Statistica Sinica **30** (2020), 1797-1827
doi:https://doi.org/10.5705/ss.202018.0044

HIGH-DIMENSIONAL LINEAR REGRESSION FOR DEPENDENT DATA WITH APPLICATIONS TO NOWCASTING

Yuefeng Han and Ruey S. Tsay

University of Chicago

Abstract: Recent research has focused on ℓ_1 penalized least squares (Lasso) estimators for high-dimensional linear regressions in which the number of covariates p is considerably larger than the sample size n. However, few studies have examined the properties of the estimators when the errors and/or the covariates are serially dependent. In this study, we investigate the theoretical properties of the Lasso estimator for a linear regression with a random design and weak sparsity under serially dependent and/or nonsubGaussian errors and covariates. In contrast to the traditional case, in which the errors are independent and identically distributed and have finite exponential moments, we show that p can be at most a power of n if the errors have only finite polynomial moments. In addition, the rate of convergence becomes slower owing to the serial dependence in the errors and the covariates. We also consider the sign consistency of the model selection using the Lasso estimator when there are serial correlations in the errors or the covariates, or both. Adopting the framework of a functional dependence measure, we describe how the rates of convergence and the selection consistency of the estimators depend on the dependence measures and moment conditions of the errors and the covariates. Simulation results show that a Lasso regression can be significantly more powerful than a mixed-frequency data sampling regression (MIDAS) and a Dantzig selector in the presence of irrelevant variables. We apply the results obtained for the Lasso method to nowcasting with mixed-frequency data, in which serially correlated errors and a large number of covariates are common. The empirical results show that the Lasso procedure outperforms the MIDAS regression and the autoregressive model with exogenous variables in terms of both forecasting and nowcasting.

Key words and phrases: Consistency, forecasting, high-dimensional time series, Lasso, mixed-frequency data, model selection, nowcasting.

1. Introduction

The past two decades have witnessed significant developments in high-dimensional linear regression analyses. Consider the following linear regression for the response variable y_i and the covariate vector $\mathbf{x}_i$:

$$y_i = \mathbf{x}_i^T \beta + e_i, \quad 1 \le i \le n, \tag{1.1}$$

where $\beta \in \mathbb{R}^p$ consists of unknown coefficients, e_i is an error term, and $\mathbf{x}_i^T$ denotes the transpose of the covariate vector $\mathbf{x}_i$. Denote the dimension of $\mathbf{x}_i$ by p. In matrix form, we can write the model as $Y = X\beta + e$, where Y is an $n \times 1$ response vector, X is an $n \times p$ design matrix, and e is an $n \times 1$ error vector. Under certain sparsity conditions on β, many studies have focused on the ℓ_1 penalized least squares (Lasso) estimator of β when the number of variables p can be much larger than the sample size n; see Efron et al. (2004), Zhao and Yu (2006), and Meinshausen and Yu (2009), among others. Other related approaches include the Dantzig selector of Candes and Tao (2007), adaptive Lasso of Zou (2006), group Lasso by Yuan and Lin (2006), and SCAD estimator of Fan and Li (2001), among others. The theoretical properties of those estimators have been established in the literature under the independence assumption; see, for example, Bickel, Ritov and Tsybakov (2009) and Bühlmann and Van De Geer (2011). Here, we focus on the Lasso estimator defined as

$$\hat{\beta} = \arg\min_{\beta} \left(\frac{1}{2} |Y - X\beta|_2^2 + \lambda |\beta|_1 \right), \tag{1.2}$$

where $\lambda \ge 0$ is a tuning parameter that controls the level of sparsity in $\hat{\beta}$.

Much of the available research dedicated to the Lasso problem examines the case of large p and small n when the design matrix is static and the errors are independent and identically distributed (i.i.d.) random variables. On the other hand, in many real applications, $\mathbf{x}_i$ consists of stochastic random variables that might be dynamically dependent, or e_i is serially dependent, or both. Despite considerable recent work on Lasso estimators, few studies examine the theoretical properties of the estimates when the observations are dependent. Wang, Li and Tsai (2007) proposed a Lasso estimator for a regression model with autoregressive errors. Gupta (2012) investigated the Lasso estimator for weakly dependent errors. Both studies concentrate on the case when n is greater than p. More recently, Basu and Michailidis (2015) investigated the theoretical properties of Lasso estimators using a random design for high-dimensional Gaussian processes. Kock and Callot (2015) established the oracle inequalities of the Lasso for Gaussian errors in stationary vector autoregressive models. Wu and Wu (2016) analyzed the Lasso estimator with a fixed design matrix, and assumed that a

restricted eigenvalue condition is satisfied. Medeiros and Mendes (2016) studied the asymptotic properties of the adaptive Lasso when the errors are nonGaussian and may be conditionally heteroskedastic. The goal of this study is to investigate the limiting properties of Lasso estimators for Model (1.1) in the presence of serial dependence in both the covariate vector $\mathbf{x}_i$ and the errors. We establish the rate of convergence and provide the sign consistency of the Lasso estimator under the weak sparsity condition. Our results extend beyond those of a fixed design and exact sparsity time series; thus, we do not assume a restricted eigenvalue condition on either the sample or on the population covariance matrix.

In practice, many important macroeconomic variables are not sampled at the same frequency. For example, gross domestic product (GDP) data are available quarterly, industrial production data are published monthly, and most interest rate data are available daily. Analyzing such data jointly is referred to as a mixed-frequency data analysis. In the econometrics literature, Ghysels, Santa-Clara and Valkanov (2004) proposed a mixed-data sampling (MIDAS) approach to analyze such data. In particular, they use newly available high-frequency data to improve the prediction of a lower-frequency macroeconomic variable of interest, and refer to such predictions as *nowcasting*. Consider, for example, the problem of predicting the quarterly GDP growth rate y_{n+1} at the forecast origin $i = n$. Here, the time interval is a quarter. Traditional forecasting methods employ quarterly data available at $i = n$ to build a model, after which, they use the fitted model to perform a prediction. In practice, some monthly and daily data become available during the quarter $i = n + 1$. Nowcasting uses newly available monthly and daily data to update its prediction of y_{n+1}. Therefore, the term nowcasting means taking advantage of high-frequency data within a given quarter to update the predictions of GDP growth rate of that quarter. In short, the basic principle of nowcasting is the exploitation of information published at higher frequencies than the target variable of interest in order to obtain an improved prediction before the official lower-frequency data become available. Because high-frequency data are relatively common in practice, employing many covariates is common in nowcasting. Therefore, Model (1.1), with dependent covariates and errors, is applicable to nowcasting, and the Lasso method is highly relevant. The mixed-data sampling approach of Ghysels, Santa-Clara and Valkanov (2004) has proven useful for various forecasting and nowcasting purposes. We compare the performance of the Lasso regression with that of the MIDAS regression and the autoregressive model with exogenous variables (ARX). To the best of our knowledge, this is the first study to apply a Lasso regression to now-

casting. Simulation studies and empirical studies show that the Lasso estimator outperforms the existing MIDAS regression and ARX model.

The rest of the paper is organized as follows. Section 2 defines the high-dimensional dependence measure, adopting the concept of Wu (2005). Section 3 deals with rates of convergence of Lasso estimators. The model selection consistency of Lasso estimators is given in Section 4, and simulation studies are carried out in Section 5. Section 6 considers real-data examples, including forecasting and nowcasting applications.

We begin with some basic definitions. Throughout the paper, for a matrix $A = (a_{ij}) \in \mathbb{R}^{p\times p}$, define the spectral norm $\rho(A) = \sup_{|x|\le 1} |Ax|_2$, the Frobenius norm $|A|_F = (\sum_{ij} a_{ij}^2)^{1/2}$, and the infinity norm $|A|_\infty = \max_{1\le i,j\le p} |a_{ij}|$. For a vector $a = (a_1, \ldots, a_p)^T \in \mathbb{R}^p$, define the vector q norm $|a|_q = (\sum_{i=1}^p |a_i|^q)^{1/q}$, for $1 \le q < \infty$. Let $|a|_\infty = \max_{1\le i\le p} |a_i|$ and $|a|_0 = \#\{i : a_i \ne 0\}$. For a random variable $\xi \in \mathcal{L}^k$, denote the q-norm by $\|\xi\|_q = (\mathsf{E}|\xi|^q)^{1/q}$, for $1 \le q \le k$. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, write $a_n = O(b_n)$ if there exists a constant C such that $|a_n| \le C|b_n|$ holds for all sufficiently large n, write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$, and write $a_n \asymp b_n$ if there are positive constants c and C, such that $c \le a_n/b_n \le C$ for all sufficiently large n. Denote $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

2. High-Dimensional Time Series

Let ε_i, for $i \in \mathbb{Z}$, be i.i.d. random vectors and the σ-field $\mathcal{F}_i = (\ldots, \varepsilon_{i-1}, \varepsilon_i)$. In our random-design setting, we assume that in Model (1.1), the covariate process $(\mathbf{x}_i, i = 1, \ldots, n)$ is high-dimensional and weakly stationary, and of the form

$$\mathbf{x}_i = (g_1(\mathcal{F}_i), \ldots, g_p(\mathcal{F}_i))^T, \tag{2.1}$$

and the error e_i satisfies

$$e_i = g_e(\mathcal{F}_i), \tag{2.2}$$

where $g_1(\cdot), \ldots, g_p(\cdot)$ and $g_e(\cdot)$ are measurable functions in $\mathbb{R}$, such that $\mathbf{x}_i$ is well defined. In the scalar case with $p = 1$, (2.1) and (2.2) include a very general class of stationary processes (see Wiener (1958), Rosenblatt (1971), Priestley (1988), Tong (1990), Tsay (2005), Wu (2005)). They also allow models with homogeneous or heteroscedastic errors; see Example 1 of Section 3. In the homogeneous case, the covariate process $(\mathbf{x}_i)$ and the errors (e_i) can be independent of each other.

Following Wu (2005), we define the functional dependence measure

$$\delta_{i,q,j} = \|x_{ij} - x_{ij}^*\|_q = \|g_j(\mathcal{F}_i) - g_j(\mathcal{F}_i^*)\|_q, \tag{2.3}$$

$$\delta_{i,q,e} = \|e_i - e_i^*\|_q = \|g_e(\mathcal{F}_i) - g_e(\mathcal{F}_i^*)\|_q, \tag{2.4}$$

where the coupled process $x_{ij}^* = g_j(\mathcal{F}_i^*)$ and $e_i^* = g_e(\mathcal{F}_i^*)$. Here $\mathcal{F}_i^* = (\ldots, \varepsilon_{-1}, \varepsilon_0', \varepsilon_1, \ldots, \varepsilon_{i-1}, \varepsilon_i)$ and $\varepsilon_0', \varepsilon_l$, for $l \in \mathbb{Z}$, are i.i.d. random variables. We assume short-range dependence, such that

$$\Delta_{m,q,j} := \sum_{i=m}^{\infty} \delta_{i,q,j} < \infty, \tag{2.5}$$

$$\Delta_{m,q,e} := \sum_{i=m}^{\infty} \delta_{i,q,e} < \infty. \tag{2.6}$$

Then, for fixed m, $\Delta_{m,q,j}$, and $\Delta_{m,q,e}$, measure the cumulative effect of ε_0 on $(x_{ij})_{i\geq m}$ and $(e_i)_{i\geq m}$. We introduce the following dependence-adjusted norm (DAN):

$$\|x_{\cdot j}\|_{q,\alpha} = \sup_{m\geq 0}(m+1)^{\alpha}\Delta_{m,q,j}, \quad \alpha \geq 0. \tag{2.7}$$

$$\|e_{\cdot}\|_{q,\alpha} = \sup_{m\geq 0}(m+1)^{\alpha}\Delta_{m,q,e}, \quad \alpha \geq 0. \tag{2.8}$$

It can happen that, owing to the dependence, $\|e_{\cdot}\|_{q,\alpha} = \infty$, while $\|e_i\|_q < \infty$. Because $e_0 = \sum_{l=-\infty}^{0}(\mathsf{E}(e_0|\mathcal{F}_l) - \mathsf{E}(e_0|\mathcal{F}_{l-1}))$, we have

$$\begin{aligned} \|e_0\|_q &\leq \sum_{l=0}^{\infty} \|\mathsf{E}(e_0|\mathcal{F}_{-l}) - \mathsf{E}(e_0|\mathcal{F}_{-l-1})\|_q = \sum_{l=0}^{\infty} \|\mathsf{E}(e_l - e_l^*|\mathcal{F}_0)\|_q \\ &\leq \sum_{l=0}^{\infty} \|e_l - e_l^*\|_q = \|e_{\cdot}\|_{q,0}, \end{aligned} \tag{2.9}$$

by stationarity. If e_i, for $i \in \mathbb{Z}$, are i.i.d., the DAN $\|e_{\cdot}\|_{q,\alpha}$ and the $\mathcal{L}^q$ norm $\|e_0\|_q$ are equivalent, in the sense that $\|e_0\|_q \leq \|e_{\cdot}\|_{q,\alpha} \leq 2\|e_0\|_q$.

To account for the cross-sectional dependence of the p-dimensional stationary process $(\mathbf{x}_i)$, we define the $\mathcal{L}^{\infty}$ functional dependence measure and its corresponding DAN (see Chen, Xu and Wu (2013), Zhang and Wu (2017)), as follows:

$$\omega_{i,q} = \| \max_{1\leq j\leq p} |x_{ij} - x_{ij}^*| \|_q,$$

$$\||\mathbf{x}_\cdot|_\infty\|_{q,\alpha} = \sup_{m\geq 0}(m+1)^\alpha \Omega_{m,q}, \quad \alpha \geq 0, \quad \text{and } \Omega_{m,q} = \sum_{i=m}^{\infty} \omega_{i,q}.$$

Additionally, we define

$$\Psi_{q,\alpha} = \max_{1\leq j\leq p} \|x_{\cdot j}\|_{q,\alpha} \quad \text{and} \quad \Upsilon_{q,\alpha} = \left(\sum_{j=1}^{p} \|x_{\cdot j}\|_{q,\alpha}^q\right)^{1/q},$$

where $\Psi_{q,\alpha}$ and $\Upsilon_{q,\alpha}$ can be viewed as the uniform and the overall DANs of $(\mathbf{x}_i)$, respectively. Clearly, $\Psi_{q,\alpha} \leq \||\mathbf{x}_\cdot|_\infty\|_{q,\alpha} \leq \Upsilon_{q,\alpha}$.

Next, we provide an example of high-dimensional time series to illustrate the univariate and multivariate DAN scale.

Example 1. Let ε_{ij}, for $i, j \in \mathbb{Z}$, be i.i.d. random variables with mean zero and variance one, and with finite qth moments, $q > 2$. Furthermore, let A_i, for $i \geq 0$, be $p \times d$ coefficient matrices with real entries, such that $\sum_{i=0}^{\infty} \operatorname{tr}(A_i A_i^T) < \infty$. Write $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{id})^T$. Then, by Kolmogorov's three-series theorem, the linear process

$$\mathbf{x}_i = \sum_{l=0}^{\infty} A_l \varepsilon_{i-l} \tag{2.10}$$

exists. Denote $A_l = (a_{l;jk})_{1\leq j\leq p, 1\leq k\leq d}$, and $A_{l,j\cdot}$ is the jth row of A_l. By Burkholder's inequality, $\|A_{l,j\cdot}\varepsilon_0\|_q \leq \sqrt{q-1}|A_{l,j\cdot}|_2 \|\varepsilon_{00}\|_q$. We assume that the linear process satisfies the decay condition

$$\max_{j\leq p} |A_{l,j\cdot}|_2 \leq K_1 (1 \vee l)^{-\theta}, \tag{2.11}$$

for all $l \geq 0$, where $\theta > 1/2$ and $K_1 > 0$. If $\theta > 1$, (2.11) implies short-range dependence (SRD) because the auto-covariance matrices $\Sigma_k = \sum_{l=0}^{\infty} A_l A_{l+k}^T$ are absolutely summable. On the other hand, if $1 > \theta > 1/2$, then $(\mathbf{x}_i)$ in (2.10) may not have summable auto-covariance matrices, thus allowing for long-range dependence (LRD). The classical literature on LRD focuses primarily on the univariate case, $p = 1$. Then, under the SRD case, the DANs have the following bounds:

$$\Psi_{q,\alpha} = \max_{1\leq j\leq p} \|x_{\cdot j}\|_{q,\alpha}$$

$$= \max_j \sup_{m\geq 0}(m+1)^\alpha \sum_{i=m}^{\infty} \|A_{i,j.}\varepsilon_0\|_q \leq K_1K_2\|\varepsilon_{00}\|_q, \tag{2.12}$$

$$\||\mathbf{x}_.|_\infty\|_{q,\alpha} = \sup_{m\geq 0}(m+1)^\alpha \sum_{i=m}^{\infty} \|\max_j |A_{i,j.}\varepsilon_0|\|_q \leq K_1K_2p^{1/q}\|\varepsilon_{00}\|_q, \tag{2.13}$$

where $\alpha = \theta - 1$ and the constant K_2 depends only on θ and q.

In this paper, we use the DANs $\||\mathbf{x}_.|_\infty\|_{q,\alpha}$, $\Psi_{q,\alpha}$, and $\Upsilon_{q,\alpha}$ to study the limiting properties of Lasso estimators in the presence of serial dependence. These adjusted norms are more convenient than the commonly used mixing conditions for handling serial dependence in high-dimensional time series.

3. Convergence Rate of the Lasso Estimator

In this section, we present the main results on convergence rate of the Lasso estimator for dependent data. In the low-dimensional case, the consistency of $\hat{\beta}$ relies on the assumption that the sample covariance matrix converges to the population covariance matrix. In the high-dimensional case ($n \ll p$), it requires that $|X(\hat{\beta} - \beta)|_2$ is small only when $|\hat{\beta} - \beta|_2$ is small. Let $\hat{\Sigma} = (\hat{\sigma}_{jk})_{1\leq j,k\leq p} = n^{-1}\sum_{i=1}^n x_i x_i^T$ be the sample covariance. Typically, researchers assume with high probability that the following restricted strong convexity condition holds:

$$u'\hat{\Sigma}u \geq \kappa_1|u|_2^2 - \kappa_2 g(n,p)|u|_1^2, \tag{3.1}$$

for all $u \in \mathbb{R}^p$, where κ_1, κ_2 are positive constants, and $g(n,p)$ is a function of the sample size n and the ambient dimension p. This can be viewed as an analogous sufficient condition in the high-dimensional case. As shown in the proof of Theorem 1, the restricted strong convexity condition for the sample covariance matrix holds with high probability under certain conditions.

To establish our theoretical results, we first impose a weak sparsity condition.

Assumption 1. *There exists some $0 \leq \theta < 1$, with a uniform radius K_θ, such that*

$$\sum_{j=1}^p |\beta_j|^\theta \leq K_\theta. \tag{3.2}$$

The following theorem shows that the L_2 and L_1 convergence rates of $\hat{\beta}$ to β depend on the moment condition and on the temporal and cross-sectional dependence conditions.

Theorem 1. *Denote the population covariance matrix by* $\Sigma = (\sigma_{jk}) = [Cov(x_{ij}, x_{ik})]$. *Suppose the minimum eigenvalue of* Σ *satisfies* $\lambda_{\min}(\Sigma) \geq \kappa > 0$. *Assume that* $\Psi_{\gamma,\alpha_X} = \max_j \|x_{\cdot j}\|_{\gamma,\alpha_X} = M_X < \infty$, *and* $\|e_{\cdot}\|_{q,\alpha_e} = M_e < \infty$, *where* $q > 2, \gamma > 4$ *and* $\alpha_X, \alpha_e > 0$. *Define*

$$\nu = \begin{cases} 1 & \text{if } \alpha_X \geq \frac{1}{2} - \frac{2}{\gamma}, \\ \frac{\gamma}{4} - \alpha_X \frac{\gamma}{2} & \text{if } \alpha_X < \frac{1}{2} - \frac{2}{\gamma}. \end{cases}$$

Assume $\tau = q\gamma/(q+\gamma) > 2$ *and let* $\alpha = \min(\alpha_X, \alpha_e)$. *Define*

$$\rho = \begin{cases} 1 & \text{if } \alpha \geq \frac{1}{2} - \frac{1}{\tau}, \\ \frac{\tau}{2} - \alpha\tau & \text{if } \alpha < \frac{1}{2} - \frac{1}{\tau}. \end{cases}$$

Denote $\omega = \sqrt{\log p/n} M_X^2 + n^{2\nu/\gamma - 1} (\log p)^{3/2} \| |\mathbf{x}_{\cdot}|_\infty \|_{\gamma,\alpha_X}^2$. *Suppose Assumption 1 holds. Then, for any* λ *such that*

$$\lambda \gtrsim \sqrt{\log p/n} M_e M_X + n^{\rho/\tau - 1} (\log p)^{3/2} M_e \| |\mathbf{x}_{\cdot}|_\infty \|_{\gamma,\alpha_X},$$

and $K_\theta \omega \lambda^{-\theta} \leq C$ *for some positive constant* C, *any Lasso solution* $\hat{\beta}$ *satisfies*

$$|\hat{\beta} - \beta|_2 \lesssim \sqrt{K_\theta} \left(\frac{\lambda}{\kappa} \right)^{1 - \theta/2}, \tag{3.3}$$

$$|\hat{\beta} - \beta|_1 \lesssim K_\theta \left(\frac{\lambda}{\kappa} \right)^{1 - \theta}, \tag{3.4}$$

with probability at least $1 - C_1 (\log p)^{-\gamma/2} - C_2 p^{-C_3} - C_4 (\log p)^{-\tau}$, *where* $C_1, \ldots, C_4$ *are positive constants.*

In the special case $\theta = 0$, the quantity of weak sparsity corresponds to an exact sparsity constraint; that is, β has at most $s := K_0$ nonzero entries. The following theorem shows the convergence rate of $\hat{\beta}$ and the prediction error $|X(\hat{\beta} - \beta)|_2^2$ for the exact sparsity case.

Theorem 2. *Suppose the same conditions of Theorem* 1 *hold. If* $|\beta|_0 = s$, $\kappa \asymp 1$, *and*

$$n \gtrsim M_X^4 s^2 \log p + s^{1/(1-2\nu/\gamma)} (\log p)^{3/(2-4\nu/\gamma)} \| |\mathbf{x}_{\cdot}|_\infty \|_{\gamma,\alpha_X}^{2/(1-2\nu/\gamma)},$$

then, for any λ *such that*

$$\lambda \gtrsim \sqrt{\log p/n} M_e M_X + n^{\rho/\tau - 1} (\log p)^{3/2} M_e \| |\mathbf{x}_{\cdot}|_\infty \|_{\gamma,\alpha_X},$$

any Lasso solution $\hat{\beta}$ *satisfies*

$$|\hat{\beta} - \beta|_2 \lesssim \frac{\lambda\sqrt{s}}{\kappa}, \tag{3.5}$$

$$|\hat{\beta} - \beta|_1 \lesssim \frac{\lambda s}{\kappa}, \tag{3.6}$$

$$|X(\hat{\beta} - \beta)|_2^2/n \lesssim \frac{\lambda^2 s}{\kappa}, \tag{3.7}$$

with probability at least $1 - C_1(\log p)^{-\gamma/2} - C_2 p^{-C_3} - C_4(\log p)^{-\tau}$.

Remark 1. In the exact sparsity case, instead of the condition $\lambda_{\min}(\Sigma) \geq \kappa > 0$, we may require that the restricted eigenvalue assumption RE(s,3) of Bickel, Ritov and Tsybakov (2009) holds for the population covariance matrix Σ; that is

$$\kappa := \min_{J \subseteq \{1,\ldots,p\}, |J|_0 \leq s} \min_{u \neq 0, |u_{J^c}|_1 \leq 3|u_J|_1} u'\Sigma u/|u|_2^2 > 0, \tag{3.8}$$

where J^c is the complement of the set J, that is, $J^c = \{1, 2, \ldots, p\} \backslash J$, and u_J is defined as a modification of u by setting its elements outside J to zero. All bounds (3.5), (3.6), and (3.7) still hold with high probability.

Remark 2. The best known convergence rate of Lasso estimators for i.i.d. sub-Gaussian data requires that $K_\theta(\log p/n)^{1-\theta/2} \leq C$, for some positive constant C. Our theorems require that $K_\theta \omega \lambda^{-\theta} \leq C$, where

$$\omega = \sqrt{\frac{\log p}{n}} M_X^2 + n^{2\nu/\gamma - 1} (\log p)^{3/2} \| |\mathbf{x}_\cdot|_\infty \|_{\gamma, \alpha_X}^2,$$

and

$$\lambda \gtrsim \sqrt{\frac{\log p}{n}} M_e M_X + n^{\rho/\tau - 1} (\log p)^{3/2} M_e \| |\mathbf{x}_\cdot|_\infty \|_{\gamma, \alpha_X}.$$

The second terms in ω and λ are introduced by the heavy tails, and thus are unavoidable. In other words, under heavy-tailed distributions in some cases, the allowed dimension p for Lasso methods can be at most a power of the sample size n.

In the exact sparsity case, we require $n \gtrsim M_X^4 s^2 \log p + s^{1/(1-2\nu/\gamma)} (\log p)^{3/(2-4\nu/\gamma)} \| |\mathbf{x}_\cdot|_\infty \|_{\gamma, \alpha_X}^{2/(1-2\nu/\gamma)}$. One may argue that the first term $M_X^4 s^2 \log p$ can be further improved to $M_X^4 s \log p$ for short-range temporal dependence data, in accordance with i.i.d. subGaussian data. However, we cannot achieve this because even the optimal Bernstein-type inequality for nonlinear weakly dependent

data is still an open problem. The best known result is proposed by Merlevède, Peligrad and Rio (2009).

Remark 3. Based on Theorem 2, we have the following cases. Assume $M_X \asymp 1$ and $M_e \asymp 1$. Under the weak cross-sectional dependence $\||\mathbf{x}_{\cdot}|_\infty\|_{\gamma,\alpha_X} \asymp p^{1/\gamma}$, which holds if the p components x_{ij} $(1 \le j \le p)$ are nearly independent, the required sample size for exact sparsity is $n \gtrsim s^2 \log p + s^{1/(1-2\nu/\gamma)}(\log p)^{3/(2-4\nu/\gamma)}$ $p^{2/(\gamma-2\nu)}$, and the regularization parameter satisfies $\lambda \gtrsim \sqrt{\log p/n} + n^{\rho/\tau-1}$ $(\log p)^{3/2}p^{1/\gamma}$. By comparison, the Bonferroni inequality and Lemma 1 in the Appendix yield $n \gtrsim s^2 \log p + s^{1/(1-2\nu/\gamma)}p^{4/(\gamma-2\nu)}$ and $\lambda \gtrsim \sqrt{\log p/n} + n^{\rho/\tau-1}p^{1/\tau}$, respectively.

In addition, under the strong cross-sectional dependence $\||\mathbf{x}_{\cdot}|_\infty\|_{\gamma,\alpha_X} \asymp 1$, which holds if the p components x_{ij} $(1 \le j \le p)$ are linear combinations of fixed random variables, the required sample size for exact sparsity is $n \gtrsim$ $s^2 \log p + s^{1/(1-2\nu/\gamma)}(\log p)^{3/(2-4\nu/\gamma)}$, and the regularization parameter satisfies $\lambda \gtrsim \sqrt{\log p/n} + n^{\rho/\tau-1}(\log p)^{3/2}$.

Next, we apply the results of Theorem 1 in an example.

Example 2. Consider the autoregressive model with exogenous variables, that is, the ARX(a, b) model:

$$y_i = \sum_{l=1}^{a} \phi_l y_{i-l} + \sum_{l=0}^{b} \psi_l' \mathbf{z}_{i-l} + e_i = \beta' \mathbf{x}_i + e_i, \tag{3.9}$$

where a and b are nonnegative integers, e_i follows a GARCH(1,1) model defined below, and $\mathbf{z}_i$ is a linear process defined by

$$\mathbf{z}_i = \sum_{l=0}^{\infty} A_l \varepsilon_{i-l}, \tag{3.10}$$

where the random variables ε_{ij} and coefficient matrices A_l are given in Example 1, with $E|\varepsilon_{ij}|^\gamma < \infty$ and $\gamma > 2$. Assume the roots of the polynomial $1 - \sum_{l=1}^{a} \phi_l B^l$ are outside the unit circle, which ensures the stationarity of the autoregressive part of the model. In addition, assume the population covariance matrix $\Sigma = \mathsf{E}\mathbf{x}_i\mathbf{x}_i'$ is positive definite.

Let

$$e_i = \sqrt{h_i}\eta_i, \qquad h_i = \pi_0 + \pi_1 e_{i-1}^2 + \pi_2 h_{i-1}, \tag{3.11}$$

with $\pi_0 > 0$, $\pi_1 \ge 0$, $\pi_2 \ge 0$, and $\mathsf{E}(\pi_1 + \pi_2\eta_{i-1}^2)^{q/2} < \infty$, $q > 4$. Then, it is easy

to show that $\|e_{\cdot}\|_{q,\alpha_e} < \infty$.

Again, by Burkholder's inequality, $\|A_{l,j\cdot}\varepsilon_0\|_\gamma \le \sqrt{\gamma-1}|A_{l,j\cdot}|_2\|\varepsilon_{00}\|_\gamma$. If there exist constants $K_1 > 1$ and $\alpha_Z > 0$, such that $\max_{j\le p}|A_{l,j\cdot}|_2 \le K_1(l+1)^{-1-\alpha_Z}$ holds for all $l \ge 0$, then we have $\max_j \|z_{\cdot j}\|_{\gamma,\alpha_Z} \le K_1K_2\|\varepsilon_{00}\|_\gamma$, where the constant K_2 depends only on α_Z and γ. Together with the assumption that the roots of the polynomial $1-\sum_{l=1}^a \phi_l B^l$ are outside the unit circle, we ensure $\max_j \|x_{\cdot j}\|_{\gamma,\alpha_Z} < \infty$.

4. Model Selection Consistency

In this section, we extend the asymptotic properties of the sign consistency for model selection, using the Lasso, to the dependent setting. The sign consistency of the Lasso was first introduced by Zhao and Yu (2006). Without loss of generality, write $\beta = (\beta_1,\dots,\beta_s,\dots,\beta_p)'$, where $\beta_j \ne 0$ if $j \le s$, and $\beta_j = 0$ if $j > s$. That is, the first s predictors are relevant variables. Denote $\beta = (\beta_{(1)}', \beta_{(2)}')'$, where $\beta_{(1)}$ is an $s\times 1$ vector. Correspondingly, for any i, denote $\mathbf{x}_i = (\mathbf{x}_{i(1)}', \mathbf{x}_{i(2)}')'$ and $X = (\mathbf{x}_1,\dots,\mathbf{x}_n)' = (X_{(1)}, X_{(2)})$, where $X_{(1)}$ is an $n\times s$ sub-matrix of relevant variables, and $X_{(2)}$ is an $n\times(p-s)$ sub-matrix of irrelevant variables. Similarly, consider the partition of the covariance matrix as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where $\Sigma_{11} = \mathsf{E}\mathbf{x}_{i(1)}\mathbf{x}_{i(1)}'$ is an $s\times s$ sub-matrix associated with the relevant variables.

We impose the following assumptions.

Assumption 2. *For any $1 \le i \le n$, $\mathsf{E}(x_{ik}|X_{(1)}, e) = \Sigma_{2k,1}\Sigma_{11}^{-1}x_{i(1)}$, where $\Sigma_{2k,1}$ is the kth row of Σ_{21}.*

Define $z_{ik} = x_{ik} - \mathsf{E}(x_{ik}|X_{(1)}, e)$, for $s+1 \le k \le p$, and $\mathbf{z}_i = (z_{i,s+1},\dots,z_{i,p})'$.

Assumption 3. *There exists $L > 0$, such that $\min_{1\le j\le s}|\beta_j| \ge L$.*

Assumption 4. *There exists a constant $N_1 > 0$, such that*

$$\inf_{|\zeta|_2=1} \zeta'\Sigma_{11}\zeta = N_1.$$

Assumption 5. *There exists a positive constant $\eta \in (0,1)$, such that*

$$|\Sigma_{21}\Sigma_{11}^{-1}sign(\beta_{(1)})|_\infty \le 1-\eta. \tag{4.1}$$

Assumption 2 explicitly defines how the irrelevant variables depend on the relevant variables and the errors. Note that $\mathsf{Cov}(\Sigma_{2k,1}\Sigma_{11}^{-1}x_{i(1)}, x_{ik}-\Sigma_{2k,1}\Sigma_{11}^{-1}x_{i(1)}) = 0$ always holds, for all $s+1 \le k \le p$. That is, $\Sigma_{2k,1}\Sigma_{11}^{-1}x_{i(1)}$ and $x_{ik} - \Sigma_{2k,1}\Sigma_{11}^{-1}x_{i(1)}$ are mutually uncorrelated. We further assume they are independent. Intuitively, $\mathbf{z}_i$ can be viewed as the unique part of irrelevant variables that cannot be explained by the relevant variables. Thus, for irrelevant variables, $\mathbf{z}_i$ is more representative than $\mathbf{x}_{i(2)}$. Assumption 3 controls the lower bound of the nonzero parameters; see, for example, Bühlmann and Van De Geer (2011). Assumption 4 imposes a lower bound, N_1, on the minimal eigenvalue of the covariance matrix of relevant variables. In practice, quantifying the rate under which N_1 decreases is difficult and problem specific, and it is frequently assumed to be constant; see, for example, Medeiros and Mendes (2016) and Kock and Callot (2015). Assumption 5 employs the strong irrepresentable condition of population covariance, which is similar to the condition in Zhao and Yu (2006).

To account for the cross-sectional dependence of the stationary process $(\mathbf{x}_{i(1)})$ and $(\mathbf{z}_i)$, we also define the $\mathcal{L}^\infty$ functional dependence measure and its corresponding DAN, as follows:

$$\begin{aligned}\omega_{i,q,1} &= \| \max_{1\le j\le s} |x_{ij} - x_{ij}^*| \|_q,\\ \| |\mathbf{x}_{\cdot(1)}|_\infty \|_{q,\alpha} &= \sup_{m\ge 0}(m+1)^\alpha \Omega_{m,q,1}, \quad \alpha \ge 0, \quad \text{and } \Omega_{m,q,1} = \sum_{i=m}^{\infty}\omega_{i,q,1}.\end{aligned}$$

Additionally, we define

$$\Psi_{q,\alpha,1} = \max_{1\le j\le s}\|x_{\cdot j}\|_{q,\alpha} \quad \text{and} \quad \Upsilon_{q,\alpha,1} = \left(\sum_{j=1}^{s}\|x_{\cdot j}\|_{q,\alpha}^q\right)^{1/q}.$$

For $(\mathbf{z}_i)$, the quantities $\| |\mathbf{z}_\cdot|_\infty \|_{q,\alpha}$, $\Psi_{q,\alpha,2}$, and $\Upsilon_{q,\alpha,2}$ can be similarly defined. Clearly, $\Psi_{q,\alpha,1} \le \| |\mathbf{x}_{\cdot(1)}|_\infty \|_{q,\alpha} \le \Upsilon_{q,\alpha,1}$ and $\Psi_{q,\alpha,2} \le \| |\mathbf{z}_\cdot|_\infty \|_{q,\alpha} \le \Upsilon_{q,\alpha,2}$.

Let $\sigma = \mathsf{E}e_i^2$. Define

$$\begin{aligned}\delta_*(\lambda, N_1, \sigma) &= \frac{\lambda^2 s}{2nN_1} + \frac{2\sigma}{n},\\ M(\delta_*, \eta, \iota, \gamma) &= \eta^{-1}\sqrt{\delta_* \log p} + \eta^{-1} n^{(\iota-1)/\gamma}\delta_*^{1/2}(\log p)^{3/2}\| |\mathbf{z}_\cdot|_\infty \|_{\gamma,\alpha_X},\\ Q(\rho,\tau) &= \sqrt{n\log s} + n^{\rho/\tau}(\log s)^{3/2}\| |\mathbf{x}_{\cdot(1)}|_\infty \|_{\gamma,\alpha_X},\\ V_1(N_1) &= \frac{s^2 \log s}{N_1},\end{aligned}$$

$$V_2(N_1) = \frac{1}{N_1} s(\log s)^{3/2} \| |\mathbf{x}_{\cdot(1)}|_\infty \|^2_{\gamma,\alpha_X}.$$

These quantities are used in the following theorem.

Theorem 3 extends the results of Zhao and Yu (2006) to a random-design linear model with dependent errors. Medeiros and Mendes (2016) derived the asymptotic properties of sign consistency for the adaptive Lasso. In contrast, our results apply to the original Lasso, and do not need any assumptions on the weights. Note that even for heavy-tail variables, our results show that if the dependence among $\mathbf{z}_i$ is strong, the allowed dimension p can be as large as some exponential of the sample size n; see Remark 4 for more details.

Theorem 3. *Suppose Assumptions* 2, 3, 4, *and* 5 *hold. Assume that* $\max_{1\le j\le p} \|x_{\cdot j}\|_{\gamma,\alpha_X} < C_\gamma < \infty$ *and* $\|e_\cdot\|_{q,\alpha_e} < C_q < \infty$, *where* $q, \gamma > 4$, $\alpha_X, \alpha_e > 0$, *and constants* C_γ, C_q *depend only on* γ, q. *Define*

$$\nu = \begin{cases} 1 & \text{if } \alpha_X > \frac{1}{2} - \frac{2}{\gamma}, \\ \frac{\gamma}{4} - \alpha_X \frac{\gamma}{2} & \text{if } \alpha_X < \frac{1}{2} - \frac{2}{\gamma}, \end{cases}$$

and

$$\iota = \begin{cases} 1 & \text{if } \alpha_X > \frac{1}{2} - \frac{1}{\gamma}, \\ \frac{\gamma}{2} - \alpha_X \gamma & \text{if } \alpha_X < \frac{1}{2} - \frac{1}{\gamma}. \end{cases}$$

Let $\alpha = \min(\alpha_X, \alpha_e)$. *Assume* $\tau = q\gamma/(q+\gamma) > 2$, *and define*

$$\rho = \begin{cases} 1 & \text{if } \alpha > \frac{1}{2} - \frac{1}{\tau}, \\ \frac{\tau}{2} - \alpha\tau & \text{if } \alpha < \frac{1}{2} - \frac{1}{\tau}. \end{cases}$$

Furthermore, suppose $s = o(n)$. *Then, for any* λ *and sample size* n, *such that*

$$n \gtrsim V_1(N_1), \tag{4.2}$$

$$n^{1-2\nu/\gamma} \gtrsim V_2(N_2), \tag{4.3}$$

$$M(\delta_*, \eta, \iota, \gamma) + Q(\rho, \tau) \lesssim \lambda \le \frac{nN_1 L}{4\sqrt{s}}, \tag{4.4}$$

the consistency probability $\mathsf{P}(\hat{\beta} =_s \beta)$ *is at least*

$$\begin{aligned} 1 \; & - C_1(\log p)^{-\gamma} - C_2(\log s)^{-\gamma/2} - C_3(\log s)^{-\tau} \\ & - C_4 p^{-C_5} - C_6 s^{-C_7} - \frac{\|e_\cdot\|^q_{q,\alpha_e}}{n^{q-1}\sigma^q} - \exp\left(-\frac{n\sigma^2}{\|e_\cdot\|^2_{2,\alpha_e}}\right). \end{aligned} \tag{4.5}$$

Remark 4. In particular, assume $N_1 \asymp 1$, $\eta \asymp 1$. In addition, assume the weak temporal dependence case $\alpha_X > 1/2 - 1/\gamma$ and $\alpha > 1/2 - 1/\tau$. If the dependence measure $\||\mathbf{x}_{\cdot(1)}|_\infty\|_{\gamma,\alpha_X} \asymp s^{1/\gamma}$ and $\||\mathbf{z}_\cdot|_\infty\|_{\gamma,\alpha_X} \asymp p^{1/\gamma}$, which hold if all components x_{ij} $(1 \le j \le s)$ and z_{ik} $(s+1 \le k \le p)$ are nearly independent, then (4.2), (4.3), and (4.4) reduce to

$$n \gtrsim s^2 \log s + s^{(1+2/\gamma)/(1-2/\gamma)} (\log s)^{3/(2-4/\gamma)} + s p^{2/\gamma} (\log p)^3$$

and

$$\sqrt{n \log s} + n^{1/\tau} s^{1/\tau} (\log s)^{3/2} \lesssim \lambda \lesssim \frac{nL}{\sqrt{s}}.$$

Additionally, if $s = O(n^{c_1})$, for some $c_1 < \min\{1/2, (\gamma-2)/(\gamma+2)\}$, then the valid regularization parameter λ has range $n^{1/2} + n^{1/\tau + c_1/\gamma} \ll \lambda \ll n^{1-c_1/2} L$. The dimension p satisfies that $p \ll n^{\gamma(1-c_1)/2}$.

On the other hand, assume $\||\mathbf{x}_{\cdot(1)}|_\infty\|_{\gamma,\alpha_X} \asymp s^{1/\gamma}$ and $\||\mathbf{z}_\cdot|_\infty\|_{\gamma,\alpha_X} \asymp 1$; that is, all components z_{ik} $(s+1 \le k \le p)$ are strongly dependent. Let $s = O(n^{c_1})$, for some $c_1 < \min\{1/2, (\gamma-2)/(\gamma+2)\}$. Then the existence of regularization parameter λ requires $n^{1/2} + n^{1/\tau + c_1/\gamma} \ll \lambda \ll n^{1-c_1/2} L$. The dimension p satisfies $p \ll \exp\{n^{(1-c_1)/3}\}$.

Furthermore, if $\||\mathbf{x}_{\cdot(1)}|_\infty\|_{\gamma,\alpha_X} \asymp 1$ and $\||\mathbf{z}_\cdot|_\infty\|_{\gamma,\alpha_X} \asymp 1$, and $s = O(n^{c_1})$ for some $c_1 < 1/2$, then the existence of regularization parameter λ requires $n^{1/2} \ll \lambda \ll n^{1-c_1/2} L$, and the dimension p satisfies $p \ll \exp\{n^{(1-c_1)/3}\}$.

In summary, the allowed dimension p varies from $n^{\gamma(1-c_1)/2}$ to $\exp\{n^{(1-c_1)/3}\}$, depending on the cross-sectional dependence of $z_{ik}, s+1 \le k \le p$.

Note that if the assumptions in Example 2 hold, then together with the strong irrepresentable condition, the results of Theorem 3 continue to apply. In general, the strong irrepresentable condition is nontrivial, particularly because we do not know $\operatorname{sign}(\beta)$ a priori. Then, we need the strong irrepresentable condition to hold for every possible combination of signs and placement of zeros. We give a simple example below in which the strong irrepresentable condition is guaranteed. All diagonal elements of Σ are assumed to be one which is equivalent to normalizing all covariates in the model to the same scale, because the strong irrepresentable condition is invariant under any common scaling of Σ.

Example 3. Consider the following autoregressive model with exogenous variables:

$$y_i = \sum_{l=1}^{a} \phi_l y_{i-l} + \psi \mathbf{z}_i + e_i = \beta' \mathbf{x}_i + e_i, \tag{4.6}$$

where a is nonnegative finite integer, $\mathbf{z}_i$ is independent of e_i, and the errors e_i are homogeneous. Assume the roots of the polynomial $1 - \sum_{l=1}^{a} \phi_l B^l$ are outside the unit circle, which ensures the stationarity of the autoregressive part of the model. In addition, assume $\Sigma = \mathsf{E}\mathbf{x}_i\mathbf{x}_i'$ is positive definite.

Furthermore, suppose β has s nonzero entries. Similarly to Corollary 2 in Zhao and Yu (2006), if Σ has ones on the diagonal and the bounded correlation $|\sigma_{jk}| \leq c/(2s-1)$, for a constant $0 < c < 1$, then the strong irrepresentable condition holds. In this case, we need the autocorrelation of y_i to be weak, and all covariates $\mathbf{z}_i$ are slightly correlated.

Remark 5. The Lasso may fail in the presence of strong serial dependence. Consider two scalar Gaussian autoregressive, AR(3), models:

$$y_i = 1.9y_{i-1} - 0.8y_{i-2} - 0.1y_{i-3} + e_i, \tag{4.7}$$

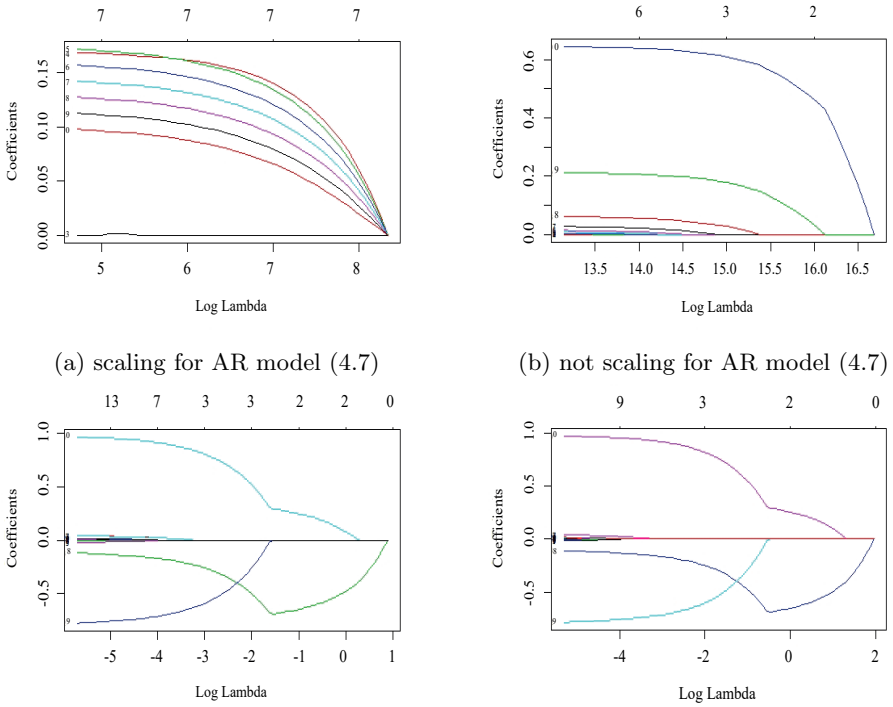
and

$$y_i = y_{i-1} - 0.8y_{i-2} - 0.1y_{i-3} + e_i, \tag{4.8}$$

where e_i follows the standard normal distribution. Then, AR(3) in model (4.7) is unit-root nonstationary, but that in model (4.8) is stationary. We generate 2,000 observations from each of the two models. We choose $y_{i-10}, y_{i-9}, \ldots, y_{i-1}$ and $x_{1i}, \ldots, x_{10,i}$ as regressors, where x_{li} are i.i.d. standard normal. Figure 1 shows the model selection results for scaling versus not scaling the predictors.

The default Lasso procedure standardizes each variable in y_i. For unit-root nonstationary time series, standardization might wash out the dependence of the stationary part; see (a) and (b) of Figure 1. In this paper, we only consider stationary time series for which scaling the predictors does not affect the estimation consistency of the Lasso estimates; see (c) and (d) of Figure 1.

The following proposition shows a necessary and sufficient condition for a stationary AR(2) model, under which the strong irrepresentable condition (Assumption 5) holds. Similar results hold for the general stationary AR(d) model.



(a) scaling for AR model (4.7)

(b) not scaling for AR model (4.7)

(c) scaling for AR model (4.8)

(d) not scaling for AR model (4.8)

Figure 1. Results of Lasso regression for the two AR(3) series in (4.7) and (4.8), created using the glmnet package of R.

Proposition 1. *Consider the stationary AR(2) model,*

$$y_i = \phi_1 y_{i-1} + \phi_2 y_{i-2} + e_i,$$

where e_i are i.i.d. random variates with mean zero and finite variance. We also normalize y_i, such that the variance of y_i is one. Then, the strong irrepresentable condition (Assumption 5) holds if and only if

$$|\phi_1| + |\phi_2| < 1. \tag{4.9}$$

5. Simulation Study

In this section, we use a simulation to demonstrate the performance of the Lasso regression for dependent data in finite samples, and to compare its efficacy with that of the mixed-frequency data sampling regression (MIDAS) commonly used in the econometric literature; see Ghysels, Santa-Clara and Valkanov (2004).

In addition, we compare the model selection consistency and parameter estimation of the Lasso estimator and the Dantzig estimator for dependent data in finite samples. We first consider the following data-generating process:

$$\begin{aligned} y_i &= \phi y_{i-1} + \boldsymbol{x}_{i-1,1}^T \beta_s + e_i, \\ \boldsymbol{x}_i &= \begin{bmatrix} \boldsymbol{x}_{i,1} \\ \boldsymbol{x}_{i,2} \end{bmatrix} = \sum_{j=1}^{m} A_j \begin{bmatrix} \boldsymbol{x}_{i-j,1} \\ \boldsymbol{x}_{i-j,2} \end{bmatrix} + \boldsymbol{\eta}_i, \end{aligned} \tag{5.1}$$

where $\phi = 0.6$, each element of β_s is given by $\beta_{s,j} = (1/\sqrt{s})(-1)^j$, and $\boldsymbol{x}_{i,1}$ is an $s \times 1$ vector of relevant variables. Let $\beta = (\beta_s, \beta_{s^c})$, where $\beta_{s^c} = \mathbf{0}$ is a $(p-s) \times 1$ vector. The errors e_i and $\boldsymbol{\eta}_{ij}$ are i.i.d. random variables from a Student-t distribution with five degrees of freedom, and e_i and $\boldsymbol{\eta}_i$ are all mutually uncorrelated. The explanatory variable process $\boldsymbol{x}_i$, which has $p-s$ irrelevant variables, follows a vector autoregressive, VAR(m), model. The following two choices of $\boldsymbol{x}_i$ are considered, denoted as Model 1 and Model 2, respectively.

(1). **Model 1**: The explanatory process $\boldsymbol{x}_i$ is a VAR(4) process, where A_1 and A_4 assume a block-diagonal structure, and $A_2 = A_3 = 0$. In particular, the first two and the last two blocks are 5×5 matrices, with all entries of the blocks of A_1 equal to 0.15, and all entries of the blocks of A_4 equal to -0.1. The other blocks are 10×10 matrices, with all elements of the blocks of A_1 equal to 0.075, and all elements of the blocks of A_4 equal to -0.05. This structure could be motivated by a model built for mixed-frequency data with some quarterly time series, often encountered in macroeconomic analysis.

(2). **Model 2**: The explanatory process $\boldsymbol{x}_i$ follows a VAR(1) model, where A_1 is block-diagonal, with the same block structure given by Model 1. The (j,k)th entry of the block is $(-1)^{|j-k|}\rho^{|j-k|+1}$, with $\rho = 0.4$. Hence, the entries decrease exponentially fast with their distances from the diagonal.

We employ sample sizes $n = 50, 100, 200$, with different choices of p and s. We set $p = 100, 200, 400$ and $s = 5, 10, 20$. For comparison, we also simulate a response series from a MIDAS model. In Model (5.1), for $s = 5, 10, 20$, let $\beta_s = \beta(1)$, $(\beta(1)^T, \beta(2)^T)^T$ or $(\beta(1)^T, \beta(2)^T, \beta(3)^T)^T$, respectively, with

$$\beta_j(l) = \frac{\exp(\delta_1 j + \delta_2 j^2)}{\sum_{k=1}^{|\beta(l)|_0} \exp(2\delta_1 k + 2\delta_2 k^2)}, \tag{5.2}$$

where $\beta(1)$ and $\beta(2)$ have five variables, $\beta(3)$ has 10 variables, and $\delta = (\delta_1, \delta_2)' = (0.5, -1)'$. All other settings remain the same. The two choices of $\boldsymbol{x}_i$ in Models

1 and 2 are used, and we denote the resulting MIDAS models as Models 3 and 4, respectively. The models estimated by the Lasso have λ selected using the BIC; see Bühlmann and Van De Geer (2011). The consistency of the Lasso estimator selected by the BIC was first proved by Zou, Hastie and Tibshirani (2007) under the case $p < n$. Then, Tibshirani and Taylor (2012) studied the effective degrees of freedom of the Lasso when $p > n$. It would be interesting to investigate the theoretical justification of the consistency of the BIC criterion for the Lasso under the time series setting. We leave this to future work. We also employed models with λ selected using cross-validation. However, cross-validation does not improve the results, and is considerably slower in its computation. For the models estimated by MIDAS, we only consider the exponential Almon lag polynomial weighting scheme (see (5.2)) for the first 100 variables, and impute the true values as initial values.

Table 1 shows the average absolute error (AE) and average root mean squared error (RMSE) for the Lasso estimators and MIDAS estimators over the 10,000 Monte Carlo simulations for the data-generating processes used. The AE and the RMSE are defined as

$$\text{AE} = \frac{1}{MC}\sum_{l=1}^{MC} |(\hat{\phi};\hat{\beta}) - (\phi;\beta)|_1,$$

$$\text{RMSE} = \sqrt{\frac{1}{MC}\sum_{l=1}^{MC} |(\hat{\phi};\hat{\beta}) - (\phi;\beta)|_2^2},$$

where MC denotes the number of Monte Carlo repetitions. From the table, it is clear that both measures show that the Lasso regression provides a substantially more accurate parameter estimation than that of the mixed-frequency data sampling regression (MIDAS) in the presence of irrelevant variables. Furthermore, as expected, the AE and the RMSE of the estimators decrease with n, but increase with s and p.

To evaluate the performance of out-of-sample forecasts, we use the estimated parameters to compute one-step-ahead forecasts, and consider 10 out-of-sample predictions, denoted by $y_{n+1}, \ldots, y_{n+10}$. Table 2 shows the average absolute forecast error (AFE) and average root mean squared forecast error (RMSFE) over the 10,000 Monte Carlo simulations, which are calculated as

$$\text{AFE} = \frac{1}{10MC}\sum_{l=1}^{MC}\sum_{k=1}^{10} |\hat{y}_{n+k} - y_{n+k}|,$$

$$\text{RMSFE} = \sqrt{\frac{1}{10MC}\sum_{l=1}^{MC}\sum_{k=1}^{10}|\hat{y}_{n+k} - y_{n+k}|^2}.$$

The forecasting results in Table 2 show that the Lasso regression has smaller AE and RMSFE than those of the MIDAS in all settings. Furthermore, the results show clearly that the performance of the Lasso regression and the MIDAS improves with the sample size, but deteriorates as the number of relevant variables s increases. Finally, both the AE and the RMSFE of the Lasso regression decrease faster than those of the MIDAS as the sample size n increases. In fact, the AE and RMSFE of the MIDAS remain high even when $n = 200$. Because we only fit the MIDAS through the first 100 variables, its performance does not change as p increases. Overall, in the presence of irrelevant variables, the Lasso regression significantly outperforms the MIDAS regression.

Next, we compare the model selection and parameter estimation of the Lasso estimator and the Dantzig estimator for dependent data. We use the same data-generating process (5.1), where $\phi = 0.6$ and $\boldsymbol{x}_{i,1}$ is an $s \times 1$ vector of relevant variables. Here, we set each element of β_s by $\beta_{s,j} = 3(-1)^j$. Model 1 and Model 2, defined previously, are chosen for $\boldsymbol{x}_i$. Table 3 shows the number of noise covariates that are selected (False Positive), number of signal covariates that are not selected (False Negative), and average root mean squared error (RMSE) for the Lasso estimators and the Dantzig estimators over the 10,000 Monte Carlo simulations for the data-generating processes used. As expected, False Positive and RMSE decrease with n, but increase with s and p. False Negative for the two methods are almost the same. In terms of False Negative and RMSE, the Lasso estimator substantially outperforms the Dantzig selector. The Dantzig selector might be more sensitive to heavy tails and outliers, because it uses the L_∞ norm. The rate of convergence of the Lasso estimator in our study is faster than that of the Dantzig selector in Wu and Wu (2016). They built an L_∞-type rate of convergence for the Dantzig estimator, which is related to the unknown L_1 norm of the true coefficients and the matrix L_1 norm of the population matrix. We overcome this weakness and achieve the same bounds for the Lasso regression under i.i.d. data, but with different requirements for the regularization parameter λ and sample size n.

Table 1. Accuracy in parameter estimation of Lasso regression and mixed-frequency data sampling regression. The results are based on 10,000 repetitions, where AE and RMSE denote the average mean absolute errors and average root mean squared errors over Monte Carlo repetitions and parameters. In the table, s, p, and n denote the number of nonzero parameters, dimension of regressors, and sample size, respectively.

		Absolute Error (AE) $\times 10^2$						Root Mean Square Error (RMSE) $\times 10^2$					
		Lasso			MIDAS			Lasso			MIDAS		
							p						
s	n	100	200	400	100	200	400	100	200	400	100	200	400
Model 1													
	50	2.44	2.64	2.84	6.63	6.64	6.67	3.08	3.75	4.55	3.73	4.44	5.29
5	100	1.89	2.07	2.22	6.24	6.26	6.28	2.79	3.44	4.21	3.64	4.33	5.15
	200	1.27	1.49	1.70	5.91	5.91	5.94	2.28	2.92	3.71	3.56	4.23	5.04
	50	4.60	4.99	5.30	8.26	8.31	8.32	3.66	4.45	5.38	4.09	4.88	5.80
10	100	3.69	4.11	4.39	7.86	7.88	7.90	3.36	4.17	5.10	4.02	4.78	5.69
	200	2.28	2.74	3.29	7.50	7.55	7.56	2.65	3.39	4.42	3.96	4.71	5.60
	50	7.83	8.81	8.93	10.76	10.82	10.83	4.08	5.00	6.00	4.42	5.26	6.26
20	100	6.56	7.33	7.70	10.38	10.43	10.44	3.84	4.75	5.77	4.35	5.18	6.16
	200	4.69	5.55	6.56	10.08	10.12	10.15	3.31	4.21	5.40	4.30	5.12	6.09
Model 2													
	50	0.95	1.14	1.38	4.95	4.97	4.99	2.02	2.53	3.19	3.31	3.94	4.70
5	100	0.54	0.60	0.67	4.55	4.58	4.58	1.56	1.92	2.36	3.18	3.79	4.50
	200	0.34	0.36	0.38	4.20	4.21	4.22	1.26	1.53	1.87	3.06	3.64	4.33
	50	1.91	2.40	2.92	5.46	5.46	5.46	2.54	3.26	4.18	3.53	4.20	4.99
10	100	1.06	1.24	1.46	5.03	5.07	5.08	1.92	2.41	3.04	3.39	4.04	4.81
	200	0.65	0.71	0.79	4.60	4.63	4.65	1.52	1.87	2.31	3.25	3.87	4.61
	50	3.19	4.21	4.94	6.12	6.15	6.18	2.95	3.85	4.96	3.76	4.48	5.34
20	100	1.75	2.14	2.59	5.68	5.69	5.70	2.23	2.85	3.64	3.63	4.32	5.15
	200	1.07	1.21	1.38	5.26	5.27	5.29	1.77	2.20	2.74	3.51	4.18	4.98
Model 3													
	50	1.71	2.05	2.43	6.57	6.60	6.62	2.62	3.29	4.11	3.74	4.46	5.30
5	100	0.93	1.06	1.21	6.27	6.31	6.33	2.03	2.54	3.18	3.65	4.35	5.18
	200	0.57	0.63	0.69	6.17	6.20	6.21	1.62	2.02	2.50	3.61	4.30	5.11
	50	3.74	4.47	5.07	8.41	8.44	8.46	3.34	4.17	5.16	4.13	4.92	5.86
10	100	2.06	2.52	3.00	8.20	8.24	8.25	2.59	3.32	4.25	4.08	4.85	5.77
	200	1.20	1.38	1.58	8.10	8.14	8.16	2.00	2.52	3.18	4.05	4.82	5.73
	50	7.23	8.77	9.38	11.02	11.07	11.09	3.90	4.88	5.95	4.47	5.32	6.32
20	100	4.45	5.81	7.01	10.92	10.97	11.00	3.22	4.16	5.32	4.43	5.28	6.28
	200	2.53	2.93	3.50	10.87	10.93	10.95	2.49	3.11	3.97	4.42	5.26	6.26
Model 4													
	50	1.39	1.58	1.78	5.14	5.16	5.16	2.49	3.10	3.83	3.47	4.13	4.90
5	100	0.96	1.05	1.12	4.58	4.59	4.59	2.12	2.63	3.22	3.31	3.95	4.69
	200	0.71	0.77	0.83	4.22	4.23	4.25	1.83	2.28	2.81	3.23	3.85	4.58
	50	2.40	2.79	3.14	6.03	6.07	6.10	2.90	3.64	4.54	3.80	4.53	5.39
10	100	1.67	1.86	2.02	5.50	5.53	5.55	2.47	3.08	3.80	3.69	4.39	5.23
	200	1.23	1.38	1.50	5.13	5.15	5.16	2.11	2.65	3.30	3.62	4.31	5.13
	50	3.68	4.58	4.97	6.88	6.93	6.93	3.22	4.09	5.14	4.06	4.84	5.75
20	100	2.43	2.77	3.06	6.38	6.42	6.45	2.71	3.41	4.25	3.96	4.72	5.62
	200	1.78	2.00	2.22	6.04	6.08	6.08	2.32	2.91	3.66	3.91	4.66	5.55

Table 2. Performance of Out-of-sample predictions of Lasso regression and mixed frequency data sampling regression (MIDAS). The results are based on 10 one-step ahead predictions and 10,000 iterations, where AFE and RMSFE denote the average absolute forecast errors and root mean squared forecast errors, respectively, and s, p, and n are the number of nonzero parameters, dimension of regressors, and sample size, respectively. For MIDAS, the maximum p is fixed at 100.

		Absolute Error (AE) $\times 10^2$						Root Mean Square Forecast Error (RMSFE) $\times 10^2$					
		Lasso			MIDAS			Lasso			MIDAS		
		p											
s	n	100	200	400	100	200	400	100	200	400	100	200	400
Model 1													
	50	120.0	125.8	130.3	169.2	161.5	162.3	147.2	153.8	158.8	206.0	197.4	198.0
5	100	102.7	106.6	110.7	162.2	156.7	156.4	127.7	132.3	136.9	197.7	191.7	191.3
	200	86.9	90.6	95.4	156.8	152.4	153.2	109.7	114.1	119.4	191.4	186.7	187.4
	50	151.6	159.6	166.4	185.0	178.8	179.9	185.2	194.3	202.0	225.9	218.6	219.9
10	100	125.6	133.9	141.7	177.6	171.6	172.3	155.1	164.9	173.8	216.9	210.1	211.4
	200	96.0	101.9	112.2	171.9	167.7	168.2	120.3	127.2	139.3	210.2	206.1	206.3
	50	177.7	188.8	195.0	205.0	200.0	199.9	216.1	229.2	236.2	250.2	244.5	244.1
20	100	150.2	162.2	170.0	195.8	191.7	191.1	184.0	198.5	207.7	239.5	235.0	234.4
	200	118.6	128.7	145.1	190.1	185.9	188.2	146.7	159.2	178.4	232.4	228.4	230.5
Model 2													
	50	96.4	101.8	107.2	147.3	148.8	148.5	119.7	125.5	131.7	179.9	181.5	180.9
5	100	84.1	85.7	88.1	142.1	142.7	142.9	106.2	108.1	110.4	173.3	174.0	174.0
	200	78.4	79.6	80.6	138.6	137.6	138.8	99.9	101.5	102.3	169.0	168.1	169.2
	50	114.1	125.7	140.0	171.5	164.2	163.9	139.9	153.7	169.8	208.0	199.6	199.5
10	100	90.9	95.3	100.8	156.7	157.9	158.0	114.1	118.7	124.9	190.6	191.9	191.9
	200	81.7	83.1	85.3	151.4	151.1	151.9	103.7	105.1	107.6	184.1	183.7	184.5
	50	126.9	144.5	167.8	178.2	173.1	173.5	155.4	175.9	202.9	216.5	211.1	211.6
20	100	97.7	105.1	113.7	169.7	164.3	164.7	121.6	130.1	139.9	206.5	200.4	200.9
	200	85.3	87.9	91.9	161.7	157.1	158.0	107.8	110.5	115.1	196.9	191.9	192.9
Model 3													
	50	117.4	128.7	140.5	152.9	153.1	153.3	143.0	155.5	168.3	187.5	187.6	187.9
5	100	89.4	92.8	97.1	144.7	145.0	145.0	112.2	116.0	120.4	144.7	178.2	178.1
	200	80.6	81.2	82.8	142.3	141.0	141.1	102.3	103.0	105.0	174.7	173.6	173.5
	50	154.4	172.5	188.9	178.9	179.1	179.8	185.9	206.0	224.6	218.4	218.8	219.7
10	100	103.1	112.9	124.3	171.2	171.3	170.6	127.9	138.7	152.2	209.7	209.5	209.1
	200	84.7	88.2	91.0	166.9	168.2	167.4	107.1	111.1	114.3	204.7	205.9	205.3
	50	197.3	224.5	244.3	206.9	205.0	205.3	236.1	266.5	288.4	251.6	249.7	249.8
20	100	130.9	150.8	172.2	196.6	197.6	196.4	160.2	182.9	207.6	240.2	240.9	239.3
	200	97.0	101.2	109.0	193.1	193.8	193.7	121.2	125.9	134.8	236.5	237.4	237.1
Model 4													
	50	103.0	108.7	113.2	131.7	131.9	130.9	126.8	133.3	138.2	162.6	162.9	161.7
5	100	88.4	90.4	92.9	121.6	122.0	121.5	110.9	113.0	115.8	150.9	151.5	150.8
	200	81.3	82.6	83.4	118.0	117.1	116.8	103.3	104.4	105.4	147.0	145.7	145.2
	50	117.6	126.6	136.2	148.6	148.5	148.5	144.1	154.4	165.7	183.7	183.0	183.0
10	100	95.8	99.8	103.4	139.4	139.5	139.3	119.8	124.2	128.2	172.5	172.6	172.3
	200	84.9	87.5	89.7	134.2	135.0	134.7	107.3	110.1	112.7	166.5	167.3	167.0
	50	132.2	148.5	162.3	163.8	164.7	163.7	161.3	180.2	196.3	201.7	202.7	201.6
20	100	102.4	108.9	115.4	154.2	154.0	154.7	127.2	134.8	142.3	190.6	190.7	190.9
	200	88.6	92.1	96.2	150.0	149.8	150.2	111.7	115.7	120.2	185.8	185.4	185.7

Table 3. Accuracy in model selection and parameter estimation of Lasso estimator and Dantzig estimator for linear regression. The results are based on 10,000 repetitions, where RMSE denote the average root mean squared error over Monte Carlo repetitions and parameters. In the table, s, p, and n denote the number of nonzero parameters, dimension of regressors, and sample size, respectively.

		Model 1						Model 2					
		Lasso			Dantzig			Lasso			Dantzig		
		p											
s	n	100	200	400	100	200	400	100	200	400	100	200	400
False Negative													
	50	0.077	0.20	0.67	0.072	0.28	0.73	0	0	0.003	0	0	0.002
5	100	0	0	0.01	0	0	0.04	0	0	0	0	0	0
	200	0	0	0	0	0	0	0	0	0	0	0	0
	50	0.67	2.07	4.28	0.81	2.50	4.04	0.011	0.14	0.79	0.045	0.17	0.95
10	100	0	0.004	0.11	0	0.006	0.15	0	0	0	0	0	0
	200	0	0	0	0	0	0	0	0	0	0	0	0
	50	4.65	7.45	9.61	4.83	7.29	1 0.1	1.94	5.70	7.18	2.48	5.45	8.34
20	100	0	0.21	2.27	0.03	0.28	2.14	0	0	0.029	0	0.002	0.052
	200	0	0	0	0	0	0	0	0	0	0	0	0
False Positive													
	50	10.9	15.0	22.5	15.85	24.20	32.0	5.25	8.01	11.5	6.45	13.3	21.7
5	100	5.30	8.97	13.6	8.03	14.11	19.8	2.05	3.02	4.00	4.14	7.43	9.91
	200	1.63	3.01	4.95	3.66	6.25	9.52	0.49	0.79	1.31	2.88	3.56	5.03
	50	13.7	23.2	29.1	19.2	32.5	37.5	10.8	18.2	23.5	13.8	23.4	32.3
10	100	8.69	15.2	23.8	12.4	23.6	30.9	4.60	7.46	9.52	9.01	13.5	18.9
	200	2.12	4.96	7.24	4.02	8.17	10.6	1.08	2.05	3.68	3.37	6.09	10.75
	50	17.6	26.9	31.8	28.9	37.4	39.8	16.5	25.3	30.0	21.3	28.8	37.3
20	100	12.0	23.1	25.0	16.6	30.2	30.5	8.21	16.2	23.9	14.0	24.7	31.1
	200	3.99	7.01	10.6	5.84	10.2	15.1	2.49	4.05	8.46	8.22	11.3	16.2
RMSE													
	50	1.78	2.63	3.88	2.06	2.76	4.07	0.80	0.98	1.17	0.88	1.04	1.21
5	100	0.87	1.04	1.19	0.95	1.02	1.27	0.44	0.48	0.54	0.50	0.52	0.57
	200	0.69	0.64	0.70	0.61	0.80	0.83	0.33	0.33	0.34	0.42	0.39	0.41
	50	4.53	7.49	9.22	5.50	7.79	9.29	1.83	3.02	5.67	2.43	3.74	6.39
10	100	1.52	1.76	2.78	1.59	2.00	2.41	0.76	0.84	0.96	0.90	0.96	1.09
	200	0.97	1.01	1.09	0.94	1.21	1.24	0.55	0.56	0.58	0.69	0.70	0.68
	50	10.6	13.3	14.4	11.1	13.3	14.5	8.48	12.8	15.1	9.58	13.0	15.3
20	100	2.61	4.06	8.25	3.53	5.55	8.95	1.46	1.81	2.71	2.21	2.67	3.98
	200	1.46	1.65	1.78	1.69	1.76	1.91	0.91	0.94	1.00	1.13	1.22	1.31

6. Empirical Analysis

6.1. Predicting GDP growth

We consider the problem of predicting the growth rate of the U.S. quarterly gross domestic product (GDP). In addition, nine macroeconomic variables with different sampling frequencies are available. The data are obtained from the St.

Louis Federal Reserve Economic Data website. The predictive regression used is

$$y_i = \phi_0 + \phi_1 y_{i-1} + \cdots + \phi_a y_{i-a} + \sum_{l=1}^{9} \sum_{b=0}^{B_l} \beta_{l,b} z_{l,i\times m_l - b} + e_i, \tag{6.1}$$

where a and B_l are nonnegative integers, y_i is the growth rate (first difference of natural logarithm) of U.S. quarterly seasonally adjusted real GDP, and $z_{l,\cdot}$ are high-frequency covariates with frequency m_l, for example, $m_l = 3$ for monthly data. The nine covariates considered in this study are as follows: $z_{1,\cdot}$ is the change of monthly civilian unemployment rates; $z_{2,\cdot}$ is the monthly growth rate of all employees' total payrolls; $z_{3,\cdot}$ is the growth rate of the monthly industrial production total index; $z_{4,\cdot}$ is the growth rate of the monthly consumer price index; $z_{5,\cdot}$ is the growth rate of the monthly Moody's Seasoned Baa Corporate Bond Yields; $z_{6,\cdot}$ is the change in the daily 3-Month Treasury Bill Secondary Market Rate; $z_{7,\cdot}$ is the change in the daily 10-Year Treasury Constant Maturity Rate; $z_{8,\cdot}$ is the change in the daily NASDAQ Composite Index; and $z_{9,\cdot}$ is the change in the daily Wilshire 5,000 Total Market Full Cap Index. The transformations of all variables are based on those of Stock and Watson (2002). Note that all data are seasonally adjusted, if necessary, and the explanatory variables are monthly or daily data. For daily variables $z_{6,\cdot}$ and $z_{7,\cdot}$, we use data of the first 16 trading days in a month. For daily variables $z_{8,\cdot}$ and $z_{9,\cdot}$, we use data of the first 15 trading days. The sampling period was January 1980 to February 2017, but the prediction origin started with the second quarter of 2013, and ended with the first quarter of 2017. There was no trading activity during weekends and holidays, and there exist some missing data in the trading activities. Trading days for each month vary. We choose the first 15 or 16 trading days, simply because they are the minimum number of trading days available for each month (mainly February).

Two types of empirical analysis are examined. First, we consider a linear model with all explanatory variables, estimated by the Lasso procedure. For comparison, we include a model with all explanatory variables except the NASDAQ Composite Index and Wilshire 5,000 Total Market Full Cap Index, estimated by the MIDAS regression (denoted by MIDAS-B model), a model with total monthly payrolls for all-employees as the only explanatory variable, also estimated by MIDAS (denoted by MIDAS-A model), and a simple ARMA model of the GDP growth rates (denoted by ARMA model). We use the BIC to select the number of autoregressive lags (a) and the lags (B_l) of the explanatory variables.

Table 4. Results of out-of-sampling prediction of U.S. quarterly real GDP growth rate. The data cover the period 1980 to February 2017, but the forecast origins start from the second quarter of 2013 to the first quarter of 2017. All measurements are multiplied by 10^3. In the table, MAD, MAE, and RMSE are the median absolute error, mean absolute error, and root mean squared error, respectively.

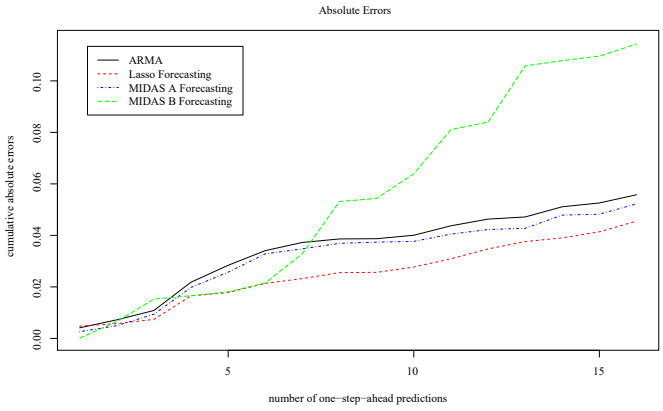
Model	MAD	MAE	RMSE
ARMA	3.175	3.486	4.319
Lasso	2.328	2.845	3.491
MIDAS-A	2.463	3.264	4.245
MIDAS-B	4.089	7.143	9.920

The Lasso tuning parameter λ is also chosen using the BIC; see Bühlmann and Van De Geer (2011). Here, we aggregate the daily explanatory variables z_6 and z_7 to a weekly frequency for the MIDAS regression.

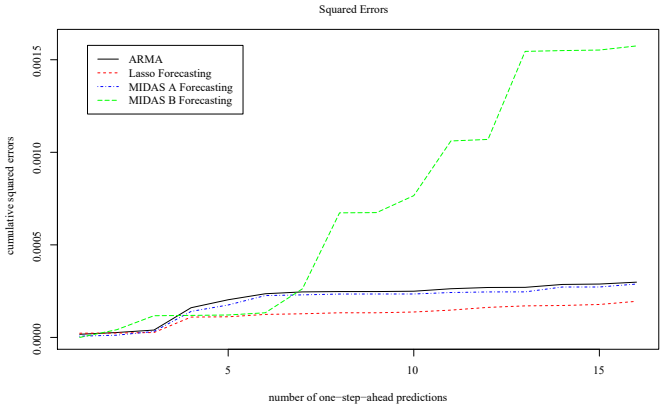
Table 4 shows the median absolute deviation (MAD), mean absolute error (MAE), and root mean squared error (RMSE) for the prediction period. From the table, it is clear that the Lasso-based model outperforms all the other models in this particular instance. The poor performance of MIDAS-B is likely due to using too many explanatory variables with multiple sampling frequencies.

Figure 2 displays the cumulative absolute errors and the cumulative squared errors for different models in predicting the GDP growth rate. It shows clearly that the Lasso model performs best. The MIDAS-A model also improves the prediction errors over those of the simple ARMA model. However, the MIDAS-B model fares poorly. Consequently, unlike the Lasso model, the MIDAS regression is not robust to the presence of irrelevant regressors. In fact, the MIDAS regression is also sensitive to the weighting schemes and the starting points of its optimization program.

Next, we compare between forecasting and nowcasting. Recall that the goal of nowcasting is to take advantage of available high-frequency data to improve the prediction of lower-frequency variables of interest. For the quarterly GDP growth rate, during the quarter of interest, some monthly macroeconomic variables, and even some daily economic variables become available. Here, nowcasting attempts to update the GDP prediction by incorporating the newly available high-frequency explanatory variables. In this exercise, we consider nowcasting using the first month's data within the quarter, and using the first two months' data.



(a)



(b)

Figure 2. Panel (a): Cumulative absolute errors. Panel (b): Cumulative squared errors. MIDAS-A represents the MIDAS regression model using all-employees' monthly total payrolls as the explanatory variable. MIDAS-B represents the MIDAS regression model with seven regressors $z_{1,\cdot}, \ldots, z_{7,\cdot}$, where $z_{6,\cdot}$ and $z_{7,\cdot}$ are aggregated into weekly data.

For comparison purposes, we employ an autoregressive (AR) model

$$y_i = \phi_0 + \phi_1 y_{i-1} + \ldots + \phi_a y_{i-a} + \epsilon_i, \tag{6.2}$$

as a benchmark for prediction. The AR order is selected using the BIC in the modeling subsample, and is assumed to be fixed in the forecasting subsample. The AR model in Equation (6.2) is estimated in two ways. First, it is estimated using the ordinary least squares method, and we denote the model by AR-OLS.

Second, assuming sparsity, we estimate the AR model via the Lasso method, with the tuning parameter λ selected using the BIC. The forecasting result of this model is denoted by AR-Lasso. These two models represent the performance of forecasting.

For nowcasting, we augment the AR model in Equation (6.2) with all explanatory variables available in the first month of the quarter, and denote the results by Nowcasting 1. Similarly, if we augment the AR model with all explanatory variables available in the first two months of the quarter, then the results are denoted by Nowcasting 2. Specifically, for nowcasting, we employ the model

$$y_i = \phi_0 + \phi_1 y_{i-1} + \cdots + \phi_a y_{i-a} + \beta^T \mathbf{x}_i + \epsilon_i,$$

where $\mathbf{x}_i$ denotes the available high-frequency explanatory variables. For Nowcasting 1, $\mathbf{x}_i$ consists of data of the first month of a given quarter, whereas for Nowcasting 2, it consists of data of the first two months of a given quarter. In this exercise, we use all monthly and daily high-frequency variables $z_{1,\cdot}, \ldots, z_{9,\cdot}$. We denote the results for the MIDAS regressions as MIDAS-C Nowcasting 1 and MIDAS-C Nowcasting 2, respectively. Finally, we employ a MIDAS regression that only uses explanatory variables $z_{1,\cdot}, \ldots, z_{7,\cdot}$ in the nowcasting and denote the results as MIDAS-D.

Table 5 summarizes the performance of nowcasting in predicting U.S. quarterly GDP growth rates in the forecast period. From the table, we make the following observations. First, as expected, now-casting fares better than forecasting. The only exception is MIDAS-D nowcasting. Second, also as expected, Nowcasting 2 shows some improvement over Nowcasting 1 for a given model. Keep in mind, however, Nowcasting 1 is available one month into a quarter, whereas Nowcasting 2 needs to wait for an additional month. Third, from the performance of MIDAS-C and MIDAS-D, the stock market indices do not seem to be helpful in predicting the GDP growth rate. In real applications, there exist many high-frequency explanatory variables, but their contributions to predicting the low-frequency variable of interest in unknown a priori. In this situation, our results suggest that the Lasso regression could be helpful.

Figure 3 shows that both the Lasso model and the MIDAS-B model improve the prediction via nowcasting. However, when irrelevant variables exist, the MIDAS regression might encounter some difficulties.

Table 5. Comparison between forecasting and nowcasting in predicting the U.S. quarterly real GDP growth rate. The data cover the period 1980 to February 2017, but the forecast origins are from the second quarter of 2013 to the first quarter of 2017. All measurements are multiplied by 10^3. In the table, MAD, MAE, RMSE are the median absolute deviation, mean absolute error, and root mean squared error, respectively.

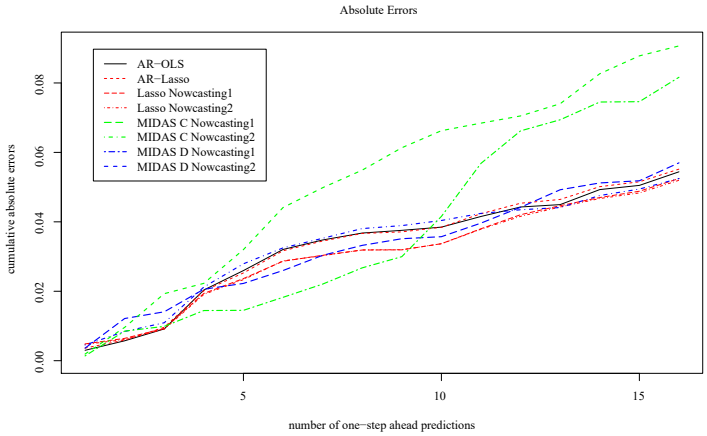
Model	MAD	MAE	RMSE
AR-OLS	2.865	3.400	4.242
AR-Lasso	3.327	3.448	4.174
Lasso Now-casting 1	2.731	3.278	3.962
Lasso Now-casting 2	2.834	3.247	3.941
MIDAS-C Now-casting 1	4.181	5.102	6.507
MIDAS-C Now-casting 2	5.108	5.666	6.430
MIDAS-D Now-casting 1	3.670	3.561	4.125
MIDAS-D Now-casting 2	2.784	3.279	4.048

6.2. Nowcasting $PM_{2.5}$

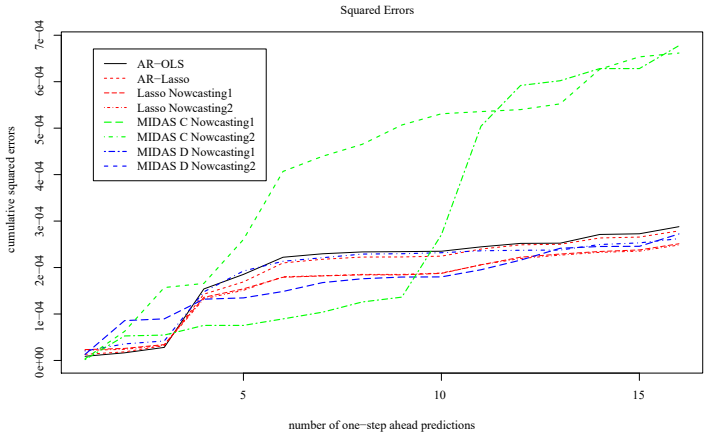
Consider next the prediction of $PM_{2.5}$. The response y is the square root transformed daily maximum of $PM_{2.5}$. Hourly data of a monitoring station in the southern part of Taiwan are used. To see the nowcasting effects, we consider adding six covariates, which are the first six hourly $PM_{2.5}$ readings of the same day, starting from midnight. The sample period is 2006 to 2015, yielding 3,650 observations. (Feb 29 was dropped.) We reserve the last 730 data points (two years) for one-step-ahead out-of-sample forecasts.

For comparison purposes, we first consider the square root $PM_{2.5}$ (i.e., response y) as a pure time series. An AR(22) model is selected. Thus, the baseline model is a univariate AR(22). We denote the model by AR-OLS. For nowcasting, we augment the AR model with the first six hourly readings. If we augment the AR model with the first hourly $PM_{2.5}$ reading, then the results are denoted by Nowcasting 1. Similarly, if we augment the AR model with the first two hourly $PM_{2.5}$ readings, then the results are denoted by Nowcasting 2, and so on. We denote the results for the autoregressive model with exogenous variables as ARX Nowcasting 1, ARX Nowcasting 2, and so on. We use the BIC to select the number of autoregressive lags. The Lasso tuning parameter λ is also chosen using the BIC.

Table 6 summarizes the performance of nowcasting in predicting the daily maximum of $PM_{2.5}$. From the table, we make the following observations. First, as expected, nowcasting outperforms forecasting. Second, also as expected, for a given model, Nowcasting 2 shows some improvement over Nowcasting 1, Now-



(a)



(b)

Figure 3. Panel (a): Cumulative absolute errors. Panel (b) Cumulative squared errors. MIDAS-D represents the MIDAS regression model with seven regressors $z_{1,\cdot}, \ldots, z_{7,\cdot}$. MIDAS-C represents the MIDAS regression model with nine regressors $z_{1,\cdot}, \ldots, z_{9,\cdot}$. Nowcasting 1 and Nowcasting 2 represent predictions of the quarterly GDP growth rate when the first month and the first two months data are available, respectively.

casting 3 shows some improvement over Nowcasting 2, and so on. Third, and of most interest, the Lasso estimator significantly outperforms the ARX model and the benchmark model. In short, the Lasso regression appears helpful in applying nowcasting to $PM_{2.5}$.

Table 6. Comparison between forecasting and nowcasting in predicting the daily maximum of $PM_{2.5}$. The data period is 2006 to 2015, and the forecast origins are from 2013 to the end of 2015. (February 29 is excluded). In the table, MAE and RMSE denote the mean absolute error and root mean squared error for one-step-ahead predictions, respectively.

Model	MAE	RMSE
AR-OLS	1619.6	73.71
ARX Now-casting 1	975.9	46.54
ARX Now-casting 2	940.9	44.92
ARX Now-casting 3	904.2	43.40
ARX Now-casting 4	879.7	42.31
ARX Now-casting 5	850.6	41.24
ARX Now-casting 6	835.3	40.31
Lasso Now-casting 1	659.3	31.74
Lasso Now-casting 2	628.4	30.58
Lasso Now-casting 3	623.2	30.72
Lasso Now-casting 4	600.7	29.63
Lasso Now-casting 5	595.0	29.49
Lasso Now-casting 6	576.3	28.47

Supplementary Material

The online Supplementary Material contains proofs of the theorems and lemmas presented in this paper.

Acknowledgements

We are grateful to the referees, Associate Editor, and Editor for their many helpful comments.

References

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **43**, 1535–1567.

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

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

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

Chen, X., Xu, M. and Wu, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *Ann. Statist.* **41**, 2994–3021.

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

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

Ghysels, E., Santa-Clara, P. and Valkanov, R. (2004). The MIDAS touch: Mixed data sampling regression models. *Finance*.

Gupta, S. (2012). A note on the asymptotic distribution of Lasso estimator for correlated data. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)* **74**, 10–28.

Jia, J., Rohe, K. and Yu, B. (2013). The Lasso under poisson-like heteroscedasticity. *Statist. Sinica* **23**, 99–118.

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

Medeiros, M. C. and Mendes, E. F. (2016). L1-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors. *J. Econom* **191**, 255–271.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37**, 246–270.

Merlevède, F., Peligrad, M. and Rio, E. (2009). *Bernstein Inequality and Moderate Deviations under Strong Mixing Conditions*, volume 5 of *Collections*, 273–292. Institute of Mathematical Statistics.

Priestley, M. (1988). *Non-linear and Non-stationary Time Series Analysis*. Academic Press.

Rosenblatt, M. (1971). *Markov Processes: Structure and Asymptotic Behavior*. Springer.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97**, 1167–1179.

Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in Lasso problems. *Ann. Statist.* **40**, 1198–1232.

Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach*. Oxford University Press.

Tsay, R. (2005). *Analysis of Financial Time Series*. John Wiley & Sons.

Wang, H., Li, G. and Tsai, C.-L. (2007). Regression coefficient and autoregressive order shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69**, 63–78.

Wiener, N. (1958). *Nonlinear Problems in Random Theory*. Wiley, New York.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 14150–14154.

Wu, W.-B. and Wu, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electron. J. Statist.* **10**, 352–379.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68**, 49–67.

Zhang, D. and Wu, W. B. (2017). Gaussian approximation for high dimensional time series. *Ann. Statist.* **45**, 1895–1919.

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

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

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the "degrees of freedom" of the Lasso. *Ann. Statist.* **5**, 2173–2192.

5747 South Ellis Avenue, Chicago, IL 60637, USA.
E-mail: yfhan@uchicago.edu
5807 South Woodlawn Avenue, Chicago, IL 60637, USA.
E-mail: Ruey.Tsay@chicagobooth.edu

(Received January 2018; accepted June 2018)