

# 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](mailto: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:

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

where  $n$  is the sample size,  $x_{t1}, \dots, x_{tp}$  are predictor variables,  $\varepsilon_t$  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  $\beta_j$ , eigenvalue conditions on the covariance (correlation) matrix of the predictor variables, and distributional conditions on  $\varepsilon_t$  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  $\varepsilon_t$  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

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

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

$$(1.3) \quad C_3 \exp(-\beta j) \leq |\beta_j| \leq C_4 \exp(-\beta j), \quad 0 < C_3 \leq C_4 < \infty, \beta > 0,$$

---

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

$$(1.4) \quad Lj^{-\gamma} \leq |\beta_{(j)}^*| \leq Uj^{-\gamma}$$

and

$$(1.5) \quad L_1 \exp(-\beta_j) \leq |\beta_{(j)}^*| \leq U_1 \exp(-\beta_j),$$

respectively, where  $0 < L \leq U < \infty$ ,  $0 < L_1 \leq U_1 < \infty$ , and  $|\beta_{(1)}^*| \geq |\beta_{(2)}^*| \geq \dots \geq |\beta_{(p)}^*|$  is a rearrangement of  $\{|\beta_j^*|\}$  in decreasing order with  $\beta_j^* = \sigma_j \beta_j$  and  $\sigma_j^2 = E(x_{tj}^2)$ . 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  $\beta_j$  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  $\beta_j$  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

$$(1.6) \quad \sum_{j=1}^p |\beta_j^*|^{1/\gamma} < M_4 \quad \text{for some } \gamma \geq 1, 0 < M_4 < \infty,$$

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  $E(\mathbf{x}_t \varepsilon_t) = \mathbf{0}$ . 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  $\hat{J}_0 = \emptyset$ , where  $\hat{J}_m$  denotes the index set of the variables chosen by OGA at the  $m$ th iteration. For  $m \geq 1$ ,  $\hat{J}_m$  is recursively updated by

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

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  $\{\mathbf{Z}_i, i \in J\}$  ( $\mathbf{H}_\emptyset = \mathbf{0}$ ).

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

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

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

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

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

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

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

$$(2.4) \quad x_{tl} = \sum_{j=0}^{\infty} \mathbf{w}_j^\top(l) \delta_{t-j}, \quad \varepsilon_t = \sum_{j=0}^{\infty} \mathbf{w}_j^\top(0) \delta_{t-j},$$

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

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

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

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

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

$$(2.7) \quad (A1) \text{ and } (A2) \text{ hold true under (2.4)–(2.6) and (2.8),}$$

where (2.8) is given by

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

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

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

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  $|z| \leq 1 + \iota$  with  $\iota$  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  $1 \leq l \leq p$ , and  $\delta_t = (\epsilon_t^{(1)}, \dots, \epsilon_t^{(p)}, \varepsilon_t)^\top$  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  $\delta_t$  is replaced by

$$(2.10) \quad \|\delta_{tk}\|_{\psi_2} \leq L, \quad k = 1, \dots, q,$$

where  $\|\cdot\|_{\psi_2}$  denotes the  $\psi_2$  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  $0 \leq s < \infty$ . 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  $\sum_{j=1}^p \beta_j^{*2} \leq \bar{M}_0$ . In addition, there exist  $\gamma \geq 1$  and  $0 < C_\gamma < \infty$  such that for any  $J \subseteq \mathcal{P}$ ,

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

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,

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

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  $\gamma$  in (2.11) can be understood as an index to describe the degree of sparseness in the underlying high-dimensional models. The larger the  $\gamma$  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  $\gamma$  (see Lemma A.2 in the Appendix), whereas (1.6) is violated due to  $L(1 + \log p) \leq \sum_{j=1}^p |\beta_j^*|^{1/\gamma} \leq U(1 + \log p)$ . 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  $\max_{1 \leq j \leq p} |\beta_j^*| \leq M_0$ . Moreover, there exists  $M_1 > 1$  such that for any  $J \subseteq \mathcal{P}$ ,

$$(2.12) \quad \sum_{j \in J} |\beta_j^*| \leq M_1 \max_{j \in J} |\beta_j^*|.$$

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  $\mathbf{\Gamma}(J) = E\{\mathbf{z}_t(J)\mathbf{z}_t^\top(J)\}$  and  $\mathbf{g}_i(J) = E(z_{ti}\mathbf{z}_t(J))$ , where  $J \subseteq \mathcal{P}$  and  $\mathbf{z}_t(J) = (z_{ti}, i \in J)^\top$ .

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

$$(2.13) \quad \max_{1 \leq \#\!(J) \leq \bar{D}(n/\log p)^{1/2}, i \notin J} \|\mathbf{\Gamma}^{-1}(J)\mathbf{g}_i(J)\|_1 < M,$$

where  $\#\!(J)$  denotes the cardinality of  $J$ .

Since  $\mathbf{\Gamma}^{-1}(J)\mathbf{g}_i(J) = \arg \min_{\mathbf{c} \in R^{\#\!(J)}} E(z_{ti} - \mathbf{c}^\top \mathbf{z}_t(J))^2$ , (2.13) essentially says that the regression coefficients for  $z_{ti}$  on  $\mathbf{z}_t(J)$  with all  $i \notin J$  and  $\#\!(J) \leq \bar{D}(n/\log p)$  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  $\mathbf{g}_y(J) = E(y_t \mathbf{z}_t(J))$  and  $\boldsymbol{\beta}^*(J) = \mathbf{\Gamma}^{-1}(J)\mathbf{g}_y(J) = \arg \min_{\mathbf{c} \in R^{\#\!(J)}} E(y_t - \mathbf{c}^\top \mathbf{z}_t(J))^2$ , 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  $\#\!(J) \leq \bar{D}(n/\log p)^{1/2}$ , there exists  $0 < C < \infty$  such that

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

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  $\#\!(J) \leq \bar{D}(n/\log p)^{1/2}$ ,

$$(2.15) \quad \|\boldsymbol{\beta}^* - \boldsymbol{\beta}^*(J)\|_1 \leq CC_\gamma \left( \sum_{j \notin J} \beta_j^{*2} \right)^{(\gamma-1)/(2\gamma-1)},$$

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  $1 \leq q \leq \bar{D}(n/\log p)^{1/2}$ , where  $j_i, i = 1, \dots, q$ , are distinct elements in  $\mathcal{P}$ . Note first that

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

where  $\beta_J^* = (\beta_{j_1}^*, \dots, \beta_{j_q}^*)^\top$ . Denote  $\mathbf{g}_y(J)$  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

$$\begin{aligned} \Gamma(J)^{-1} \Gamma(J) (\beta^*(J) - \beta_J^*) &= \Gamma(J)^{-1} \left( \sum_{l \notin J} \rho_{j_1 l} \beta_l^*, \dots, \sum_{l \notin J} \rho_{j_q l} \beta_l^* \right)^\top \\ &= \sum_{l \notin J} \beta_l^* \Gamma(J)^{-1} \mathbf{g}_l(J). \end{aligned}$$

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  $J \subseteq \mathcal{P}$ ,  $\sum_{j \in J} |\beta_j^*| \leq M_1 \times (\sum_{j \in J} \beta_j^{*2})^{\lim_{\gamma \rightarrow \infty} (\gamma-1)/(2\gamma-1)}$ . In addition, the strong sparsity condition,

$$(2.17) \quad k_0 = \sharp(N_n) < M_7,$$

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

**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  $\hat{y}_m(\mathbf{x}) = \mathbf{x}^\top (\hat{J}_m) \hat{\beta}(\hat{J}_m)$ , where  $\mathbf{x}(J) = (x_i, i \in J)$  and  $\hat{\beta}(J) = (\sum_{t=1}^n \mathbf{x}_t(J) \mathbf{x}_t^\top(J))^{-1} \times \sum_{t=1}^n \mathbf{x}_t(J) y_t$ . Note also that  $\hat{y}_m(\mathbf{x}) = \mathbf{z}^\top (\hat{J}_m) \hat{\beta}^*(\hat{J}_m)$ , where  $\mathbf{z}(J) = (z_i, i \in J)$  with  $z_i = x_i/\sigma_i$ , and  $\hat{\beta}^*(\hat{J}_m) = (\sum_{t=1}^n \mathbf{z}_t(J) \mathbf{z}_t^\top(J))^{-1} \sum_{t=1}^n \mathbf{z}_t(J) y_t$ . One of the most natural performance measures for  $\hat{y}_m(\mathbf{x})$  is the conditional mean squared prediction error (CMSPE),

$$(2.18) \quad E_n \{y(\mathbf{x}) - \hat{y}_m(\mathbf{x})\}^2 = E_n \{y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})\}^2 + E_n \{y_{\hat{J}_m}(\mathbf{x}) - \hat{y}_m(\mathbf{x})\}^2,$$

where  $E_n(\cdot) = E(\cdot | y_1, \mathbf{x}_1, \dots, y_n, \mathbf{x}_n)$ , and  $y_J(\mathbf{x}) = \beta^{*\top}(J) \mathbf{z}(J)$ . 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),*

$$(2.19) \quad \lambda_{\min}(\Gamma) \geq \lambda_1 > 0$$

and

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

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

$$(2.21) \quad \begin{aligned} &\bar{\tau} = \sup \tau \\ &\equiv \sup \left\{ \tau : \tau > 0, \limsup_{n \rightarrow \infty} \frac{\tau c_2^*}{\min_{\sharp(J) \leq \tau(n/\log p)^{1/2}} \lambda_{\min}(\Gamma(J))} \leq 1 \right\}. \end{aligned}$$

Then

$$(2.22) \quad \max_{1 \leq m \leq K_n} \left( \frac{E_n \{y(\mathbf{x}) - \hat{y}_m(\mathbf{x})\}^2}{m^{-2\gamma+1} + n^{-1} m \log p} \right) = O_p(1).$$

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

$$(2.23) \quad \max_{1 \leq m \leq K_n} \left( \frac{E_n \{y(\mathbf{x}) - \hat{y}_m(\mathbf{x})\}^2}{\exp(-G_3 m) + n^{-1} m \log p} \right) = O_p(1),$$

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

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

and

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

where  $\mu_{J,i} = E[(y(\mathbf{x}) - y_J(\mathbf{x}))z_i]$ ,  $s > 0$  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  $1 \leq q \leq m$ ,

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

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

$$(2.24) \quad E_n(y(\mathbf{x}) - y_{\hat{j}_m}(\mathbf{x}))^2 \leq G_1 m^{-2\gamma+1} \quad \text{on } A_n(m) \cap B_n(m).$$

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

$$\begin{aligned} & E_n(y(\mathbf{x}) - y_{\hat{j}_i}(\mathbf{x}))^2 \\ & \leq \max_{1 \leq j \leq p} |\mu_{\hat{j}_i,j}| \sum_{l=1, l \notin \hat{j}_i}^p |\beta_l^*| \\ & \leq C_\gamma \max_{1 \leq j \leq p} |\mu_{\hat{j}_i,j}| \left( \sum_{l=1, l \notin \hat{j}_i}^p \beta_l^{*2} \right)^{(\gamma-1)/(2\gamma-1)} \\ & \leq C_\gamma \max_{1 \leq j \leq p} |\mu_{\hat{j}_i,j}| \lambda_1^{-(\gamma-1)/(2\gamma-1)} (E_n(y(\mathbf{x}) - y_{\hat{j}_i}(\mathbf{x}))^2)^{(\gamma-1)/(2\gamma-1)}, \end{aligned}$$

and hence

$$(2.25) \quad E_n(y(\mathbf{x}) - y_{\hat{j}_i}(\mathbf{x}))^2 \leq \left( C_\gamma \max_{1 \leq j \leq p} |\mu_{\hat{j}_i,j}| \right)^{2-\gamma^{-1}} \lambda_1^{-1+\gamma^{-1}}.$$

By (2.25),

$$\begin{aligned} & E_n(y(\mathbf{x}) - y_{\hat{j}_m}(\mathbf{x}))^2 \\ & \leq \min_{0 \leq i \leq m-1} E_n(y(\mathbf{x}) - y_{\hat{j}_i}(\mathbf{x}))^2 \\ (2.26) \quad & \leq C_\gamma^{2-\gamma^{-1}} \lambda_1^{-1+\gamma^{-1}} \left( \min_{0 \leq i \leq m-1} \max_{1 \leq j \leq p} |\mu_{\hat{j}_i,j}| \right)^{2-\gamma^{-1}} \\ & \leq C_\gamma^{2-\gamma^{-1}} \lambda_1^{-1+\gamma^{-1}} (\tilde{\xi} s)^{2-\gamma^{-1}} (n^{-1} \log p)^{1-(2\gamma)^{-1}} \quad \text{on } B_n^c(m). \end{aligned}$$

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

$$(2.27) \quad E_n(y(\mathbf{x}) - y_{\hat{j}_m}(\mathbf{x}))^2 I_{A_n(K_n)} \leq C_2 \max\{m^{-2\gamma+1}, \{n^{-1} \log p\}^{1-(2\gamma)^{-1}}\}.$$

We show in Section S1 of the Supplementary Material that

$$(2.28) \quad P(\|\hat{\Gamma}^{-1}(\hat{J}_{K_n})\| \leq \bar{B}) = 1 + o(1),$$

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

$$(2.29) \quad \bar{B} > \frac{1}{\liminf_{n \rightarrow \infty} \min_{\#(J) \leq K_n} \lambda_{\min}(\Gamma(J)) - c_2^* \delta},$$

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

$$(2.30) \quad \lim_{n \rightarrow \infty} P(A_n(K_n)) = 1,$$

which, together with (2.27), yields

$$(2.31) \quad \max_{1 \leq m \leq K_n} \frac{E_n(y(\mathbf{x}) - y_{\hat{j}_m}(\mathbf{x}))^2}{\max\{m^{-2\gamma+1}, (n^{-1} \log p)^{1-(2\gamma)^{-1}}\}} = O_p(1).$$

Moreover, we have

$$(2.32) \quad \max_{1 \leq m \leq K_n} \frac{n E_n[\{\hat{y}_m(\mathbf{x}) - y_{\hat{j}_m}(\mathbf{x})\}^2]}{m \log p} = O_p(1),$$

which is also proved in Section S1 of the Supplementary Material. In view of (2.31), (2.32) and the fact that  $(\log p/n)^{1-(2\gamma)^{-1}} \leq m^{-2\gamma+1}$  if  $m \leq (n/\log p)^{(2\gamma)^{-1}}$  and  $(\log p/n)^{1-(2\gamma)^{-1}} \leq n^{-1} m \log p$  if  $m \geq (n/\log p)^{(2\gamma)^{-1}}$ , 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.  $\square$

REMARK 1. It is easy to see that the  $\tau$  defined in (2.21) is nonempty. In particular,  $x \in \tau$  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  $\gamma$  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

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

and

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

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

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

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

$$(2.36) \quad \max_{1 \leq m \leq K_n} \left( \frac{E_n \{y(\mathbf{x}) - \hat{y}_m(\mathbf{x})\}^2}{m^{-2\gamma+1} + n^{-1}m(\log p)^{1+\bar{c}}} \right) = O_p(1)$$

and

$$(2.37) \quad \max_{1 \leq m \leq K_n} \left( \frac{E_n \{y(\mathbf{x}) - \hat{y}_m(\mathbf{x})\}^2}{\exp(-G_3m) + n^{-1}m(\log p)^{1+\bar{c}}} \right) = O_p(1),$$

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

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

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

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

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*

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

*hold. Then, for*

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

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

(i)

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

*provided (A3) is true;*

(ii)

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

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

(iii)

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

provided  $E(y_t^2)$  is bounded above by a finite constant and

$$(3.7) \quad \min_{j \in N_n} |\beta_j^*| \geq \underline{\theta}, \quad \text{for some } \underline{\theta} > 0,$$

$$k_0 \left( \sum_{j \in N_n} |\beta_j^*| \right)^2 = o(n/\log p).$$

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

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

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:

$$(3.9) \quad \mathbf{y}_t = \sum_{l=1}^p \mathbf{b}_l x_{tl} + \mathbf{e}_t, \quad t = 1, \dots, n,$$

where  $\mathbf{y}_t$ ,  $\mathbf{e}_t$ , and  $\mathbf{b}_j$  are  $d$ -dimensional vectors,  $d$  is allowed to grow to infinity with  $n$  and  $\{(\mathbf{e}_t^\top, \mathbf{x}_t^\top)^\top\}$  is a zero-mean stationary time series satisfying  $E(\mathbf{x}_t \mathbf{e}_t^\top) = \mathbf{0}$ . 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  $\mathcal{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)^\top$ , and  $\mathbf{H}_J$  and  $\mathbf{Z}_i$  are defined as in Section 2.1. A multivariate version of OGA, MOGA, is initialized by  $\hat{L}_0 = \emptyset$ . 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$(3.11) \quad \max_{1 \leq m \leq K_n} \left(\frac{d^{-1} E_n \|\mathbf{y}(\mathbf{x}) - \hat{\mathbf{y}}_m(\mathbf{x})\|^2}{m^{-2\gamma+1} + n^{-1} m \log pd}\right) = O_p(1),$$

and for some  $G_4 > 0$ ,

$$(3.12) \quad \max_{1 \leq m \leq K_n} \left(\frac{d^{-1} E_n \|\mathbf{y}(\mathbf{x}) - \hat{\mathbf{y}}_m(\mathbf{x})\|^2}{\exp(-G_4 m) + n^{-1} m \log pd}\right) = O_p(1).$$

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

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

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

$$\hat{m}_n = \arg \min_{1 \leq m \leq K_n} \text{MHDAIC}(\hat{\mathbf{L}}_m).$$

We conjecture that  $d^{-1} E_n \|\mathbf{y}(\mathbf{x}) - \hat{\mathbf{y}}_{\hat{m}_n}(\mathbf{x})\|^2$  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:

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

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

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

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

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

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

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

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  $p \rightarrow \infty$  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:

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

and

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

where  $p \rightarrow \infty$ ,  $k_0 = O(n/\log p)$ , and  $\lambda_n \geq c^*(\log p/n)^{1/2}$  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),

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

and

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

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  $\{\epsilon_t\}$  are independent, and hence are applicable to ARX models. Moreover, (3.17) and (3.18) are established under

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

and

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

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

$$\begin{aligned} \lambda_{\min}(\Gamma) &= \lambda_{\min}(\Gamma_{\mathbf{x}}(0)) = \lambda_{\min}\left(\int_{-\pi}^{\pi} f_{\mathbf{x}}(\theta) d\theta\right) \\ &\geq 2\pi \left\{ \operatorname{ess\,inf}_{\theta \in [-\pi, \pi]} \lambda_{\min}(f_{\mathbf{x}}(\theta)) \right\}. \end{aligned}$$

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  $\lambda_{\max}(\Gamma) \rightarrow \infty$  as  $p \rightarrow \infty$ , and hence  $\operatorname{ess\,sup}_{\theta \in [-\pi, \pi]} \lambda_{\max}(f_{\mathbf{x}}(\theta)) \geq [1/(2\pi)] \lambda_{\max}(\Gamma)$ . 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  $\hat{J}(m)$ ,  $1 \leq m \leq K_n$ , 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  $E_n(y(\mathbf{x}) - \hat{y}_{\hat{J}(m)}(\mathbf{x}))^2 = E_n(y(\mathbf{x}) - y_{\hat{J}(m)}(\mathbf{x}))^2 + E_n(\hat{y}_{\hat{J}(m)}(\mathbf{x}) - y_{\hat{J}(m)}(\mathbf{x}))^2$ , where  $\hat{y}_{\hat{J}(m)}(\mathbf{x}) = \mathbf{x}^\top(\hat{J})\hat{\beta}(\hat{J})$ . It is not difficult to show that the squared bias terms obey

$$(3.23) \quad E_n(y(\mathbf{x}) - y_{\hat{J}(m)}(\mathbf{x}))^2 \geq E(y(\mathbf{x}) - y_{J_m^*}(\mathbf{x}))^2,$$

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

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

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

$$(3.25) \quad \max_{1 \leq m \leq K_n} \frac{nE_n(\hat{y}_{\hat{J}(m)}(\mathbf{x}) - y_{\hat{J}(m)}(\mathbf{x}))^2}{m \log p} = O_p(1).$$

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

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

According to Lemma A.3, (A.11) and  $E(y(\mathbf{x}) - y_{J_m^*}(\mathbf{x}))^2 = 0$  if  $m \geq k_0$ , 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  $\text{AR}(1), \dots, \text{AR}(K_n)$ , with  $K_n$  approaching  $\infty$  at a rate slower than  $n$ . Unlike  $\hat{J}(m)$ ,  $1 \leq m \leq$

$K_n$ , 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

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

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

$$P\left(\min_{1 \leq m \leq K_n} E_n(y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x}))^2 = 0\right) \rightarrow 1 \quad \text{as } n \rightarrow \infty,$$

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,

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

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  $1 \leq m \leq K_n$ ,

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

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),  $R_{1,p} = \max_{1 \leq i,l \leq p} |n^{-1} \sum_{t=1}^n z_{ti} z_{tl} - \rho_{il}|$ ,  $\varepsilon_t(J) = y_t - \varepsilon_t - \beta^{*\top}(J) \mathbf{z}_t(J)$  and  $\varepsilon(J) = y(\mathbf{x}) - y_J(\mathbf{x}) = y(\mathbf{x}) - \beta^{*\top}(J) \mathbf{z}(J)$ . In addition, it is shown in Section S2 of the Supplementary Material that

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

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

and

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

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

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

$$(3.33) \quad \tilde{k}_n = \min\{k : 1 \leq k \leq K_n, E_n(\varepsilon^2(\hat{J}_k)) \leq G m_n^{*-2\gamma+1}\} \quad (\min \emptyset = K_n),$$

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

$$(3.34) \quad \lim_{n \rightarrow \infty} P(\hat{k}_n \leq \tilde{k}_n - 1) = 0.$$

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

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

where

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

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

$$(3.36) \quad \lim_{n \rightarrow \infty} P(W_n) = 1,$$

where

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

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

$$(3.37) \quad \begin{aligned} & n^{-1} \sum_{t=1}^n \varepsilon_t^2(\hat{J}_k) \\ & \geq E_n(\varepsilon^2(\hat{J}_k)) \\ & \quad \times \left\{ 1 - \frac{C_{M,\gamma,\lambda_1} c_2^*}{G^{1/(2\gamma-1)}} \left( I_{\{\gamma=1\}} + \left( \frac{\log p}{n} \right)^{(\gamma-1)/2\gamma} I_{\{\gamma>1\}} \right) \right\} \quad \text{on } W_n, \end{aligned}$$

$$(3.38) \quad \begin{aligned} & n^{-1} \left| \sum_{t=1}^n \varepsilon_t(\hat{J}_k)\varepsilon_t \right| \leq E_n(\varepsilon^2(\hat{J}_k)) \frac{C_{M,\gamma,\lambda_1} c_1^*}{G^{\gamma/(2\gamma-1)}} \\ & \quad \text{on } W_n, \end{aligned}$$

$$(3.39) \quad \begin{aligned} & n^{-1} \left| \sum_{t=1}^n \varepsilon_t(\hat{J}_{m_n^*})\varepsilon_t \right| \leq E_n(\varepsilon^2(\hat{J}_k)) \frac{C_{M,\gamma,\lambda_1} c_1^*}{G^{\gamma/(2\gamma-1)}} \\ & \quad \text{on } W_n \cap A_n(K_n), \end{aligned}$$

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

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

and

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

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  $n$ ,

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

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

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

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

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

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  $\gamma > 1$ ,

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

and for  $\gamma = 1$ ,

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

To show (3.46), note first that

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

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

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

Moreover, we have

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

and

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

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.  $\square$

**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  $0 < \xi \leq 1$  be given. The algorithm initializes  $J_{\xi,0} = \emptyset$ . For  $m \geq 1$ ,  $J_{\xi,m}$  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

$$|E(u_{m-1} z_l)| \geq \xi \max_{1 \leq j \leq p} |E(u_{m-1} z_j)|,$$

with  $u_0 = y(\mathbf{x})$  and  $u_m = y(\mathbf{x}) - y_{J_{\xi,m}}(\mathbf{x})$  if  $m \geq 1$ . 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  $E(u_m^2)$  under (A3) or (A4).

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

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

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

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

PROOF. Straightforward calculations yield

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

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

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

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

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

and hence

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

In view of (A.6), one has

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

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

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

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

$$(A.8) \quad E(u_{m+1}^2) \leq E(u_m^2) - \xi^2 \lambda_1 M_1^{-2} E(u_m^2) \\ = E(u_m^2) \{1 - \xi^2 \lambda_1 M_1^{-2}\}.$$

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

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  $\gamma$ .*

PROOF. See Section S1 of the Supplementary Material.  $\square$

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

$$(A.9) \quad \mathbb{E}(y(\mathbf{x}) - y_{J_{\xi,m}}(\mathbf{x}))^2 \leq D_1 \mathbb{E}(y(\mathbf{x}) - y_{J_m^*}(\mathbf{x}))^2,$$

and

$$(A.10) \quad D_2 m^{-2\gamma+1} \leq \mathbb{E}(y(\mathbf{x}) - y_{J_m^*}(\mathbf{x}))^2 \leq D_3 m^{-2\gamma+1}.$$

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

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

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

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

$$\lambda_1 \sum_{j \notin J_m^o} \beta_j^{*2} \leq \mathbb{E}(y(\mathbf{x}) - y_{J_m^*}(\mathbf{x}))^2 \leq \mathbb{E}\left(\sum_{j \notin J_m^o} \beta_j^* z_j\right)^2,$$

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

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

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

$$(A.12) \quad \mathbb{E}(y(\mathbf{x}) - y_{J_{\xi,m}}(\mathbf{x}))^2 = O(\exp(-mf_{\text{oga}})),$$

where  $f_{\text{oga}} = \xi^2 \lambda_1 (L_1/U_1)^2 (1 - \exp(-\beta))^2 < 2\beta$ . 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  $m(2\beta/f_{\text{oga}} - 1)$  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

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

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  $\Gamma$  obeys  $\lambda_{\max}(\Gamma) \rightarrow \infty$  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 MICHAELIDIS, 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  $\ell_q$ -hulls for  $0 < q \leq 1$ . *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>
- NEGABAN, 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](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  $\ell_q$ -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