ζ_w grows while short sales are not allowed if $\zeta_w = 1$ ($w_0^- = 0$). We assume the fund manager holds $s/2$ stocks for long and $s/2$ stocks for short sales respectively, and she employs equal weights among long and short sales; that is, we assume $\omega_{0i} = w_0^+/(s/2)$ for $i \in \boldsymbol{\omega}_{0A+}$, $-w_0^-/(s/2)$ for $i \in \boldsymbol{\omega}_{0A-}$, and 0 for $i \in \boldsymbol{\omega}_{0B}$, where ω_{0i} is ith element of $\boldsymbol{\omega}_0$, and $\boldsymbol{\omega}_{0A+}$, $\boldsymbol{\omega}_{0A-}$, and $\boldsymbol{\omega}_{0B}$ are sets of stocks of long, short, and no sales, respectively.

4.2.2 Data and evaluation strategy

We retrieve weekly stock price data of the NYSE from *Yahoo! Finance*. Our dataset contains 1853 adjusted stock prices ($p = 1853$) with starting from the 1st week of January in 2009 to the 4th week of April in 2016. We apply the log-difference and standardize them so that the data are converted to rates of returns with zero means and unit variances. We investigate the cases of $s = 34$ and 40 with $a = 14$, SNR $= 10$, and $\zeta_w = 10$. Non-zero s stocks are drawn randomly from p candidates with equal probabilities. Furthermore, we assume the fund manager does not rebalance the portfolio. Hence it remain unchanged in all sample period. Brodie et al. (2009) argued a possibility of estimating a weight vector for a portfolio in the presence of rebalancing, but for simplicity we do not consider the case here.

The purpose of this application is to unseal the hidden stocks in which the fund manager invested from a large number of potential stocks. We examine how well the penalized estimator $\hat{\boldsymbol{\omega}}$ can distinguish the nonzeros from zero elements of $\boldsymbol{\omega}_0$ in finite samples. Then we evaluate the finite sample performances of $\hat{\boldsymbol{\omega}}$ to focus on SC-$A = P\left(\text{sgn}(\hat{\boldsymbol{\omega}}_A) = \text{sgn}(\boldsymbol{\omega}_{0A})\right)$ and SC-$B = P\left(\text{sgn}(\hat{\boldsymbol{\omega}}_B) = \text{sgn}(\boldsymbol{\omega}_{0B})\right)$; the SC-$A$ refers to the success rate of detecting nonzero elements of $\boldsymbol{\omega}_0$ with the correct sign while the SC-B that of detecting zero elements. We anticipate that the SCAD-type penalized regression estimator can have high SC-A and SC-B values asymptotically thanks to the oracle property. The SC-A and SC-B are sequentially computed for 172 evaluation periods, where the endpoint gradually grows by one with the start point fixed; the initial evaluation period starts from the 2nd week of January 2009 and ends in 1st week of December 2010 ($T = 209$). The 2nd evaluation period runs from the 2nd week of January 2009 to the 2nd week of December 2010 ($T = 210$), and so on. The terminal evaluation period is from the 2nd week of January 2009 to the 4th week of April 2016 ($T = 381$).

4.2.3 Empirical results

Figures 1–2 and 3–4 show the SC-A and SC-B of the MCP, SCAD, and Lasso for 172 evaluation periods with $s = 34$ and 40, respectively. To begin with, we consider the SC-A. At a glance, both Figures 1 and 2 reveal two characteristics of $\hat{\boldsymbol{\omega}}$. First, the SC-$A$ increases toward one as T grows, irrespective of the penalties. Although the SC-A of $s = 40$ seems uniformly lower than that of $s = 34$ for all T, this is due to the fact that many nonzero elements require a greater search cost. Second, the SC-A of the Lasso tends to be higher than that of the MCP and SCAD when T is relatively small, while it seems reversed when T is large. This is consistent with the theory because the Lasso tends to have many false positive estimates. That is, it overestimates the total number of nonzero elements since it rarely satisfies the assumptions for model selection consistency, while the MCP and SCAD may satisfy these assumptions. The SC-A of the Lasso is not expected to be higher than that of the MCP and SCAD when T is large.

Next, we focus on the SC-B. Figures 3 and 4 show that SC-B of the MCP and SCAD are successfully nearly equal to one with dominating that of the Lasso for all T. The results are again consistent with the theory because the MCP and SCAD have the oracle property, which means that they can generally detect true zero parameters more precisely than the Lasso can.

In summary, our empirical results reveal that the model selection consistency of the SCAD-type penalty works well in a large stock price dataset. This implies that the penalized regression is effective when we want to detect the composition of fund manager's portfolio from large financial datasets. However, we should keep in mind that conditions for the variable selection consistency do not necessarily hold in macroeconomic data as discussed in Section 3.2.

5 Conclusion

We have studied macroeconomic forecasting and variable selection using a folded-concave penalized regression with a very large number of predictors. The contributions include both theoretical and empirical results. The first half of the paper developed the theory for a folded-concave penalized regression in ultrahigh dimensions when the model exhibits time series dependences. Specifically, we have proved the oracle inequality under appropriate con-

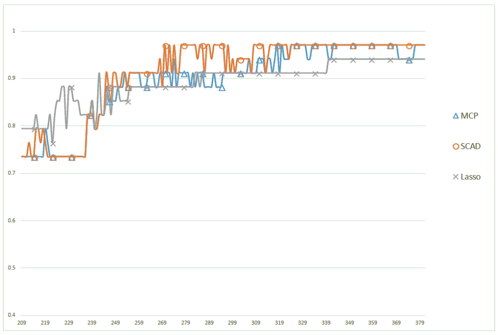

Figure 1: SC-A when $s = 34$ (from $T = 209$ to $T = 381$)

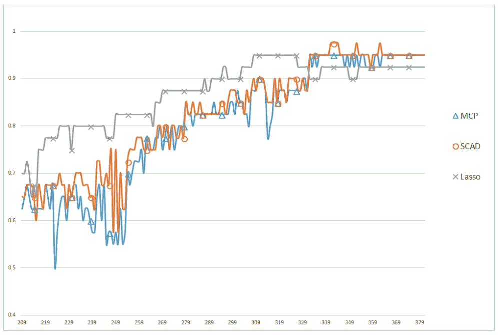

Figure 2: SC-A when $s = 40$ (from $T = 209$ to $T = 381$)

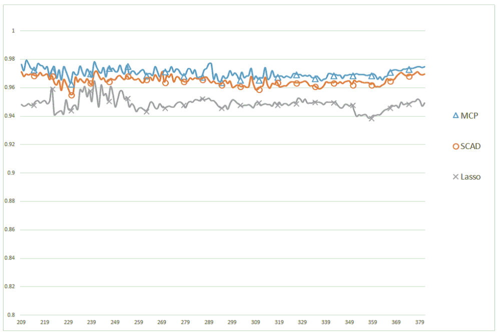

Figure 3: SC-B when $s = 34$ (from $T = 209$ to $T = 381$)

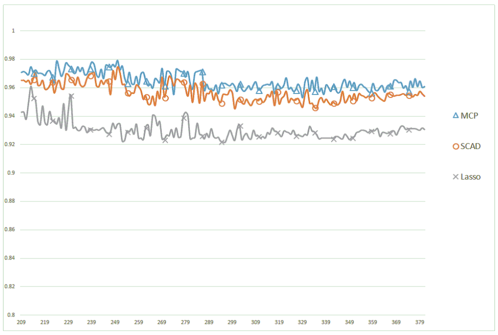

Figure 4: SC-B when $s = 40$ (from $T = 209$ to $T = 381$)

ditions for macroeconomic time series. A limitation of the oracle property was discussed as well. The latter half of the paper provided two empirical applications that motivated us to use the penalized regression for a large macroeconomic dataset. The first was the forecasting of quarterly U.S. real GDP with a large amount of monthly macroeconomic data taken from the FRED-MD through the MIDAS regression framework; the forecasting model consisted of more than 1000 monthly predictors including lags while the sample size was much smaller than the total number of predictors. The forecasting performance of the penalized regression is promising one compared to that of the factor MIDAS proposed by Marcellino and Schumacher (2010), the OLS post-Lasso proposed by Belloni and Chernozhukov (2013) and the state-space (nowcasting) model of Banbura and Modugno (2014). The second application screened a portfolio that contained about 40 stocks from more than 1800 stocks using NYSE stock price data. The oracle property ensured the variable selection consistency, that is, the penalized regression with the SCAD-type penalty could detect the portfolio from the data theoretically. In fact, we observed that the variable selection consistency worked properly when screening the portfolio.

Acknowledgements

The authors are grateful for the invaluable comments of the editor Victor Chernozhukov and two anonymous referees, and for the important suggestion for the earlier version by Jinchi Lv, Ryo Okui, Yohei Yamamoto, and the participants of the econometric workshop in Kyoto University. All the remaining errors are ours. The authors also thank Michelle Modugno for providing MATLAB codes used in Banbura and Modugno (2014), which were quite helpful when coding our R codes. Uematsu acknowledges the financial support from a Grant-in-Aid for JSPS Fellows No.26-1905 and No.29-60. Tanaka acknowledges the financial supports from a JSPS Grant-in-Aid for Young Scientists (B) No.16K17100 and the Joint Usage and Research Center, Institute of Economic Research, Hitotsubashi University.

References

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

Andreou, E., E. Ghysels and A. Kourtellos (2010). Regression models with mixed sampling frequencies. *Journal of Econometrics*, *158*, 246–261.

Bai, J. and Y. Liao (2016). Efficient estimation of approximate factor models via penalized maximum likelihood. *Journal of Econometrics*, *191*, 1–18.

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

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

Bańbura, M., D. Giannone, and L. Reichlin (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, *25*, 71–92.

Bańbura, M. and M. Modugno (2014). Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *Journal of Applied Econometrics*, *29*, 133–160.

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

Belloni, A. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, *19*, 521–547.

Belloni, A., V. Chernozhukov, and C. Hansen (2014a). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, *81*, 608–650.

Belloni, A., V. Chernozhukov, and L. Wang (2014b). Pivotal estimation via square-root lasso in nonparametric regression. *Annals of Statistics*, *42*, 757–788.

Breheny, P. and J. Huang (2011). Coordinate descent algorithm for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, *5*, 232–253.

Brodie, J., I. Daubechies, C. De Mol, D. Giannone and I. Loris (2009). Sparse and stable markowitz portfolios. *Proceedings of the National Academy of Sciences*, *106*, 12267–12272.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Callot, L. and A. B. Kock (2014). *On the oracle property of the grouped adaptive lasso for vector autoregressions*. Nonlinear Time Series Econometrics, Festschrift in Honor of Timo Teräsvirta. Oxford University Press.

Carrasco, M. and N. Noumon (2012). Optimal portfolio detection using regularization. mimeo.

Chernozhukov, V., C. Hansen, and M. Spindler (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, *7*, 649–688.

Chetverikov, D., Z. Liao, and V. Chernozhukov (2017). On cross-validated Lasso arXiv:1605.02214.

Clements, M. and A. B. Galvão (2008). Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the United States. *Journal of Business and Economic Statistics*, *26*, 546–554.

Davidson, J. (1994). *Stochastic Limit Theory.* Oxford University Press.

Davis, R. A., P. Zang and T. Zheng (2016). Sparse Vector Autoregressive Modeling. *Journal of Computational and Graphical Statistics*, *25*, 1077–1096.

De Mol, C., D. Giannone, and L. Reichlin (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, *146*, 318–328.

Fan, J., Y. Ke, and Y. Liao (2016). Robust factor models with explanatory proxies. *mimeo.*

Fan, J., Y. Liao, and W. Wang (2016). Projected principal component analysis in factor models. *Annals of Statistics*, *44*, 219–254.

Fan, J., L. Xue, and J. Yao (2017). Sufficient forecasting using factor models. *Journal of Econometrics*, *201*, 292–306.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Fan, J. and J. Lv (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, *57*, 5467–5484.

Fan, J. and Y. Liao (2014). Endogeneity in high dimensions. *Annals of Statistics*, *42*, 872–917.

Fan, J. and J. Lv (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, *57*, 5467–5484.

Fan, J., J. Zhang and K. Yu (2012). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, *107*, 592–606.

Fan, Y. and J. Lv (2013). Asymptotic equivalence of regularization methods in thresholded parameter space. *Journal of the American Statistical Association*, *108*, 1044–1061.

Frank, Ildiko E., and Jerome H. Friedman (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, *35*, 109–135.

Gefang, D. (2014). Bayesian doubly adaptive elastic-net Lasso for VAR shrinkage. *International Journal of Forecasting*, *30*, 1–11.

Giannone, D., L. Reichlin and D. Small (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, *55*, 665–676.

Ghysels, E., A. Sinko and R. Valkanov (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews*, *26*, 53–90.

Hansen, C. and Y. Liao (2016). The factor-lasso and k-step bootstrap approach for inference in high-dimensional economic applications. *mimeo.*

Javanmard, A. and A. Montanari (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, *15*, 2869–2909.

Kallestrup-Lamb, M., A. B. Kock, J. T. Kristensen (2016). Lassoing the determinants of retirement. *Econometric Reviews*, *35*, 1522–1561.

Kim H. H. and N. R. Swanson (2018). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting*, *34*, 339–354.

Kock, A. B. (2013). Oracle Efficient Variable Selection in Random and Fixed Effects Panel Data Models. *Econometric Theory*, *29*, 115–152.

Kock, A. B. (2016). Consistent and conservative model selection with the adaptive lasso in stationary and nonstationary autoregressions. *Econometric Theory*, *32*, 243–259.

Kock, A. B. and L. Callot (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, *186*, 325–344.

Kock, A. B. and H. Tang (2018). Inference in high-dimensional dynamic panel data models. *Econometric Theory, forthcoming.*

Kock, A. B. and T. Teräsvirta (2014). Forecasting performances of three automated modelling techniques during the economic crisis 2007 寠 · 009. *International Journal of Forecasting*, *30*, 616–631.

Kock, A. B. and T. Teräsvirta (2016). Forecasting macroeconomic variables using neural network models and three automated model selection techniques. *Econometric Reviews*, *35*, 1753–1779.

Konzen, E. and F. A. Ziegelmann (2016). LASSO-type penalties for covariate selection and forecasting in time series. *Journal of Forecasting*, *35*, 592–612.

Koop, G. M. (2013). Forecasting with medium and large Bayesian VARs. *Journal of Applied Econometrics*, *28*, 177–203.

Koop, G and D. Korobilis (2018). Forecasting with High-Dimensional Panel VARs. *Essex Finance Centre Working Papers 21329*, University of Essex, Essex Business School.

Li, J. and W. Chen (2014). Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models. *International Journal of Forecasting*, *30*, 996–1015.

Li, X., L. Zbonakova, and W. K. Härdle (2017). Penalized adaptive method in forecasting with large information set and structure change. *SFB 649 Discussion Papers SFB649DP2017-023, Sonderforschungsbereich 649*, Humboldt University, Berlin, Germany.

Marcellino, M. H. and C. Schumacher (2010). Factor MIDAS for nowcasting and forecasting with ragged-edge data: A model comparison for German GDP. *Oxford Bulletin of Economics and Statistics*, *72*, 518–550.

Marsilli, C. (2014). Variable selection in predictive MIDAS models. Banque de France Working Paper 520.

Medeiros, M. C. and E. F. Mendes (2016). ℓ_1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors. *Journal of Econometrics*, *191*, 255–271.

McCracken, M. W. and S. Ng (2015). FRED-MD: A monthly database for macroeconomic research. Federal Reserve Bank of ST. Louis Working Paper Series 2015-012A.

Negahban, S. N., P. Ravikumar, M. J. Wainwright and B. Yu (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, *27*, 538–557.

Nicholson, W. B., D. S. Matteson, and J. Bien (2015). VARX-L: Structured regularization for large vector autoregressions with exogenous variables. arXiv:1508.07497.

Ng, S. (2013). Variable Selection in Predictive Regressions. *Chapter Chapter 14 in Handbook of Economic Forecasting*, Elsevier, 752–789.

Leeb, H. and B. M. Pötscher (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics*, *142*, 201–211.

Pötscher, B. M. and H. Leeb (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis*, *100*, 2065–2082.

Schnücker, A. (2017). Penalized estimation of panel vector autoregressive models: a lasso approach. mimeo.

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

Song, S. and P. J. Bickel (2011). Large vector auto regressions. arXiv:1106.3915v1.

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

Stock, J. H. and M. W. Watson (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, *30*, 481–493.

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

Uematsu, Y. and S. Tanaka (2015). Regularization parameter selection via cross-validation in the presence of dependent regressors: A simulation study. *Economics Bulletin*, *36*, 313–319.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, *38*, 894–9421.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of Royal Statistical Society Series B*, *76*, 217–242.

Zhao, P. and B. Yu (2006). On model selection consistency of LASSO. *Journal of Machine Learning Research*, *7*, 2541–2563.

Appendix

A Lemmas for the Main Proofs

Lemma 1. *Assume $X_i \sim$ ind. subG(α_i^2) and $Y_i \sim$ ind. subE(γ_i). Then, for any deterministic sequences (ϕ_i) and (ψ_i), the following statements are true:*

(a) $X_iX_j \sim$ subE$(4e\alpha_i\alpha_j)$ for $i \neq j$.

(b) $\sum_{i=1}^n \phi_i X_i \sim$ subG$(\sum_{i=1}^n \phi_i^2\alpha_i^2)$.

(c) $\sum_{i=1}^n \psi_i Y_i \sim$ subE$((\sum_{i=1}^n \psi_i^2\gamma_i^2)^{1/2}, \max_i |\psi_i|\gamma_i)$.

Proof. (a) Since X_i is subG(α_i^2), we obtain $\mathrm{E}|X|^k \leq (2\alpha_i^2)^{k/2}k\Gamma(k/2)$; see Rigollet and Hütter (2017), for instance. Then we see from the dominated convergence theorem and independence that

$$\begin{aligned}
\mathrm{E}\exp(sX_iX_j) &= 1 + \sum_{k=2}^{\infty}\frac{s^k\mathrm{E}(X_iX_j)^k}{k!} \leq 1 + \sum_{k=2}^{\infty}\frac{s^k\mathrm{E}|X_i|^k\mathrm{E}|X_j|^k}{k!} \\
&\leq 1 + \sum_{k=2}^{\infty}\frac{s^k(2\alpha_i\alpha_j)^k k^2\Gamma(k/2)^2}{k!} \leq 1 + \sum_{k=2}^{\infty}\frac{s^k(2\alpha_i\alpha_j)^k k^2(k/2)^k}{k!} \\
&= 1 + \sum_{k=2}^{\infty}\frac{s^k(\alpha_i\alpha_j)^k k^{k+2}}{k!} \leq 1 + \sum_{k=2}^{\infty}(2e\alpha_i\alpha_j s)^k \\
&= 1 + (2e\alpha_i\alpha_j s)^2\sum_{k=0}^{\infty}(2e\alpha_i\alpha_j s)^k,
\end{aligned}$$

where we have used $\Gamma(k/2) \leq (k/2)^{k/2}$ and $k^{k+2} \leq (2\pi)^{-1/2}k!e^k k^{3/2} \leq k!(2e)^k$. Therefore, for any $|s| \leq (4e\alpha_i\alpha_j)^{-1}$, it holds that

$$\mathrm{E}\exp(sX_iX_j) \leq 1 + 8(e\alpha_i\alpha_j)^2 s^2 \leq \exp((4e\alpha_i\alpha_j)^2 s^2/2).$$

This means that the product X_iX_j is subE$(4e\alpha_i\alpha_j)$.

(b) By the definition of subG, we have

$$\begin{aligned}
\mathrm{E}\exp\left(s\sum_{i=1}^n \phi_i X_i\right) &= \prod_{i=1}^n \mathrm{E}\exp\left(s\phi_i X_i\right) \\
&\leq \prod_{i=1}^n \exp\left(s^2\phi_i^2\alpha_i^2/2\right) = \exp\left(s^2\sum_{i=1}^n \phi_i^2\alpha_i^2/2\right),
\end{aligned}$$

which yields the result.

(c) First note that $\psi_i Y_i \sim \text{subE}(\psi_i\gamma_i, |\psi_i|\gamma_i)$ because $\mathrm{E}\exp(s\psi_i Y_i) \leq \exp(s^2\psi_i^2\gamma_i^2/2)$ holds for all $|s| \leq (|\psi_i|\gamma_i)^{-1}$. Thus, we can see that

$$
\begin{aligned}
\mathrm{E}\exp\left(s\sum_{i=1}^{n}\psi_i Y_i\right) &= \prod_{i=1}^{n}\mathrm{E}\exp\left(s\psi_i Y_i\right) \\
&\leq \prod_{i=1}^{n}\exp\left(s^2\psi_i^2\gamma_i^2/2\right) = \exp\left(s^2\sum_{i=1}^{n}\psi_i^2\gamma_i^2/2\right),
\end{aligned}
$$

where the inequality holds for all $|s| \leq (\max_i|\psi_i|\gamma_i)^{-1}$. This gives the result by the definition of subE, and completes all the proofs. □

□

Lemma 2. *Under Assumption 4, we have*

$$
T^{-1}\sum_{u,v=1}^{T}(\sum_{t=1}^{T}r_{tu}r_{tv})^2 + T^{-1}\sum_{u,v=1}^{T}(\sum_{t=1}^{T}r_{tu}s_{tv})^2 = O(1).
$$

Proof. We prove the boundedness of the first term. We have

$$
\begin{aligned}
T^{-1}\sum_{u,v=1}^{T}(\sum_{t=1}^{T}r_{tu}r_{tv})^2 &\leq T^{-1}\sum_{u,v=1}^{T}\max_t|r_{tu}|^2(\sum_{t=1}^{T}|r_{tv}|)^2 \\
&\leq \sum_{u=1}^{T}\max_t|r_{tu}|^2\max_v(\sum_{t=1}^{T}|r_{tv}|)^2 = \max_t\|\boldsymbol{r}_{t\cdot}\|_2^2\max_v\|\boldsymbol{r}_{\cdot v}\|_1^2,
\end{aligned}
$$

which is bounded for all (large) T by Assumption 4. The same result holds for the second term under Assumption 4 as well. □

□

B Proofs of the Main Results

B.1 Proof of Theorem 1

Proof. For any $\hat{\boldsymbol{\beta}}$ that minimizes $Q_T(\boldsymbol{\beta})$, we have

$$
(2T)^{-1}\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2 + \|p_\lambda(\hat{\boldsymbol{\beta}})\|_1 \leq (2T)^{-1}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_0\|_2^2 + \|p_\lambda(\boldsymbol{\beta}_0)\|_1.
$$

By model (1) and Hölder's inequality, this can be rewritten and bounded as

$$
\begin{aligned}
(2T)^{-1}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|_2^2 &\leq T^{-1}\boldsymbol{u}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) + \|p_\lambda(\boldsymbol{\beta}_0)\|_1 - \|p_\lambda(\hat{\boldsymbol{\beta}})\|_1 \\
&\leq \|T^{-1}\boldsymbol{X}^\top\boldsymbol{u}\|_\infty\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1 + \|p_\lambda(\boldsymbol{\beta}_0)\|_1 - \|p_\lambda(\hat{\boldsymbol{\beta}})\|_1. \qquad (1)
\end{aligned}
$$

The Mean value theorem, Assumption 1, and the triangle inequality give

$$\begin{aligned} \|p_\lambda(\boldsymbol{\beta}_0)\|_1 - \|p_\lambda(\hat{\boldsymbol{\beta}})\|_1 &= \sum_{j=1}^{p} \Big(|p_\lambda(\beta_{0j})| - |p_\lambda(\hat{\beta}_j)| \Big) = \sum_{j=1}^{p} p'_\lambda(b_j) \Big(|\beta_{0j}| - |\hat{\beta}_j| \Big) \\ &\le p'_\lambda(0+) \sum_{j=1}^{p} \Big| |\beta_{0j}| - |\hat{\beta}_j| \Big| \le \lambda \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1, \end{aligned} \tag{2}$$

where b_j is a point between $|\beta_{0j}|$ and $|\hat{\beta}_j|$. In what follows, we have only to work on event $\mathcal{E}_1$ defined in Assumption 2. On the event, we have $\|T^{-1}\boldsymbol{X}^\top \boldsymbol{u}\|_\infty \le \lambda/2$, so that (1) and (2) entail

$$\begin{aligned} (2T)^{-1}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 &\le 2^{-1}\lambda \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 + \|p_\lambda(\boldsymbol{\beta}_0)\|_1 - \|p_\lambda(\hat{\boldsymbol{\beta}})\|_1 \\ &\le 2^{-1}\lambda \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 + \lambda \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \\ &= (3/2)\lambda \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1. \end{aligned} \tag{3}$$

Lemma 1 of Negahban, Ravikumar, Wainwright, and Yu (2012) yields

$$\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_{0B}\|_1 \le 3 \|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\|_1. \tag{4}$$

Thus, Assumption 3 gives

$$T^{-1}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 \ge \gamma \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2. \tag{5}$$

By (3)–(5) and the Cauchy–Schwarz inequality, we have

$$\begin{aligned} \gamma \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 &\le 3\lambda \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \\ &= 3\lambda \|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\|_1 + 3\lambda \|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_{0B}\|_1 \\ &\le 12\lambda \|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\|_1 \\ &\le 12 s^{1/2}\lambda \|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\|_2 \\ &\le 12 s^{1/2}\lambda \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2. \end{aligned}$$

This concludes the error bound in the ℓ_2-norm

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \le 12 s^{1/2}\lambda/\gamma. \tag{6}$$

Using (6), we can obtain the error bound in the ℓ_1-norm as well. We have

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 = \|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\|_1 + \|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_{0B}\|_1$$

$$\leq 4\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\|_1 \leq 4s^{1/2}\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\|_2 \leq 4s^{1/2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \leq 48s\lambda/\gamma. \tag{7}$$

Finally, we derive the prediction error bound from (7)and (3). We obtain

$$T^{-1}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 \leq 3\lambda\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq 144s\lambda^2/\gamma. \tag{8}$$

Results (6)–(8) hold with probability at least $1 - O(p^{-c_1}) - O(p^{-c_2})$ by Assumptions 2 and 3. □ □

B.2 Proof of Proposition 1

Proof. We are interested in the probabilistic behavior of $\|\boldsymbol{G}_{0T}\|_\infty = \|T^{-1}\boldsymbol{X}^\top\boldsymbol{u}\|_\infty$; specifically we want to find a positive sequence $\lambda = \lambda_{pT}$ and some positive constant c_1 such that

$$P(\|\boldsymbol{G}_{0T}\|_\infty > \lambda/2) = O(p^{-c_1}). \tag{9}$$

We observe from the construction that

$$\begin{aligned}\left\|T^{-1}\boldsymbol{X}^\top\boldsymbol{u}\right\|_\infty &= \left\|\tau T^{-1}\boldsymbol{\Sigma}^{1/2\top}\boldsymbol{Z}^\top\boldsymbol{R}^{1/2\top}\boldsymbol{S}^{1/2}\boldsymbol{e}\right\|_\infty \\ &= \tau\max_j\left|T^{-1}\sum_{t,u=1}^T(\sum_{s=1}^T r_{st}s_{su})e_u\sum_{i=1}^p z_{ti}\sigma_{ij}\right|.\end{aligned}$$

Lemma 1 entails

$$\tilde{z}_{tj} := \sum_{i=1}^p z_{ti}\sigma_{ij} \sim \text{subG}\left(\alpha^2\sum_{i=1}^p\sigma_{ij}^2\right) = \text{subG}\left(\alpha^2\|\boldsymbol{\sigma}_{\cdot j}\|_2^2\right)$$

and

$$\tilde{e}_u := (\sum_{s=1}^T r_{st}s_{su})e_u \sim \text{subG}\left(\alpha^2(\sum_{s=1}^T r_{st}s_{su})^2\right),$$

which furthermore imply $\tilde{z}_{tj}\tilde{e}_u \sim$ ind. $\text{subE}(4e\alpha^2|\sum_{s=1}^T r_{st}s_{su}|\|\boldsymbol{\sigma}_{\cdot j}\|_2)$ for each j. Therefore, we can find by Lemma 1(c) that

$$\begin{aligned}&T^{-1}\sum_{t,u=1}^T\tilde{z}_{tj}\tilde{e}_u \\ &\sim \text{subE}\left(4e\alpha^2\|\boldsymbol{\sigma}_{\cdot j}\|_2T^{-1}\left(\sum_{t,u}\left(\sum_{s=1}^T r_{st}s_{su}\right)^2\right)^{1/2}, 4e\alpha^2\|\boldsymbol{\sigma}_{\cdot j}\|_2T^{-1}\max_{t,u}\left|\sum_{s=1}^T r_{st}s_{su}\right|\right)\end{aligned}$$

for each j. Thus, Bernstein's inequality of a sub-exponential random variable together with the union bound yields

$$P\left(\max_j \left|T^{-1}\sum_{t,u=1}^{T} \tilde{z}_{tj}\tilde{e}_u\right| > x\right) \leq p \max_j P\left(\left|T^{-1}\sum_{t,u=1}^{T} \tilde{z}_{tj}\tilde{e}_u\right| > x\right)$$

$$\leq 2p \max_j \exp\left(-\frac{x^2}{32e^2\alpha^4\|\boldsymbol{\sigma}_{\cdot j}\|_2^2 T^{-2}\sum_{t,u=1}^{T}\left(\sum_{s=1}^{T} r_{st}s_{su}\right)^2}\right)$$

$$\leq 2p \exp\left(-\frac{x^2}{32e^2\alpha^4 T^{-1}\max_j\|\boldsymbol{\sigma}_{\cdot j}\|_2^2 \max_t\|\boldsymbol{r}_{\cdot t}\|_1^2 \max_s\|\boldsymbol{s}_{s\cdot}\|_2^2}\right)$$

for all $x \in [0, I_1]$, where

$$I_1 = 4e\alpha^2 \min_j \|\boldsymbol{\sigma}_{\cdot j}\|_2 T^{-1}\sum_{t,u=1}^{T}\left(\sum_{s=1}^{T} r_{st}s_{su}\right)^2 \Big/ \left(\max_{t,u}|\sum_{s=1}^{T} r_{st}s_{su}|\right).$$

Note that $I_1 = O(1)$ and $\liminf_T I_1 > 0$ due to Assumption 4 and Lemma 2. Thus, if we can set $x = (\tilde{c}_1(1+\nu)T^{-1}\log p)^{1/2}$ with

$$\tilde{c}_1 = 32e^2\alpha^4 \max_j \|\boldsymbol{\sigma}_{\cdot j}\|_2^2 \max_t \|\boldsymbol{r}_{\cdot t}\|_1^2 \max_s \|\boldsymbol{s}_{s\cdot}\|_2^2 = O(1)$$

and an arbitrary fixed $\nu > 0$, this falls into $[0, I_1]$ eventually. Then, the upper bound of the probability reduces to

$$2p \exp\left(-\frac{\tilde{c}_1(1+\nu)\log p}{32e^2\alpha^4 \max_j\|\boldsymbol{\sigma}_{\cdot j}\|_2^2 \max_t\|\boldsymbol{r}_{\cdot t}\|_1^2 \max_s\|\boldsymbol{s}_{s\cdot}\|_2^2}\right)$$

$$= 2p\exp(-(1+\nu)\log p) = 2p^{-\nu}.$$

This means that $\|T^{-1}\boldsymbol{X}^\top\boldsymbol{u}\|_\infty \leq \tau(\tilde{c}_1(1+\nu)T^{-1}\log p)^{1/2}$ holds with probability at least $1-2p^{-\nu}$ for any fixed $\nu > 0$. Thus, (9) holds with $\lambda = c_0(\log p/T)^{1/2}$ and $c_1 = \nu$, where $c_0 = 2\tau(\tilde{c}_1(1+\nu))^{1/2}$ and ν arbitrary but positive fixed constant. □ □

B.3 Proof of Proposition 2

Proof. To bound $T^{-1}\|\boldsymbol{X}\boldsymbol{v}\|_2^2/\|\boldsymbol{v}\|_2^2$ from below, it is helpful to find the convergence rate of $\|T^{-1}\boldsymbol{X}^\top\boldsymbol{X} - \mathrm{E}[T^{-1}\boldsymbol{X}^\top\boldsymbol{X}]\|_\infty$. By the construction, the (i,j)th element of $\boldsymbol{X}^\top\boldsymbol{X}$ is given by

$$(\boldsymbol{X}^\top\boldsymbol{X})_{ij} = \sum_{t=1}^{T}\sum_{u=1}^{T}\sum_{v=1}^{T}\sum_{k=1}^{p}\sum_{\ell=1}^{p} r_{tu}r_{tv}\sigma_{ki}\sigma_{\ell j}z_{uk}z_{v\ell}.$$

This summation is divided into to parts: (i) $(u,k)=(v,\ell)$ and (ii) $(u,k)\neq(v,\ell)$. We bound each and then combine these results.

We first consider (i). In this case, we have

$$(T^{-1}\boldsymbol{X}^\top\boldsymbol{X})_{ij}-\mathrm{E}(T^{-1}\boldsymbol{X}^\top\boldsymbol{X})_{ij}=T^{-1}\sum_{u=1}^{T}\sum_{k=1}^{p}\left(\sum_{t=1}^{T}r_{tu}^2\right)\sigma_{ki}\sigma_{kj}\left(z_{uk}^2-\mathrm{E}z_{uk}^2\right).$$

By Lemma 1.12 of Rigollet and Hütter (2017), we have $z_{uk}^2-\mathrm{E}z_{uk}^2\sim\mathrm{subE}(16\alpha^2)$. Hence, Lemma 1 yields

$$\begin{aligned}&(T^{-1}\boldsymbol{X}^\top\boldsymbol{X})_{ij}-\mathrm{E}(T^{-1}\boldsymbol{X}^\top\boldsymbol{X})_{ij}\sim\\&\mathrm{subE}\left(16\alpha^2\left(\sum_{k=1}^{p}\sigma_{ki}^2\sigma_{kj}^2\right)^{1/2}T^{-1}\left(\sum_{u=1}^{T}\left(\sum_{t=1}^{T}r_{tu}^2\right)^2\right)^{1/2},16\alpha^2\max_k\sigma_{ki}\sigma_{kj}T^{-1}\max_u\sum_{t=1}^{T}r_{tu}^2\right)\end{aligned}$$

for each i and j. This implies that

$$\begin{aligned}&P\left(\|T^{-1}\boldsymbol{X}^\top\boldsymbol{X}-\mathrm{E}[T^{-1}\boldsymbol{X}^\top\boldsymbol{X}]\|_\infty>x\right)\\&\quad\le p^2\max_{i,j}P\left(\left|(T^{-1}\boldsymbol{X}^\top\boldsymbol{X})_{ij}-\mathrm{E}(T^{-1}\boldsymbol{X}^\top\boldsymbol{X})_{ij}\right|>x\right)\\&\quad\le 2p^2\max_{i,j}\exp\left(-\frac{x^2}{2\cdot16^2\alpha^4\sum_{k=1}^{p}\sigma_{ki}^2\sigma_{kj}^2T^{-2}\sum_{u=1}^{T}\left(\sum_{t=1}^{T}r_{tu}^2\right)^2}\right)\\&\quad\le 2p^2\exp\left(-\frac{x^2}{512\alpha^4T^{-1}\max_i\|\boldsymbol{\sigma}_{\cdot i}^2\|_2^2\max_u\|\boldsymbol{r}_{\cdot u}\|_2^4}\right)\end{aligned}$$

for all $x\in[0,I_2]$, where

$$I_2=16\alpha^2\min_i\|\boldsymbol{\sigma}_{\cdot i}^2\|_2^2T^{-1}\sum_{u=1}^{T}\left(\sum_{t=1}^{T}r_{tu}^2\right)^2\Big/\left(\max_{k,i}\sigma_{ki}^2\max_u\sum_{t=1}^{T}r_{tu}^2\right).$$

This is shown to be bounded and $\liminf_T I_2>0$. Setting $x=(\tilde{c}_2(2+\nu)T^{-1}\log p)^{1/2}$ with $\tilde{c}_2=512\alpha^4\max_i\|\boldsymbol{\sigma}_{\cdot i}^2\|_2^2\max_u\|\boldsymbol{r}_{\cdot u}\|_2^4$ and an arbitrary fixed $\nu>0$ makes the upper bound be equal to

$$2p^2\exp\left(-\frac{\tilde{c}_2(2+\nu)\log p}{512\alpha^4\max_i\|\boldsymbol{\sigma}_{\cdot i}^2\|_2^2\max_u\|\boldsymbol{r}_{\cdot u}\|_2^4}\right)=2p^{-\nu}.$$

This establishes the bound $\|T^{-1}\boldsymbol{X}^\top\boldsymbol{X}-\mathrm{E}[T^{-1}\boldsymbol{X}^\top\boldsymbol{X}]\|_\infty\le(\tilde{c}_2(2+\nu)T^{-1}\log p)^{1/2}$, which holds with probability at least $1-2p^{-\nu}$.

Next we consider (ii). In this case, we have

$$(T^{-1}\boldsymbol{X}^\top\boldsymbol{X})_{ij}-\mathrm{E}(T^{-1}\boldsymbol{X}^\top\boldsymbol{X})_{ij}=T^{-1}\sum_{u=1}^{T}\sum_{v=1}^{T}\sum_{k=1}^{p}\sum_{\ell=1}^{p}(\sum_{t=1}^{T}r_{tu}r_{tv})\sigma_{ki}\sigma_{\ell j}z_{uk}z_{v\ell},$$

where $\sigma_{ki}\sigma_{\ell j}z_{uk}z_{v\ell} \sim$ ind. subG$(4e\alpha^2\sigma_{ki}\sigma_{\ell j})$ for each i and j by Lemma 1. Thus, we obtain

$$\begin{aligned}
&(T^{-1}\boldsymbol{X}^\top\boldsymbol{X})_{ij} - \mathrm{E}(T^{-1}\boldsymbol{X}^\top\boldsymbol{X})_{ij}\\
&\sim \mathrm{subE}\left(4e\alpha^2T^{-1}\left(\sum_{u,v=1}^{T}\sum_{k,\ell=1}^{p}\left(\sum_{t=1}^{T}r_{tu}r_{tv}\right)^2\sigma_{ki}^2\sigma_{\ell j}^2\right)^{1/2}, T^{-1}\max_{u,v,k,\ell}\left|\sum_{t=1}^{T}r_{tu}r_{tv}\right|\sigma_{ki}\sigma_{\ell j}\right)\\
&= \mathrm{subE}\left(4e\alpha^2\|\boldsymbol{\sigma}_{\cdot i}\|_2\|\boldsymbol{\sigma}_{\cdot j}\|_2T^{-1}\left(\sum_{u,v=1}^{T}\left(\sum_{t=1}^{T}r_{tu}r_{tv}\right)^2\right)^{1/2}, \max_{k,\ell}\sigma_{ki}\sigma_{\ell j}\max_{u,v}T^{-1}\left|\sum_{t=1}^{T}r_{tu}r_{tv}\right|\right)
\end{aligned}$$

for each i and j. By using the same inequality as in (i), we have a similar inequality

$$\begin{aligned}
&P\left(\|T^{-1}\boldsymbol{X}^\top\boldsymbol{X} - \mathrm{E}[T^{-1}\boldsymbol{X}^\top\boldsymbol{X}]\|_\infty > x\right)\\
&\quad\le 2p^2\exp\left(-\frac{x^2}{2\cdot 16e^2\alpha^4\max_i\|\boldsymbol{\sigma}_{\cdot i}\|_2^2T^{-2}\sum_{u,v=1}^{T}(\sum_{t=1}^{T}r_{tu}r_{tv})^2}\right)\\
&\quad\le 2p^2\exp\left(-\frac{x^2}{32e^2\alpha^4\max_i\|\boldsymbol{\sigma}_{\cdot i}\|_2^2T^{-1}\max_t\|\boldsymbol{r}_{t\cdot}\|_2^2\max_v\|\boldsymbol{r}_{\cdot v}\|_1^2}\right)
\end{aligned}$$

for all $x \in [0, I_3]$, where

$$I_3 = 4e\alpha^2\min_i\|\boldsymbol{\sigma}_{\cdot i}\|_2^2T^{-1}\sum_{u,v=1}^{T}\left(\sum_{t=1}^{T}r_{tu}r_{tv}\right)^2 \Big/ \max_{k,i}\sigma_{ki}^2\max_{u,v}\left|\sum_{t=1}^{T}r_{tu}r_{tv}\right|$$

and is shown to be bounded and $\liminf_T I_3 > 0$. Setting $x = (\tilde{c}_3(2+\nu)T^{-1}\log p)^{1/2}$ with $\tilde{c}_3 = 32e^2\alpha^4\max_i\|\boldsymbol{\sigma}_{\cdot i}\|_2^2\max_t\|\boldsymbol{r}_{t\cdot}\|_2^2\max_v\|\boldsymbol{r}_{\cdot v}\|_1^2$ and an arbitrary fixed $\nu > 0$ makes the upper bound be equal to

$$2p^2\exp\left(-\frac{\tilde{c}_3(2+\nu)\log p}{32e^2\alpha^4\max_i\|\boldsymbol{\sigma}_{\cdot i}\|_2^2\max_t\|\boldsymbol{r}_{t\cdot}\|_2^2\max_v\|\boldsymbol{r}_{\cdot v}\|_1^2}\right) = 2p^{-\nu}.$$

This establishes the bound $\|T^{-1}\boldsymbol{X}^\top\boldsymbol{X} - \mathrm{E}[T^{-1}\boldsymbol{X}^\top\boldsymbol{X}]\|_\infty \le (\tilde{c}_3(2+\nu)T^{-1}\log p)^{1/2}$, which holds with probability at least $1 - 2p^{-\nu}$. Finally, combining (i) and (ii) with setting $\tilde{c}_2 \vee \tilde{c}_3 \le \tilde{c}_4 := 512\alpha^4\max_i\|\boldsymbol{\sigma}_{\cdot i}\|_2^2\max_t\|\boldsymbol{r}_{t\cdot}\|_2^2\max_v\|\boldsymbol{r}_{\cdot v}\|_1^2$ leads to the result

$$P\left(\|T^{-1}\boldsymbol{X}^\top\boldsymbol{X} - \mathrm{E}[T^{-1}\boldsymbol{X}^\top\boldsymbol{X}]\|_\infty \le \gamma_0(\log p/T)^{1/2}\right) \ge 1 - 4p^{-\nu}, \tag{10}$$

where $\gamma_0 = 2\tilde{c}_4^{1/2}(2+\nu)^{1/2}$.

Finally, we bound $T^{-1}\|\boldsymbol{X}\boldsymbol{v}\|_2^2/\|\boldsymbol{v}\|_2^2$ from below by some positive constant. Let $\boldsymbol{W} = \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}$. We see that its expectation is bounded from below as

$$\mathrm{E}T^{-1}\|\boldsymbol{X}\boldsymbol{v}\|_2^2/\|\boldsymbol{v}\|_2^2 = \mathrm{E}T^{-1}\left(\frac{\boldsymbol{v}^\top\boldsymbol{W}^\top\boldsymbol{R}\boldsymbol{W}\boldsymbol{v}}{\boldsymbol{v}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{v}}\right)\left(\frac{\boldsymbol{v}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{v}}{\boldsymbol{v}^\top\boldsymbol{v}}\right)$$

$$
\begin{aligned}
&\geq \min_{\boldsymbol{h}\in\mathbb{R}^T}\left(\frac{\boldsymbol{h}^\top\boldsymbol{R}\boldsymbol{h}}{\boldsymbol{h}^\top\boldsymbol{h}}\right)\min_{\boldsymbol{v}\in\mathbb{R}^p}\left(\frac{\boldsymbol{v}^\top\boldsymbol{\Sigma}\boldsymbol{v}}{\boldsymbol{v}^\top\boldsymbol{v}}\right)\\
&\geq c_{\min}^2, \qquad (11)
\end{aligned}
$$

where the last inequalities hold by Assumption 4. Since $\boldsymbol{v}$ is in $\mathbb{V} = \{\boldsymbol{v} \in \mathbb{R}^p : \|\boldsymbol{v}_B\|_1 \leq 3\|\boldsymbol{v}_A\|_1\}$, we have

$$
\begin{aligned}
\boldsymbol{v}^\top\mathrm{E}[T^{-1}\boldsymbol{X}^\top\boldsymbol{X}]\boldsymbol{v} - \boldsymbol{v}^\top T^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v} &\leq \|\boldsymbol{v}\|_1\left\|(\mathrm{E}[T^{-1}\boldsymbol{X}^\top\boldsymbol{X}] - T^{-1}\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{v}\right\|_\infty\\
&\leq \|\boldsymbol{v}\|_1^2\left\|\mathrm{E}[T^{-1}\boldsymbol{X}^\top\boldsymbol{X}] - T^{-1}\boldsymbol{X}^\top\boldsymbol{X}\right\|_\infty\\
&\leq (\|\boldsymbol{v}_A\|_1 + \|\boldsymbol{v}_B\|_1)^2\gamma_0(\log p/T)^{1/2}\\
&\leq 16\|\boldsymbol{v}_A\|_1^2\gamma_0(\log p/T)^{1/2}\\
&\leq 16\|\boldsymbol{v}\|_2^2 s\gamma_0(\log p/T)^{1/2},
\end{aligned}
$$

where the third inequality follows from Lemma 10 with probability at least $1 - 4p^{-\nu}$. Rearranging the terms and using (11) yield

$$
\begin{aligned}
T^{-1}\boldsymbol{v}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}/\|\boldsymbol{v}\|_2^2 &\geq T^{-1}\boldsymbol{v}^\top\mathrm{E}[\boldsymbol{X}^\top\boldsymbol{X}]\boldsymbol{v}/\|\boldsymbol{v}\|_2^2 - 16\gamma_0 s(\log p/T)^{1/2}\\
&\geq c_{\min}^2 - 16\gamma_0 s(\log p/T)^{1/2} =: \gamma.
\end{aligned}
$$

Taking infimum over $\boldsymbol{v} \in \mathbb{V}$ gives the result. □ □

C Extension to Heavy-Tailed Random Variables

We have derived the oracle inequality (Theorem 1) under the assumption that $\boldsymbol{X}$ and $\boldsymbol{u}$ possess sub-Gaussian tails; see (3) and Assumption 4 with Definition 1. Taking recent attention on heavy-tailed phenomena into consideration, however, this tail assumption is sometimes too restrictive. In this section, we present a basic idea to extend it to heavy-tailed situation. We first introduce a formal definition of a class of heavy-tailed random variables called *semi-exponential* (denoted by semiE) class.

Definition 3. A random variable $X \in \mathbb{R}$ is said to be *semi-exponential* if $\mathrm{E}[X] = 0$ and it satisfies $\mathrm{E}[\exp(a|X|^\xi)] \leq K$ for some $a > 0$, $\xi \in (0, 1)$, and $K > 0$. In this case, we write $X \sim \mathrm{semiE}(a, \xi, K)$. (We just write semiE($\xi$) below since a and K are not so important.)

This forms a family of random variables that have fatter tails than subE. In fact, subE corresponds to the boundary case, $\xi = 1$, in Definition 3. It is noteworthy that for a sum of independent semiE random variables, Bernstein's inequality is available as in the case of subE though it has a little different formulation. As is explained in Section 3, Propositions 1 and 2 provide complete information for obtaining Theorem 1. More specifically, Bernstein's inequality (for subE) controls the entire stochastic behaviors in their proofs. In this sense, it is conceptually straightforward to achieve the oracle inequality even under semiE conditions since Bernstein's inequality for semiE is applicable. We observe a specific way of extension.

For comparison, suppose first the case where $\boldsymbol{X}$ and $\boldsymbol{u}$ are such that $(x_{ti}u_t)_t \sim$ i.i.d. subE(α) for all $i \in \{1, \ldots, p\}$. Note that this holds when x_{ti} and u_t are both assumed subG; see Lemma 1. Here, we suppressed any time dependence in the model to elucidate an essential point. Then, by Bernstein's inequality together with the union bound, we have

$$\begin{aligned} P\left(\left\|T^{-1}\boldsymbol{X}^\top \boldsymbol{u}\right\|_\infty > x\right) &\leq pP\left(\left|T^{-1}\sum_{t=1}^{T} x_{t1}u_t\right| > x\right) \\ &\lesssim p\exp\left(-C_1 T x^2\right) + p\exp\left(-C_2 T x\right), \end{aligned}$$

where $\lesssim$ denotes $\leq$ up to a positive constant factor and C_1 and C_2 are some positive constants that depend only on α. We can see that, if we assign $x = \lambda \sim (\log p/T)^{1/2}$, we only consider the first term in the upper bound because it dominates the second asymptotically. This is also true for all λ converging to zero.

Next, suppose the case where $\boldsymbol{X}$ and $\boldsymbol{u}$ satisfy $(x_{ti}u_t)_t \sim$ i.i.d. semiE(ξ) for all $i \in \{1, \ldots, p\}$. Then, from Nerlevede, Peligrad and Rio (2011), Bernstein's inequality for semi-exponential random variables is available:

$$P\left(\left\|T^{-1}\boldsymbol{X}^\top \boldsymbol{u}\right\|_\infty > x\right) \lesssim p\exp\left(-C_1 T x^2\right) + p\exp\left(-C_2 T^\xi x^\xi\right).$$

Contrary to the case of $x_{t1}u_t \sim$ subE(α) above, the second term in the upper bound will dominate the first for some class of λ to be assigned, depending on its convergence rate to zero. This also depends how large p and ξ the model has. In general, we should take

$$\lambda \sim (\log p/T)^{1/2} \vee (\log p/T^\xi)^{1/\xi},$$

which will make the upper bound of the probability be $O(p^{-C})$ for some positive constant C as long as we appropriately choose a large constant factor of λ. We are now interested in which value becomes larger in response to p and ξ. To see this, suppose $\log p = T^\delta$ for some

$\delta \in (0,1)$. By a simple algebra, we have

$$\lambda \sim \begin{cases} (\log p/T)^{1/2} = T^{(\delta-1)/2} & \text{for} \quad 2\delta/(\delta+1) \le \xi < 1, \\ (\log p/T^{\xi})^{1/\xi} = T^{(\delta-\xi)/\xi} & \text{for} \quad 0 < \xi < 2\delta/(\delta+1). \end{cases}$$

Note that $\xi = 2\delta/(\delta+1)$ is a concave and increasing function in $\delta \in (0,1)$ that connects the origin and $(1,1)$. Roughly speaking, a combination of small δ and/or large ξ (i.e., lower dimensionality and/or closer to the subE tails) leads to the usual bound; $\|T^{-1}\boldsymbol{X}^\top \boldsymbol{u}\|_\infty \le \lambda/2$ holds with probability at least $1 - O(p^{-C})$ for some $\lambda \sim (\log p/T)^{1/2}$ and C. On the other hand, when δ is large and/or ξ is small (i.e., higher dimensionality and/or fatter tails than subE), the bound holds with replacing by $\lambda \sim (\log p/T^{\xi})^{1/\xi}$. As a result, the rates of convergence implied by Theorem 1 change in the latter case.

Of course, we have to check the lower bound of $\inf_{\boldsymbol{d}\in\mathbb{V}} \|\boldsymbol{X}\boldsymbol{d}\|_2^2/\|\boldsymbol{d}\|_2^2$ as in Proposition 2 at the same time, but we omit the discussion here because the problem is essentially the same. In fact, it is sufficient to consider a further upper bound of the probability

$$\begin{aligned} &P\left(\left\|T^{-1}\boldsymbol{X}^\top\boldsymbol{X} - \mathrm{E}T^{-1}\boldsymbol{X}^\top\boldsymbol{X}\right\|_\infty > x\right) \\ &\quad \le \sum_{i,j=1}^{p} \max_{i,j} P\left(\left|T^{-1}\sum_{t=1}^{T}(x_{ti}x_{tj} - \mathrm{E}x_{ti}x_{tj})\right| > x\right), \end{aligned}$$

where $(x_{ti}x_{tj} - \mathrm{E}x_{ti}x_{tj})_t \sim$ i.i.d. semiE, by using Bernstein's inequality for semiE; see the proof of Proposition 2. This may result in some rate change in (10) that will be caused by a combination of given ξ and p.

D Collinearity

Another important assumption to achieve the oracle property (Property 1) is

$$\|\boldsymbol{H}_{BAT}\|_{2,\infty} \equiv \max_{\|\boldsymbol{v}\|_2=1} \|\boldsymbol{H}_{BAT}\boldsymbol{v}\|_\infty = O_p(1); \tag{12}$$

see Fan and Lv (2011). This condition controls how much collinearity is allowed. In this section, we investigate how collinearity between $\boldsymbol{X}_B$ and $\boldsymbol{X}_A$ affects condition (12). Recall that $\boldsymbol{H}_{BAT} = T^{-1}\boldsymbol{X}_B^\top\boldsymbol{X}_A$ for $\boldsymbol{X}_A \in \mathbb{R}^{T\times s}$ and $\boldsymbol{X}_B \in \mathbb{R}^{T\times(p-s)}$. We are interested in the behavior of

$$\|\boldsymbol{H}_{BAT}\|_{2,\infty} \equiv \max_{\|\boldsymbol{v}\|_2=1} \|\boldsymbol{H}_{BAT}\boldsymbol{v}\|_\infty = \max_{b\in B}\max_{\|\boldsymbol{v}\|_2=1} \left|T^{-1}\boldsymbol{x}_b^\top\boldsymbol{X}_A\boldsymbol{v}\right|,$$

where we write $\boldsymbol{X}_A \boldsymbol{v} = \sum_{a \in A} v_a \boldsymbol{x}_a$. This value is expected to become unbounded (and hence condition (12) is violated) under strong collinearity.

To obtain understandable results, we make the following simplified assumptions: the regressors are deterministic, and for any $b \in B$ and $a \in A$, $T^{-1}\boldsymbol{x}_b^\top \boldsymbol{x}_a \to \rho_{ba} \geq 0$. Moreover, we assume either of the two conditions:

1. $\max_{b \in B} \rho_{ba} \geq c > 0$ for all $a \in A$,

2. $\max_{b \in B} \rho_{ba} \leq c a^{-q/2}$ for some $q > 1$.

Condition 1 describes a highly correlated case. The correlation between $\boldsymbol{x}_b$ and $\boldsymbol{x}_a$ always exists even if s increases. On the other hand, condition 2 models weaker correlations than condition 1 does. Specifically, most of the correlations become small as q becomes large, meaning that the effect of collinearity is limited in this case. In fact, it is not difficult to see that $\|\boldsymbol{H}_{BAT}\|_{2,\infty}$ diverges at least as fast as $s^{1/2}$ under condition 1 while $\|\boldsymbol{H}_{BAT}\|_{2,\infty}$ is uniformly bounded under condition 2. First, we suppose condition 1 and let $\bar{\boldsymbol{v}} = (s^{-1/2}, \dots, s^{-1/2})^\top$. We then observe that

$$\max_{b \in B} \max_{\|\boldsymbol{v}\|_2 = 1} \left| T^{-1}\boldsymbol{x}_b^\top \boldsymbol{X}_A \boldsymbol{v} \right| \geq \max_{b \in B} \left| T^{-1}\boldsymbol{x}_b^\top \boldsymbol{X}_A \bar{\boldsymbol{v}} \right| = \max_{b \in B} \left| s^{-1/2} \sum_{a \in A} T^{-1}\boldsymbol{x}_b^\top \boldsymbol{x}_a \right|.$$

By condition 1, the last term is bounded from below by

$$s^{-1/2} \max_{b \in B} \left| \sum_{a \in A} (\rho_{ba} + o(1)) \right| \geq s^{1/2}(c - o(1)),$$

which goes to infinity as $s \to \infty$. Next, we suppose condition 2. By the Cauchy-Schwarz inequality, we observe that

$$\begin{aligned}
\max_{b \in B} \max_{\|\boldsymbol{v}\|_2 = 1} \left| T^{-1}\boldsymbol{x}_b^\top \boldsymbol{X}_A \boldsymbol{v} \right| &= \max_{b \in B} \max_{\|\boldsymbol{v}\|_2 = 1} \left| \sum_{a \in A} v_a T^{-1}\boldsymbol{x}_b^\top \boldsymbol{x}_a \right| \\
&= \max_{b \in B} \max_{\|\boldsymbol{v}\|_2 = 1} \left| \sum_{a \in A} v_a \left(\rho_{ba} + o(1)\right) \right| \\
&\leq \max_{b \in B} \left(\sum_{a \in A} \rho_{ba}^2 \left(1 + o(1)\right) \right)^{1/2} \leq c \left(\sum_{a \in A} a^{-q} \left(1 + o(1)\right) \right)^{1/2}.
\end{aligned}$$

The last term converges since $q > 1$ under condition 2.

The following simulation shows that the strong collinearity (condition 1) affects the oracle property. Table D.1 shows the *relative* finite sample success rates of the MCP detecting nonzero (*SC-A*) coefficients and zero coefficients (*SC-B*) that are defined as

$$\mathit{SC\text{-}A} = P\left(\operatorname{sgn}(\hat{\boldsymbol{\beta}}_A) = \operatorname{sgn}(\boldsymbol{\beta}_{0A})\right),$$

$$SC\text{-}B = P\left(\operatorname{sgn}(\hat{\boldsymbol{\beta}}_B) = \operatorname{sgn}(\boldsymbol{\beta}_{0B})\right),$$

respectively, and (average) mean squared error for estimates of non-zero coefficients (MSE($\hat{\boldsymbol{\beta}}_A$)) under condition 1 with 5000 repetitions compared to that of condition 2, when $T = 300, 500, 1000$ and $c = 0.5, 0.98$ with $q = 4$, $p = 1.5\exp(T^{0.31})$ and $s = 20T^{0.3}$. Then, the finite sample properties of estimators under condition 1 are equivalent to those of condition 2 if the values in the table are 1. We can confirm facts from Table D.1 that (i) the values of $SC\text{-}A$ are relatively low under condition 1 irrespective of the degree of collinearity (c) and (ii) the MSE of condition 1 is expected to be much worse than the that of condition 2 asymptotically especially when the degree of collinearity is high. These facts are consistent to the theoretical results because the condition 1 violates assumption (12) so that the oracle property is no longer proved under condition 1.

E Finite Sample Forecasting Performance with Simulated Data

In this section, we examine finite sample performances of forecasts based on the MCP, SCAD, Lasso, and OLS post-Lasso estimators using simulated data. We assume $\boldsymbol{x}_t \sim$ i.i.d. $N(\mathbf{0}, \boldsymbol{\Sigma}_x)$, where $\boldsymbol{\Sigma}_x = \{\sigma_{x,ij}\}$ in this experiment and set the DGP as follows:

$$\begin{aligned}
\sigma_{x,ij} &= \rho^{|i-j|}, \quad i, j = 1, 2, \ldots, p, \\
y_t &= \boldsymbol{x}^\top \boldsymbol{\beta}_0 + \boldsymbol{u}_t \\
&= \boldsymbol{x}_{At}^\top \boldsymbol{\beta}_{0A} + \boldsymbol{x}_{Bt}^\top \boldsymbol{\beta}_{0B} + \boldsymbol{u}_t,
\end{aligned}$$

where $\boldsymbol{\beta}_{0,A}$ is the s-dimensional unit vector, $\boldsymbol{\beta}_{0,B}$ is the $(p-s)$-dimensional zero vector, and $\boldsymbol{u}_t \sim$ i.i.d. $N(0, \boldsymbol{\Sigma}_u)$. Here we set $\boldsymbol{\Sigma}_u = \boldsymbol{\beta}_0^\top \boldsymbol{\Sigma}_x \boldsymbol{\beta}_0/\text{SNR}$, where SNR $\in (0, \infty)$ is the "signal-to-noise ratio" of the model. Note that the covariates allow for collinearity but are assumed to be time-independence. We set SNR $= 9$, $p = 1000$ and $s = [T^{0.5}]$ throughout the experiment. Under the setting, we see forecast performances of the MCP, SCAD, MCP, and OLS post-Lasso when $T = 200, 500, 700, 1000$ and $\rho = 0.3, 0.5, 0.8$, respectively. Following Section 4.1, we compare the out-of-sample forecast performances measured by the MSFEs. The MSFEs are evaluated over 30 repetitions: we make the start point in the sample for each estimation be fixed while the end point increases for each repetition, so that the size of the estimation sample is ranging from $T - 30$ to $T - 1$.

Table E.1 shows the MSFEs of the MCP, SCAD, Lasso, and OLS post-Lasso. The values in the table are standardized by that of the MCP, so that the value less than one means a forecast based on the corresponding estimator performs well compared to that of the MCP. As a whole, the table shows the OLS post-Lasso dominates the others for all T and ρ. It seems to perform much better when T is small in particular. Note that the OLS post-Lasso estimator does not necessarily perform well in Section 4.1. We can conjecture from the two conflicting results that time-dependent data distorts finite sample performance of the OLS post-selection estimators.

F Forecast performance in real-time data

In this section, we investigate how well the forecast with penalized regression works with the real-time data. It should be mentioned that in our experiment, strictly speaking, we consider "pseudo" real-time forecasting; we suppose each monthly data for all evaluation periods have the same jagged (ragged)-edge pattern as of the 2018-02 version of the FRED-MD. For example, Real manufacturing and trade industry sales (CMRMTSPLx) and the Help-wanted index (HWI) have two and one month missing values owing to publication lags in the 2018-02 version, respectively. Then we suppose the data for all estimation periods have the same jagged-edge patterns even if our dataset contains complete data for those periods. Moreover, we assume no data revisions occur in our dataset.

Tables F.1–F.3 show the relative mean and median squared forecast errors of the penalized regression and the state-space ML estimator proposed by Banbura and Modugno (2014) in the real-time overall sample (2000Q1–2017Q4), 1st subsample (2000Q1–2007Q4), and 2nd subsample (2008Q1–2017Q4), respectively. The tables omit the results for $h > 1$ and concentrate on the nowcast situation ($0 \leq h \leq 1$) because the real-time forecasting is meaningful only in a very short horizon. The state-space ML estimation enables us to handle real-time mixed frequency data by embedding missing patterns of data in the model; see Banbura and Modugno (2014) for details. On the other hand, the penalized regression requires an interpolated dataset to obtain the forecast values. Thus, we employ an interpolation method based on the EM algorithm proposed by Stock and Watson (2002).

From the tables, we first find the effects of the jagged-edge and interpolation on the forecast accuracy of the penalized regression are small and they do not essentially affect the mean/median squared forecast errors values compared with the results in Tables 1–3. Second, we see that the penalized regression performs well in terms of the median squared errors although the state-space ML tends to performs better in terms of the MSFE. The state-space ML is expected to have dominating forecasting performance compared to the penalized regression, because the state-space ML is based on a system equation with richer information while the penalized regression relies only on a single equation. However, this would not be true when a model misspecification is present, as Bai, Ghysels, and Wright (2013) claimed. Then, our results imply that the system equation may contain a certain level of the misspecification. Moreover, it should be mentioned that the penalized regression is much simpler and rapid than the state-space ML in obtaining the forecast values. Since the dimension of the state-space model can be very large when we forecast with mixed frequency (117 dimensional state-space models with 40 latent factors in our case), the estimation is much computationally demanding and time consuming (roughly eight times longer than the penalized regression). Furthermore, the estimated values can be unstable if we consider to apply the state-space ML to a dataset with larger N and/or r.

Although we do not examine them here, we should also note that the Bayesian VAR (BVAR) would be potential alternatives to the state-space ML; see e.g., Banbura, Giannone and Reichlin (2010), Koop (2013), Schorfheide and Song (2015). The BVAR is expected to have promising forecasting performance, however, it also seems much computationally demanding than our univariate penalized regression. Moreover, its theoretical properties would not have been well explored under ultrahigh dimensionality.

References

Bai, J., E. Ghysels, and J. H. Wright (2013). State Space Models and MIDAS Regressions. *Econometric Reviews* , *32*, 779–813.

Bańbura, M., D. Giannone and L. Reichlin (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, *25*, 71–92.

Bańbura, M. and M. Modugno (2014). Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *Journal of Applied Econometrics*, *29*, 133–160.

Bercu, B., B. Delyon, and E. Rio (2015). *Concentration inequalities for sums and martingales.* New York: Springer.

Davidson, J. (1994). *Stochastic Limit Theory.* Oxford University Press.

Fan, J. and J. Lv (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE*

Transactions on Information Theory, *57*, 5467–5484.

Koop, G. M. (2013). Forecasting with medium and large Bayesian VARs. *Journal of Applied Econometrics*, *28*, 177–203.

Lv, J. and Y. Fan (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Annals of Statistics*, *37*, 3498–3528.

Merlevède, F., M. Peligrad, and E. Rio (2011). A Bernstein type inequality and moderate deviations for weakly dependent sequences *Probability Theory and Related Fields* , *151*, 435–474.

Rigollet, P. and J.-C. Hütter (2017). *High Dimensional Statistics.* Lecture notes, Massachusetts Institute of Technology, MIT Open CourseWare.

Schorfheide, F. and D. Song (2015). Real-time forecasting with a mixed-frequency VAR. *Journal of Business & Economic Statistics*, *33*, 366–380.

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

Zhao, P. and B. Yu (2006). On model selection consistency of LASSO. *Journal of Machine Learning Research*, *7*, 2541–2563.

Table 1: Mean/Median Forecast Errors of the forecasts in complete data [Overall Sample]

without y_{t-1}	$h=0$	$h=1/3$	$h=2/3$	$h=1$	$h=4/3$	$h=5/3$	$h=2$
MCP	0.60	0.57	0.66	0.79	1.16	1.41	1.42
(median)	(0.80)	(0.63)	(0.74)	(0.69)	(1.06)	(1.59)	(1.70)
SCAD	0.59	0.58	0.68	0.81	1.15	1.39	1.40
(median)	(0.72)	(0.55)	(0.81)	(0.68)	(1.09)	(1.48)	(1.36)
Lasso	0.60	0.59	0.65	0.81	1.15	1.39	1.41
(median)	(0.90)	(0.68)	(0.66)	(0.63)	(1.08)	(1.48)	(1.48)
Factor	0.76	2.41	0.65	0.95	2.06	1.15	1.49
(median)	(0.68)	(0.90)	(0.74)	(1.09)	(2.29)	(2.38)	(1.85)
post-MCP	0.75	0.71	0.63	0.75	1.26	1.87	1.86
(median)	(1.12)	(0.94)	(1.01)	(0.98)	(1.49)	(2.00)	(2.38)
post-SCAD	0.78	0.78	0.62	0.79	1.24	1.65	1.87
(median)	(0.77)	(0.88)	(0.91)	(0.75)	(1.50)	(2.57)	(2.51)
post-Lasso	0.79	0.75	0.63	0.79	1.24	1.65	1.86
(median)	(0.88)	(0.73)	(0.91)	(0.75)	(1.50)	(2.57)	(2.51)

with y_{t-1}	$h=0$	$h=1/3$	$h=2/3$	$h=1$	$h=4/3$	$h=5/3$	$h=2$
MCP	0.61	0.57	0.55	0.78	1.17	1.45	1.46
(median)	(0.79)	(0.84)	(0.76)	(0.76)	(1.32)	(1.79)	(1.88)
SCAD	0.58	0.58	0.57	0.79	1.15	1.46	1.54
(median)	(0.85)	(0.77)	(0.69)	(0.82)	(1.42)	(1.65)	(1.84)
Lasso	0.64	0.60	0.61	0.79	1.14	1.44	1.44
(median)	(0.93)	(0.73)	(0.61)	(0.71)	(1.41)	(1.50)	(1.86)
Factor	0.63	2.50	0.59	1.12	2.48	1.27	1.66
(median)	(0.73)	(0.93)	(0.64)	(0.96)	(1.30)	(2.10)	(1.70)
post-MCP	0.74	0.77	0.49	0.73	1.27	1.81	1.94
(median)	(0.94)	(0.61)	(0.68)	(0.93)	(1.43)	(2.45)	(2.57)
post-SCAD	0.77	0.73	0.55	0.77	1.42	1.59	1.79
(median)	(0.92)	(0.90)	(0.70)	(0.95)	(1.19)	(2.27)	(2.44)
post-Lasso	0.79	0.73	0.54	0.77	1.40	1.58	1.79
(median)	(0.92)	(0.90)	(0.65)	(0.97)	(1.19)	(2.27)	(2.44)

Note) The entries are the ratios of mean/median forecast errors to that of AR(4) forecast. Values in parentheses are median forecast errors.

Table 2: Mean/Median Forecast Errors of the forecasts in complete data [1st Subsample]

without y_{t-1}	$h = 0$	$h = 1/3$	$h = 2/3$	$h = 1$	$h = 4/3$	$h = 5/3$	$h = 2$
MCP	0.73	0.73	0.79	0.92	1.03	1.32	1.35
(median)	(0.60)	(0.73)	(0.88)	(0.87)	(1.30)	(1.86)	(1.97)
SCAD	0.75	0.76	0.80	0.92	1.01	1.27	1.27
(median)	(0.74)	(0.63)	(1.11)	(0.88)	(1.14)	(1.80)	(1.76)
Lasso	0.74	0.75	0.76	0.93	1.01	1.27	1.27
(median)	(0.57)	(0.54)	(0.80)	(0.88)	(1.14)	(1.80)	(1.76)
Factor	0.73	0.79	0.97	1.11	1.41	1.62	1.66
(median)	(0.82)	(0.94)	(1.11)	(1.47)	(3.11)	(3.18)	(3.88)
post-MCP	0.76	0.74	1.02	0.99	1.17	1.66	1.66
(median)	(1.33)	(1.33)	(2.18)	(1.84)	(1.78)	(3.36)	(3.36)
post-SCAD	0.73	0.77	1.00	1.00	1.23	1.65	1.65
(median)	(1.07)	(1.12)	(1.50)	(1.62)	(1.95)	(3.03)	(3.03)
post-Lasso	0.73	0.77	1.00	1.00	1.23	1.65	1.65
(median)	(1.07)	(1.12)	(1.50)	(1.62)	(1.95)	(3.03)	(3.03)

with y_{t-1}	$h = 0$	$h = 1/3$	$h = 2/3$	$h = 1$	$h = 4/3$	$h = 5/3$	$h = 2$
MCP	0.81	0.79	0.75	0.99	1.11	1.33	1.38
(median)	(0.82)	(0.86)	(1.11)	(1.08)	(1.69)	(2.03)	(2.25)
SCAD	0.80	0.82	0.77	0.96	1.08	1.34	1.38
(median)	(0.97)	(0.69)	(1.03)	(0.97)	(1.68)	(1.97)	(2.36)
Lasso	0.86	0.79	0.78	0.96	1.08	1.35	1.32
(median)	(0.97)	(0.71)	(0.83)	(0.97)	(1.70)	(1.92)	(2.09)
Factor	0.64	0.96	0.89	0.94	1.03	1.95	1.79
(median)	(1.04)	(1.20)	(0.91)	(1.33)	(1.63)	(3.73)	(3.23)
post-MCP	0.78	0.75	0.80	1.04	1.27	1.61	1.60
(median)	(1.49)	(1.08)	(1.53)	(1.64)	(1.75)	(3.43)	(3.24)
post-SCAD	0.84	0.86	0.74	0.99	1.24	1.60	1.60
(median)	(1.15)	(1.36)	(0.99)	(1.56)	(1.96)	(2.97)	(2.97)
post-Lasso	0.84	0.85	0.75	0.99	1.24	1.60	1.60
(median)	(1.15)	(1.27)	(1.24)	(1.56)	(1.96)	(2.97)	(2.97)

Note) The entries are the ratios of mean/median forecast errors to that of AR(4) forecast. Values in parentheses are median forecast errors.

Table 3: Mean/Median Forecast Errors of the forecasts in complete data [2nd Subsample]

without y_{t-1}	$h = 0$	$h = 1/3$	$h = 2/3$	$h = 1$	$h = 4/3$	$h = 5/3$	$h = 2$
MCP	0.54	0.50	0.60	0.73	1.23	1.45	1.46
(median)	(0.83)	(0.70)	(0.70)	(0.60)	(1.02)	(1.65)	(1.61)
SCAD	0.51	0.49	0.63	0.76	1.22	1.45	1.46
(median)	(0.71)	(0.55)	(0.72)	(0.65)	(1.15)	(1.40)	(1.29)
Lasso	0.53	0.52	0.59	0.75	1.21	1.45	1.48
(median)	(0.88)	(0.77)	(0.62)	(0.59)	(1.21)	(1.40)	(1.40)
Factor	0.77	3.16	0.49	0.87	2.37	0.92	1.41
(median)	(0.77)	(0.94)	(0.74)	(1.91)	(2.58)	(1.56)	(1.91)
post-MCP	0.74	0.69	0.44	0.64	1.30	1.98	1.95
(median)	(0.73)	(0.86)	(0.55)	(0.65)	(1.06)	(1.58)	(2.05)
post-SCAD	0.81	0.79	0.44	0.69	1.25	1.65	1.97
(median)	(0.73)	(0.62)	(0.40)	(0.70)	(0.89)	(2.44)	(2.38)
post-Lasso	0.82	0.75	0.45	0.69	1.25	1.66	1.97
(median)	(0.81)	(0.66)	(0.77)	(0.69)	(0.84)	(2.44)	(2.38)

with y_{t-1}	$h = 0$	$h = 1/3$	$h = 2/3$	$h = 1$	$h = 4/3$	$h = 5/3$	$h = 2$
MCP	0.51	0.47	0.45	0.68	1.20	1.51	1.49
(median)	(0.91)	(0.84)	(0.69)	(0.41)	(1.05)	(1.75)	(1.78)
SCAD	0.49	0.47	0.48	0.71	1.18	1.51	1.61
(median)	(0.84)	(0.78)	(0.58)	(0.66)	(1.23)	(1.47)	(1.72)
Lasso	0.53	0.51	0.53	0.71	1.18	1.48	1.50
(median)	(0.95)	(0.75)	(0.57)	(0.61)	(1.20)	(1.37)	(1.82)
Factor	0.62	3.22	0.44	1.20	3.17	0.95	1.60
(median)	(0.57)	(1.45)	(0.57)	(1.54)	(1.64)	(1.49)	(1.94)
post-MCP	0.71	0.78	0.34	0.58	1.27	1.90	2.10
(median)	(0.76)	(0.55)	(0.51)	(0.70)	(0.95)	(1.48)	(2.13)
post-SCAD	0.73	0.67	0.46	0.67	1.50	1.59	1.87
(median)	(0.88)	(0.50)	(0.67)	(0.77)	(1.04)	(1.93)	(2.29)
post-Lasso	0.77	0.68	0.44	0.67	1.48	1.57	1.87
(median)	(0.90)	(0.59)	(0.57)	(0.77)	(1.04)	(1.77)	(2.26)

Note) The entries are the ratios of mean/median forecast errors to that of AR(4) forecast. Values in parentheses are median forecast errors.

Table D.1: Relative SC-A, SC-B and MSE (cond.1/cond.2)

		$c = 0.5$			$c = 0.98$	
	$SC-A$	$SC-B$	MSE	$SC-A$	$SC-B$	MSE
$T = 300$	0.80	1.01	1.17	0.97	1.00	0.99
$T = 500$	0.80	1.01	1.36	0.98	0.99	0.94
$T = 1000$	1.00	1.00	1.06	0.94	1.00	2.36

Table E.1: Mean squared forecast errors of the forecasts in simulated data. Values less than one indicate better performance than MCP.

$\rho = 0.8$	$T = 200$	$T = 500$	$T = 700$	$T = 1000$
MCP	1.00	1.00	1.00	1.00
SCAD	1.00	1.00	1.00	1.00
Lasso	0.92	0.93	0.94	0.95
post-Lasso	0.86	0.90	0.92	0.92

$\rho = 0.5$	$T = 200$	$T = 500$	$T = 700$	$T = 1000$
MCP	1.00	1.00	1.00	1.00
SCAD	1.02	1.01	1.00	1.00
Lasso	1.02	1.02	1.02	1.02
post-Lasso	0.84	0.92	0.94	0.95

$\rho = 0.3$	$T = 200$	$T = 500$	$T = 700$	$T = 1000$
MCP	1.00	1.00	1.00	1.00
SCAD	1.01	1.00	1.00	1.00
Lasso	1.11	1.05	1.05	1.05
post-Lasso	0.84	0.93	0.94	0.95

Note) The entries are the ratios of the MSFEs to that of MCP.

Table F.1: Mean/Median Forecast Errors of the forecasts in jagged-edge data [Overall sample]

	$h = 0$	$h = 1/3$	$h = 2/3$	$h = 1$
MCP	0.62	0.60	0.56	0.79
(median)	(0.83)	(0.64)	(0.81)	(0.66)
SCAD	0.61	0.59	0.60	0.81
(median)	(0.76)	(0.57)	(0.94)	(0.67)
Lasso	0.61	0.59	0.54	0.79
(median)	(0.94)	(0.71)	(0.64)	(0.65)
State-Space ML	0.51	0.50	0.60	0.72
(median)	(0.76)	(0.74)	(1.08)	(0.87)

Note) The entries are the ratios of mean/median forecast errors to that of AR(4) forecast. Values in parentheses are median forecast errors.

Table F.2: Mean/Median Forecast Errors of the forecasts in jagged-edge data [1st Subsample]

	$h = 0$	$h = 1/3$	$h = 2/3$	$h = 1$
MCP	0.73	0.75	0.72	0.92
(median)	(0.73)	(0.94)	(1.07)	(0.87)
SCAD	0.75	0.80	0.72	0.92
(median)	(0.74)	(1.02)	(1.11)	(0.89)
Lasso	0.73	0.78	0.71	0.92
(median)	(0.60)	(1.05)	(0.77)	(0.89)
State-Space ML	0.65	0.67	0.71	0.92
(median)	(0.71)	(1.10)	(0.69)	(1.20)

Note) The entries are the ratios of mean/median forecast errors to that of AR(4) forecast. Values in parentheses are median forecast errors.

Table F.3: Mean/Median Forecast Errors of the forecasts in jagged-edge data [2nd Subsample]

	$h = 0$	$h = 1/3$	$h = 2/3$	$h = 1$
MCP	0.57	0.54	0.48	0.73
(median)	(0.99)	(0.61)	(0.61)	(0.47)
SCAD	0.54	0.49	0.54	0.75
(median)	(0.85)	(0.53)	(0.73)	(0.58)
Lasso	0.55	0.51	0.46	0.74
(median)	(0.98)	(0.65)	(0.56)	(0.56)
State-Space ML	0.44	0.43	0.54	0.63
(median)	(0.93)	(0.69)	(1.13)	(0.60)

Note) The entries are the ratios of mean/median forecast errors to that of AR(4) forecast. Values in parentheses are median forecast errors.