The Annals of Statistics
2020, Vol. 48, No. 4, 1959–1980
https://doi.org/10.1214/19-AOS1872

MODEL SELECTION FOR HIGH-DIMENSIONAL LINEAR REGRESSION WITH DEPENDENT OBSERVATIONS

BY CHING-KANG ING

Institute of Statistics, National Tsing Hua University, cking@stat.nthu.edu.tw

We investigate the prediction capability of the orthogonal greedy algorithm (OGA) in high-dimensional regression models with dependent observations. The rates of convergence of the prediction error of OGA are obtained under a variety of sparsity conditions. To prevent OGA from overfitting, we introduce a high-dimensional Akaike's information criterion (HDAIC) to determine the number of OGA iterations. A key contribution of this work is to show that OGA, used in conjunction with HDAIC, can achieve the optimal convergence rate without knowledge of how sparse the underlying high-dimensional model is.

1. Introduction. Model selection for high-dimensional regression models has been one of the most vibrant topics in statistics over the past decade. It also has broad applications in a variety of important fields such as bioinformatics, quantitative finance, image processing and advanced manufacturing; see Negahban et al. (2012) and Ing et al. (2017) for further discussion. A typical high-dimensional regression model takes the following form:

$$y_t = \sum_{j=1}^{p} \beta_j x_{tj} + \varepsilon_t, \quad t = 1, \dots, n, \tag{1.1}$$

where n is the sample size, $x_{t1}, \dots, x_{tp}$ are predictor variables, ε_t are mean-zero random disturbance terms and $p = p_n$ is allowed to be much larger than n. There are computational and statistical difficulties in estimating the regression function by standard regression methods owing to $p \gg n$. However, by assuming sparsity conditions on β_j, eigenvalue conditions on the covariance (correlation) matrix of the predictor variables, and distributional conditions on ε_t or x_{tj}, it has been shown that consistent estimation of the regression function or optimal prediction is still possible either through penalized least squares methods (see Zhao and Yu (2006), Candes and Tao (2007), Bickel, Ritov and Tsybakov (2009) and Zhang (2010)) or through greedy forward selection algorithms (see Bühlmann (2006), Chen and Chen (2008), Wang (2009), Fan and Lv (2008) and Ing and Lai (2011)).

The vast majority of studies on model (1.1), however, have focused on situations where $\mathbf{x}_t = (x_{t1}, \dots, x_{tp})^\top$ are nonrandom and ε_t are independently and identically distributed (i.i.d.) or $(\mathbf{x}_t, \varepsilon_t)$ are i.i.d., which regrettably preclude most serially correlated data. In fact, (1.1) can encompass a broad array of time series models if these restrictions are relaxed. For example, it becomes the well-known autoregressive (AR) model when $x_{tj} = y_{t-j}$. Since the predictor variables in AR models have a natural ordering, a commonly used sparsity condition is

$$C_1 j^{-\gamma} \le |\beta_j| \le C_2 j^{-\gamma}, \quad 0 < C_1 \le C_2 < \infty, \gamma > 1, \tag{1.2}$$

in which $|\beta_j|$ decay polynomially, or

$$C_3 \exp(-\beta j) \le |\beta_j| \le C_4 \exp(-\beta j), \quad 0 < C_3 \le C_4 < \infty, \beta > 0, \tag{1.3}$$

Received November 2018; revised May 2019.
MSC2020 subject classifications. Primary 63M30; secondary 62F07, 62F12.
Key words and phrases. Best m-term approximations, high-dimensional Akaike's information criterion, orthogonal greedy algorithm, sparsity conditions, time series.

in which $|\beta_j|$ decay exponentially (see Shibata (1980) and Ing (2007)). Moreover, the model selection problem in the AR case is simplified to an order selection one, which has been well explored in the literature (see Shibata (1980)). When x_{tj}, $j = 1, \dots, p$, do not have a natural ordering, for example, the autoregressive exogenous (ARX) model, (1.2) and (1.3) can be generalized as

$$Lj^{-\gamma} \le |\beta^*_{(j)}| \le Uj^{-\gamma} \tag{1.4}$$

and

$$L_1 \exp(-\beta j) \le |\beta^*_{(j)}| \le U_1 \exp(-\beta j), \tag{1.5}$$

respectively, where $0 < L \le U < \infty$, $0 < L_1 \le U_1 < \infty$, and $|\beta^*_{(1)}| \ge |\beta^*_{(2)}| \ge \cdots \ge |\beta^*_{(p)}|$ is a rearrangement of $\{|\beta^*_j|\}$ in decreasing order with $\beta^*_j = \sigma_j \beta_j$ and $\sigma_j^2 = \mathrm{E}(x_{tj}^2)$. However, unlike the order selection problem, the model selection problem in (1.1) with dependent observations and with coefficients satisfying (1.4) or (1.5) seems to be seldom investigated. The problem becomes more challenging when β_j may obey either one of (1.4), (1.5) or $k_0 \ll n$, but it is unclear which of the three is true. Here, k_0 denotes the number of nonzero coefficients in model (1.1), and $k_0 \ll n$ is referred to as the strong sparsity condition.

In this paper, we assume that the $(\mathbf{x}_t, \varepsilon_t)$ in model (1.1) is a time series obeying concentration inequalities (2.2) and (2.3). We also assume that the β_j in model (1.1) follow one of the following sparsity conditions: (i) (A3), (ii) (A4), or (iii) $k_0 \ll n$, where (A3) and (A4) are defined in Section 2.1. Note that (A3) includes (1.4) and

$$\sum_{j=1}^{p} |\beta^*_j|^{1/\gamma} < M_4 \quad \text{for some } \gamma \ge 1, 0 < M_4 < \infty, \tag{1.6}$$

as special cases, whereas (A4) contains (1.5). We use the orthogonal greedy algorithm (OGA) (Temlyakov (2000)) to sequentially include candidate variables and introduce a high-dimensional Akaike's information criterion (HDAIC) to determine the number of OGA iterations. This model selection procedure is denoted by OGA+HDAIC. A key contribution of this paper is to show that OGA+HDAIC achieves the optimal convergence rate without knowing which sparsity condition among (i), (ii) and (iii) would follow, thereby alleviating the dilemma mentioned in the previous paragraph.

Following this Introduction, the rest of the paper is organized as follows. In Section 2.1, we introduce OGA and the assumptions required for our asymptotic analysis of the algorithm. Section 2.2 derives an error bound for OGA, which is the sum of an approximation error and a term accounting for the sampling variability. Since the approximation error decreases as the number m of iterations increases and the sampling variability increases with m, the optimal m can be determined by equating the two terms in the error bound for OGA. This approach, however, is infeasible because not only does the solution involve the unknown parameters in (A3) or (A4), but it is unknown which kind of sparsity among (i), (ii) and (iii) holds true. To overcome this difficulty, Theorem 3.1 in Section 3.1 proposes using HDIC to determine the number of iterations, and shows that OGA+HDAIC is rate optimal regardless of which sparsity condition is true. In Section 3.2, we offer a comprehensive comparison of our results with those in Negahban et al. (2012) and Ing and Lai (2011), in which the statistical properties of Lasso (Tibshirani (1996)) and OGA, respectively, are explored under model (1.1) with independent observations. In this connection, Section 3.2 also discusses the papers by Basu and Michailidis (2015) and Wu and Wu (2016), which investigate the performance of Lasso under sparse high-dimensional time series models. The proof of Theorem 3.1 is given in Section 3.3. We conclude in Section 4. An Appendix consisting of some technical results is given at the end of the paper. A simulation study to illustrate the performance of OGA+HDAIC, along with further technical details, is deferred to the Supplementary Material (Ing (2020)).

2. Asymptotic theory of OGA in weakly sparse models. This section aims at establishing the convergence rate of OGA under sparse high-dimensional regression models with dependent observations. The definition of OGA and the assumptions required for our analysis of OGA are given in Section 2.1. The main result of this section is stated and proved in Section 2.2.

2.1. *Models and assumptions.* We assume that $\{(\mathbf{x}_t, \varepsilon_t)\}$ in model (1.1) is a zero-mean stationary time series satisfying $\mathrm{E}(\mathbf{x}_t\varepsilon_t) = \mathbf{0}$. The OGA is a recursive procedure that selects variables from the set of predictor variables in (1.1) one at a time. Define $\mathbf{X}_i = (x_{1i}, \dots, x_{ni})^\top$, $\mathbf{Z}_i = (z_{1i}, \dots, z_{ni})^\top = \mathbf{X}_i/\sigma_i$, and $\mathbf{Y} = (y_1, \dots, y_n)^\top$. The algorithm is initialized by setting $\hat{J}_0 = \varnothing$, where $\hat{J}_m$ denotes the index set of the variables chosen by OGA at the mth iteration. For $m \geq 1$, $\hat{J}_m$ is recursively updated by

$$\hat{J}_m = \hat{J}_{m-1} \cup \{\hat{j}_m\}, \tag{2.1}$$

where

$$\hat{j}_m = \arg \max_{1\leq j\leq p, j\notin \hat{J}_{m-1}} |\hat{\boldsymbol{\mu}}_{\hat{J}_{m-1},j}|,$$

with $\hat{\boldsymbol{\mu}}_{J,i} = \mathbf{Z}_i^\top(\mathbf{I} - \mathbf{H}_J)\mathbf{Y}/(n^{1/2}\|\mathbf{Z}_i\|)$, $\|\mathbf{a}\|$ denoting the L_2-norm of vector $\mathbf{a}$, and $\mathbf{H}_J$, $J \subseteq \mathcal{P} \equiv \{1, \dots, p\}$, being the orthogonal projection matrix onto the linear span of $\{\mathbf{Z}_i, i \in J\}$ $(\mathbf{H}_\varnothing = \mathbf{0})$.

To investigate the performance of OGA, we make the following distributional assumptions:

(A1) There exists $c_1^* > 0$ such that

$$P\left(\max_{1\leq j\leq p}\left|n^{-1}\sum_{i=1}^n z_{ij}\varepsilon_i\right| \geq c_1^*(\log p)^{1/2}/n^{1/2}\right) = o(1). \tag{2.2}$$

(A2) There exists $c_2^* > 0$ such that

$$P\left(\max_{1\leq k,l\leq p}\left|n^{-1}\sum_{i=1}^n z_{ik}z_{il} - \rho_{kl}\right| \geq c_2^*(\log p)^{1/2}/n^{1/2}\right) = o(1), \tag{2.3}$$

where $\rho_{kl} = \mathrm{E}(z_{1k}z_{1l})$.

The following examples help illustrate (A1) and (A2). Let $\lambda_{\min}(A)$ $(\lambda_{\max}(A))$ denote the minimum (maximum) eigenvalue of matrix A and $\|\mathbf{a}\|_1$, the L_1-norm of vector $\mathbf{a}$.

EXAMPLE 1 (Gaussian linear processes). Let

$$x_{tl} = \sum_{j=0}^{\infty} \boldsymbol{w}_j^\top(l)\boldsymbol{\delta}_{t-j}, \qquad \varepsilon_t = \sum_{j=0}^{\infty} \boldsymbol{w}_j^\top(0)\boldsymbol{\delta}_{t-j}, \tag{2.4}$$

where $\boldsymbol{\delta}_t = (\delta_{t1}, \dots, \delta_{tq})^\top$ are i.i.d. Gaussian random vectors satisfying

$$\mathrm{E}(\boldsymbol{\delta}_t) = \mathbf{0}, \qquad \max_{1\leq i\leq q} \mathrm{E}(\delta_{ti}^2) < \bar{c} < \infty, \qquad \lambda_{\min}(\mathrm{E}(\boldsymbol{\delta}_t\boldsymbol{\delta}_t^\top)) > \underline{c}_0 > 0, \tag{2.5}$$

and $\boldsymbol{w}_j(l)$ obey

$$\max_{0\leq l\leq p} \sum_{j=0}^{\infty} \|\boldsymbol{w}_j(l)\|_1 < M_2 < \infty, \qquad \min_{0\leq l\leq p} \|\boldsymbol{w}_0(l)\| > \underline{c}_1 > 0. \tag{2.6}$$

Then, by making use of the Hanson–Wright inequality (see Theorem 1.1 of Rudelson and Vershynin (2013)), it is shown in Section S1 of the Supplementary Material that

(2.7) (A1) and (A2) hold true under (2.4)–(2.6) and (2.8),

where (2.8) is given by

$$p \to \infty \quad \text{as } n \to \infty, \quad \frac{\log p}{n} = o(1). \tag{2.8}$$

As an application, we consider a high-dimensional ARX model,

$$y_t = \sum_{j=1}^{q_0} a_j y_{t-j} + \sum_{l=1}^{p_1} \sum_{j=1}^{r_l} \beta_j^{(l)} x_{t-j+1}^{(l)} + \varepsilon_t, \tag{2.9}$$

in which $p = q_0 + \sum_{l=1}^{p_1} r_l$ satisfies (2.8), $1 - \sum_{j=1}^{q_0} a_j z^j \neq 0$ for all $|z| \le 1 + \iota$ with ι being some positive constant, $\sum_{j=1}^{q_0} |a_j| + \sum_{l=1}^{p_1} \sum_{j=1}^{r_l} |\beta_j^{(l)}| < M_5 < \infty$, $x_t^{(l)} = \epsilon_t^{(l)} + \sum_{j=1}^{\infty} b_j^{(l)} \epsilon_{t-j}^{(l)}$, with $\sum_{j=1}^{\infty} |b_j^{(l)}| < M_6 < \infty$ for all $1 \le l \le p$, and $\boldsymbol{\delta}_t = (\epsilon_t^{(1)}, \ldots, \epsilon_t^{(p)}, \varepsilon_t)^\top$ are i.i.d. $(p+1)$-dimensional Gaussian random vectors obeying (2.5) with $q = p + 1$. It is not difficult to see that (2.4) and (2.6) are fulfilled by the regressor variables and the error term in (2.9). Hence (A1) and (A2) are applicable to model (2.9).

EXAMPLE 2 (Linear processes with sub-Gaussian innovations). Suppose that (2.4)–(2.6) and (2.8) are satisfied except that the Gaussianity of $\boldsymbol{\delta}_t$ is replaced by

$$\|\delta_{tk}\|_{\psi_2} \le L, \quad k = 1, \ldots, q, \tag{2.10}$$

where $\|\cdot\|_{\psi_2}$ denotes the ψ_2 Orlicz norm and L is some positive number. We note that (2.10) is fulfilled by sub-Gaussian random variables. Assume $q = p^s$ for some $0 \le s < \infty$. Then by making use of the concentration inequality given in Theorem 1.4 of Adamczak and Wolff (2015), it can be shown that (A1) and (A2) hold for some large c_1^* and c_2^*. For more details, see Huang and Ing (2019). In addition, the regressor variables and the error term in (2.9) still obey (A1) and (A2), provided assumption (2.10) is used in place of the Gaussian assumption in Example 1.

We also need a sparsity condition on regression coefficients:

(A3) There is $0 < \bar{M}_0 < \infty$ such that $\sum_{j=1}^{p} \beta_j^{*2} \le \bar{M}_0$. In addition, there exist $\gamma \ge 1$ and $0 < C_\gamma < \infty$ such that for any $J \subseteq \mathcal{P}$,

$$\sum_{j \in J} |\beta_j^*| \le C_\gamma \left(\sum_{j \in J} \beta_j^{*2} \right)^{(\gamma-1)/(2\gamma-1)}. \tag{2.11}$$

When $\gamma = 1$, (2.11) and (1.6) are equivalent. However, (2.11) is weaker than (1.6) for $\gamma > 1$. To see this, note that if (1.6) is true for some $\gamma > 1$, then by Hölder's inequality,

$$\begin{aligned} \sum_{j \in J} |\beta_j^*| &\le \left(\sum_{j \in J} |\beta_j^*|^{1/\gamma} \right)^{\gamma/(2\gamma-1)} \left(\sum_{j \in J} \beta_j^{*2} \right)^{(\gamma-1)/(2\gamma-1)} \\ &\le M_4^{\gamma/(2\gamma-1)} \left(\sum_{j \in J} \beta_j^{*2} \right)^{(\gamma-1)/(2\gamma-1)}, \end{aligned}$$

implying that (2.11) holds for $C_\gamma = M_4^{\gamma/(2\gamma-1)}$. In view of the connection between (2.11) and (1.6), the parameter γ in (2.11) can be understood as an index to describe the degree of sparseness in the underlying high-dimensional models. The larger the γ is, the sparser the model is. Although assumptions similar to (1.6) are quite popular for high-dimensional regression analysis (see, e.g., Wang et al. (2014)), there is a subtle difference between (2.11) and (1.6). To see this, assume that (1.4) holds for some $\gamma > 1$. Then (2.11) holds for the same γ (see Lemma A.2 in the Appendix), whereas (1.6) is violated due to $L(1+\log p) \le \sum_{j=1}^{p} |\beta_j^*|^{1/\gamma} \le U(1+\log p)$. It is worth mentioning that (1.4) not only plays an important role in time series modeling, it also allows us to demonstrate that the approximation error of the population counterpart of OGA (which is defined at the beginning of Appendix A and is referred to as the population OGA) is almost as small as that of the best m-term approximation (see (3.24) and Lemma A.3 in the Appendix). In the sequel, we refer to (2.11) as the "polynomial decay" case, owing to its connection with (1.4). To broaden OGA's applications, we also consider a coefficient condition sparser than (2.11):

(A4) There exists $0 < M_0 < \infty$ such that $\max_{1\le j\le p} |\beta_j^*| \le M_0$. Moreover, there exists $M_1 > 1$ such that for any $J \subseteq \mathcal{P}$,

$$\sum_{j\in J} |\beta_j^*| \le M_1 \max_{j\in J} |\beta_j^*|. \tag{2.12}$$

Assumption (A4) is referred to as the "exponential decay" case because (1.5) is included by (2.12).

The following assumption on the covariance structure of $\mathbf{z}_t = (z_1, \dots, z_p)^\top$ is frequently used throughout the paper. Define $\boldsymbol{\Gamma}(J) = \mathrm{E}\{\mathbf{z}_t(J)\mathbf{z}_t^\top(J)\}$ and $\boldsymbol{g}_i(J) = \mathrm{E}(z_{ti}\mathbf{z}_t(J))$, where $J \subseteq \mathcal{P}$ and $\mathbf{z}_t(J) = (z_{ti}, i \in J)^\top$.

(A5) For some positive numbers $\bar{D}$ and M,

$$\max_{1\le \sharp(J)\le \bar{D}(n/\log p)^{1/2}, i\notin J} \|\boldsymbol{\Gamma}^{-1}(J)\boldsymbol{g}_i(J)\|_1 < M, \tag{2.13}$$

where $\sharp(J)$ denotes the cardinality of J.

Since $\boldsymbol{\Gamma}^{-1}(J)\boldsymbol{g}_i(J) = \arg\min_{\boldsymbol{c}\in R^{\sharp(J)}} \mathrm{E}(z_{ti} - \boldsymbol{c}^\top \mathbf{z}_t(J))^2$, (2.13) essentially says that the regression coefficients for z_{ti} on $\mathbf{z}_t(J)$ with all $i \notin J$ and $\sharp(J) \le \bar{D}(n/\log p)$ are L_1 bounded. This condition holds even when $z_{t1}, \dots, z_{tp}$ are highly correlated; see Section S3 of the supplementary document. Let $\boldsymbol{g}_y(J) = \mathrm{E}(y_t\mathbf{z}_t(J))$ and $\boldsymbol{\beta}^*(J) = \boldsymbol{\Gamma}^{-1}(J)\boldsymbol{g}_y(J) = \arg\min_{\boldsymbol{c}\in R^{\sharp(J)}} \mathrm{E}(y_t - \boldsymbol{c}^\top\mathbf{z}_t(J))^2$, which is the regression coefficients for y_t on $\mathbf{z}_t(J)$. By making use of (2.13), we will show later that for any $J \subseteq \mathcal{P}$ with $\sharp(J) \le \bar{D}(n/\log p)^{1/2}$, there exists $0 < C < \infty$ such that

$$\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^*(J)\|_1 \le C \sum_{j\notin J} |\beta_j^*|, \tag{2.14}$$

where $\boldsymbol{\beta}^* = (\beta_1^*, \dots, \beta_p^*)^\top$ and $\boldsymbol{\beta}^*(J)$ here is regarded as a p-dimensional vector with undefined entries set to 0. Inequality (2.14) is referred to as the uniform Baxter's inequality. (For more details on Baxter's inequality in autoregressive modeling, see Baxter (1962), Berk (1974) and Pourahmadi (1989).) This inequality can be used together with (2.11) to yield, for all $\sharp(J) \le \bar{D}(n/\log p)^{1/2}$,

$$\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^*(J)\|_1 \le C C_\gamma \left(\sum_{j\notin J} \beta_j^{*2}\right)^{(\gamma-1)/(2\gamma-1)}, \tag{2.15}$$

which is one of the key ingredients in our asymptotic analysis of OGA+HDAIC.

To derive (2.14) from (2.13), we may assume without loss of generality that $J = \{j_1, \dots, j_q\}$ for some $1 \le q \le \bar{D}(n/\log p)^{1/2}$, where $j_i, i = 1, \dots, q$, are distinct elements in $\mathcal{P}$. Note first that

$$\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^*(J)\|_1 \le \|\boldsymbol{\beta}^*(J) - \boldsymbol{\beta}_J^*\|_1 + \sum_{j\notin J} |\beta_j^*|, \tag{2.16}$$

where $\boldsymbol{\beta}_J^* = (\beta_{j_1}^*, \dots, \beta_{j_q}^*)^\top$. Denote $\boldsymbol{g}_y(J)$ by $(\gamma_{y,j_1}, \dots, \gamma_{y,j_q})^\top$. Then it follows that $\gamma_{y,j_i} = \sum_{l=1}^{p} \rho_{j_i l}\beta_l^*$, $i = 1, \dots, q$, and hence

$$\begin{aligned}\boldsymbol{\Gamma}(J)^{-1}\boldsymbol{\Gamma}(J)\big(\boldsymbol{\beta}^*(J) - \boldsymbol{\beta}_J^*\big) &= \boldsymbol{\Gamma}(J)^{-1}\bigg(\sum_{l\notin J}\rho_{j_1 l}\beta_l^*, \dots, \sum_{l\notin J}\rho_{j_q l}\beta_l^*\bigg)^\top \\ &= \sum_{l\notin J}\beta_l^*\boldsymbol{\Gamma}(J)^{-1}\boldsymbol{g}_l(J).\end{aligned}$$

Taking the L_1-norm on both sides, (2.14) (with $C = M + 1$) follows from (2.16) and (2.13).

Before closing this section, we remark that (2.12) can be viewed as a limiting case of (2.11). To see this, note that (2.12) implies that for any $J \subseteq \mathcal{P}$, $\sum_{j\in J}|\beta_j^*| \le M_1 \times (\sum_{j\in J}\beta_j^{*2})^{\lim_{\gamma\to\infty}(\gamma-1)/(2\gamma-1)}$. In addition, the strong sparsity condition,

$$k_0 = \sharp(N_n) < M_7, \tag{2.17}$$

where $N_n = \{j : \beta_j^* \ne 0, 1 \le j \le p\}$ and M_7 is some positive integer, is also a limiting case of (2.11) because (2.17) yields that for any $J \subseteq \mathcal{P}$, $\sum_{j\in J}|\beta_j^*| \le M_7^{1/2} \times (\sum_{j\in J}\beta_j^{*2})^{\lim_{\gamma\to\infty}(\gamma-1)/(2\gamma-1)}$.

2.2. *Rates of convergence of the OGA.* Let $\mathbf{x} = (x_1, \dots, x_p)^\top$ be independent of and have the same covariance structure as $\{\mathbf{x}_t\}$ and $y(\mathbf{x}) = \sum_{j=1}^{p}\beta_j x_j$. Then $y(\mathbf{x})$ can be predicted by $\hat{y}_m(\mathbf{x}) = \mathbf{x}^\top(\hat{J}_m)\hat{\boldsymbol{\beta}}(\hat{J}_m)$, where $\mathbf{x}(J) = (x_i, i \in J)$ and $\hat{\boldsymbol{\beta}}(J) = (\sum_{t=1}^{n}\mathbf{x}_t(J)\mathbf{x}_t^\top(J))^{-1} \times \sum_{t=1}^{n}\mathbf{x}_t(J)y_t$. Note also that $\hat{y}_m(\mathbf{x}) = \mathbf{z}^\top(\hat{J}_m)\hat{\boldsymbol{\beta}}^*(\hat{J}_m)$, where $\mathbf{z}(J) = (z_i, i \in J)$ with $z_i = x_i/\sigma_i$, and $\hat{\boldsymbol{\beta}}^*(\hat{J}_m) = (\sum_{t=1}^{n}\mathbf{z}_t(J)\mathbf{z}_t^\top(J))^{-1}\sum_{t=1}^{n}\mathbf{z}_t(J)y_t$. One of the most natural performance measures for $\hat{y}_m(\mathbf{x})$ is the conditional mean squared prediction error (CMSPE),

$$\mathrm{E}_n\{y(\mathbf{x}) - \hat{y}_m(\mathbf{x})\}^2 = \mathrm{E}_n\{y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})\}^2 + \mathrm{E}_n\{y_{\hat{J}_m}(\mathbf{x}) - \hat{y}_m(\mathbf{x})\}^2, \tag{2.18}$$

where $\mathrm{E}_n(\cdot) = \mathrm{E}(\cdot|y_1, \mathbf{x}_1, \dots, y_n, \mathbf{x}_n)$, and $y_J(\mathbf{x}) = \boldsymbol{\beta}^{*\top}(J)\mathbf{z}(J)$. A convergence rate of the left-hand side of (2.18) is established in the next theorem.

THEOREM 2.1. *Suppose that* (1.1), (A1)–(A3), (A5),

$$\lambda_{\min}(\boldsymbol{\Gamma}) \ge \lambda_1 > 0 \tag{2.19}$$

and

$$\log p = o(n), \quad K_n = \bar{\delta}\bigg(\frac{n}{\log p}\bigg)^{1/2} \tag{2.20}$$

hold, where $\boldsymbol{\Gamma} = \mathrm{E}(\mathbf{z}\mathbf{z}^\top)$ *and* $0 < \bar{\delta} < \min\{\bar{\tau}, \bar{D}\}$, *with* $\bar{D}$ *defined in assumption* (A5) *and*

$$\begin{aligned}\bar{\tau} &= \sup \boldsymbol{\tau} \\ &\equiv \sup\bigg\{\tau : \tau > 0, \limsup_{n\to\infty}\frac{\tau c_2^*}{\min_{\sharp(J)\le\tau(n/\log p)^{1/2}}\lambda_{\min}(\boldsymbol{\Gamma}(J))} \le 1\bigg\}.\end{aligned} \tag{2.21}$$

Then

$$\max_{1\le m\le K_n}\bigg(\frac{\mathrm{E}_n\{y(\mathbf{x}) - \hat{y}_m(\mathbf{x})\}^2}{m^{-2\gamma+1} + n^{-1}m\log p}\bigg) = O_p(1). \tag{2.22}$$

Moreover, if (A4) *holds instead of* (A3), *then*

$$\max_{1\le m\le K_n}\bigg(\frac{\mathrm{E}_n\{y(\mathbf{x}) - \hat{y}_m(\mathbf{x})\}^2}{\exp(-G_3 m) + n^{-1}m\log p}\bigg) = O_p(1), \tag{2.23}$$

where G_3 *is some positive constant given in* (A.2) *in the Appendix.*

PROOF. We first prove (2.22). Recall $\hat{J}_k = \{\hat{j}_1, \dots, \hat{j}_k\}$ and define

$$\hat{\mu}_{J,i} = \frac{n^{-1}\sum_{t=1}^{n}(y_t - \hat{y}_{t;J})x_{ti}}{(n^{-1}\sum_{t=1}^{n}x_{ti}^2)^{1/2}} = \frac{n^{-1}\sum_{t=1}^{n}(y_t - \hat{y}_{t;J})z_{ti}}{(n^{-1}\sum_{t=1}^{n}z_{ti}^2)^{1/2}},$$

where $(\hat{y}_{1;J}, \dots, \hat{y}_{n;J})^\top = \mathbf{H}_J\mathbf{Y}$. Moreover, let

$$A_n(m) = \Big\{\max_{(J,i):\sharp(J)\le m-1, i\notin J} |\hat{\mu}_{J,i} - \mu_{J,i}| \le s(\log p/n)^{1/2}\Big\}$$

and

$$B_n(m) = \Big\{\min_{0\le i\le m-1}\max_{1\le j\le p} |\mu_{\hat{J}_i,j}| > \tilde{\xi}s(\log p/n)^{1/2}\Big\},$$

where $\mu_{J,i} = \mathrm{E}[(y(\mathbf{x}) - y_J(\mathbf{x}))z_i]$, $s > 0$ is some large constant, and $\tilde{\xi} = 2/(1-\xi)$ with $0 < \xi < 1$ being arbitrarily given.

By an argument similar to that of (3.10) in Ing and Lai (2011), it follows that for all $1 \le q \le m$,

$$|\mu_{\hat{J}_{q-1},\hat{j}_q}| \ge \xi \max_{1\le i\le p} |\mu_{\hat{J}_{q-1},i}| \quad \text{on } A_n(m)\cap B_n(m).$$

This and (A.1) in the Appendix, which gives an error bound for the population OGA under (A3), lead to

$$\mathrm{E}_n\big(y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})\big)^2 \le G_1 m^{-2\gamma+1} \quad \text{on } A_n(m)\cap B_n(m). \tag{2.24}$$

Moreover, (A3) and (2.19) imply that for any $0 \le i \le m-1$,

$$\begin{aligned}
&\mathrm{E}_n\big(y(\mathbf{x}) - y_{\hat{J}_i}(\mathbf{x})\big)^2 \\
&\quad\le \max_{1\le j\le p}|\mu_{\hat{J}_i,j}| \sum_{l=1,l\notin\hat{J}_i}^{p} |\beta_l^*| \\
&\quad\le C_\gamma \max_{1\le j\le p}|\mu_{\hat{J}_i,j}| \bigg(\sum_{l=1,l\notin\hat{J}_i}^{p} \beta_l^{*2}\bigg)^{(\gamma-1)/(2\gamma-1)} \\
&\quad\le C_\gamma \max_{1\le j\le p}|\mu_{\hat{J}_i,j}| \lambda_1^{-(\gamma-1)/(2\gamma-1)} \big(\mathrm{E}_n\big(y(\mathbf{x}) - y_{\hat{J}_i}(\mathbf{x})\big)^2\big)^{(\gamma-1)/(2\gamma-1)},
\end{aligned}$$

and hence

$$\mathrm{E}_n\big(y(\mathbf{x}) - y_{\hat{J}_i}(\mathbf{x})\big)^2 \le \Big(C_\gamma \max_{1\le j\le p}|\mu_{\hat{J}_i,j}|\Big)^{2-\gamma^{-1}} \lambda_1^{-1+\gamma^{-1}}. \tag{2.25}$$

By (2.25),

$$\begin{aligned}
&\mathrm{E}_n\big(y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})\big)^2 \\
&\quad\le \min_{0\le i\le m-1} \mathrm{E}_n\big(y(\mathbf{x}) - y_{\hat{J}_i}(\mathbf{x})\big)^2 \\
&\quad\le C_\gamma^{2-\gamma^{-1}} \lambda_1^{-1+\gamma^{-1}} \Big(\min_{0\le i\le m-1}\max_{1\le j\le p}|\mu_{\hat{J}_i,j}|\Big)^{2-\gamma^{-1}} \\
&\quad\le C_\gamma^{2-\gamma^{-1}} \lambda_1^{-1+\gamma^{-1}} (\tilde{\xi}s)^{2-\gamma^{-1}} \big(n^{-1}\log p\big)^{1-(2\gamma)^{-1}} \quad \text{on } B_n^c(m).
\end{aligned} \tag{2.26}$$

Since $A_n(m)$ decreases as m increases, (2.24) and (2.26) yield that for all $1 \le m \le K_n$ and some $C_2 > 0$,

$$\mathrm{E}_n\big(y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})\big)^2 I_{A_n(K_n)} \le C_2 \max\big\{m^{-2\gamma+1}, \{n^{-1}\log p\}^{1-(2\gamma)^{-1}}\big\}. \tag{2.27}$$

We show in Section S1 of the Supplementary Material that

$$(2.28)\qquad P\big(\|\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_{K_n})\| \le \bar{B}\big) = 1 + o(1),$$

where $\hat{\boldsymbol{\Gamma}}(J) = n^{-1}\sum_{t=1}^{n}\mathbf{z}_t(J)\mathbf{z}_t^{\top}(J)$ and

$$(2.29)\qquad \bar{B} > \frac{1}{\liminf_{n\to\infty}\min_{\sharp(J)\le K_n}\lambda_{\min}(\boldsymbol{\Gamma}(J)) - c_2^*\bar{\bar{\delta}}},$$

noting that the positiveness of the denominator is ensured by (2.20) and (2.21). With the help of (2.28), (A1), (A2) and (A5), it is shown in the same section that there exists a sufficiently large s such that

$$(2.30)\qquad \lim_{n\to\infty} P\big(A_n(K_n)\big) = 1,$$

which, together with (2.27), yields

$$(2.31)\qquad \max_{1\le m\le K_n}\frac{\mathrm{E}_n(y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x}))^2}{\max\{m^{-2\gamma+1}, (n^{-1}\log p)^{1-(2\gamma)^{-1}}\}} = O_p(1).$$

Moreover, we have

$$(2.32)\qquad \max_{1\le m\le K_n}\frac{n\mathrm{E}_n[\{\hat{y}_m(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})\}^2]}{m\log p} = O_p(1),$$

which is also proved in Section S1 of the Supplementary Material. In view of (2.31), (2.32) and the fact that $(\log p/n)^{1-(2\gamma)^{-1}} \le m^{-2\gamma+1}$ if $m \le (n/\log p)^{(2\gamma)^{-1}}$ and $(\log p/n)^{1-(2\gamma)^{-1}} \le n^{-1}m\log p$ if $m \ge (n/\log p)^{(2\gamma)^{-1}}$, the desired conclusion (2.22) follows. Equation (2.23) follows from (A.2) in the Appendix (which gives an error bound for the population OGA under (A4)) and an argument similar to that used to prove (2.22). We skip the details in order to save space. □

REMARK 1. It is easy to see that the $\boldsymbol{\tau}$ defined in (2.21) is nonempty. In particular, $x \in \boldsymbol{\tau}$ for any $x \in (0, \lambda_1/c_2^*]$. It is also not difficult to see that $\bar{\tau} < a/c_2^*$ for any $a > 1$.

In view of (2.22), to strike a suitable balance between squared bias and variance, one should choose $m \approx (n/\log p)^{1/2\gamma}$ in the polynomial decay case, which yields a rate of convergence, $(n^{-1}\log p)^{1-(2\gamma)^{-1}}$. Similarly, (2.23) suggests that the best convergence rate one can expect in the exponential decay case is $n^{-1}\log n\log p$, which is ensured by selecting $m \approx \log n/G_3$. The optimality of the rates, $(n^{-1}\log p)^{1-(2\gamma)^{-1}}$ and $n^{-1}\log n\log p$, will be discussed further in Section 3.2. In most practical situations, however, not only do we not know what γ or G_3 is, we do not even know which of (A3) and (A4) is true. To attain the aforementioned optimal convergence rates without knowing the degree of sparseness, a data-driven method to determine the number of OGA iterations is called for. In the next section, we show that HDAIC (see (3.1)) can fulfill this need.

Finally, we note that if (2.2) and (2.3) are weakened to

$$(2.33)\qquad P\left(\max_{1\le j\le p}\left|n^{-1}\sum_{i=1}^{n} z_{ij}\varepsilon_j\right| \ge c_1^*(\log p)^{(1+\bar{c}_1)/2}/n^{1/2}\right) = o(1)$$

and

$$(2.34)\qquad P\left(\max_{1\le k,l\le p}\left|n^{-1}\sum_{i=1}^{n} z_{ik}z_{il} - \rho_{kl}\right| \ge c_2^*(\log p)^{(1+\bar{c}_2)/2}/n^{1/2}\right) = o(1),$$

respectively, where $0 \le \bar{c}_1, \bar{c}_2 < \infty$ are some constants, and (2.20) is strengthened to

$$(\log p)^{1+\bar{c}} = o(n) \quad \text{and} \quad K_n = \underline{\delta} \frac{n^{1/2}}{(\log p)^{(1+\bar{c})/2}}, \tag{2.35}$$

where $\bar{c} = \max\{\bar{c}_1, \bar{c}_2\}$ and $\underline{\delta}$ is some positive constant, then (2.22) and (2.23) become

$$\max_{1 \le m \le K_n} \left(\frac{\mathrm{E}_n\{y(\mathbf{x}) - \hat{y}_m(\mathbf{x})\}^2}{m^{-2\gamma+1} + n^{-1}m(\log p)^{1+\bar{c}}} \right) = O_p(1) \tag{2.36}$$

and

$$\max_{1 \le m \le K_n} \left(\frac{\mathrm{E}_n\{y(\mathbf{x}) - \hat{y}_m(\mathbf{x})\}^2}{\exp(-G_3 m) + n^{-1}m(\log p)^{1+\bar{c}}} \right) = O_p(1), \tag{2.37}$$

respectively. While (2.33) and (2.34) are satisfied by a broader class of time series models (see Wu and Wu (2016) for a detailed discussion), to determine the optimal m in (2.36) or (2.37), the HDAIC must also be corrected according to the value of $\bar{c}$. This kind of correction, however, is hardly implemented in practice because $\bar{c}$ is in general unknown.

3. Analysis of OGA+HDAIC.

In Section 3.1, the rate of convergence of OGA+HDAIC is established under various sparsity conditions; see Theorem 3.1. Comparisons of Theorem 3.1 and related existing results are given in Section 3.2. The proof of Theorem 3.1 is provided in Section 3.3.

3.1. *Error bounds for OGA+HDAIC.*

Define

$$\mathrm{HDAIC}(J) = \left(1 + \frac{s_a \sharp(J) \log p}{n}\right) \hat{\sigma}_J^2, \tag{3.1}$$

where $\hat{\sigma}_J^2 = n^{-1}\mathbf{Y}^\top(\mathbf{I} - \mathbf{H}_J)\mathbf{Y}$ and s_a is some positive constant, and define

$$\hat{k}_n = \arg \min_{1 \le k \le K_n} \mathrm{HDAIC}(\hat{J}_k),$$

noting that $\hat{J}_k$ is defined in (2.1).

THEOREM 3.1. *Suppose that* (1.1), (A1), (A2), (A5), (2.19), (2.20) *and*

$$n^{-1} \sum_{t=1}^{n} \varepsilon_t^2 = \sigma^2 + o_p(1) \tag{3.2}$$

hold. Then, for

$$s_a > \bar{V}_0 \equiv \frac{2\bar{B}(c_1^{*^2} + c_2^{*^2})}{\sigma^2}, \tag{3.3}$$

where $\bar{B}$ is defined in (2.29), *we have*:

(i)

$$\frac{\mathrm{E}_n(y(\mathbf{x}) - \hat{y}_{\hat{k}_n}(\mathbf{x}))^2}{(\frac{\log p}{n})^{1-1/2\gamma}} = O_p(1), \tag{3.4}$$

provided (A3) *is true*;

(ii)

$$\frac{\mathrm{E}_n(y(\mathbf{x})-\hat{y}_{\hat{k}_n}(\mathbf{x}))^2}{\frac{\log n \log p}{n}} = O_p(1), \tag{3.5}$$

provided (A4) *is true and* $\log p = o(n/(\log n)^2)$;

(iii)

$$\frac{\mathrm{E}_n(y(\mathbf{x})-\hat{y}_{\hat{k}_n}(\mathbf{x}))^2}{\frac{k_0 \log p}{n}} = O_p(1), \tag{3.6}$$

provided $\mathrm{E}(y_t^2)$ *is bounded above by a finite constant and*

$$\min_{j\in N_n} |\beta_j^*| \geq \underline{\theta}, \quad \textit{for some } \underline{\theta} > 0,$$
$$k_0\Big(\sum_{j\in N_n} |\beta_j^*|\Big)^2 = o(n/\log p). \tag{3.7}$$

REMARK 2. The sparsity condition (3.7) implies $k_0 = o((n/\log p)^{1/3})$, allowing k_0 to grow to ∞ slowly with n. Moreover, (3.6) also holds when (2.19) is weakened to

$$\min_{\sharp(J)\leq \eta(n/\log p)^{1/2}} \lambda_{\min}(\mathbf{\Gamma}(J)) \geq \lambda_1, \tag{3.8}$$

for some $\eta > 0$; see Section S2 in the Supplementary Material. However, since it is unknown which kind of sparsity condition is true among those described in (i), (ii) and (iii) of Theorem 3.1, and since (2.19) appears to be indispensable for the proofs of (3.4) and (3.5), the latter assumption is still adopted in our unified theory.

REMARK 3. We briefly discuss extensions of Theorems 2.1 and 3.1 to the following multivariate time series models:

$$\boldsymbol{y}_t = \sum_{l=1}^{p} \boldsymbol{b}_j x_{tj} + \boldsymbol{e}_t, \quad t = 1, \ldots, n, \tag{3.9}$$

where $\boldsymbol{y}_t$, $\boldsymbol{e}_t$, and $\boldsymbol{b}_j$ are d-dimensional vectors, d is allowed to grow to infinity with n and $\{(\boldsymbol{e}_t^\top, \mathbf{x}_t^\top)^\top\}$ is a zero-mean stationary time series satisfying $\mathrm{E}(\mathbf{x}_t\boldsymbol{e}_t^\top) = \mathbf{0}$. Define $\hat{\boldsymbol{\psi}}_{J,i} = \|\mathcal{Y}^\top(\mathbf{I}-\mathbf{H}_J)\mathbf{Z}_i^\top\|/(n^{1/2}\|\mathbf{Z}_i\|)$, where $\mathcal{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^\top$, and $\mathbf{H}_J$ and $\mathbf{Z}_i$ are defined as in Section 2.1. A multivariate version of OGA, MOGA, is initialized by $\hat{L}_0 = \varnothing$. For $m \geq 1$, $\hat{L}_m$ is recursively updated by

$$\hat{L}_m = \hat{L}_{m-1} \cup \{\hat{l}_m\},$$

where $\hat{l}_m = \arg\max_{1\leq l\leq p, l\notin \hat{L}_{m-1}} \hat{\boldsymbol{\psi}}_{\hat{L}_{m-1},l}$. Consider multivariate extensions of the sparsity conditions (A3) and (A4):

(A3′) There is $0 < \bar{M}_0 < \infty$ such that

$$d^{-1}\sum_{j=1}^{p} \|\boldsymbol{b}_j^*\|^2 < \bar{M}_0.$$

Moreover, there exist $\gamma \geq 1$ and $0 < C_\gamma < \infty$ such that for any $J \subseteq \mathcal{P}$,

$$\sum_{j\in J} \|\boldsymbol{b}_j^*\|/d^{1/2} \leq C_\gamma \Big\{\sum_{j\in J} \|\boldsymbol{b}_j^*\|^2/d\Big\}^{(\gamma-1)/(2\gamma-1)},$$

where $\boldsymbol{b}_j^* = \sigma_j \boldsymbol{b}_j$.

(A4′) There is $0 < M_0 < \infty$ such that

$$\max_{1\le j\le p} \|\boldsymbol{b}_j^*\| < d^{1/2} M_0.$$

Moreover, there exists $M_1 > 1$ such that for any $J \subseteq \mathcal{P}$,

$$\sum_{j\in J} \|\boldsymbol{b}_j^*\| \le M_1 \max_{j\in J} \|\boldsymbol{b}_j^*\|.$$

Moreover, a natural generalization of (A1) under model (3.9) is

(A1′) There exists $c_1^* > 0$ such that

$$P\left(\max_{1\le j\le p, 1\le l\le d} \left| n^{-1} \sum_{t=1}^{n} z_{tj}\varepsilon_{tl} \right| \ge c_1^* (\log pd)^{1/2}/n^{1/2}\right) = o(1),$$

where $(\varepsilon_{t1}, \dots, \varepsilon_{td})^\top = \boldsymbol{e}_t$.

Let $\mathbf{x} = (x_1, \dots, x_p)^\top$ be defined as in Section 2.2 and $\boldsymbol{y}(\mathbf{x}) = \sum_{j=1}^{p} \boldsymbol{b}_j x_j$. Then, $\boldsymbol{y}(\mathbf{x})$ can be predicted by $\hat{\boldsymbol{y}}_m(\mathbf{x}) = \hat{\boldsymbol{B}}(\hat{L}_m)^\top \mathbf{x}(\hat{L}_m)$, where $\hat{\boldsymbol{B}}(J) = (\sum_{t=1}^{n} \mathbf{x}_t(J) \times \mathbf{x}_t^\top(J))^{-1} \sum_{t=1}^{n} \mathbf{x}_t(J) \times \boldsymbol{y}_t^\top$. Suppose that

$$\log pd = o(n) \quad \text{and} \quad K_n = \zeta(n/\log pd)^{1/2}, \tag{3.10}$$

for some $\zeta > 0$. Then, under (3.10) and the assumptions of Theorem 2.1, with (A1), (A3) and (A4) replaced by (A1′), (A3′) and (A4′), it can be shown that

$$\max_{1\le m\le K_n} \left(\frac{d^{-1}\mathrm{E}_n \|\boldsymbol{y}(\mathbf{x}) - \hat{\boldsymbol{y}}_m(\mathbf{x})\|^2}{m^{-2\gamma+1} + n^{-1} m \log pd}\right) = O_p(1), \tag{3.11}$$

and for some $G_4 > 0$,

$$\max_{1\le m\le K_n} \left(\frac{d^{-1}\mathrm{E}_n \|\boldsymbol{y}(\mathbf{x}) - \hat{\boldsymbol{y}}_m(\mathbf{x})\|^2}{\exp(-G_4 m) + n^{-1} m \log pd}\right) = O_p(1). \tag{3.12}$$

To choose a suitable number of MOGA iterations, one may consider a multivariate extension of HDAIC (MHDAIC),

$$\mathrm{MHDAIC}(J) = \left(1 + \frac{\iota_a \sharp(J) \log pd}{n}\right) \hat{\boldsymbol{\Sigma}}_J,$$

where $\hat{\boldsymbol{\Sigma}}_J = (nd)^{-1} \operatorname{tr}(\mathcal{Y}^\top (\mathbf{I} - \mathbf{H}_J)\mathcal{Y})$ and ι_a is some positive constant, and define

$$\hat{m}_n = \arg \min_{1\le m\le K_n} \mathrm{MHDAIC}(\hat{L}_m).$$

We conjecture that $d^{-1}\mathrm{E}_n \|\boldsymbol{y}(\mathbf{x}) - \hat{\boldsymbol{y}}_{\hat{m}_n}(\mathbf{x})\|^2$ is of order $O_p((\log pd/n)^{1-1/(2\gamma)})$, $O_p(\log n \log pd/n)$, or $O_p(k_0 \log pd/n)$ under (A3′), (A4′), or a strong sparsity condition resembling (3.7), respectively. However, the rigorous proof of this result and those of (3.11) and (3.12) are out of the scope of this paper, and are left for future work.

3.2. *Some comparisons with existing results.* It would be interesting to compare (3.4) with Corollary 3 of Negahban et al. (2012), which provides an error bound for Lasso in the following high-dimensional regression model:

$$y_t = \sum_{j=1}^{p} \beta_j^* x_{tj} + \epsilon_t, \quad t = 1, \dots, n, \tag{3.13}$$

where $\{\epsilon_t\}$ is a sequence of i.i.d. $N(0,\sigma^2)$ random variables and $\{x_{tj}\}$ are nonrandom constants satisfying $n^{-1}\sum_{t=1}^n x_{tj}^2 \le 1$, $1 \le j \le p$, and the restricted eigenvalue condition defined in (31) of their paper. When

$$\sum_{j=1}^{p} |\beta_j^*|^{1/\gamma} \le (n/\log p)^{1-1/(2\gamma)}, \tag{3.14}$$

for some $\gamma \ge 1$, it is shown in the corollary that

$$\|\hat{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^*\|^2 = O_p\left(\sum_{j=1}^{p} |\beta_j^*|^{1/\gamma} \left(\frac{\log p}{n}\right)^{1-1/(2\gamma)}\right), \tag{3.15}$$

where $\hat{\boldsymbol{\beta}}_{\lambda_n}$ is the Lasso estimate of $\boldsymbol{\beta}^*$ with $\lambda_n = 4\sigma(\log p/n)^{1/2}$. On the other hand, (3.4) implies that under model (1.1),

$$\begin{aligned} \|\hat{\boldsymbol{\beta}}_n(\hat{J}_{\hat{k}_n}) - \boldsymbol{\beta}^*\|^2 &\le \lambda_1^{-1} \mathrm{E}_n\big(y(\mathbf{x}) - \hat{y}_{\hat{k}_n}(\mathbf{x})\big)^2 \\ &= O_p\left(\left(\frac{\log p}{n}\right)^{1-1/(2\gamma)}\right). \end{aligned} \tag{3.16}$$

In addition to allowing for serially correlated data, (3.16) may lead to a faster convergence rate than (3.15). In particular, the bound on the right-hand side of (3.15) is larger than that on the right-hand side of (3.16) by a factor of $\log p$ as $p \to \infty$ when (1.4), with $\gamma > 1$, and (3.14) follows.

Assuming that the $\{\mathbf{x}_t\}$ and $\{\epsilon_t\}$ in (3.13) are generated according to independent, centered, Gaussian stationary time series, Proposition 3.3 of Basu and Michailidis (2015) establishes for Lasso the following bounds:

$$\|\hat{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^*\|^2 = O_p\left(\frac{k_0 \log p}{n}\right) \tag{3.17}$$

and

$$n^{-1}\sum_{t=1}^{n} \big(\mathbf{x}_t^\top(\hat{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^*)\big)^2 = O_p\left(\frac{k_0 \log p}{n}\right), \tag{3.18}$$

where $p \to \infty$, $k_0 = O(n/\log p)$, and $\lambda_n \ge c^*(\log p/n)^{1/2}$ for some $c^* > 0$. By (3.6) and an argument used in Section S2 of the Supplementary Material, it can be shown that under model (1.1),

$$\|\hat{\boldsymbol{\beta}}(\hat{J}_{\hat{k}_n}) - \boldsymbol{\beta}^*\|^2 = O_p\left(\frac{k_0 \log p}{n}\right) \tag{3.19}$$

and

$$n^{-1}\sum_{t=1}^{n} \big(\mathbf{x}_t^\top(\hat{\boldsymbol{\beta}}(\hat{J}_{\hat{k}_n}) - \boldsymbol{\beta}^*)\big)^2 = O_p\left(\frac{k_0 \log p}{n}\right). \tag{3.20}$$

Although (3.17)–(3.20) suggest that Lasso and OGA+HDAIC share the same error rate in the case of $k_0 \ll n$, they are obtained under somewhat different assumptions. Note first that unlike (3.17) and (3.18), (3.19) and (3.20) do not require that $\{\mathbf{x}_t\}$ and $\{\varepsilon_t\}$ are independent, and hence are applicable to ARX models. Moreover, (3.17) and (3.18) are established under

$$\operatorname*{ess\,sup}_{\theta\in[-\pi,\pi]} \lambda_{\max}\big(f_{\mathbf{x}}(\theta)\big) < \bar{S}, \tag{3.21}$$

and

$$\operatorname*{ess\,inf}_{\theta\in[-\pi,\pi]} \lambda_{\min}\big(f_{\mathbf{x}}(\theta)\big) > \underline{s}, \tag{3.22}$$

where $0 < \underline{s} \le \bar{S} < \infty$ and $f_{\mathbf{x}}(\theta) = [1/(2\pi)]\sum_{l=-\infty}^{\infty} \boldsymbol{\Gamma}_{\mathbf{x}}(l)\exp(-il\theta)$ with $\boldsymbol{\Gamma}_{\mathbf{x}}(l) = \mathrm{E}(\mathbf{x}_t \times \mathbf{x}_{t+l}^{\top})$. Assumption (3.22) is comparable to (2.19) (which assumes that $\lambda_{\min}(\boldsymbol{\Gamma})$ is bounded away from zero and is needed for proving (3.19) and (3.20)), but is more stringent than the latter because

$$\begin{aligned}\lambda_{\min}(\boldsymbol{\Gamma}) = \lambda_{\min}\big(\boldsymbol{\Gamma}_{\mathbf{x}}(0)\big) &= \lambda_{\min}\bigg(\int_{-\pi}^{\pi} f_{\mathbf{x}}(\theta)\,d\theta\bigg) \\ &\ge 2\pi\Big\{\operatorname*{ess\,inf}_{\theta\in[-\pi,\pi]} \lambda_{\min}\big(f_{\mathbf{x}}(\theta)\big)\Big\}.\end{aligned}$$

Maximum eigenvalue assumptions like (3.21) are not required for (3.19) and (3.20). This type of assumption can be easily violated when the components of $\mathbf{x}_t$ are highly correlated, as illustrated by an ARX example in Section S3 of the Supplementary Material, in which $\lambda_{\max}(\boldsymbol{\Gamma}) \to \infty$ as $p \to \infty$, and hence $\operatorname{ess\,sup}_{\theta\in[-\pi,\pi]}\lambda_{\max}(f_{\mathbf{x}}(\theta)) \ge [1/(2\pi)]\lambda_{\max}(\boldsymbol{\Gamma})$. On the other hand, while (3.19) and (3.20) are obtained under the beta-min condition given in (3.7), (3.17) and (3.18) do not assume any beta-min condition. Wu and Wu (2016) also investigate the performance of Lasso under (3.13) with $k_0 \ll n$ and $\{x_{ti}\}$ being nonrandom and obeying the restricted eigenvalue condition defined in (4.2) of their paper. They allow $\{\epsilon_t\}$ to be a stationary process following some general moment and dependence conditions. The error rates that they derive for Lasso, however, are usually larger than those in (3.17)–(3.20).

In fact, it can be argued that all error bounds obtained in Theorem 3.1 are rate optimal. To see this, let $\hat{J}(m)$, $1 \le m \le K_n$, be a sequence of nested models chosen from p candidate variables in a data-driven fashion, where $\sharp(\hat{J}(m)) = m$. The CMSPE of model $\hat{J}(m)$ is $\mathrm{E}_n(y(\mathbf{x}) - \hat{y}_{\hat{J}(m)}(\mathbf{x}))^2 = \mathrm{E}_n(y(\mathbf{x}) - y_{\hat{J}(m)}(\mathbf{x}))^2 + \mathrm{E}_n(\hat{y}_{\hat{J}(m)}(\mathbf{x}) - y_{\hat{J}(m)}(\mathbf{x}))^2$, where $\hat{y}_J(\mathbf{x}) = \mathbf{x}^{\top}(J)\hat{\boldsymbol{\beta}}(J)$. It is not difficult to show that the squared bias terms obey

$$\mathrm{E}_n\big(y(\mathbf{x}) - y_{\hat{J}(m)}(\mathbf{x})\big)^2 \ge \mathrm{E}\big(y(\mathbf{x}) - y_{J_m^*}(\mathbf{x})\big)^2, \tag{3.23}$$

where $y_{J_m^*}(\mathbf{x})$, satisfying $\sharp(J_m^*) = m$ and

$$\mathrm{E}\big(y(\mathbf{x}) - y_{J_m^*}(\mathbf{x})\big)^2 = \min_{\sharp(J)=m} \mathrm{E}\big(y(\mathbf{x}) - y_J(\mathbf{x})\big)^2, \tag{3.24}$$

is called the best m-term approximation of $y(\mathbf{x})$. In addition, an argument similar to that used to prove (2.32) implies that the variance terms satisfy

$$\max_{1\le m\le K_n} \frac{n\mathrm{E}_n(\hat{y}_{\hat{J}(m)}(\mathbf{x}) - y_{\hat{J}(m)}(\mathbf{x}))^2}{m\log p} = O_p(1). \tag{3.25}$$

In view of (3.23) and (3.25), the best possible rate that can be achieved by a forward inclusion method accompanied by a stopping criterion is the same as that of

$$\bar{L}_n(m_n^*) \equiv \min_{1\le m\le K_n} \bar{L}_n(m) = \min_{1\le m\le K_n} \big\{\mathrm{E}\big(y(\mathbf{x}) - y_{J_m^*}(\mathbf{x})\big)^2 + m\log p/n\big\}. \tag{3.26}$$

According to Lemma A.3, (A.11) and $\mathrm{E}(y(\mathbf{x}) - y_{J_m^*}(\mathbf{x}))^2 = 0$ if $m \ge k_0$, the convergence rate of $\bar{L}_n(m_n^*)$ under (1.4), (1.5) or (2.17) is $(\log p/n)^{1-1/2\gamma}$, $\log n \log p/n$, or $k_0\log p/n$, which coincides with that of (3.4), (3.5) or (3.6), respectively. We therefore conclude that the bounds obtained in Theorem 3.1 are rate optimal. In this connection, we also note that when (1.1) is a stationary AR(p) model with $p \gg n$, the set of candidate models are usually given by AR(1), …, AR(K_n), with K_n approaching ∞ at a rate slower than n. Unlike $\hat{J}(m)$, $1 \le m \le$

K_n, the candidate set in this case is not determined by any data-driven methods, and hence the corresponding variance terms can get rid of the variance inflation factor $\log p$ (see (3.25)), which is introduced by data-dependent selection of the candidate set from all p variables. As a result, the optimal rate that can be attained by an order selection criterion is equivalent to that of

$$\min_{1\le m\le K_n}\{\mathrm{E}\big(y(\mathbf{x})-y_{J_m^*}(\mathbf{x})\big)^2+m/n\}; \tag{3.27}$$

see Shibata (1980) for more details. Under (1.2), (1.3) or (2.17) with $N_n=\{1,\dots,k_0\}$, the convergence rate of (3.27) is $(1/n)^{1-1/2\gamma}$, $\log n/n$ or k_0/n, which differs by a factor of $(\log p)^{1-1/2\gamma}$ from that of $\bar{L}_n(m_n^*)$ under (1.4), (1.5) or (2.17), respectively.

We would also like to point out the differences between the current paper and the paper by Ing and Lai (2011), which investigates the performance of OGA under (1.1) with $(\mathbf{x}_t,\varepsilon_t)$ being i.i.d. and obeying sub-Gaussian or subexponential distributions. Note first that Theorem 1 of Ing and Lai (2011) can be understood as a special case of Theorem 2.1 when $\gamma=1$ and observations are independent over time. However, since the former theorem only focuses on the case of $\gamma=1$, its proof does not involve the approximation errors of the population OGA under general sparsity conditions such as those given in Lemma A.1 in the Appendix. Moreover, when $\gamma=1$ is known, the optimal rate, $(\log p/n)^{1/2}$, can be achieved by choosing $m=(n/\log p)^{1/2}$, without recourse to any data-driven method to help determine the number of iterations. Alternatively, Theorem 3.1 encompasses a much wider class of sparsity conditions, and demonstrates that HDAIC can automatically choose a suitable m, leading to the optimal balance between the squared bias term and the variance term, without knowing the degree of sparseness. Indeed, Theorem 4 of Ing and Lai (2011) has suggested using a high-dimensional information criterion (whose penalty is heavier than that of HDAIC) to decide the number of OGA iterations when the regression coefficients satisfy the strong sparsity condition, (2.17), and a beta-min condition. Theorem 5 of Ing and Lai (2011) further introduces a backward elimination method based on the aforementioned information criterion to remove possible redundant variables surviving the first two (variable) screening stages, and shows that the resultant set of variables is equivalent to N_n with probability tending to 1. Although the approaches adopted in both papers can be considered similar to a certain extent, their goals are entirely different. In particular, whereas Ing and Lai (2011) aim to establish selection consistency under the strong sparsity condition, this paper focuses on prediction efficiency under much more general sparsity conditions, which include the strong sparsity one as a special case. From a technical point of view, the main differences between the two papers are: (i) serial correlation is not allowed in Ing and Lai (2011); and (ii) the squared bias term in Theorems 4 (or Theorem 5) of Ing and Lai (2011) completely vanishes along the OGA path in the sense that

$$P\Big(\min_{1\le m\le K_n}\mathrm{E}_n\big(y(\mathbf{x})-y_{\hat{J}_m}(\mathbf{x})\big)^2=0\Big)\to 1\quad\text{as } n\to\infty,$$

which is ensured by the sure screening property of OGA under the strong sparsity condition (see Theorem 3 of Ing and Lai (2011)), but the squared bias term in Theorem 3.1 decays at a variety of unknown rates and can never be zero along the OGA path, making it much harder to pursue the bias-variance tradeoff along this data-driven path.

We close this section by mentioning that while condition (3.3) on s_a involves unknown parameters, we have introduced a data-driven method for determining s_a in Section S3 of the Supplementary Material, which is of practical relevance.

3.3. *Proof of Theorem* 3.1. We only prove (3.4). The proof of (3.5) is similar to that of (3.4), and hence is omitted. The proof of (3.6) is slightly different, and is deferred to the Supplementary Material because of space constraints. In the rest of the proof, a weaker restriction on the penalty term,

$$s_a > \bar{V}^* \equiv \frac{2\bar{B}c_1^{*^2}}{\sigma^2}, \tag{3.28}$$

is used instead of (3.3), although the latter one is required in the proof of (3.6).

By making use of (2.11), (2.14) (which is ensured by (2.13)) and (2.19), we show in Section B in the Appendix that for any $1 \le m \le K_n$,

$$\begin{aligned} &-C_{M,\gamma,\lambda_1} R_{1,p}\{\mathrm{E}_n(\varepsilon^2(\hat{J}_m))\}^{(2\gamma-2)/(2\gamma-1)} \\ &\quad \le n^{-1}\sum_{t=1}^n \varepsilon_t^2(\hat{J}_m) - \mathrm{E}_n(\varepsilon^2(\hat{J}_m)) \\ &\quad \le C_{M,\gamma,\lambda_1} R_{1,p}\{\mathrm{E}_n(\varepsilon^2(\hat{J}_m))\}^{(2\gamma-2)/(2\gamma-1)}, \end{aligned} \tag{3.29}$$

where $C_{M,\gamma,\lambda_1} = (M+1)^2 C_\gamma^2 \lambda_1^{-(2\gamma-2)/(2\gamma-1)}$, with M defined in (2.13), $R_{1,p} = \max_{1 \le i,l \le p} |n^{-1}\sum_{t=1}^n z_{ti} z_{tl} - \rho_{il}|$, $\varepsilon_t(J) = y_t - \varepsilon_t - \boldsymbol{\beta}^{*^\top}(J)\mathbf{z}_t(J)$ and $\varepsilon(J) = y(\mathbf{x}) - y_J(\mathbf{x}) = y(\mathbf{x}) - \boldsymbol{\beta}^{*^\top}(J)\mathbf{z}(J)$. In addition, it is shown in Section S2 of the Supplementary Material that

$$\begin{aligned} &\left| n^{-1}\sum_{t=1}^n \varepsilon_t \varepsilon_t(\hat{J}_m) \right| \\ &\quad \le C_{M,\gamma,\lambda_1}^{1/2} R_{2,p}\{\mathrm{E}_n(\varepsilon^2(\hat{J}_m))\}^{(\gamma-1)/(2\gamma-1)}, \end{aligned} \tag{3.30}$$

$$\begin{aligned} &\max_{1 \le m \le K_n} \frac{\|n^{-1}\sum_{t=1}^n \mathbf{z}_t(\hat{J}_m)\varepsilon_t(\hat{J}_m)\|^2_{\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_m)}}{m\{\mathrm{E}_n(\varepsilon^2(\hat{J}_m))\}^{(2\gamma-2)/(2\gamma-1)}} \\ &\quad \le C_{M,\gamma,\lambda_1} \|\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_{K_n})\| R_{1,p}^2, \end{aligned} \tag{3.31}$$

and

$$\max_{1 \le m \le K_n} \frac{\|n^{-1}\sum_{t=1}^n \mathbf{z}_t(\hat{J}_m)\varepsilon_t\|^2_{\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_m)}}{m} \le \|\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_{K_n})\| R_{2,p}^2, \tag{3.32}$$

where $R_{2,p} = \max_{1 \le i \le p} |n^{-1}\sum_{t=1}^n z_{ti}\varepsilon_t|$ and $\|\boldsymbol{v}\|_{\boldsymbol{A}}^2 = \boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v}$ for vector $\boldsymbol{v}$ and nonnegative definite matrix $\boldsymbol{A}$.

Let $m_n^* = \min\{(n/\log p)^{1/2\gamma}, K_n\}$ and

$$\tilde{k}_n = \min\{k : 1 \le k \le K_n, \mathrm{E}_n(\varepsilon^2(\hat{J}_k)) \le G m_n^{*^{-2\gamma+1}}\} \quad (\min \varnothing = K_n), \tag{3.33}$$

in which $G \gg C_2$ and C_2 is defined in (2.27). Using (3.29)–(3.32), we next show that

$$\lim_{n\to\infty} P(\hat{k}_n \le \tilde{k}_n - 1) = 0. \tag{3.34}$$

Since (2.27) implies $\mathrm{E}_n(\varepsilon^2(\hat{J}_{m_n^*})) \le C_2 m_n^{*^{-2\gamma+1}} \le G m_n^{*^{-2\gamma+1}}$ on $A_n(K_n)$, it follows that $m_n^* \ge \tilde{k}_n$ on $A_n(K_n)$. By (2.30), one obtains

$$\begin{aligned} &P(\hat{k}_n \le \tilde{k}_n - 1) \\ &\quad \le P(\hat{k}_n \le \tilde{k}_n - 1, A_n(K_n)) + P(A_n^c(K_n)) \\ &\quad \le P\left(\min_{1 \le k \le \tilde{k}_n - 1} Q_n(k) \le s_a m_n^*\left(n^{-1}\sum_{t=1}^n y_t^2\right)\log p/n, A_n(K_n)\right) + o(1), \end{aligned} \tag{3.35}$$

where

$$Q_n(k) = n^{-1}\sum_{t=1}^n \varepsilon_t^2(\hat{J}_k) + 2n^{-1}\sum_{t=1}^n \varepsilon_t(\hat{J}_k)\varepsilon_t - 2n^{-1}\sum_{t=1}^n \varepsilon_t(\hat{J}_{m_n^*})\varepsilon_t - n^{-1}\sum_{t=1}^n \varepsilon_t^2(\hat{J}_{m_n^*}) - \left\| n^{-1}\sum_{t=1}^n \mathbf{z}_t(\hat{J}_k)\big(\varepsilon_t + \varepsilon_t(\hat{J}_k)\big)\right\|^2_{\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_k)}.$$

By (2.2), (2.3) and (2.28),

$$\lim_{n\to\infty} P(W_n) = 1, \tag{3.36}$$

where

$$W_n = \{R_{1,p} \le c_2^*(\log p)^{1/2}/n^{1/2}\} \cap \{R_{2,p} \le c_1^*(\log p)^{1/2}/n^{1/2}\} \cap \{\|\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_{K_n})\| \le \bar{B}\}.$$

Moreover, (3.29)–(3.32), (2.28) and (2.27) imply that for $1 \le k \le \tilde{k}_n - 1$ and all large n,

$$n^{-1}\sum_{t=1}^n \varepsilon_t^2(\hat{J}_k) \ge \mathrm{E}_n\big(\varepsilon^2(\hat{J}_k)\big) \times \left\{1 - \frac{C_{M,\gamma,\lambda_1}c_2^*}{G^{1/(2\gamma-1)}}\left(I_{\{\gamma=1\}} + \left(\frac{\log p}{n}\right)^{(\gamma-1)/2\gamma} I_{\{\gamma>1\}}\right)\right\} \quad \text{on } W_n, \tag{3.37}$$

$$n^{-1}\left|\sum_{t=1}^n \varepsilon_t(\hat{J}_k)\varepsilon_t\right| \le \mathrm{E}_n\big(\varepsilon^2(\hat{J}_k)\big)\frac{C_{M,\gamma,\lambda_1}^{1/2}c_1^*}{G^{\gamma/(2\gamma-1)}} \quad \text{on } W_n, \tag{3.38}$$

$$n^{-1}\left|\sum_{t=1}^n \varepsilon_t(\hat{J}_{m_n^*})\varepsilon_t\right| \le \mathrm{E}_n\big(\varepsilon^2(\hat{J}_k)\big)\frac{C_{M,\gamma,\lambda_1}^{1/2}c_1^*}{G^{\gamma/(2\gamma-1)}} \quad \text{on } W_n \cap A_n(K_n), \tag{3.39}$$

$$n^{-1}\sum_{t=1}^n \varepsilon_t^2(\hat{J}_{m_n^*}) \le \mathrm{E}_n\big(\varepsilon^2(\hat{J}_k)\big) \times \left\{\frac{C_2}{G} + \frac{C_{M,\gamma,\lambda_1}c_2^*}{G^{1/(2\gamma-1)}}\left(I_{\{\gamma=1\}} + \left(\frac{\log p}{n}\right)^{(\gamma-1)/2\gamma} I_{\{\gamma>1\}}\right)\right\} \quad \text{on } W_n \cap A_n(K_n), \tag{3.40}$$

$$\left\| n^{-1}\sum_{t=1}^n \mathbf{z}_t(\hat{J}_k)\varepsilon_t(\hat{J}_k)\right\|^2_{\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_k)} \le \mathrm{E}_n\big(\varepsilon^2(\hat{J}_k)\big) \times \frac{C_{M,\gamma,\lambda_1}\bar{B}c_2^{*2}\bar{\delta}}{G^{1/(2\gamma-1)}}\left(I_{\{\gamma=1\}} + \left(\frac{\log p}{n}\right)^{(\gamma-1)/2\gamma} I_{\{\gamma>1\}}\right) \quad \text{on } W_n \cap A_n(K_n), \tag{3.41}$$

and

$$\left\| n^{-1} \sum_{t=1}^{n} \mathbf{z}_t(\hat{J}_k)\varepsilon_t \right\|^2_{\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_k)} \leq \mathrm{E}_n\big(\varepsilon^2(\hat{J}_k)\big)\frac{\bar{B}c_1^{*2}}{G} \quad \text{on } W_n \cap A_n(K_n). \tag{3.42}$$

By (3.37)–(3.42), it follows that for large enough G in (3.33), there exists $0 < \iota < 1/2$ such that for all large n,

$$\begin{aligned} \min_{1\leq k\leq \tilde{k}_n-1} Q_n(k) &\geq \min_{1\leq k\leq \tilde{k}_n-1} \mathrm{E}_n\big(\varepsilon^2(\hat{J}_k)\big)(1-\iota) \\ &\geq G m_n^{*^{-2\gamma+1}}(1-\iota) \quad \text{on } W_n \cap A_n(K_n). \end{aligned} \tag{3.43}$$

In addition, (2.2), (2.3), (A3) and $\log p/n \leq m_n^{*^{-2\gamma+1}}$ ensure that there exists $\bar{M}_2 > 0$ such that

$$\lim_{n\to\infty} P\left(\frac{s_a m_n^* n^{-1}\sum_{t=1}^n y_t^2}{n}\log p \leq \bar{M}_2 m_n^{*^{-2\gamma+1}}\right) = 1. \tag{3.44}$$

By (2.30), (3.36), (3.44) and selecting G in (3.43) larger than $2\bar{M}_2$, we obtain the desired conclusion (3.34).

Using (3.29)–(3.32) again, it is shown in Section S2 of the Supplementary Material that

$$\lim_{n\to\infty} P\big(\hat{k}_n \geq V m_n^*\big) = 0, \quad \gamma > 1, \tag{3.45}$$

where V is a sufficiently large constant to be specified in the proof of (3.45). With the help of (3.34) and (3.45), the desired conclusion follows if one can show that for $\gamma > 1$,

$$\mathrm{E}_n\big\{y(\mathbf{x}) - \hat{y}_{\hat{k}_n}(\mathbf{x})\big\}^2 I_{\{\tilde{k}_n \leq \hat{k}_n < V m_n^*\}} = O_p\big(m_n^{*^{-2\gamma+1}}\big), \tag{3.46}$$

and for $\gamma = 1$,

$$\mathrm{E}_n\big\{y(\mathbf{x}) - \hat{y}_{\hat{k}_n}(\mathbf{x})\big\}^2 I_{\{\tilde{k}_n \leq \hat{k}_n \leq K_n\}} = O_p\big((\log p/n)^{1/2}\big). \tag{3.47}$$

To show (3.46), note first that

$$\begin{aligned} &\mathrm{E}_n\big(y(\mathbf{x}) - \hat{y}_{\hat{k}_n}(\mathbf{x})\big)^2 I_{\{\tilde{k}_n \leq \hat{k}_n \leq V m_n^*\}} \\ &\quad \leq \mathrm{E}_n\varepsilon^2(\hat{J}_{\tilde{k}_n}) + \|\mathbf{L}(\hat{J}_{\hat{k}_n})\|^2 \|\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_{\hat{k}_n})\| I_{\{\tilde{k}_n \leq \hat{k}_n \leq V m_n^*\}} \\ &\qquad + \|\mathbf{L}(\hat{J}_{\hat{k}_n})\|^2 \|\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_{\hat{k}_n})\|^2 \|\hat{\mathbf{\Gamma}}(\hat{J}_{\hat{k}_n}) - \mathbf{\Gamma}(\hat{J}_{\hat{k}_n})\| I_{\{\tilde{k}_n \leq \hat{k}_n \leq V m_n^*\}}, \end{aligned} \tag{3.48}$$

where $\mathbf{L}(J) = n^{-1}\sum_{t=1}^n \mathbf{z}_t(J)(\varepsilon_t + \varepsilon_t(J))$. By (A3), (2.2), (2.3), (2.13) and straightforward algebraic manipulations, it holds that

$$\begin{aligned} &\|\mathbf{L}(\hat{J}_{\hat{k}_n})\|^2 I_{\{\tilde{k}_n \leq \hat{k}_n \leq V m_n^*\}} \\ &\quad \leq 2V m_n^* \max_{1\leq i\leq p}\left(n^{-1}\sum_{t=1}^n \varepsilon_t z_{ti}\right)^2 + 2V m_n^* \max_{1\leq i,j\leq p}\left(n^{-1}\sum_{t=1}^n z_{ti}z_{tj} - \rho_{ij}\right)^2 \\ &\qquad \times \left(\sum_{j=1}^p |\beta_j^*|\right)^2 \Big(1 + \max_{1\leq \sharp(J)\leq K_n, 1\leq l\leq p} \|\mathbf{\Gamma}^{-1}(J)\boldsymbol{g}_l(J)\|_1\Big)^2 \\ &\quad = O_p\left(\frac{m_n^*\log p}{n}\right) \\ &\quad = O_p\big(m_n^{*^{-2\gamma+1}}\big). \end{aligned} \tag{3.49}$$

Moreover, we have

$$\mathrm{E}_n \varepsilon^2(\hat{J}_{\tilde{k}_n}) \le \mathrm{E}_n \varepsilon^2(\hat{J}_{m_n^*}) \le C_2 m_n^{*^{-2\gamma+1}} \quad \text{on } A(K_n), \tag{3.50}$$

and

$$\begin{aligned} \|\hat{\boldsymbol{\Gamma}}(\hat{J}_{\hat{k}_n}) - \boldsymbol{\Gamma}(\hat{J}_{\hat{k}_n})\| I_{\{\tilde{k}_n \le \hat{k}_n \le V m_n^*\}} &\le K_n \max_{1 \le i,j \le p} \left| n^{-1} \sum_{t=1}^{n} z_{ti} z_{tj} - \rho_{ij} \right| \\ &= O_p(1), \end{aligned} \tag{3.51}$$

where the equality is ensured by (2.3) and (2.20). Consequently, (3.46) follows from (3.48)–(3.51), (2.30) and (2.28). The proof of (3.47) is similar to that of (3.46). The details are omitted. □

4. Conclusions. This paper has addressed the important problem of selecting high-dimensional linear regression models with dependent observations when knowledge is lacking about the degree of sparseness of the true model. When the true model is known to be an AR model or a regression model whose predictor variables have been ranked a priori based on their importance, this type of problem has been tackled in the past; see, for example, Ing (2007), Yang (2007), Zhang and Yang (2015) and Ding, Tarokh and Yang (2018). These authors have proposed various ways to combine the strengths of AIC and BIC and shown that their methods achieve the optimal rate without knowing whether (1.2), (1.3) or (2.17), with $N_n = \{1, \dots, k_0\}$, is true. Their approaches, however, are not applicable to situations where the predictor variables have no natural ordering or their importance ranks are unknown. To alleviate this difficulty, we first use OGA to rank predictor variables, and then choose along the OGA path the model that has the smallest HDAIC value. Our approach is not only computationally feasible, but also rate optimal without the need for knowing how sparse the underlying time series model is.

Compared to a similar attempt made in Negahban et al. (2012), in which Lasso is used instead of OGA+HDAIC, the novelty of this paper is threefold: first, the validity of OGA+HDAIC is established not only for independent data, but also for time series data; second, the advantage of OGA+HDAIC is obtained in the important special case (1.5), which is seldom discussed in the high-dimensional literature; third, in another important special case (1.4), it is shown that OGA+HDAIC can have a faster convergence rate than Lasso. Finally, we note that OGA is exclusive for linear models. The counterpart of OGA in nonlinear models is the Chebyshev greedy algorithm (CGA) (Temlyakov (2015)). Investigating the performance of CGA+HDAIC in high-dimensional nonlinear time series models would be an interesting topic for future research.

APPENDIX A: RATES OF CONVERGENCE OF THE POPULATION OGA

In this section, we consider the population counterpart of OGA, whose convergence rate plays a crucial role in the analysis of the first term on the right-hand side of (2.18). Let $0 < \xi \le 1$ be given. The algorithm initializes $J_{\xi,0} = \varnothing$. For $m \ge 1$, $J_{\xi,m}$ is recursively updated by

$$J_{\xi,m} = J_{\xi,m-1} \cup \{j_{\xi,m}\},$$

where $j_{\xi,m}$ is any element l in $\mathcal{P}$ satisfying

$$\left|\mathrm{E}(u_{m-1} z_l)\right| \ge \xi \max_{1 \le j \le p} \left|\mathrm{E}(u_{m-1} z_j)\right|,$$

with $u_0 = y(\mathbf{x})$ and $u_m = y(\mathbf{x}) - y_{J_{\xi,m}}(\mathbf{x})$ if $m \ge 1$. Because the algorithm is implemented based on the "population" correlations of $\mathbf{x}$, it is referred to as the population OGA when $\xi = 1$, and the population weak OGA when $0 < \xi < 1$. The following lemma provides a rate of convergence of the $\mathrm{E}(u_m^2)$ under (A3) or (A4).

LEMMA A.1. *Assume* (2.11) *and* (2.19). *Then there exists* $G_1 > 0$ *such that*

$$\mathrm{E}(u_m^2) = \mathrm{E}(y(\mathbf{x}) - y_{J_{\xi,m}}(\mathbf{x}))^2 \leq G_1 m^{-2\gamma+1}. \tag{A.1}$$

Moreover, if (2.12) *holds instead of* (2.11), *then there exist* $G_2, G_3 > 0$ *such that*

$$\mathrm{E}(u_m^2) = \mathrm{E}(y(\mathbf{x}) - y_{J_{\xi,m}}(\mathbf{x}))^2 \leq G_2 \exp(-G_3 m). \tag{A.2}$$

PROOF. Straightforward calculations yield

$$\begin{aligned} \mathrm{E}(u_m^2) &= \mathrm{E}\left[(y(\mathbf{x}) - y_{J_{\xi,m}}(\mathbf{x})) \sum_{j=1}^{p} \beta_j^* z_j\right] \\ &\leq \max_{1\leq j\leq p} |\mu_{J_{\xi,m},j}| \sum_{j=1, j\notin J_{\xi,m}}^{p} |\beta_j^*|, \end{aligned} \tag{A.3}$$

recalling $\mu_{J,i} = \mathrm{E}[(y(\mathbf{x}) - y_J(\mathbf{x}))z_i]$. In addition, (2.19) implies

$$\mathrm{E}(u_m^2) \geq \lambda_1 \sum_{j=1, j\notin J_{\xi,m}}^{p} \beta_j^{*2}. \tag{A.4}$$

By (A.3), (A.4) and (2.11), it follows that

$$\begin{aligned} \mathrm{E}(u_m^2) &\leq C_\gamma \max_{1\leq j\leq p} |\mu_{J_{\xi,m},j}| \left(\sum_{j=1, j\notin J_{\xi,m}}^{p} \beta_j^{*2}\right)^{(\gamma-1)/(2\gamma-1)} \\ &\leq C_\gamma \lambda_1^{-(\gamma-1)/(2\gamma-1)} \max_{1\leq j\leq p} |\mu_{J_{\xi,m},j}| [\mathrm{E}(u_m^2)]^{(\gamma-1)/(2\gamma-1)}, \end{aligned} \tag{A.5}$$

and hence

$$[\mathrm{E}(u_m^2)]^{\gamma/(2\gamma-1)} \leq C_\gamma \lambda_1^{-(\gamma-1)/(2\gamma-1)} \max_{1\leq j\leq p} |\mu_{J_{\xi,m},j}|. \tag{A.6}$$

In view of (A.6), one has

$$\begin{aligned} \mathrm{E}(u_{m+1}^2) &\leq \mathrm{E}(u_m - \mu_{J_{\xi,m},j_{\xi,m+1}} z_{j_{\xi,m+1}})^2 \\ &\leq \mathrm{E}(u_m^2) - \xi^2 \max_{1\leq j\leq p} \mu_{J_{\xi,m},j}^2 \\ &\leq \mathrm{E}(u_m^2) - \xi^2 \lambda_1^{2(\gamma-1)/(2\gamma-1)} C_\gamma^{-2} [\mathrm{E}(u_m^2)]^{2\gamma/(2\gamma-1)} \\ &= \mathrm{E}(u_m^2)\{1 - \xi^2 \lambda_1^{2(\gamma-1)/(2\gamma-1)} C_\gamma^{-2} [\mathrm{E}(u_m^2)]^{1/(2\gamma-1)}\}. \end{aligned} \tag{A.7}$$

The desired conclusion (A.1) follows from (A.7) and Lemma 1 of Gao, Ing and Yang (2013).

To show (A.2), note first that (2.12), (A.3) and (A.4) yield

$$\mathrm{E}(u_m^2)^{1/2} \leq \lambda_1^{-1/2} M_1 \max_{1\leq j\leq p} |\mu_{J_{\xi,m},j}|.$$

This and an argument similar to that used in (A.7) imply

$$\begin{aligned} \mathrm{E}(u_{m+1}^2) &\leq \mathrm{E}(u_m^2) - \xi^2 \lambda_1 M_1^{-2} \mathrm{E}(u_m^2) \\ &= \mathrm{E}(u_m^2)\{1 - \xi^2 \lambda_1 M_1^{-2}\}. \end{aligned} \tag{A.8}$$

Since $M_1 > 1$, $0 < \lambda_1 \leq 1$, and $0 < \xi \leq 1$, (A.8) leads directly to (A.2). □

Lemma A.2 shows that (1.4) is a special case of (2.11). Using Lemmas A.1 and A.2, Lemma A.3 demonstrates that the rate $m^{-2\gamma+1}$ obtained in (A.1) cannot be improved under (1.4). More specifically, recall the best m-term approximation, $y_{J_m^*}(\mathbf{x})$, of $y(\mathbf{x})$ (see (3.24)). Lemma A.3 asserts that when (1.4) and (2.19) hold true, the approximation errors of $y_{J_{\xi,m}}(\mathbf{x})$ and $y_{J_m^*}(\mathbf{x})$ only differ by a positive constant.

LEMMA A.2. *Suppose that* (1.4) *is true for some* $\gamma > 1$. *Then* (2.11) *holds for the same* γ.

PROOF. See Section S1 of the Supplementary Material. □

LEMMA A.3. *Suppose that* (1.4) *holds for some* $\gamma > 1$ *and* (2.19) *is true*. *Then*, *for all* $1 \le m \le (1-\epsilon)p$, *where* ϵ *is an arbitrarily small positive constant*, *there exist* D_1, D_2 *and* D_3 *such that*

$$\mathrm{E}\big(y(\mathbf{x}) - y_{J_{\xi,m}}(\mathbf{x})\big)^2 \le D_1 \mathrm{E}\big(y(\mathbf{x}) - y_{J_m^*}(\mathbf{x})\big)^2, \tag{A.9}$$

and

$$D_2 m^{-2\gamma+1} \le \mathrm{E}\big(y(\mathbf{x}) - y_{J_m^*}(\mathbf{x})\big)^2 \le D_3 m^{-2\gamma+1}. \tag{A.10}$$

PROOF. By Lemmas A.1 and A.2, it follows that for all $1 \le m \le (1-\epsilon)p$,

$$\begin{aligned} G_1 m^{-2\gamma+1} &\ge \mathrm{E}\big(y(\mathbf{x}) - y_{J_{\xi,m}}(\mathbf{x})\big)^2 \ge \mathrm{E}\big(y(\mathbf{x}) - y_{J_m^*}(\mathbf{x})\big)^2 \\ &\ge \lambda_1 \sum_{j \notin J_m^*} \beta_j^{*^2} \ge \lambda_1 \sum_{j \notin J_m^o} \beta_j^{*^2} \ge \lambda_1 L^2 \sum_{j=m+1}^{p} j^{-2\gamma} \\ &\ge \lambda_1 L^2 \underline{d} m^{-2\gamma+1}, \end{aligned}$$

where J_m^o is the index set corresponding to $\{\beta_{(1)}^2, \ldots, \beta_{(m)}^2\}$ and $\underline{d} > 0$ depends only on γ and ϵ. These inequalities lead immediately to (A.9) and (A.10). □

REMARK A.1. Theorem 2.1 of Temlyakov (1998) shows that a near best m-term approximation can be realized by a greedy-type algorithm under a basis L_p-equivalent to the Haar basis. Since the Haar basis yields an identity correlation matrix, our correlation assumption, (2.19), appears to be substantially weaker. The performance of the m-term approximation of OGA has been investigated by Tropp (2004) under a noise-free underdetermined system and a condition on the cumulative coherence function, which requires that the atoms in the dictionary are "nearly" uncorrelated. His approximation error for OGA is larger than that of the best m-term approximation by a factor of $(1+6m)^{1/2}$. Suppose that (1.5) holds. Then

$$\lambda_1 \sum_{j \notin J_m^o} \beta_j^{*^2} \le \mathrm{E}\big(y(\mathbf{x}) - y_{J_m^*}(\mathbf{x})\big)^2 \le \mathrm{E}\bigg(\sum_{j \notin J_m^o} \beta_j^* z_j\bigg)^2,$$

which, together with (1.5) and Minkowski's inequality, yields

$$C_{1,\beta}\lambda_1 L_1^2 \exp(-2\beta m) \le \mathrm{E}\big(y(\mathbf{x}) - y_{J_m^*}(\mathbf{x})\big)^2 \le C_{2,\beta} U_1^2 \exp(-2\beta m), \tag{A.11}$$

where $C_{1,\beta} \le C_{2,\beta}$ are some positive constants depending on β. On the other hand, the argument used to prove (A.2) leads to

$$\mathrm{E}\big(y(\mathbf{x}) - y_{J_{\xi,m}}(\mathbf{x})\big)^2 = O\big(\exp(-m f_{\mathrm{oga}})\big), \tag{A.12}$$

where $f_{\mathrm{oga}} = \xi^2 \lambda_1 (L1/U_1)^2 (1-\exp(-\beta))^2 < 2\beta$. Equations (A.11) and (A.12) suggest that the population OGA and the best m-term approximation in general do not share the same convergence rate in the exponential decay case. To be as efficient as the best m-term approximation, the population OGA needs to run for another $m(2\beta/f_{\mathrm{oga}} - 1)$ iterations, which is still of order m.

APPENDIX B: PROOF OF (3.29)

Recall that (2.13) implies (2.14) with $C = M + 1$. This, (2.11) and (2.19) yield

$$\begin{aligned}
&\left| n^{-1} \sum_{t=1}^{n} \varepsilon_t^2(\hat{J}_m) - \mathrm{E}_n\big(\varepsilon^2(\hat{J}_m)\big) \right| \\
&= \left| \sum_{\sharp(J)=m} \left\{ n^{-1} \sum_{t=1}^{n} \varepsilon_t^2(J) - \mathrm{E}\big(\varepsilon^2(J)\big) \right\} I_{\{\hat{J}_m = J\}} \right| \\
&\leq \sum_{\sharp(J)=m} \left\{ \sum_{i=1}^{p} \sum_{l=1}^{p} |\beta_i^* - \beta_i^*(J)| |\beta_l^* - \beta_l^*(J)| \left| n^{-1} \sum_{t=1}^{n} z_{ti} z_{tl} - \rho_{il} \right| \right\} I_{\{\hat{J}_m = J\}} \\
&\leq (M+1)^2 \max_{1 \leq i,l \leq p} \left| n^{-1} \sum_{t=1}^{n} z_{ti} z_{tl} - \rho_{il} \right| \sum_{\sharp(J)=m} \left(\sum_{j \notin J} |\beta_i^*| \right)^2 I_{\{\hat{J}_m = J\}} \\
&\leq C_\gamma^2 (M+1)^2 R_{1,p} \sum_{\sharp(J)=m} \left(\sum_{j \notin J} \beta_i^{*2} \right)^{(2\gamma-2)/(2\gamma-1)} I_{\{\hat{J}_m = J\}} \\
&\leq C_{M,\gamma,\lambda_1} R_{1,p} \sum_{\sharp(J)=m} \big\{ \mathrm{E}\big(\varepsilon^2(J)\big) \big\}^{(2\gamma-2)/(2\gamma-1)} I_{\{\hat{J}_m = J\}} \\
&= C_{M,\gamma,\lambda_1} R_{1,p} \big\{ \mathrm{E}_n\big(\varepsilon^2(\hat{J}_m)\big) \big\}^{(2\gamma-2)/(2\gamma-1)}.
\end{aligned}$$

Thus, (3.29) follows.

Acknowledgments. I thank Hai-Tang Chiou, Hsueh-Han Huang and Tze Leung Lai for helpful suggestions on earlier versions of this paper. I would also like to thank an Associate Editor and two anonymous referees for insightful and constructive comments.

This work was supported in part by the Science Vanguard Research Program of the Ministry of Science and Technology, Taiwan.

SUPPLEMENTARY MATERIAL

Supplement to "Model selection for high-dimensional linear regression with dependent observations" (DOI: 10.1214/19-AOS1872SUPP; .pdf). The Supplementary Material contains the proofs of (2.7), (2.28), (2.30), (2.32), (3.6), (3.30)–(3.32), (3.45) and Lemma A.2, and a simulation study to demonstrate the performance of OGA+HDAIC under a high-dimensional ARX model whose $\mathbf{\Gamma}$ obeys $\lambda_{\max}(\mathbf{\Gamma}) \to \infty$ and (2.19).

REFERENCES

ADAMCZAK, R. and WOLFF, P. (2015). Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order. *Probab. Theory Related Fields* **162** 531–586. MR3383337 https://doi.org/10.1007/s00440-014-0579-3

BASU, S. and MICHAILIDIS, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **43** 1535–1567. MR3357870 https://doi.org/10.1214/15-AOS1315

BAXTER, G. (1962). An asymptotic result for the finite predictor. *Math. Scand.* **10** 137–144. MR0149584 https://doi.org/10.7146/math.scand.a-10520

BERK, K. N. (1974). Consistent autoregressive spectral estimates. *Ann. Statist.* **2** 489–502. MR0421010

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

BÜHLMANN, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist.* **34** 559–583. MR2281878 https://doi.org/10.1214/009053606000000092

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644 https://doi.org/10.1214/009053606000001523

CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189 https://doi.org/10.1093/biomet/asn034

DING, J., TAROKH, V. and YANG, Y. (2018). Bridging AIC and BIC: A new criterion for autoregression. *IEEE Trans. Inform. Theory* **64** 4024–4043. MR3809725 https://doi.org/10.1109/TIT.2017.2717599

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322 https://doi.org/10.1111/j.1467-9868.2008.00674.x

GAO, F., ING, C.-K. and YANG, Y. (2013). Metric entropy and sparse linear approximation of ℓ_q-hulls for $0 < q \le 1$. *J. Approx. Theory* **166** 42–55. MR3003947 https://doi.org/10.1016/j.jat.2012.10.002

HUANG, H.-H. and ING, C.-K. (2019). Concentration inequalities for sample higher order moments of stationary linear processes. Technical Report, Institute of Statistics, National Tsing Hua Univ.

ING, C.-K. (2007). Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *Ann. Statist.* **35** 1238–1277. MR2341705 https://doi.org/10.1214/009053606000001550

ING, C.-K. (2020). Supplement to "Model selection for high-dimensional linear regression with dependent observations." https://doi.org/10.1214/19-AOS1872SUPP.

ING, C.-K. and LAI, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statist. Sinica* **21** 1473–1513. MR2895106 https://doi.org/10.5705/ss.2010.081

ING, C.-K., LAI, T. L., SHEN, M., TSANG, K. and YU, S.-H. (2017). Multiple testing in regression models with applications to fault diagnosis in the big data era. *Technometrics* **59** 351–360. MR3677966 https://doi.org/10.1080/00401706.2016.1236755

NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133 https://doi.org/10.1214/12-STS400

POURAHMADI, M. (1989). On the convergence of finite linear predictors of stationary processes. *J. Multivariate Anal.* **30** 167–180. MR1015366 https://doi.org/10.1016/0047-259X(89)90033-X

RUDELSON, M. and VERSHYNIN, R. (2013). Hanson–Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.* **18** no. 82, 9. MR3125258 https://doi.org/10.1214/ECP.v18-2865

SHIBATA, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8** 147–164. MR0557560

TEMLYAKOV, V. N. (1998). The best m-term approximation and greedy algorithms. *Adv. Comput. Math.* **8** 249–265. MR1628182 https://doi.org/10.1023/A:1018900431309

TEMLYAKOV, V. N. (2000). Weak greedy algorithms. *Adv. Comput. Math.* **12** 213–227. MR1745113 https://doi.org/10.1023/A:1018917218956

TEMLYAKOV, V. N. (2015). Greedy approximation in convex optimization. *Constr. Approx.* **41** 269–296. MR3315679 https://doi.org/10.1007/s00365-014-9272-0

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TROPP, J. A. (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory* **50** 2231–2242. MR2097044 https://doi.org/10.1109/TIT.2004.834793

WANG, H. (2009). Forward regression for ultra-high dimensional variable screening. *J. Amer. Statist. Assoc.* **104** 1512–1524. MR2750576 https://doi.org/10.1198/jasa.2008.tm08516

WANG, Z., PATERLINI, S., GAO, F. and YANG, Y. (2014). Adaptive minimax regression estimation over sparse ℓ_q-hulls. *J. Mach. Learn. Res.* **15** 1675–1711. MR3225246

WU, W.-B. and WU, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electron. J. Stat.* **10** 352–379. MR3466186 https://doi.org/10.1214/16-EJS1108

YANG, Y. (2007). Prediction/estimation with simple linear models: Is it really that simple? *Econometric Theory* **23** 1–36. MR2338950 https://doi.org/10.1017/S0266466607070016

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

ZHANG, Y. and YANG, Y. (2015). Cross-validation for selecting a model selection procedure. *J. Econometrics* **187** 95–112. MR3347297 https://doi.org/10.1016/j.jeconom.2015.02.006

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

Supplement to "Model selection for high-dimensional linear regression with dependent observations"

Ching-Kang Ing

National Tsing Hua University

This supplementary document contains three sections. The proofs of (2.7), (2.28), (2.30), (2.32), and Lemma A1.2 in Ing (2019) are provided in Section S1, and the proofs of (3.6), (3.30)–(3.32), and (3.45) in Ing (2019) are offered in Section S2. In Section S3, a simulation study is given to demonstrate the performance of OGA+HDAIC under a high-dimensional ARX model whose $\mathbf{\Gamma}$ obeys $\lambda_{\max}(\mathbf{\Gamma}) \to \infty$ and $\lambda_{\min}(\mathbf{\Gamma}) > \lambda_1 > 0$.

S1. Proofs of (2.7), (2.28), (2.30), (2.32), and Lemma A1.2.

Proof of (2.7). By (2.4), the second relation of (2.5), and the first relation of (2.6), it holds that for some $0 < M_3 < \infty$,

$$\max_{0 \le l \le p} \lambda_{\max}(\mathbf{\Sigma}_l) < M_3, \tag{S1.1}$$

where $\mathbf{\Sigma}_l = \mathrm{E}(\mathbf{X}_l \mathbf{X}_l^\top)$ with $\mathbf{X}_0 = (x_{10}, \ldots, x_{n0})^\top \equiv (\varepsilon_1, \ldots, \varepsilon_n)^\top$. Then, by the Hanson-Wright inequality (see Theorem 1.1 of Rudelson and Vershynin (2013)), (2.8), (S1.1), and $\|A\|_{\mathrm{F}}^2 \le n\lambda_{\max}^2(A)$, there exists some $c > 0$ such that for any $s > 0$, any $0 \le l \le p$, and all large n,

$$\begin{aligned} &P(|n^{-1}\sum_{t=1}^n x_{tl}^2 - \sigma_l^2| > s(\log p/n)^{1/2}) \\ &\le 2\exp\left[-c\min\left\{\frac{s^2 n\log p}{K^4\|\mathbf{\Sigma}_l\|_{\mathrm{F}}^2}, \frac{s(n\log p)^{1/2}}{K^2\lambda_{\max}(\mathbf{\Sigma}_l)}\right\}\right] \\ &\le 2\exp\left[\frac{-cs^2\log p}{K^4\lambda_{\max}^2(\mathbf{\Sigma}_l)}\right] \le 2\exp[-cs^2(K^2M_3)^{-2}\log p], \end{aligned} \tag{S1.2}$$

where $K = (8/3)^{1/2}$, $\|A\|_F$ denotes the Frobenius norm of A, and $\sigma_0^2 = \mathrm{E}\varepsilon_t^2$. Similarly, using $\lambda_{\max}(\mathbf{\Sigma}_l + \mathbf{\Sigma}_0) \le 2M_3$ and $\lambda_{\max}\{\mathrm{E}(\mathbf{X}_{l_1} + \mathbf{X}_{l_2})(\mathbf{X}_{l_1} + \mathbf{X}_{l_2})^\top\} \le 4M_3$, where $1 \le l, l_1, l_2 \le p$, we obtain

$$\begin{aligned} &P(|n^{-1}\sum_{t=1}^n (x_{tl} + \varepsilon_t)^2 - (\sigma_l^2 + \sigma_0^2)| > s(\log p/n)^{1/2}) \\ &\le 2\exp[-cs^2(2K^2M_3)^{-2}\log p], \end{aligned} \tag{S1.3}$$

and

$$P(|n^{-1}\sum_{t=1}^{n}(x_{tl_1}+x_{tl_2})^2 - \mathrm{E}(x_{tl_1}+x_{tl_2})^2| > s(\log p/n)^{1/2}) \tag{S1.4}$$
$$\leq 2\exp[-cs^2(4K^2M_3)^{-2}\log p].$$

By (S1.2)–(S1.4), it holds that for all large n,

(S1.5)

$$P(\max_{1\leq l_1,l_2\leq p}|n^{-1}\sum_{t=1}^{n}x_{tl_1}x_{tl_2} - \mathrm{E}(x_{tl_1}x_{tl_2})| > s(\log p/n)^{1/2})$$
$$\leq p^2\max_{1\leq l_1,l_2\leq p}P(|n^{-1}\sum_{t=1}^{n}(x_{tl_1}+x_{tl_2})^2 - \mathrm{E}(x_{tl_1}+x_{tl_2})^2| > (2s/3)(\log p/n)^{1/2})$$
$$+\,2p^2\max_{1\leq l\leq p}P(|n^{-1}\sum_{t=1}^{n}x_{tl}^2-\sigma_l^2| > (2s/3)(\log p/n)^{1/2})$$
$$\leq 6p^2\exp[-c(2s/3)^2(4K^2M_3)^{-2}\log p] = o(1),$$

provided $s = s_1^* > (72K^4M_3^2/c)^{1/2}$, and

(S1.6)

$$P(\max_{1\leq i\leq p}|n^{-1}\sum_{t=1}^{n}x_{ti}\varepsilon_t| > s(\log p/n)^{1/2})$$
$$\leq p\max_{1\leq i\leq p}P(|n^{-1}\sum_{t=1}^{n}(x_{ti}+\varepsilon_t)^2-(\sigma_i^2+\sigma_0^2)| > (2s/3)(\log p/n)^{1/2})$$
$$+\,2p\max_{0\leq i\leq p}P(|n^{-1}\sum_{t=1}^{n}x_{ti}^2-\sigma_i^2| > (2s/3)(\log p/n)^{1/2})$$
$$\leq 6p\exp[-c(2s/3)^2(2M_3)^{-2}\log p] = o(1),$$

provided $s = s_2^* > (9K^4M_3^2/c)^{1/2}$. By the third relation of (2.5) and the second relation of (2.6), one obtains

$$\min_{1\leq i\leq p}\sigma_i^2 > \underline{c} > 0. \tag{S1.7}$$

Combining (S1.7), (S1.5), and (S1.6) yields (2.2) with $c_1^* = s_2^*/\underline{c}^{1/2}$ and (2.3) with $c_2^* = s_1^*/\underline{c}$.

PROOF OF (2.28). Note first that (2.21) implies

$$\liminf_{n\to\infty} \min_{\sharp(J)\le K_n} \lambda_{\min}(\mathbf{\Gamma}(J)) - c_2^*\bar{\delta} > 0. \tag{S1.8}$$

Define $F_n = \{\|\hat{\mathbf{\Gamma}}(\hat{J}_{K_n}) - \mathbf{\Gamma}(\hat{J}_{K_n})\| \le K_n c_2^*(\log p)^{1/2}/n^{1/2}\}$. Then it follows from (2.3) and some algebraic manipulations that

$$\lim_{n\to\infty} P(F_n) = 1. \tag{S1.9}$$

Equations (S1.8) and (S1.9) ensure

$$\lim_{n\to\infty} P(Q_n) = 1, \tag{S1.10}$$

where $Q_n = \{\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_{K_n}) \text{ exists}\}$.

In view of (2.29), one has for all large n,

$$\begin{aligned}
&P(\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_{K_n}) \le \bar{B}, Q_n, F_n)\\
&\ge P\left((1 - \frac{c_2^*\bar{\delta}}{\min_{\sharp(J)\le K_n}\lambda_{\min}(\mathbf{\Gamma}(J))})\|\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_{K_n})\| \le \frac{1}{\min_{\sharp(J)\le K_n}\lambda_{\min}(\mathbf{\Gamma}(J))}, Q_n, F_n\right)\\
&\ge P\left((1 - \|\hat{\mathbf{\Gamma}}(\hat{J}_{K_n}) - \mathbf{\Gamma}(\hat{J}_{K_n})\|\|\mathbf{\Gamma}^{-1}(\hat{J}_{K_n})\|)\|\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_{K_n})\| \le \|\mathbf{\Gamma}^{-1}(\hat{J}_{K_n})\|, Q_n, F_n\right)\\
&= P(Q_n, F_n),
\end{aligned}$$

which, together with (S1.9) and (S1.10), leads to (2.28). □

PROOF OF (2.30). Since (A2) implies $\max_{1\le i\le p}|n^{-1}\sum_{t=1}^n z_{ti}^2 - 1| = o_p(1)$, (2.30) is ensured by

$$P(\max_{\sharp(J)\le K_n-1, i\notin J}|n^{-1}\sum_{t=1}^n \varepsilon_t \hat{z}_{ti;J}^{\perp}| > s(\log p_n/n)^{1/2}) = o(1), \tag{S1.11}$$

and

(S1.12)

$$P(\max_{i,j\notin J, \sharp(J)\le K_n-1}|n^{-1}\sum_{t=1}^n z_{tj}\hat{z}_{ti;J}^{\perp} - E(z_j z_{i;J}^{\perp})| > s(\log p_n/n)^{1/2}) = o(1),$$

for some large s, where $(\hat{z}^{\perp}_{1i;J}, \dots, \hat{z}^{\perp}_{ni;J})^{\top} = (\mathbf{I} - \mathbf{H}_J)\mathbf{Z}_i$ and $z^{\perp}_{i;J} = z_i - \boldsymbol{g}_i^{\top}(J)\boldsymbol{\Gamma}^{-1}(J)\mathbf{z}(J)$. Straightforward calculations yield

$$
\begin{aligned}
&\max_{\sharp(J)\le K_n-1, i\notin J} |n^{-1}\sum_{t=1}^{n} \varepsilon_t \hat{z}^{\perp}_{ti;J}| \le \max_{1\le i\le p} |n^{-1}\sum_{t=1}^{n} z_{ti}\varepsilon_t| \\
&+ \max_{\sharp(J)\le K_n} \|\hat{\boldsymbol{\Gamma}}^{-1}(J)\| K_n (1 + \max_{\sharp(J)\le K_n} \|\boldsymbol{\Gamma}^{-1}(J)\boldsymbol{g}_i(J)\|_1) \\
&\times \max_{1\le i,j\le p} |n^{-1}\sum_{t=1}^{n} z_{ti}z_{tj} - \rho_{ij}| \max_{1\le i\le p} |n^{-1}\sum_{t=1}^{n} z_{ti}\varepsilon_t| \\
&+ \max_{\sharp(J)\le K_n} \|\boldsymbol{\Gamma}^{-1}(J)\boldsymbol{g}_i(J)\|_1 \max_{1\le i\le p} |n^{-1}\sum_{t=1}^{n} z_{ti}\varepsilon_t|.
\end{aligned} \tag{S1.13}
$$

By (S1.13), (A1), (A2), (A5), and (2.28), (S1.11) holds with $s > c_1^*(1+M)$.

To prove (S1.12), we note that (A5) and some algebraic manipulations give

$$
\begin{aligned}
&\max_{\sharp(J)\le K_n-1, i,j\notin J} |n^{-1}\sum_{t=1}^{n} z_{tj}\hat{z}^{\perp}_{ti;J} - E(z_j z^{\perp}_{i;J})| \\
&\le \max_{1\le i,j\le p} |n^{-1}\sum_{t=1}^{n} z_{ti}z_{tj} - \rho_{ij}|(1+2M+M^2) \\
&+ \max_{\sharp(J)\le K_n} \|\hat{\boldsymbol{\Gamma}}^{-1}(J)\| K_n (1+M)^2 \max_{1\le i,j\le p} |n^{-1}\sum_{t=1}^{n} z_{ti}z_{tj} - \rho_{ij}|.
\end{aligned} \tag{S1.14}
$$

Combining (S1.14), (A2), (2.20), and (2.28), one obtains (S1.12) when $s > c_2^*(1 + 2M + M^2 + (1+M)^2\bar{\delta}c_2^*\bar{B})$. Thus, the proof of (2.30) is complete. □

Proof of (2.32). Note first that

$$
\begin{aligned}
&\mathrm{E}_n(\hat{y}_m(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x}))^2 \\
&\le \|n^{-1}\sum_{t=1}^{n} \mathbf{z}_t(\hat{J}_m)(\varepsilon_t + \varepsilon_t(\hat{J}_m))\|^2 \|\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_{K_n})\| \\
&+ \|\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_{K_n})\|^2 \max_{1\le m\le K_n} \|\hat{\boldsymbol{\Gamma}}(\hat{J}_m) - \boldsymbol{\Gamma}(\hat{J}_m)\| \|n^{-1}\sum_{t=1}^{n} \mathbf{z}_t(\hat{J}_m)(\varepsilon_t + \varepsilon_t(\hat{J}_{K_n}))\|^2.
\end{aligned} \tag{S1.15}
$$

Since $K_n(\log p/n)^{1/2} \leq \bar{\delta}$, by (A2),

$$P(\max_{1\leq m\leq K_n} \|\hat{\mathbf{\Gamma}}(\hat{J}_m) - \mathbf{\Gamma}(\hat{J}_m)\| \leq \bar{\delta}c_2^*) = 1. \tag{S1.16}$$

Some algebraic manipulations yield

$$\|n^{-1}\sum_{t=1}^{n}\mathbf{z}_t(\hat{J}_m)\varepsilon_t\|^2 \leq m(\max_{1\leq i\leq p}|n^{-1}\sum_{t=1}^{n} z_{ti}\varepsilon_t|)^2, \tag{S1.17}$$

and

$$\begin{aligned}
&\leq \|n^{-1}\sum_{t=1}^{n}\mathbf{z}_t(\hat{J}_m)\varepsilon_t(\hat{J}_m)\|^2 \\
&\leq m(\max_{1\leq i,j\leq p}|n^{-1}\sum_{t=1}^{n} z_{ti}z_{tj} - \rho_{ij}|)^2\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^*(\hat{J}_m)\|_1^2 \qquad \text{(S1.18)}\\
&\leq m(\max_{1\leq i,j\leq p}|n^{-1}\sum_{t=1}^{n} z_{ti}z_{tj} - \rho_{ij}|)^2\|\boldsymbol{\beta}^*\|_1^2(M+1)^2,
\end{aligned}$$

where the last inequality is ensured by (2.14). Consequently, the desired conclusion (2.32) follows from (S1.15)–(S1.18), (2.28), (A1), (A2), and (A4). □

PROOF OF LEMMA A1.2. We first show that for any $J \subseteq \{1, \cdots, p\}$,

$$D_1(\sum_{j\in J} j^{-\gamma}) \leq (\sum_{j\in J} j^{-2\gamma})^{(\gamma-1)/(2\gamma-1)}, \tag{S1.19}$$

where $D_1 = \min\{2^{(\gamma-1)/(2\gamma-1)}-1, (\gamma-1)/(4\gamma-2)\}$. We prove (S1.19) by induction. It is easy to see that (S1.19) is true for all J with $\sharp(J) = 1$. Assume that (S1.19) holds for all J with $1 < \sharp(J) = k < p$. We are going to show that (S1.19) is also true for all J with $\sharp(J) = k + 1$. To this end, define $L = \min\{i : i \in J\}$ and $B = J - \{L\}$. Since $\sharp(B) = k$, by the induction hypothesis, it follows that

$$D_1(\sum_{j\in B} j^{-\gamma}) \leq (\sum_{j\in B} j^{-2\gamma})^{(\gamma-1)/(2\gamma-1)}. \tag{S1.20}$$

If $L^{-2\gamma} \leq \sum_{j\in B} j^{-2\gamma}$, then (S1.20) and $\sum_{j\in B} j^{-2\gamma} \leq L^{-2\gamma+1}$ yield

(S1.21)

$$(\sum_{j\in J} j^{-2\gamma})^{(\gamma-1)/(2\gamma-1)} = (\sum_{j\in B} j^{-2\gamma})^{(\gamma-1)/(2\gamma-1)} \left(1 + \frac{1}{L^{2\gamma}\sum_{j\in B} j^{-2\gamma}}\right)^{(\gamma-1)/(2\gamma-1)}$$
$$\geq (\sum_{j\in B} j^{-2\gamma})^{(\gamma-1)/(2\gamma-1)} \left(1 + \frac{\gamma-1}{(4\gamma-2)L^{2\gamma}\sum_{j\in B} j^{-2\gamma}}\right)$$
$$\geq D_1 \sum_{j\in B} j^{-\gamma} + \frac{\gamma-1}{(4\gamma-2)L^{\gamma}\{L^{-2\gamma+1}\}^{\gamma/(2\gamma-1)}} L^{-\gamma}$$
$$\geq D_1 \sum_{j\in J} j^{-\gamma}.$$

On the other hand, assume $L^{-2\gamma} > \sum_{j\in B} j^{-2\gamma}$. Then, by (S1.20),

(S1.22)

$$L^{-\gamma} + \sum_{j\in B} j^{-\gamma} \leq L^{-\gamma} + D_1^{-1}(\sum_{j\in B} j^{-2\gamma})^{(\gamma-1)/(2\gamma-1)}$$
$$= D_1^{-1}(\sum_{j\in J} j^{-2\gamma})^{(\gamma-1)/(2\gamma-1)} \left\{\frac{(\sum_{j\in B} j^{-2\gamma})^{(\gamma-1)/(2\gamma-1)} + D_1 L^{-\gamma}}{(\sum_{j\in J} j^{-2\gamma})^{(\gamma-1)/(2\gamma-1)}}\right\}$$
$$\leq D_1^{-1}(\sum_{j\in J} j^{-2\gamma})^{(\gamma-1)/(2\gamma-1)} \left\{\frac{(L^{2\gamma}\sum_{j\in B} j^{-2\gamma})^{(\gamma-1)/(2\gamma-1)} + D_1}{(1 + L^{2\gamma}\sum_{j\in B} j^{-2\gamma})^{(\gamma-1)/(2\gamma-1)}}\right\}.$$

Since $D_1 \leq 2^{(\gamma-1)/(2\gamma-1)} - 1$, it can be shown that

$$\frac{\theta^{(\gamma-1)/(2\gamma-1)} + D_1}{(1+\theta)^{(\gamma-1)/(2\gamma-1)}} \leq 1, \text{ for all } 0 < \theta \leq 1,$$

which, together with (S1.22) and (S1.21), yields that (S1.19) holds with $\sharp(J) = k+1$.

Let $\tilde{j}$ be defined implicitly by $|\beta_j^*| = |\beta_{(\tilde{j})}^*|$. For $J = \{j_1, \cdots, j_l\} \subseteq \{1, \cdots, p\}$, also define $\tilde{J} = \{\tilde{j}_1, \cdots, \tilde{j}_l\}$. It follows from (1.4) and (S1.19) that

$$\sum_{j\in J}|\beta_j^*| \le U\sum_{j\in\tilde{J}} j^{-\gamma} \le UD_1^{-1}(\sum_{j\in\tilde{J}} j^{-2\gamma})^{(\gamma-1)/(2\gamma-1)}$$
$$\le UD_1^{-1}L^{-(2\gamma-2)/(2\gamma-1)}(\sum_{j\in\tilde{J}}\beta_{(j)}^{*^2})^{(\gamma-1)/(2\gamma-1)}$$
$$\le UD_1^{-1}L^{-(2\gamma-2)/(2\gamma-1)}(\sum_{j\in J}\beta_j^{*^2})^{(\gamma-1)/(2\gamma-1)}.$$

Therefore, (2.11) holds for $C_\gamma = UD_1^{-1}L^{-(2\gamma-2)/(2\gamma-1)}$. □

S2. Proofs of (3.6), (3.30)–(3.32), and (3.45). To prove (3.6), we need an auxiliary lemma, which is a counterpart of Lemma A1.1 in the case of strong sparsity.

LEMMA S2.1. *Assume*

$$\min_{\sharp(J)\le k_0+m}\lambda_{\min}(\mathbf{\Gamma}(J)) \ge \lambda_0 > 0. \tag{S2.1}$$

Then,

$$\mathrm{E}(u_m^2) = \mathrm{E}(y(\mathbf{x}) - y_{J_{\xi,m}}(\mathbf{x}))^2 \le \mathrm{E}(y^2(\mathbf{x}))\exp(-\xi^2\lambda_0 m/k_0). \tag{S2.2}$$

Proof. By (A1.3), the Cauchy-Schwarz inequality, and (S2.1),

$$\mathrm{E}(y(\mathbf{x}) - y_{J_{\xi,m}}(\mathbf{x}))^2 \le \max_{1\le j\le p}|\mu_{J_{\xi,m},j}|k_0^{1/2}(\sum_{j=1,j\notin J_{\xi,m}}^{p}\beta_j^{*^2})^{1/2},$$

and

$$\mathrm{E}(y(\mathbf{x}) - y_{J_{\xi,m}}(\mathbf{x}))^2 \ge \lambda_0\sum_{j=1,j\notin J_{\xi,m}}^{p}\beta_j^{*^2}.$$

Combining these two inequalities gives

$$\mathrm{E}(y(\mathbf{x}) - y_{J_{\xi,m}}(\mathbf{x}))^2 \le \frac{k_0}{\lambda_0}\max_{1\le j\le p}\mu^2_{J_{\xi,m},j},$$

and hence

$$
\begin{aligned}
&\mathrm{E}(u_{m+1}^2) \le \mathrm{E}(u_m - \mu_{J_{\xi,m},j_{\xi,m+1}} z_{j_{\xi,m+1}})^2 \\
&\le \mathrm{E}(u_m^2) - \xi^2 \max_{1\le j\le p} \mu_{J_{\xi,m},j}^2 \\
&\le \mathrm{E}(u_m^2)(1-\xi^2\lambda_0/k_0),
\end{aligned}
$$

leading to the desired conclusion (S2.2). □

In this section, we prove (3.6) in situations where the minimum eigenvalue assumption (2.19) is weaken to (3.8). With the help of Lemma S2.1, the next theorem provides a counterpart of (2.31):

(S2.3)

$$
\max_{1\le m\le K_n} \frac{\mathrm{E}_n(y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x}))^2}{\max\{\exp(-\xi^2\lambda_1^* m/k_0), k_0(\sum_{j\in N_n}|\beta_j^*|)^2 \log p/n\}} = O_p(1),
$$

which plays a crucial role in the proof of (3.6). Here ξ is any constant lying between 0 and 1, and $0 < \lambda_1^* \le \bar{\delta} c_2^*$.

Theorem S2.1. *Suppose that* (A1), (A2), (A5), (3.7), (3.8), *and* (2.20) *hold, and* $\mathrm{E}(y_t^2)$ *is bounded above by a constant. Then,* (S2.3) *follows. Moreover,*

(S2.4)
$$
\lim_{n\to\infty} P(N_n \subseteq \hat{J}_{k_0 \bar{R}}) = 1,
$$

where $\bar{R}$ *is some large constant.*

Proof. Note first that (3.8) implies

(S2.5)
$$
\limsup_{n\to\infty} \frac{\tau c_2^*}{\min_{\sharp(J)\le\tau(n/\log p)^{1/2}} \lambda_{\min}(\mathbf{\Gamma}(J))} \le 1,
$$

for any $0 < \tau \le \min\{\lambda_1/c_2^*, \eta\}$. This guarantees that the $\boldsymbol{\tau}$ in (2.21) associated with the definition of K_n is nonempty. It follows from (3.7) that $k_0 = o\big((n/\log p)^{1/3}\big)$, which together with (2.20) and (S2.5), yields

(S2.6)
$$
\min_{\sharp(J)\le k_0+K_n} \lambda_{\min}(\mathbf{\Gamma}(J)) > \bar{\delta} c_2^*,
$$

noting that $0 < \bar{\delta} c_2^* < 1$. Define

$$
C_n(m) = \left\{ \max_{(J,i):\sharp(J)\le m-1, i\notin J} |\hat{\mu}_{J,i} - \mu_{J,i}| \le s(\sum_{j\in N_n} |\beta_j^*|)(\log p/n)^{1/2} \right\},
$$

and

$$D_n(m) = \left\{ \min_{0\le i\le m-1} \max_{1\le j\le p} |\mu_{\hat{J}_i,j}| > \tilde{\xi} s(\sum_{j\in N_n} |\beta_j^*|)(\log p/n)^{1/2} \right\},$$

where $s > 0$ is some large constant and $\tilde{\xi} = 2/(1-\xi)$ with $0 < \xi < 1$ being arbitrarily given. Because we do not assume that $\sum_{j\in N_n} |\beta_j^*|$ is bounded above by a finite constant, the definitions of $C_n(m)$ and $D_n(m)$ are slightly different from those of $A_n(m)$ and $B_n(m)$ in the proof of Theorem 2.1.

By an argument similar to that used to prove (2.24), it holds that

$$|\mu_{\hat{J}_{q-1},\hat{j}_q}| \ge \xi \max_{1\le i\le p} |\mu_{\hat{J}_{q-1},i}| \text{ on } C_n(m) \bigcap D_n(m),$$

which, together with Lemma S2.1, (S2.6), and the boundedness of $\mathrm{E}(y_t^2)$, gives

(S2.7)
$$\mathrm{E}_n(y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x}))^2 \le \bar{G}\exp(-\xi^2\lambda_1^* m/k_0) \text{ on } C_n(m) \bigcap D_n(m),$$

where $\bar{G}$ is some positive constant and $0 < \lambda_1^* \le \bar{\delta} c_2^*$.

By (S2.6) and an argument similar to that used to prove Lemma S2.1, we also have

$$\begin{aligned} \mathrm{E}_n(y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x}))^2 &\le \min_{0\le i\le m-1} \mathrm{E}_n(y(\mathbf{x}) - y_{\hat{J}_i}(\mathbf{x}))^2 \\ &\le \frac{k_0}{\lambda_1^*} \min_{0\le i\le m-1} \max_{1\le j\le p} |\mu_{\hat{J}_i,j}|^2 \\ &\le \frac{k_0}{\lambda_1^*} (\tilde{\xi} s)^2 (\sum_{j\in N_n} |\beta_j^*|)^2 \frac{\log p}{n} \text{ on } D_n^c(m). \end{aligned} \tag{S2.8}$$

Combining (S2.7), (S2.8), and $C_n(K_n) \subseteq C_n(m)$ for $1 \le m \le K_n$ yields

(S2.9)
$$\max_{1\le m\le K_n} \frac{\mathrm{E}_n(y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x}))^2}{\max\{\exp(-\xi^2\lambda_1^* m/k_0), k_0(\sum_{j\in N_n} |\beta_j^*|)^2 \log p/n\}} \le \bar{C} \text{ on } C_n(K_n),$$

where $\bar{C}$ is some positive constant. In view of (S2.9), the desired conclusion (S2.3) follows if

$$\lim_{n\to\infty} P(C_n(K_n)) = 1. \tag{S2.10}$$

Since (2.28) is still valid when (3.8) is used in place of (2.19), (S2.10) can be proved in a similar fashion as in the proof of (2.30). The details are omitted.

To prove (S2.4), let $\bar{R}$ be large enough such that $\bar{C}\exp(-\xi^2\lambda_1^*\bar{R}) < \lambda_1^*\underline{\theta}^2/2$, where $\underline{\theta}$ is defined in (3.7). By (S2.9), (S2.10), and (3.7),

$$\lim_{n\to\infty} P(\mathrm{E}_n(y(\mathbf{x}) - y_{\hat{J}_{\bar{R}k_0}}(\mathbf{x}))^2 \le \lambda_1^*\underline{\theta}^2/2) = 1. \tag{S2.11}$$

Define $G_n = \{N_n \bigcap \hat{J}^c_{\bar{R}k_0} \neq \emptyset\}$. Then, on the set G_n,

$$\mathrm{E}_n(y(\mathbf{x}) - y_{\hat{J}_{\bar{R}k_0}}(\mathbf{x}))^2 \ge \lambda_1^* \sum_{j\notin \hat{J}_{\bar{R}k_0}} \beta_j^{*^2} \ge \lambda_1^*\underline{\theta}^2,$$

and hence by (S2.11),

$$P(G_n) \le P(\mathrm{E}_n(y(\mathbf{x}) - y_{\hat{J}_{\bar{R}k_0}}(\mathbf{x}))^2 \ge \lambda_1^*\underline{\theta}^2) = o(1).$$

Thus (S2.4) is proved. □

We are in the position to prove (3.6).

Theorem S2.2. *Suppose that* (3.2) *and the same assumptions as in Theorem S2.1 hold. Then, for s_a satisfying* (3.3), (3.6) *follows.*

Proof. Define $E_n = \{N_n \subseteq \hat{J}_{\bar{R}k_0}\}$ and $\tilde{k}_n = \min_{1\le k\le K_n}\{k : 1 \le k \le K_n, N_n \subseteq \hat{J}_k\}$ ($\min\emptyset = K_n$). We first show that

$$\lim_{n\to\infty} P(\hat{k}_n < \tilde{k}_n, E_n) = 0. \tag{S2.12}$$

Straightforward calculations yield on the set $\{\hat{k}_n < \tilde{k}_n\}\bigcap E_n$,

$$\beta_{\hat{j}_{\tilde{k}_n}}^{*^2}\hat{A}_n - 2|\beta^*_{\hat{j}_{\tilde{k}_n}}\hat{B}_n| \le \frac{s_a k_0 \bar{R}\log p}{n}\hat{\sigma}^2_{\hat{J}_{\tilde{k}_n}}, \tag{S2.13}$$

where

$$\hat{A}_n = n^{-1}\mathbf{Z}^\top_{\hat{j}_{\tilde{k}_n}}(\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}_n-1}})\mathbf{Z}_{\hat{j}_{\tilde{k}_n}} \text{ and } \hat{B}_n = n^{-1}\mathbf{Z}^\top_{\hat{j}_{\tilde{k}_n}}(\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}_n-1}})\boldsymbol{\varepsilon},$$

with $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$. Note also that

$$\hat{A}_n \ge \min_{\sharp(J)\le\bar{R}k_0} \lambda_{\min}(\boldsymbol{\Gamma}(J)) - \max_{\sharp(J)\le\bar{R}k_0} \|\hat{\boldsymbol{\Gamma}}(J) - \boldsymbol{\Gamma}(J)\|,$$

$$\hat{B}_n \leq \max_{1\leq i\leq p} |n^{-1}\sum_{t=1}^{n} z_{ti}\varepsilon_t|$$
$$+ \max_{\sharp(J)\leq \bar{R}k_0} \|\hat{\boldsymbol{\Gamma}}^{-1}(J)\| \bar{R}k_0 (1 + \max_{\sharp(J)\leq \bar{R}k_0} \|\boldsymbol{\Gamma}^{-1}(J)\boldsymbol{g}_i(J)\|_1)$$
$$\times \max_{1\leq i,j\leq p} |n^{-1}\sum_{t=1}^{n} z_{ti}z_{tj} - \rho_{ij}| \max_{1\leq i\leq p} |n^{-1}\sum_{t=1}^{n} z_{ti}\varepsilon_t|$$
$$+ \max_{\sharp(J)\leq \bar{R}k_0} \|\boldsymbol{\Gamma}^{-1}(J)\boldsymbol{g}_i(J)\|_1 \max_{1\leq i\leq p} |n^{-1}\sum_{t=1}^{n} z_{ti}\varepsilon_t|,$$

and

$$|\hat{\sigma}^2_{\hat{J}_{\tilde{k}_n}} - \sigma^2| \leq |n^{-1}\sum_{t=1}^{n} \varepsilon_t^2 - \sigma^2|$$
$$+ \max_{\sharp(J)\leq \bar{R}k_0} \|\hat{\boldsymbol{\Gamma}}^{-1}(J)\| \bar{R}k_0 \big(\max_{1\leq i\leq p} |n^{-1}\sum_{t=1}^{n} z_{ti}\varepsilon_t| \big)^2.$$

Therefore, by (A1), (A2), (A5), (2.28), (2.20), (3.8), (3.2), and (3.7) (which yields $|\beta^*_{\hat{j}_{\tilde{k}_n}}| = o\big((n/(k_0 \log p))^{1/2}\big)$), one obtains,

$$|\beta^*_{\hat{j}_{\tilde{k}_n}} \hat{B}_n| = o_p(1), \tag{S2.14}$$

and

$$\hat{\sigma}^2_{\hat{J}_{\tilde{k}_n}} - \sigma^2 = o_p(1). \tag{S2.15}$$

Moreover, there exists some positive constant $\underline{v}$ such that

$$\lim_{n\to\infty} P(\beta^{*2}_{\hat{j}_{\tilde{k}_n}} \hat{A}_n \geq \underline{\theta}^2 \hat{A}_n \geq \underline{v}) = 1. \tag{S2.16}$$

Consequently, (S2.12) follows from (S2.13)–(S2.16).

We next show that

$$\lim_{n\to\infty} P(\hat{k}_n > \tilde{k}_n, E_n) = 0. \tag{S2.17}$$

Some algebraic manipulations imply that on the set $\{\hat{k}_n > \tilde{k}_n\} \bigcap E_n$,

(S2.18)
$$\left(1 + \frac{s_a K_n \log p}{n}\right) n^{-1}\boldsymbol{\varepsilon}^\top (\mathbf{H}_{\hat{J}_{\hat{k}_n}} - \mathbf{H}_{\hat{J}_{\tilde{k}_n}})\boldsymbol{\varepsilon} \geq \frac{s_a(\hat{k}_n - \tilde{k}_n)\log p}{n} \hat{\sigma}^2_{\hat{J}_{\tilde{k}_n}},$$

and

$$\begin{aligned} n^{-1}\boldsymbol{\varepsilon}^\top(\mathbf{H}_{\hat{J}_{\hat{k}_n}} - \mathbf{H}_{\hat{J}_{\tilde{k}_n}})\boldsymbol{\varepsilon} &\leq 2(\hat{k}_n - \tilde{k}_n) \max_{\sharp(J)\leq K_n} \|\hat{\boldsymbol{\Gamma}}^{-1}(J)\| \\ &\times \left\{ (\max_{1\leq i,j\leq p} |n^{-1}\sum_{t=1}^n z_{ti}z_{tj} - \rho_{ij}|)^2 + (\max_{1\leq i\leq p} |n^{-1}\sum_{t=1}^n z_{ti}\varepsilon_t|)^2 \right\} \end{aligned} \tag{S2.19}$$

As a result, (S2.17) follows from (S2.19), (S2.18), (2.20), (S2.15), (3.3), (2.28), (A1), and (A2). Thus,

$$\lim_{n\to\infty} P(\hat{k}_n = \tilde{k}_n, E_n) = 0. \tag{S2.20}$$

Now, by (A1), (A2), (2.28), and

$$\lim_{n\to\infty} P(\|\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_{\tilde{k}_n}) - \boldsymbol{\Gamma}^{-1}(\hat{J}_{\tilde{k}_n})\| \leq k_0\bar{R}c_2^*(\log p/n)^{1/2}, \tilde{k}_n \leq \bar{R}k_0) = 1,$$

one gets

$$\begin{aligned} &\mathrm{E}_n(\hat{y}_{\hat{k}_n}(\mathbf{x}) - y(\mathbf{x}))^2 I_{\{\hat{k}_n=\tilde{k}_n,E_n\}} \\ &\leq \|n^{-1}\sum_{t=1}^n \mathbf{z}_t(\hat{J}_{\tilde{k}_n})\varepsilon_t\|^2 \|\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_{\tilde{k}_n})\| I_{\{\tilde{k}_n\leq\bar{R}k_0\}} \\ &+ \|n^{-1}\sum_{t=1}^n \mathbf{z}_t(\hat{J}_{\tilde{k}_n})\varepsilon_t\|^2 \|\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_{\tilde{k}_n})\|^2 \|\hat{\boldsymbol{\Gamma}}(\hat{J}_{\tilde{k}_n}) - \boldsymbol{\Gamma}(\hat{J}_{\tilde{k}_n})\| I_{\{\tilde{k}_n\leq\bar{R}k_0\}} \\ &= O_p\left(\frac{k_0 \log p}{n}\right), \end{aligned}$$

which, together with (S2.20), leads to the desired conclusion (3.6). □

In the rest of this section, we prove (3.30)–(3.32) and (3.45).

Proof of (3.30). By an argument similar to that used to prove

(3.29), one obtains

$$
\begin{aligned}
&\left|n^{-1}\sum_{t=1}^{n}\varepsilon_t(\hat{J}_m)\varepsilon_t\right| \\
&= \left|\sum_{\sharp(J)=m}\left\{n^{-1}\sum_{t=1}^{n}\varepsilon_t(J)\varepsilon_t\right\} I_{\{\hat{J}_m=J\}}\right| \\
&\le R_{2,p}\sum_{\sharp(J)=m}\sum_{j=1}^{p}|\beta_j^* - \beta_j^*(J)| I_{\{\hat{J}_m=J\}} \\
&\le C_{M,\gamma,\lambda_1}^{1/2} R_{2,p}\big\{\mathrm{E}_n(\varepsilon^2(\hat{J}_m))\big\}^{(\gamma-1)/(2\gamma-1)}.
\end{aligned}
$$

Thus, (3.30) follows. □

Proof of (3.31). By (2.14) with $C = M+1$, (2.11), (2.19), and for any $J \subseteq \{1,\dots,p\}$ and $i \in J$,

$$
\sum_{j=1}^{p}(\beta_j^* - \beta_j^*(J))\rho_{ji} = 0,
$$

it follows that for any $1 \le m \le K_n$,

$$
\begin{aligned}
&\|n^{-1}\sum_{t=1}^{n}\mathbf{z}_t(\hat{J}_m)\varepsilon_t(\hat{J}_m)\|^2_{\hat{\mathbf{\Gamma}}^{-1}(m)} \le \|\hat{\mathbf{\Gamma}}^{-1}(K_n)\|\|n^{-1}\sum_{t=1}^{n}\mathbf{z}_t(\hat{J}_m)\varepsilon_t(\hat{J}_m)\|^2 \\
&\le \|\hat{\mathbf{\Gamma}}^{-1}(K_n)\|\sum_{\sharp(J)=m}\left\{\sum_{i\in J}[n^{-1}\sum_{t=1}^{n}\sum_{j=1}^{p}(\beta_j^* - \beta_j^*(J))(z_{tj}z_{ti} - \rho_{ji})]^2\right\} I_{\{\hat{J}_m=J\}} \\
&\le \|\hat{\mathbf{\Gamma}}^{-1}(K_n)\| m R_{1,p}^2 \sum_{\sharp(J)=m}(\sum_{j=1}^{p}|\beta_j^* - \beta_j^*(J)|)^2 I_{\{\hat{J}_m=J\}} \\
&= \|\hat{\mathbf{\Gamma}}^{-1}(K_n)\| C_{M,\gamma,\lambda_1} m R_{1,p}^2\big\{\mathrm{E}_n(\varepsilon^2(\hat{J}_m))\big\}^{(2\gamma-2)/(2\gamma-1)},
\end{aligned}
$$

yielding (3.31). □

Proof of (3.32). Equation (3.32) follows directly from

$$\begin{aligned}
&\|n^{-1}\sum_{t=1}^{n}\mathbf{z}_t(\hat{J}_m)\varepsilon_t\|^2_{\hat{\mathbf{\Gamma}}^{-1}(m)} \le \|\hat{\mathbf{\Gamma}}^{-1}(K_n)\|\|n^{-1}\sum_{t=1}^{n}\mathbf{z}_t(\hat{J}_m)\varepsilon_t\|^2 \\
&\le \|\hat{\mathbf{\Gamma}}^{-1}(K_n)\| \sum_{\sharp(J)=m}\left\{\sum_{i\in J}(n^{-1}\sum_{t=1}^{n} z_{ti}\varepsilon_t)^2\right\} I_{\{\hat{J}_m=J\}} \\
&\le \|\hat{\mathbf{\Gamma}}^{-1}(K_n)\| m R^2_{2,p},
\end{aligned}$$

where $1 \le m \le K_n$. □

Proof of (3.45). Since for $\gamma > 1$ and all large n, $m_n^* = (n/\log p)^{1/2\gamma}$, $K_n > V m_n^*$, and $s_a m_n^* \log p/n < V^{-1}$, it follows that

(S2.21)

$$\begin{aligned}
&P(\hat{k}_n > V m_n^*) \le P(\hat{k}_n > V m_n^*, A_n(K_n)) + P(A_n^c(K_n)) \\
&\le P\left(\min_{V m_n^* \le k \le K_n} \frac{s_a(k - m_n^*)\log p}{(1+V^{-1})n}\hat{\sigma}^2_{\hat{J}_{K_n}} - Z_n(k) \le 0, A_n(K_n)\right) + P(A_n^c(K_n)),
\end{aligned}$$

as n is sufficiently large, where

$$\begin{aligned}
Z_n(k) &= n^{-1}\sum_{t=1}^{n}\varepsilon_t^2(\hat{J}_{m_n^*}) + 2n^{-1}|\sum_{t=1}^{n}\varepsilon_t(\hat{J}_{m_n^*})\varepsilon_t| + 2n^{-1}|\sum_{t=1}^{n}\varepsilon_t(\hat{J}_k)\varepsilon_t| \\
&+ 2\|n^{-1}\sum_{t=1}^{n}\mathbf{z}_t(\hat{J}_k)\varepsilon_t(\hat{J}_k)\|^2_{\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_k)} + 2\|n^{-1}\sum_{t=1}^{n}\mathbf{z}_t(\hat{J}_k)\varepsilon_t\|^2_{\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_k)}.
\end{aligned}$$

By (3.29)–(3.32), (2.28), (3.2), and (2.27), one obtains for $V m_n^* \le k \le K_n$ and all large n,

$$\begin{aligned}
&n^{-1}\sum_{t=1}^{n}\varepsilon_t^2(\hat{J}_{m_n^*}) \le \left\{C2 + C_{M,\gamma,\lambda_1} c_2^*(C2+1)\left(\frac{\log p}{n}\right)^{(\gamma-1)/2\gamma}\right\} \\
&\times \left(\frac{\log p}{n}\right)^{1-1/(2\gamma)} \le \frac{Q_1^*}{V}\frac{k\log p}{n} \text{ on } W_n \bigcap A_n(K_n),
\end{aligned} \tag{S2.22}$$

where $Q_1^* = (C_2 + \iota_1)$ with ι_1 being arbitrarily small positive constant,

$$n^{-1}|\sum_{t=1}^{n}\varepsilon_t(\hat{J}_{m_n^*})\varepsilon_t| \le \frac{Q_2^*}{V}\frac{k\log p}{n} \text{ on } W_n \bigcap A_n(K_n), \tag{S2.23}$$

where $Q_2^* = \{C_{M,\gamma,\lambda_1}^{1/2} c_1^* [1 + C_2^{1/2}]\}$,

$$n^{-1} \Big| \sum_{t=1}^{n} \varepsilon_t(\hat{J}_k) \varepsilon_t \Big| \leq \frac{Q_2^*}{V} \frac{k \log p}{n} \text{ on } W_n \bigcap A_n(K_n), \tag{S2.24}$$

$$\begin{aligned} \| n^{-1} \sum_{t=1}^{n} \mathbf{z}_t(\hat{J}_k) \varepsilon_t(\hat{J}_k) \|^2_{\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_k)} &\leq \big(1 + C_2\big) \big(\frac{\log p}{n}\big)^{1-1/\gamma} \\ \times C_{M,\gamma,\lambda_1} \bar{B} c_2^{*^2} \frac{k \log p}{n} &\leq \frac{k \log p}{Vn} \text{ on } W_n \bigcap A_n(K_n), \end{aligned} \tag{S2.25}$$

$$\| n^{-1} \sum_{t=1}^{n} \mathbf{z}_t(\hat{J}_k) \varepsilon_t \|^2_{\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_k)} \leq \frac{\bar{B} c_1^{*^2} k \log p}{n} \text{ on } W_n, \tag{S2.26}$$

and

$$\lim_{n \to \infty} P(-\sigma^2/V \leq \hat{\sigma}^2_{\hat{J}_{K_n}} - \sigma^2 \leq \sigma^2/V) = 1. \tag{S2.27}$$

Let V be large enough such that

$$(Q_1^* + 4Q_2^* + 2)/(V\sigma^2) + \bar{V}^* < s_a(1 - V^{-1})^2/(1 + V^{-1}), \tag{S2.28}$$

where $\bar{V}^*$ is defined in (3.28). Then, by (S2.22)–(S2.28),

$$\begin{aligned} \min_{Vm_n^* \leq k \leq K_n} \frac{s_a(k - m_n^*) \log p}{(1 + V^{-1})n} \hat{\sigma}^2_{\hat{J}_{K_n}} - Z_n(k) &\geq \frac{\sigma^2 V m_n^* \log p}{n} \\ \times \left[\frac{(1 - V^{-1})^2 s_a}{1 + V^{-1}} - \bar{V}^* - \frac{Q_1^* + 4Q_2^* + 2}{V\sigma^2} \right] &\text{ on } W_n \bigcap A_n(K_n), \end{aligned}$$

which, together with (2.30), (3.36), and (S2.21), gives (3.45). □

S3. A Simulation Study based on high-dimensional ARX models.

S3.1. *Some numerical results.* In this section, we report the finite-sample performance of OGA+HDAIC under the following ARX model,

$$y_t = a y_{t-1} + \sum_{j=1}^{p} \alpha_j x_t^{(j)} + \epsilon_t, \quad t = 1, \ldots, n, \tag{S3.1}$$

in which $|a| < 1$, α_j satisfies (A3), (A4) or (2.17),

$$x_t^{(j)} = \phi x_{t-1}^{(j)} + \varepsilon_{tj}, \tag{S3.2}$$

with $|\phi| < 1$,

$$\varepsilon_{tj} = d_{tj} + \eta w_t, \tag{S3.3}$$

$$|\eta\phi| \neq 0, \tag{S3.4}$$

and $(d_{t1}, \ldots, d_{tp}, w_t)^\top$ being i.i.d. $(p+1)$-dimensional standard normal vector, and $\{\epsilon_t\}$ is a sequence of i.i.d. standard normal variables independent of $\{(d_{t1}, \ldots, d_{tp}, w_t)^\top\}$. Relation (S3.4) implies that the regressor vector $\mathbf{x}_t = (x_{t1}, \ldots, x_{t,p+1})^\top \equiv (y_{t-1}, x_t^{(1)}, \ldots, x_t^{(p)})^\top$ are not only contemporaneously dependent, but also time dependent. Moreover, it is not difficult to see that $\lambda_{\max}(\mathbf{\Gamma})$ grows to ∞ as p does, where $\mathbf{\Gamma} = \mathrm{E}(\mathbf{z}_t\mathbf{z}_t^\top)$ with $\mathbf{z}_t = (z_{t1}, \ldots, z_{t,p+1})^\top$ and $z_{ti} = x_{ti}/(var(x_{ti}))^{1/2}$.

On the other hand, we show in the Section S3.2 that

$$\lambda_{\min}(\mathbf{\Gamma}) \text{ satisfies (2.19).} \tag{S3.5}$$

With an argument similar to that used to prove (S3.5), it also can be shown that there exists some positive constant $C_1 < \infty$ such that for any $J \subset \{1, \ldots, p+1\}$ and $1 \leq i \leq p+1$, $\|\mathbf{\Gamma}^{-1}(J)\boldsymbol{g}_i(J)\| \leq C_1$, and hence (A5) follows. Consequently, it is valid to apply OGA+HDAIC to model (S3.1), according to Theorem 3.1.

In our numerical study, (n,p)=(100, 1000), (200, 2000), or (400, 4000), $\phi = -0.8, -0.5, 0.5$ or 0.8, $\eta = 1$ or 2, and $a = -0.7, 0$ or 0.8. The maximum number K_n of the OGA iterations is given by $10\sqrt{n/\log p}$, and the penalty term s_a in HDAIC ((3.1)) is chosen from $\bar{S} = \{s_{a,1} = 1.75, s_{a,2} = 2.75, s_{a,3} = 3.75, s_{a,4} = 4.75\}$ in a data-driven manner. More specifically, we first select model $\hat{J}_{\hat{k}_n,j}$ using OGA+HDAIC$_j$, $j = 1, \ldots, 4$, where HDAIC$_j$ is HDAIC with $s_a = s_{a,j}$. We then choose $s_{a,\hat{j}}$ from a prediction point of view, where $\hat{j}$ is the minimizer of the prediction loss,

$$\frac{1}{n - \lfloor 4n/5 \rfloor} \sum_{t=\lfloor 4n/5 \rfloor + 1}^{n} \left(y_t - \hat{y}_{t,\hat{J}_{\hat{k}_n,j}} \right)^2,$$

over $1 \le j \le 4$. Here $\hat{y}_{t,\hat{J}_{\hat{k}_{n,j}}} = \mathbf{x}_t^\top(\hat{J}_{\hat{k}_{n,j}})\hat{\boldsymbol{\beta}}_{t-1}(\hat{J}_{\hat{k}_{n,j}})$ and

$$\hat{\boldsymbol{\beta}}_t(J) = (\sum_{l=1}^{t} \mathbf{x}_l(J)\mathbf{x}_l^\top(J))^{-1} \sum_{l=1}^{t} \mathbf{x}_l(J) y_l.$$

While $\bar{S}$ is allowed to vary from one user to another, we have found that the $\bar{S}$ given above usually leads to reliable results.

Let $(y_t^{(i)}, \mathbf{x}_t^{(i)}), t = 1, \ldots, n$, be observations generated from model (S3.1) and $\hat{J}^{(i)}_{\hat{k}_{n,\hat{j}^{(i)}}}$ denote the model selected by OGA+HDAIC at the ith simulation run. In view of Theorem 3.1, the performance of OGA+HDAIC is evaluated by the empirical mean squared prediction error (EMSPE),

$$\text{EMSPE} = \frac{1}{1000} \sum_{i=1}^{1000} \left[y(\mathbf{x}^{(i)}) - \hat{y}^{(i)}_{\hat{k}_{n,\hat{j}^{(i)}}} \right]^2, \tag{S3.6}$$

where $\mathbf{x}^{(i)}$ is an independent copy of $\{\mathbf{x}_t^{(i)}\}$ obtained at the ith simulation run, $y(\mathbf{x}^{(i)}) = \boldsymbol{\beta}^\top \mathbf{x}^{(i)}$ with $\boldsymbol{\beta} = (a, \alpha_1, \ldots, \alpha_p)^\top$, and

$$\hat{y}^{(i)}_{\hat{k}_{n,\hat{j}^{(i)}}} = \mathbf{x}^{(i)^\top}(\hat{J}^{(i)}_{\hat{k}_{n,\hat{j}^{(i)}}})\hat{\boldsymbol{\beta}}^{(i)}(\hat{J}^{(i)}_{\hat{k}_{n,\hat{j}^{(i)}}})$$

with $\hat{\boldsymbol{\beta}}^{(i)}(J) = (\sum_{t=1}^n \mathbf{x}_t^{(i)}(J)\mathbf{x}_t^{(i)^\top}(J))^{-1} \sum_{t=1}^n \mathbf{x}_t^{(i)}(J) y_t^{(i)}$. In addition to OGA+HDAIC, we also evaluate the performance of Lasso and ISIS-SCAD (Fan and Lv (2008)) via (S3.6) with $\hat{y}^{(i)}_{\hat{k}_{n,\hat{j}^{(i)}}}$ replaced by the predictors obtained from these two methods. We use the `Glmnet` and `SIS` packages in `R` to implement Lasso and ISIS-SCAD, respectively. Note that it does not seem appropriate to choose the tuning parameter λ_n for Lasso by cross validation because time series data have a natural temporal ordering. We therefore set $\lambda_n = (\log p/n)^{1/2}$, which is suggested in the numerical section of Basu and Michailidis (2015). The following three examples compare the aforementioned three model selection methods under different sparsity conditions.

Example 1. We set $\alpha_j = 15j^{-\gamma}, \gamma = 1.5, 2.5$ or 3.5, which satisfy (A3), the polynomial decay case. We compute the EMSPEs of OGA+HDAIC, Lasso, and ISIS-SCAD, and summarize the results in Table S1. The table reveals that when $\gamma \ge 2.5$, OGA+HDAIC outperforms the other two methods in most cases, except for $(\gamma, \eta) = (2.5, 1)$ and $(a, \phi) = (0.8, 0.8)$ or (-0.7, -0.8) (in which ISIS-SCAD has the smallest EMSPEs) and for $(\gamma, \eta) = (2.5, 2)$ and $(a, \phi) = (0.8, 0.8)$ or (-0.7, -0.8) (in which Lasso has the smallest EMSPEs when $n < 400$). On the other hand, Lasso usually works better than OGA+HDAIC and ISIS-SCAD in the case of $(\gamma, \eta) = (1.5, 1)$. This is particularly true for $a = 0$. However, the advantage of Lasso under this pair of (γ, η) vanishes when η increases to 2. More specifically, when $(\gamma, \eta) = (1.5, 2)$, Lasso is still better than the other two methods if $(n, p) = (100, 1000)$, but OGA+HDAIC tends to surpass Lasso when $(n, p) = (200, 2000)$ or (400, 4000).

Example 2. We set $\alpha_j = 15\exp(-\beta j), \beta = 1, 1.5$ or 2, which satisfy (A4), the exponential decay case. The EMSPEs of OGA+HDAIC, Lasso, and ISIS+SCAD are summarized in Table S2. The table shows that OGA+HDAIC outperforms Lasso and ISIS-SCAD in the majority of combinations of $(\eta, \beta, n, p, a, \phi)$. Since the previous example suggests that OGA+HDIC tends to surpass the other two methods when γ becomes large and since it is shown in Section 2.1 that the exponential decay case can be viewed as the case of $\gamma \to \infty$, this phenomenon does not seem counterintuitive. For most (a, ϕ) pairs, Lasso demonstrates a clear advantage over ISIS-SCAD. The only exception is $(a, \phi) = (-0.7, 0.8)$, in which both methods are comparable.

Example 3 We set $(\alpha_1, \ldots, \alpha_5) = (3, -3.5, 4, -2.8, 3.2)$ and $\alpha_j = 0$ for $j \geq 6$. Therefore, (2.17) (which is stronger than (3.7)) is satisfied. The EMSPEs of OGA+HDAIC, Lasso, and ISIS-SCAD are summarized in Table S3. Since (2.17) can also be viewed as the case of $\gamma \to \infty$ (see Section 2.1), it is not surprising to see that OGA+HDAIC has smaller EMSPEs than those of Lasso in all cases of Table S3. Actually, OGA+HDAIC also works better than ISIS-SCAD, except for $(n, p) = (400, 4000)$, $\eta = 1$, and $\phi = 0.5$ or -0.5. In this connection, it is interesting to note that ISIS-SCAD is better than Lasso as long as $n \geq 200$. For $n = 100$, Lasso outperforms ISIS-SCAD when $\phi = -0.8$ or 0.8, and is comparable to the latter when $\phi = -0.5$ or 0.5.

The above examples suggest that the finite sample performance of our method compares favorably to Lasso and ISIS-SCAD regardless of what kind of sparsity condition is being imposed on model (S3.1). This is in particular the case when the regression coefficients decay relatively quickly, the correlations among the regressors are relatively high, or (n, p) is relatively large.

TABLE S1

The EMSPEs of OGA+HDAIC (M1), Lasso (M2), and ISIS-SCAD (M3) in Example 1 with the smalleest one marked in blue

EMSPE, $a = 0$

			$\phi = -0.8$			$\phi = -0.5$			$\phi = 0.5$			$\phi = 0.8$		
η	γ	(n, p)	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
1	1.5	(100, 1000)	22.1961	10.7045	36.8100	13.2042	8.7596	20.4714	14.7278	9.2787	21.6220	23.2504	10.3417	34.3955
		(200, 2000)	6.7092	4.8439	7.5811	5.3121	3.3460	5.0037	5.2053	3.8885	5.8881	7.0679	4.8641	7.6271
		(400, 4000)	3.4712	2.6017	2.7204	2.5430	1.5501	2.0693	2.4968	1.7043	2.0720	3.4137	2.8289	2.9407
	2.5	(100, 1000)	1.0024	1.1157	1.2848	0.6626	0.6784	1.2904	0.6183	0.6804	1.3824	0.9866	0.9630	1.3702
		(200, 2000)	0.4761	0.5051	0.8689	0.3427	0.3459	1.0195	0.3356	0.3680	1.0319	0.4566	0.5218	0.9558
		(400, 4000)	0.2498	0.2775	0.5752	0.1858	0.2089	0.6094	0.1960	0.2085	0.6584	0.2472	0.3021	0.5855
	3.5	(100, 1000)	0.3119	0.4929	0.9786	0.2938	0.3436	0.9638	0.3029	0.3399	1.0584	0.3317	0.4347	0.9692
		(200, 2000)	0.1527	0.2241	0.6367	0.1432	0.1771	0.6018	0.1399	0.1894	0.6403	0.1548	0.2494	0.6642
		(400, 4000)	0.0862	0.1363	0.3458	0.0787	0.1091	0.1289	0.0715	0.1054	0.1440	0.0932	0.1349	0.3235
2	1.5	(100, 1000)	22.3940	7.5179	23.3027	14.7638	8.3148	18.5156	14.0111	9.0098	19.3890	24.0272	7.1240	22.8202
		(200, 2000)	7.5428	6.8712	8.4175	5.5065	5.6629	4.9733	4.9934	5.3248	4.3016	6.7764	6.7898	7.8329
		(400, 4000)	3.2673	5.0939	4.9644	2.3035	2.7283	2.6599	2.5927	2.9722	2.6172	3.4509	5.4842	5.1715
	2.5	(100, 1000)	1.0731	0.9132	0.9160	0.6608	0.7221	1.1767	0.7367	0.7143	1.1288	1.0629	0.8335	1.0533
		(200, 2000)	0.4757	0.4890	0.5110	0.3454	0.3886	0.7700	0.3324	0.4041	0.8044	0.4887	0.5265	0.5016
		(400, 4000)	0.2457	0.3507	0.2462	0.2030	0.2274	0.4057	0.1926	0.2383	0.4129	0.2606	0.3456	0.2646
	3.5	(100, 1000)	0.3644	0.4364	0.6999	0.2922	0.3558	0.8992	0.2928	0.3496	0.8521	0.3770	0.3889	0.6784
		(200, 2000)	0.1829	0.2111	0.3687	0.1405	0.1842	0.6152	0.1446	0.1899	0.5898	0.1934	0.2184	0.3909
		(400, 4000)	0.0890	0.1306	0.1243	0.0797	0.1090	0.4512	0.0743	0.1036	0.4463	0.0945	0.1245	0.1278

EMSPE, $a = 0.8$

			$\phi = -0.8$			$\phi = -0.5$			$\phi = 0.5$			$\phi = 0.8$		
η	γ	(n, p)	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
1	1.5	(100, 1000)	23.9721	17.5310	85.7529	14.0960	13.7620	38.3318	14.8493	11.6079	36.4437	26.0687	6.8995	28.5434
		(200, 2000)	7.5498	6.9242	11.1260	5.1280	5.0385	8.4774	5.6938	5.1562	5.1509	7.9694	3.5413	9.6480
		(400, 4000)	3.4080	3.0575	2.7114	2.4409	2.0277	1.9751	2.5443	2.3934	2.5755	3.4371	2.9683	6.5900
	2.5	(100, 1000)	1.0967	1.7889	1.3915	0.6926	1.0308	1.2704	0.6651	1.1098	1.4305	1.1701	1.2033	0.9691
		(200, 2000)	0.4877	0.8384	0.8278	0.3717	0.5015	0.9865	0.3906	0.5398	0.7374	0.4768	0.5393	0.3866
		(400, 4000)	0.2519	0.4177	0.6411	0.2036	0.2771	0.6515	0.1860	0.2896	0.2040	0.2358	0.2957	0.2099
	3.5	(100, 1000)	0.3134	0.9253	0.8893	0.2798	0.5682	1.1478	0.3371	0.5347	1.0402	0.4100	0.6334	0.5409
		(200, 2000)	0.1571	0.4552	0.6569	0.1454	0.2801	0.6570	0.1401	0.2657	0.4705	0.1623	0.3063	0.1391
		(400, 4000)	0.0902	0.2518	0.3456	0.0802	0.1672	0.1522	0.0820	0.1453	0.0870	0.0836	0.1713	0.0880
2	1.5	(100, 1000)	22.9579	11.8524	92.0668	14.5482	11.7189	123.6786	15.3556	9.9395	17.7668	24.9125	9.1306	36.8341
		(200, 2000)	7.2616	8.1428	8.7994	5.5308	6.6218	8.8938	5.3159	6.4400	8.0841	8.0254	9.8247	25.2122
		(400, 4000)	3.3697	5.4927	3.7112	2.6246	3.2296	2.2848	2.7131	4.7056	5.9056	3.4238	10.8819	19.7596
	2.5	(100, 1000)	1.0808	1.8080	1.2866	0.7967	1.1999	5.8798	0.7378	0.9207	0.7579	1.1800	0.9973	1.4335
		(200, 2000)	0.5025	0.8315	0.7196	0.3777	0.6013	0.8206	0.3642	0.4394	0.3981	0.5745	0.5116	0.9175
		(400, 4000)	0.2518	0.4815	0.3871	0.2117	0.3293	0.6060	0.2063	0.2533	0.2122	0.2490	0.3913	0.5946
	3.5	(100, 1000)	0.4047	0.9779	0.7136	0.3067	0.6546	3.6636	0.3522	0.5108	0.6496	0.4470	0.6484	0.4173
		(200, 2000)	0.1786	0.4693	0.5181	0.1464	0.3093	0.6508	0.1486	0.2189	0.1695	0.1827	0.3207	0.2165
		(400, 4000)	0.0885	0.2478	0.3413	0.0807	0.1807	0.5232	0.0779	0.0877	0.0678	0.0924	0.1354	0.1309

EMSPE, $a = -0.7$

			$\phi = -0.8$			$\phi = -0.5$			$\phi = 0.5$			$\phi = 0.8$		
η	γ	(n, p)	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
1	1.5	(100, 1000)	23.7873	6.9571	27.5542	16.4657	11.7806	31.5113	15.0582	15.0576	38.8196	25.7137	19.6362	74.7621
		(200, 2000)	7.8205	3.4614	8.6765	5.6263	5.1266	5.6158	5.0666	5.5067	6.8292	8.1514	7.6958	9.6657
		(400, 4000)	3.6674	3.1707	5.3638	2.5397	2.2089	2.2186	2.5357	2.2711	2.0741	3.3660	3.3883	2.6878
	2.5	(100, 1000)	1.1392	1.2737	0.9707	0.7059	1.0189	1.4391	0.7087	1.1831	1.3936	1.1557	2.1703	1.4029
		(200, 2000)	0.5040	0.5659	0.4494	0.3767	0.4872	0.9395	0.3763	0.5687	1.0103	0.5651	0.9838	0.9273
		(400, 4000)	0.2312	0.3013	0.1812	0.2097	0.2644	0.3598	0.2009	0.2995	0.6883	0.2627	0.4622	0.6653
	3.5	(100, 1000)	0.3593	0.5783	0.6178	0.3092	0.4888	1.0893	0.3473	0.6419	1.1307	0.3789	1.1312	0.9047
		(200, 2000)	0.1649	0.2953	0.1986	0.1399	0.2400	0.6399	0.1393	0.3347	0.6228	0.1677	0.5503	0.6058
		(400, 4000)	0.0896	0.1648	0.0807	0.0821	0.1361	0.1279	0.0836	0.1732	0.1582	0.0906	0.2590	0.4024
2	1.5	(100, 1000)	20.9301	8.4471	30.1870	16.6701	8.5779	24.6006	16.0150	13.2857	89.8765	23.2190	11.2563	89.8992
		(200, 2000)	7.5029	8.2639	19.2839	5.7050	6.3007	6.4398	5.3648	6.7807	10.0675	7.3821	8.2874	12.6981
		(400, 4000)	3.4153	7.6315	13.2952	2.6129	4.4953	4.5728	2.3785	3.0696	2.1969	3.3456	5.2534	3.8305
	2.5	(100, 1000)	1.1858	0.8933	1.2094	0.7702	0.8371	6.6903	0.6787	1.3412	7.3425	1.2542	2.0541	1.2137
		(200, 2000)	0.5244	0.4930	0.6567	0.3834	0.4255	0.4109	0.3563	0.6768	0.8503	0.6031	0.9921	0.7274
		(400, 4000)	0.2607	0.3435	0.4033	0.1943	0.2472	0.1933	0.1992	0.3494	0.6045	0.2594	0.5429	0.3693
	3.5	(100, 1000)	0.3961	0.6252	0.3350	0.3107	0.4713	0.6952	0.3588	0.6951	0.8830	0.4216	1.1544	0.8793
		(200, 2000)	0.1686	0.2738	0.1913	0.1481	0.2115	0.3070	0.1424	0.3743	0.7121	0.1792	0.5662	0.5236
		(400, 4000)	0.0914	0.1361	0.1077	0.0779	0.1021	0.0823	0.0834	0.1936	0.5124	0.0951	0.2682	0.3333

Table S2

The EMSPEs of OGA+HDAIC (M1), Lasso (M2), and ISIS-SCAD (M3) in Example 2 with the smallest one marked in blue

			EMSPE											
			$a = 0$											
			$\phi = -0.8$			$\phi = -0.5$			$\phi = 0.5$			$\phi = 0.8$		
η	β	(n, p)	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
1	1.0	(100, 1000)	0.3281	0.5610	0.8980	0.2708	0.4023	0.9951	0.2863	0.4624	1.0768	0.4253	0.6136	0.9255
		(200, 2000)	0.1185	0.2842	0.5572	0.1256	0.2280	0.5828	0.1403	0.2233	0.6203	0.1446	0.2868	0.6510
		(400, 4000)	0.0722	0.1429	0.2416	0.0644	0.1202	0.1232	0.0671	0.1277	0.1217	0.0757	0.1604	0.2755
	1.5	(100, 1000)	0.2006	0.2988	0.7511	0.1803	0.2393	0.8078	0.1791	0.2665	0.7383	0.1966	0.3350	0.9051
		(200, 2000)	0.0982	0.1736	0.5659	0.1085	0.1481	0.3177	0.1055	0.1378	0.3189	0.1204	0.1711	0.5717
		(400, 4000)	0.0427	0.0876	0.2202	0.0672	0.0806	0.0835	0.0648	0.0831	0.0911	0.0443	0.0955	0.2491
	2.0	(100, 1000)	0.1670	0.2238	0.6848	0.1732	0.1848	0.7434	0.1890	0.2021	0.8594	0.1785	0.2552	0.8219
		(200, 2000)	0.0460	0.1279	0.3439	0.0880	0.1152	0.4739	0.0926	0.1061	0.4497	0.0564	0.1199	0.3975
		(400, 4000)	0.0234	0.0595	0.0781	0.0301	0.0553	0.0551	0.0322	0.0597	0.0587	0.0262	0.0677	0.0788
2	1.0	(100, 1000)	0.4442	0.5674	0.8024	0.2749	0.4263	0.8466	0.3146	0.4785	0.8068	0.4256	0.6297	0.8531
		(200, 2000)	0.1411	0.3065	0.5099	0.1337	0.2487	0.6774	0.1443	0.2432	0.6935	0.1518	0.3003	0.5392
		(400, 4000)	0.0701	0.1570	0.3425	0.0603	0.1340	0.5275	0.0691	0.1427	0.5243	0.0716	0.1753	0.3731
	1.5	(100, 1000)	0.2250	0.3177	0.6630	0.1783	0.2569	0.7564	0.1803	0.2819	0.7266	0.2169	0.3496	0.7209
		(200, 2000)	0.1097	0.1862	0.4902	0.1093	0.1604	0.6618	0.1073	0.1484	0.6252	0.1323	0.1815	0.5291
		(400, 4000)	0.0448	0.0968	0.3387	0.0642	0.0878	0.5137	0.0660	0.0911	0.4982	0.0459	0.1041	0.3426
	2.0	(100, 1000)	0.2086	0.2396	0.5690	0.2028	0.1983	0.7341	0.1985	0.2179	0.7381	0.2121	0.2707	0.6386
		(200, 2000)	0.0564	0.1393	0.4326	0.0956	0.1246	0.6158	0.0991	0.1150	0.5950	0.0688	0.1307	0.4436
		(400, 4000)	0.0235	0.0645	0.3247	0.0303	0.0604	0.5002	0.0324	0.0651	0.5013	0.0254	0.0748	0.3223

			EMSPE											
			$a = 0.8$											
			$\phi = -0.8$			$\phi = -0.5$			$\phi = 0.5$			$\phi = 0.8$		
η	β	(n, p)	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
1	1.0	(100, 1000)	0.3124	1.0860	0.8638	0.3001	0.7034	1.1454	0.2838	0.6581	1.1340	0.4351	0.9258	1.1025
		(200, 2000)	0.1427	0.5321	0.5437	0.1356	0.3443	0.5742	0.1542	0.2917	0.6306	0.1492	0.3829	0.4467
		(400, 4000)	0.0689	0.2706	0.2814	0.0601	0.1653	0.1070	0.0654	0.1440	0.1408	0.0761	0.1730	0.1083
	1.5	(100, 1000)	0.1847	0.7151	0.7371	0.1750	0.4652	0.7263	0.1997	0.3776	0.7978	0.2518	0.4736	0.9103
		(200, 2000)	0.1152	0.3766	0.5056	0.1084	0.2355	0.3644	0.1152	0.1835	0.3840	0.1236	0.2320	0.6219
		(400, 4000)	0.0401	0.1995	0.2280	0.0614	0.1298	0.0875	0.0636	0.1001	0.0802	0.0437	0.1155	0.2435
	2.0	(100, 1000)	0.1851	0.5590	0.6781	0.2066	0.3691	0.8591	0.2295	0.2906	0.9937	0.2333	0.3385	0.9691
		(200, 2000)	0.0503	0.2772	0.3668	0.0934	0.1836	0.4480	0.1010	0.1450	0.5301	0.0651	0.1657	0.4763
		(400, 4000)	0.0279	0.1569	0.0664	0.0299	0.1071	0.0605	0.0317	0.0727	0.0612	0.0241	0.0783	0.1161
2	1.0	(100, 1000)	0.4650	1.1777	0.8390	0.3184	0.7600	5.6650	0.3166	0.7468	0.9741	0.5277	0.6918	0.5469
		(200, 2000)	0.1528	0.5744	0.4820	0.1444	0.3701	0.6382	0.1654	0.3310	0.6212	0.1599	0.2864	0.1612
		(400, 4000)	0.0705	0.2724	0.3573	0.0583	0.1747	0.5431	0.0634	0.1589	0.3118	0.0710	0.1825	0.0705
	1.5	(100, 1000)	0.2424	0.7881	0.7280	0.1678	0.5070	0.7480	0.1861	0.4234	0.8527	0.2502	0.4947	0.7544
		(200, 2000)	0.1277	0.4061	0.4740	0.1122	0.2488	0.6323	0.1192	0.2044	0.6216	0.1221	0.2320	0.4798
		(400, 4000)	0.0405	0.2071	0.3364	0.0584	0.1368	0.4770	0.0639	0.1080	0.5334	0.0486	0.1001	0.1590
	2.0	(100, 1000)	0.2275	0.6298	0.6724	0.2149	0.4086	0.7517	0.2360	0.3154	0.8320	0.2691	0.3613	0.7145
		(200, 2000)	0.0680	0.3139	0.4204	0.0936	0.1984	0.6189	0.1032	0.1584	0.6089	0.0828	0.1868	0.4681
		(400, 4000)	0.0261	0.1715	0.2989	0.0307	0.1145	0.4921	0.0304	0.0786	0.4835	0.0225	0.0847	0.3055

			EMSPE											
			$a = -0.7$											
			$\phi = -0.8$			$\phi = -0.5$			$\phi = 0.5$			$\phi = 0.8$		
η	β	(n, p)	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
1	1.0	(100, 1000)	0.3923	0.8555	0.8500	0.2603	0.6613	1.0998	0.2793	0.7804	1.0841	0.3710	1.3601	1.1162
		(200, 2000)	0.1470	0.3876	0.5197	0.1357	0.3111	0.6854	0.1433	0.3805	0.7145	0.1443	0.6393	0.6618
		(400, 4000)	0.0695	0.1696	0.1920	0.0638	0.1292	0.1274	0.0624	0.1775	0.1580	0.0718	0.2828	0.2523
	1.5	(100, 1000)	0.2061	0.4892	0.8020	0.1837	0.4066	0.7410	0.1795	0.5265	0.7851	0.2176	0.9117	0.8910
		(200, 2000)	0.1180	0.2392	0.5618	0.1150	0.1950	0.3468	0.1091	0.2671	0.3806	0.1230	0.4565	0.6260
		(400, 4000)	0.0437	0.1074	0.2071	0.0615	0.0927	0.0847	0.0640	0.1331	0.0850	0.0449	0.2042	0.2539
	2.0	(100, 1000)	0.2111	0.3675	0.7299	0.2275	0.3222	0.9492	0.2056	0.4218	0.9569	0.2017	0.6927	0.8491
		(200, 2000)	0.0610	0.1681	0.4017	0.0946	0.1482	0.4252	0.0941	0.2137	0.4932	0.0603	0.3432	0.3904
		(400, 4000)	0.0264	0.0766	0.0747	0.0295	0.0717	0.0751	0.0300	0.1053	0.0780	0.0267	0.1609	0.0861
2	1.0	(100, 1000)	0.4524	0.6927	0.5352	0.3052	0.7218	0.8919	0.2963	0.8245	0.9141	0.4763	1.4264	0.9073
		(200, 2000)	0.1606	0.3163	0.2322	0.1424	0.3474	0.6142	0.1454	0.3932	0.6633	0.1568	0.6549	0.5642
		(400, 4000)	0.0762	0.1887	0.0775	0.0606	0.1523	0.5302	0.0624	0.1876	0.5374	0.0697	0.2979	0.3853
	1.5	(100, 1000)	0.2795	0.5065	0.6255	0.1729	0.4228	0.7275	0.1834	0.5488	0.8646	0.2494	0.9429	0.7724
		(200, 2000)	0.1383	0.2539	0.4527	0.1152	0.2104	0.6273	0.1116	0.2722	0.5771	0.1238	0.4693	0.5333
		(400, 4000)	0.0452	0.1085	0.2962	0.0599	0.1019	0.4993	0.0636	0.1383	0.5239	0.0496	0.2163	0.3605
	2.0	(100, 1000)	0.2500	0.3842	0.6739	0.2251	0.3318	0.7768	0.2218	0.4457	0.8517	0.2295	0.7436	0.6365
		(200, 2000)	0.0723	0.1852	0.4351	0.0984	0.1588	0.5822	0.1001	0.2263	0.5809	0.0739	0.3772	0.4849
		(400, 4000)	0.0274	0.0853	0.3129	0.0319	0.0772	0.4474	0.0296	0.1108	0.4816	0.0248	0.1747	0.2951

TABLE S3
The EMSPEs of OGA+HDAIC (M1), Lasso (M2), and ISIS-SCAD (M3) in Example 3 with the smallest one marked in blue

		EMSPE											
		$a=0$											
		$\phi=-0.8$			$\phi=-0.5$			$\phi=0.5$			$\phi=0.8$		
η	(n,p)	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
1	(100, 1000)	2.1666	2.3035	9.7631	0.1702	1.0979	0.7898	0.1913	1.2168	0.7542	1.1725	2.7214	7.8962
	(200, 2000)	0.0585	0.7577	0.3773	0.0908	0.4943	0.2118	0.0699	0.4845	0.2231	0.0565	0.7205	0.3690
	(400, 4000)	0.0310	0.2657	0.0581	0.0329	0.2337	0.0283	0.0314	0.2240	0.0227	0.0268	0.2920	0.1071
2	(100, 1000)	2.2190	5.3889	15.0595	0.1528	2.3749	0.8041	0.1609	2.5926	0.7337	1.3191	5.6142	9.6182
	(200, 2000)	0.0621	1.6642	0.4467	0.0858	1.1228	0.5852	0.0640	1.0948	0.5789	0.0522	1.6282	0.5101
	(400, 4000)	0.0288	0.5757	0.3323	0.0335	0.5277	0.4776	0.0332	0.5602	0.5140	0.0267	0.6405	0.3429
		EMSPE											
		$a=0.8$											
		$\phi=-0.8$			$\phi=-0.5$			$\phi=0.5$			$\phi=0.8$		
η	(n,p)	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
1	(100, 1000)	0.9335	3.7597	14.6948	0.1845	1.6253	0.9614	0.2225	1.2694	2.1679	2.4628	2.5611	16.1690
	(200, 2000)	0.0598	1.2985	0.3585	0.0750	0.6996	0.2331	0.0807	0.4857	0.2549	0.0620	0.7099	0.3885
	(400, 4000)	0.0303	0.4977	0.0693	0.0374	0.3174	0.0230	0.0362	0.2508	0.0281	0.0309	0.2997	0.0671
2	(100, 1000)	1.4416	7.6038	42.5595	0.1753	3.1684	15.8637	0.1792	2.6573	2.6631	1.1706	5.7290	18.0536
	(200, 2000)	0.0569	2.4468	0.4382	0.0726	1.3914	0.5536	0.0806	1.0938	0.6285	0.0625	1.6736	0.3374
	(400, 4000)	0.0266	0.9251	0.3533	0.0397	0.6386	0.4703	0.0398	0.5764	0.4867	0.0308	0.7217	0.0824
		EMSPE											
		$a=-0.7$											
		$\phi=-0.8$			$\phi=-0.5$			$\phi=0.5$			$\phi=0.8$		
η	(n,p)	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
1	(100, 1000)	1.5272	2.6511	13.7363	0.1978	1.2996	1.7059	0.1878	1.5092	1.8538	1.5593	3.3836	13.1523
	(200, 2000)	0.0651	0.7409	0.3473	0.0772	0.4967	0.2240	0.0834	0.6907	0.2669	0.0592	1.3005	0.3997
	(400, 4000)	0.0316	0.2900	0.0866	0.0383	0.2411	0.0288	0.0397	0.3484	0.0394	0.0301	0.6243	0.0812
2	(100, 1000)	1.9373	6.3218	16.9112	0.1730	2.7961	2.1322	0.1801	2.9404	12.9745	1.9536	6.8037	35.9484
	(200, 2000)	0.0613	1.7077	0.4350	0.0709	1.1145	0.5607	0.0787	1.3475	0.6728	0.0568	2.3924	0.5076
	(400, 4000)	0.0300	0.6915	0.2209	0.0421	0.5516	0.5122	0.0397	0.6976	0.5253	0.0306	1.0985	0.3441

S3.2. *Proof of* (S3.5). Let $A(z)=\sum_{i=0}^{\infty} a_i z^i \neq 0$ for all $|z| \leq 1$, where $a_0 = 1$ and $\sum_{i=0}^{\infty} |a_i| < \infty$. Denote by B the backshift operator. In this section, (S3.5) is proved under the ARX model,

$$A(B)y_t = \sum_{j=1}^{p} \alpha_j x_t^{(j)} + \epsilon_t, \tag{S3.7}$$

which includes (S3.1) as a special case. We also relax (S3.2) to

$$x_t^{(j)} = \sum_{l=0}^{\infty} \phi_l \varepsilon_{t-l,j}, \tag{S3.8}$$

where $\phi_0 = 1$ and $\sum_{l=0}^{\infty} |\phi_l| < \infty$. On the other hand, the assumptions on $\varepsilon_{tj}, d_{tj}, w_t$ and ϵ_t remain unchanged. In view of (S3.3) and (S3.8), it is easy to show that

$$\text{(S3.9)} \qquad \mathrm{E}(x_t^{(i)} x_t^{(j)}) = \begin{cases} \eta^2 \sum_{l=0}^{\infty} \phi_l^2 & i \neq j\,, \\ (\eta^2+1) \sum_{l=0}^{\infty} \phi_l^2 & i = j, \end{cases}$$

and

$$\text{(S3.10)} \qquad \mathrm{corr}(x_t^{(i)}, x_t^{(j)}) = \begin{cases} \rho \equiv \eta^2/(1+\eta^2) & i \neq j\,, \\ 1 & i = j. \end{cases}$$

Moreover, straightforward calculations give for $i \geq 0$,

$$\text{(S3.11)} \qquad \mathrm{E}(y_{t-i} x_{tj}) = \alpha_j V_i + G_i,$$

where $V_i = \sum_{s=0}^{\infty} c_s \phi_{s+i}$ and $G_i = \eta^2 (\sum_{l=0}^{p} \alpha_l) V_i$ with $c_s = \sum_{k=0}^{s} \phi_{s-k} b_k$ and $\sum_{k=0}^{\infty} b_k z^k = \phi^{-1}(z)$. Let $\mathbf{x}_t = (y_{t-1}, \ldots, y_{t-m}, x_t^{(1)}, \ldots, x_t^{(p)})^\top$ and $\mathbf{\Sigma} = \mathrm{E}(\mathbf{x}_t \mathbf{x}_t^\top)$, where m is any positive integer. Then,

$$\text{(S3.12)} \qquad \boldsymbol{Q} \mathbf{\Sigma} \boldsymbol{Q}^\top = \begin{pmatrix} \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12} \mathbf{\Sigma}_{22}^{-1} \mathbf{\Sigma}_{21} & \mathbf{0}_{m \times p} \\ \mathbf{0}_{p \times m} & \mathbf{\Sigma}_{22} \end{pmatrix},$$

where

$$\text{(S3.13)} \qquad \boldsymbol{Q} = \begin{pmatrix} \boldsymbol{I}_{m \times m} & -\mathbf{\Sigma}_{12} \mathbf{\Sigma}_{22}^{-1} \\ \mathbf{0}_{p \times m} & \boldsymbol{I}_{p \times p} \end{pmatrix},$$

$\mathbf{0}_{r \times s}$ denotes the $r \times s$ matrix of zeros, $\boldsymbol{I}_{r \times r}$ denotes the r-dimensional identity matrix, $\mathbf{\Sigma}_{11} = \mathrm{E}(\mathbf{y}_{t-1}(m) \mathbf{y}_{t-1}^\top(m))$ with $\mathbf{y}_s(m) = (y_s, \ldots, y_{s-m+1})^\top$, $\mathbf{\Sigma}_{12} = \mathrm{E}(\mathbf{y}_{t-1}(m) \mathbf{x}_{t,ex}^\top)$ with $\mathbf{x}_{t,ex}^\top = (x_t^{(1)}, \ldots, x_t^{(p)})$, $\mathbf{\Sigma}_{21} = \mathbf{\Sigma}_{12}^\top$, and $\mathbf{\Sigma}_{22} = \mathrm{E}(\mathbf{x}_{t,ex} \mathbf{x}_{t,ex}^\top)$. Since $\{\epsilon_t\}$ is independent of $\{x_t^{(j)}\}$ for all $1 \leq j \leq p$, it follows that

$$\text{(S3.14)} \qquad \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12} \mathbf{\Sigma}_{22}^{-1} \mathbf{\Sigma}_{21} \geq \mathrm{E}(\boldsymbol{r}_{t-1}(m) \boldsymbol{r}_{t-1}^\top(m)),$$

with $\boldsymbol{r}_{t-1}^\top(m) = (\phi^{-1}(B)\epsilon_{t-1}, \ldots, \phi^{-1}(B)\epsilon_{t-m})$. Moreover, one has

$$\text{(S3.15)} \qquad \lambda_{\min}\big(\mathrm{E}(\boldsymbol{r}_{t-1}(m) \boldsymbol{r}_{t-1}^\top(m))\big) \geq \underline{\eta}_1,$$

and

$$\text{(S3.16)} \qquad \lambda_{\min}(\mathbf{\Sigma}_{22}) \geq \underline{\eta}_2,$$

for some positive constants $\underline{\eta}_1$ and $\underline{\eta}_2$. Combining (S3.12)–(S3.16) gives

$$\lambda_{\min}(\boldsymbol{\Sigma}) \geq \underline{\eta}_1 \underline{\eta}_2 \lambda_{\max}^{-1}(\boldsymbol{Q}^\top \boldsymbol{Q}). \tag{S3.17}$$

Let $\boldsymbol{\nu} = (\boldsymbol{\nu}_1^\top, \boldsymbol{\nu}_2^\top)^\top$, where $\boldsymbol{\nu}_1 \in R^m, \boldsymbol{\nu}_2 \in R^p$ and $\|\boldsymbol{\nu}\| = 1$. By (S3.9), (S3.10), (S3.11), (S3.13) and some algebraic manipulations, it holds that

$$\begin{aligned} &\boldsymbol{Q}\boldsymbol{\nu} = \begin{pmatrix} \boldsymbol{\nu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\nu}_2 \\ \boldsymbol{\nu}_2 \end{pmatrix}, \\ &\boldsymbol{\Sigma}_{21} = \big(\boldsymbol{\alpha} V_i + \mathbf{1} G_i, i = 1, \ldots, m\big), \\ &|\boldsymbol{\nu}_2^\top \boldsymbol{\Sigma}_{22}^{-1}\big(\boldsymbol{\alpha} V_i + \mathbf{1} G_i\big)| \leq (\sum_{l=0}^{\infty} \phi_l^2)^{-1} (\sum_{j=1}^{p} \alpha_j^2)^{1/2} |V_i| + o(p^{-1/2}), \end{aligned} \tag{S3.18}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top$ and $\mathbf{1} = (1, \ldots, 1)^\top$. The first and the third equations of (S3.18) and $\sum_{i=1}^m V_i^2 \leq m(\sum_{l=0}^\infty \phi_l^2) \sum_{s=0}^\infty c_s^2$ further imply that there exists $0 < \bar{C} < \infty$ such that for any $\|\boldsymbol{\nu}\| = 1$,

$$\begin{aligned} \|\boldsymbol{Q}\boldsymbol{\nu}\|^2 \leq\ & 2 + 4m(\sum_{l=0}^{\infty} \phi_l^2)^{-1} \sum_{j=1}^{p} \alpha_j^2 \sum_{s=0}^{\infty} c_s^2 \\ & + o(p^{-1}) \leq \bar{C}. \end{aligned} \tag{S3.19}$$

By (S3.17) and (S3.19), it holds that

$$\lambda_{\min}(\boldsymbol{\Sigma}) \geq \underline{\eta}_1 \underline{\eta}_2 \bar{C}^{-1},$$

which, together with (S3.9) and

$$\begin{aligned} \mathrm{E}(y_t^2) &= \{\sum_{i=1}^{p} \alpha_i^2 + \eta^2 (\sum_{i=1}^{p} \alpha_i)^2\} \sum_{s=0}^{\infty} c_s^2 + \mathrm{E}(\phi^{-1}(B)\epsilon_{t-1})^2 \\ &\leq \bar{C}_1 \quad \text{for some } 0 < \bar{C}_1 < \infty, \end{aligned}$$

yields the desired conclusion (S3.5). □

References.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *J. Roy. Statist. Soc. Ser. B* **70**, 849-911.

Ing, C.-K. (2019). Model selection for high-dimensional linear regression with dependent observations. *Technical Report*, Institute of Statistics, National Tsing Hua University.

Rudelson, M. and Vershynin, R. (2013). Hanson-Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.* **18**, no. 82, 9.