Variable Selection for High-Dimensional Regression Models with Time Series and Heteroscedastic Errors

Hai-Tang Chiou[a], Meihui Guo[b], Ching-Kang Ing[a*]
[a]Institute of Statistics, National Tsing Hua University, Hsinchu 30013, Taiwan
[b]Department of Applied Mathematics, National Sun Yat-sen University, Kaohsiung 80424, Taiwan

December 13, 2019

Abstract

Although existing literature on high-dimensional regression models is rich, the vast majority of studies have focused on independent and homogeneous error terms. In this article, we consider the problem of selecting high-dimensional regression models with heteroscedastic and time series errors, which have broad applications in economics, quantitative finance, environmental science, and many other fields. The error term in our model is the product of two components: one time series component, allowing for a short-memory, long-memory, or conditional heteroscedasticity effect, and a high-dimensional dispersion function accounting for exogenous heteroscedasticity. By making use of the orthogonal greedy algorithm and the high-dimensional information criterion, we propose a new model selection procedure that consistently chooses the relevant variables in both the regression and the dispersion functions. The finite sample performance of the proposed procedure is also illustrated via simulations and real data analysis.

Keywords: Heteroscedasticity, High-dimensional information criterion, Orthogonal greedy algorithm, Long-range dependence

*Corresponding author.
E-mail addresses: htchiou1@gmail.com (H.-T. Chiou), guomh@math.nsysu.edu.tw (M. Guo), cking@stat.nthu.edu.tw (C.-K. Ing).

1 Introduction

Consider a multiple-input transfer function model,

$$y_t = C + \sum_{j=1}^{k} \frac{w_j(B)}{\delta_j(B)} \xi_{t,j} + \eta_t, \quad t = 1, \ldots, n, \tag{1}$$

where n is the sample size, C is a constant, $\{\xi_{t,j}\}, j = 1, \ldots, k$, are input series (or exogenous variables), $\{y_t\}$ is the output series, $\{\eta_t\}$ is a mean-zero stationary noise series independent of $\{\xi_{t,j}\}$, the polynomials in B, $\delta_j(B) = 1 - \delta_{1,j}B - \cdots - \delta_{r_j,j}B^{r_j}$ and $w_j(B) = w_{0,j} - w_{1,j}B - \cdots - w_{s_j,j}B^{s_j}$, are of degrees r_j and s_j, respectively, and B denotes the backshift operator. Model (1) encompasses not only the classical regression and time series models, but also the celebrated intervention model proposed by Box and Tiao (1975); see Tsay (1984) and Tiao (1985) for a more detailed discussion. When $\delta_j(z) \neq 0$ for all $|z| \leq 1$, $w_j(z)/\delta_j(z)$ can be approximated by $\sum_{l=0}^{p_j} c_{l,j} z^l$, where p_j is a sufficiently large integer and $\{c_{l,j}\}$ satisfies $\sum_{l=0}^{\infty} c_{l,j} z^l = w_j(z)/\delta_j(z)$. Therefore, model (1) can be approximated by

$$y_t = C + \sum_{j=1}^{k} \sum_{l=0}^{p_j} c_{l,j} \xi_{t-l,j} + \eta_t, \tag{2}$$

which, in turn, is a special case of the linear regression model,

$$y_t = \beta_0 + \sum_{j=1}^{p_{\mathrm{L}}} \beta_j x_{tj} + \eta_t, \tag{3}$$

where p_{L}, corresponding to $\sum_{j=1}^{k}(p_j + 1)$ in model (2), can be large compared to n, $x_{tj}, j = 1, \ldots, p_{\mathrm{L}}$, are exogenous variables, and $\beta_j, 0 \leq j \leq p_{\mathrm{L}}$, are regression coefficients. In the case of $p_{\mathrm{L}} \gg n$, there are computational and statistical difficulties in estimating the regression coefficients by standard regression methods. In particular, it is no longer feasible to use the classical model selection techniques to estimate $N_{\mathrm{L},n} = \{1 \leq j \leq p_{\mathrm{L}} : \beta_j \neq 0\}$, the set of relevant variables. However, by imposing sparsity conditions on β_j, eigenvalue conditions on the covariance (correlation) matrix of x_{tj}, and distributional conditions on η_t or x_{tj}, it has been shown that $N_{\mathrm{L},n}$ can still be consistently estimated either using penalized least squares methods (see, e.g., Basu and Michailidis, 2015; Wu and Wu, 2016) or greedy forward selection algorithms (see, e.g., Ing and Lai, 2011; Hsu et al., 2019; Ing, 2019).

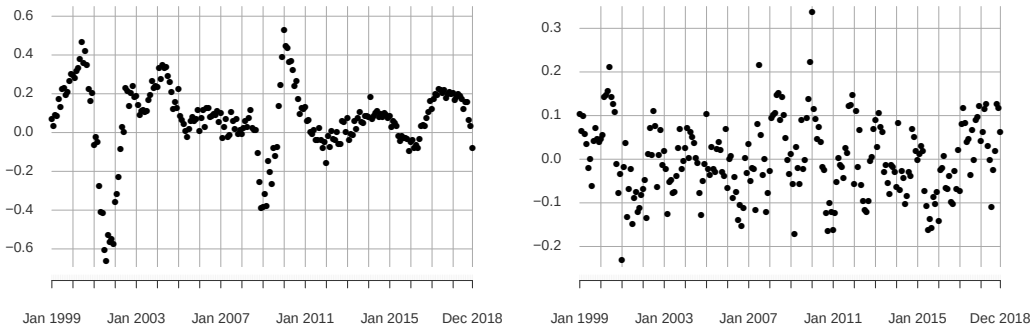

Fig. 1. The left panel is the time plot of the monthly growth rate of the semiconductor worldwide market billings defined in (36), and the right one is the time plot of the estimated residuals, $\hat{\eta}_t$, obtained from OGA+HDIC+Trim of Ing and Lai (2011).

On the other hand, model (3), assuming homogenous errors, can be restrictive in terms of economic applications. To illustrate the need to consider heteroscedastic errors, we provide a preliminary analysis of the monthly growth rate of the semiconductor industry based on model (3). In our analysis, the output variable, $\{y_t\}$, is the monthly growth rate of the semiconductor world market billings from the dataset of the World Semiconductor Trade Statistics (see (36)) and the sample size is $n = 240$. The input variables include quite a large number of macroeconomic, financial, and semiconductor variables and their lagged values, leading to $p_{\mathrm{L}} = 1584 \gg n$. The time plot of $\{y_t\}$ given in the left panel of Fig. 1 offers a clear indication of heteroscedasticity. Using the high-dimensional model selection method suggested in Ing and Lai (2011), we select and estimate the non-zero β_j; the time plot of the resultant residuals, $\{\hat{\eta}_t\}$, is provided in the right panel of Fig. 1, revealing that the heterogeneity in variance of $\{y_t\}$ carries over to that of $\{\hat{\eta}_t\}$. Since the heteroscedastic variance may result from the exogenous variables, we are led to consider

$$\eta_t = \sigma_t \epsilon_t \quad \text{and} \quad \sigma_t^2 = \exp\left\{\alpha_0 + \sum_{j=1}^{p_{\mathrm{D}}} \alpha_j z_{tj}\right\}, \tag{4}$$

in which $z_{tj}, 1 \le j \le p_{\mathrm{D}}$, denote the exogenous variables that may influence the variance of η_t, p_{D} is allowed to be larger than n, and $\alpha_j, 0 \le j \le p_{\mathrm{D}}$, are unknown coefficients. In fact, in the case of independent observations, model (3) with the error term satisfying (4) has

been used in several studies on expression quantitative trait loci (eQTLs) and production engineering where heteroscedasticity is present in the data; see Daye et al. (2012) and Chien et al. (2016). The $\{\epsilon_t\}$ in (4), playing a role similar to the stationary error in (1), is modeled by

$$\epsilon_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}, \tag{5}$$

where $\{w_t\}$ is a sequence of i.i.d. random variables independent of $\{\boldsymbol{x}_t = (x_{t1}, \ldots x_{tp_{\rm L}})^\top\}$ and $\{\boldsymbol{z}_t = (z_{t1}, \ldots z_{tp_{\rm D}})^\top\}$, $E(w_1) = 0$, $E(w_1^2) = 1$, $\psi_0 = 1$, and $\{\psi_j\}$ obeys either

$$\sum_{j=0}^{\infty} |\psi_j| < \infty \quad \text{and} \quad \sum_{j=0}^{\infty} \psi_j z^j \neq 0, \ |z| \leq 1, \tag{6}$$

or

$$\psi_j = O(j^{-1+d}), \text{ for some } 0 < d < 1/2. \tag{7}$$

Condition (6) implies that $\{\epsilon_t\}$ is a short-memory process, whereas (7) allows $\{\epsilon_t\}$ to be long-memory.

An alternative way to model the monthly growth rate of the semiconductor world market billings is to include the past values of y_t as explanatory variables, that is, $y_{t-j}, 1 \leq j \leq q$, are included in $\boldsymbol{x}_t$, for some prescribed positive integer q. Since endogenous variables have entered the regression function, $\beta_0 + \sum_{j=1}^{p_{\rm L}} \beta_j x_{tj}$, instead of assuming that $\{\epsilon_t\}$ in (4) is a short- or long-memory process, we now postulate that

$$\{\epsilon_t, \mathcal{F}_t\} \text{ is a martingale difference sequence,} \tag{8}$$

where $\{\mathcal{F}_t\}$ is an increasing sequence of σ-fields and $E(\epsilon_t^2) = 1$ for all t. While (8) implies that $\{\epsilon_t\}$ is no longer serially correlated, it permits $\{\epsilon_t\}$ to have the ARCH/GARCH effects. In addition, we assume that

$$(\boldsymbol{x}_t, \boldsymbol{z}_t) \text{ is } \mathcal{F}_{t-1}\text{-measurable and } \{\boldsymbol{z}_t\} \text{ is independent of } \{\epsilon_t\}. \tag{9}$$

When $N_{{\rm D},n} = \{1 \leq j \leq p_{\rm D} : \alpha_j \neq 0\}$ is an empty set (or σ_t^2 is a constant), these specifications include as the special case the high-dimensional autoregressive exogenous

(ARX) model with GARCH/ARCH errors; see Han and Tsay (2019). When $N_{\mathrm{D},n} \neq \emptyset$, these specifications allow heteroscedasticity of η_t to arise either from the exogenous variables $\boldsymbol{z}_t$ or from the ARCH/GARCH effects of ϵ_t. Since the assumptions considered in this paragraph are somewhat different from those in the previous one, in the sequel, we refer to

> model (3) with $\boldsymbol{x}_t, \eta_t, \boldsymbol{z}_t$, and ϵ_t satisfying (4), (5), and (6) (or (7)) as **Model I**,
> model (3) with $\boldsymbol{x}_t, \eta_t, \boldsymbol{z}_t$, and ϵ_t satisfying (4), (8), and (9) as **Model II**.

Our goal is to consistently estimate the relevant sets $N_{\mathrm{L},n}$ and $N_{\mathrm{D},n}$ in the regression function and the dispersion function, respectively, when Model I or II holds true, noting that the dispersion function is defined by $\sigma_t^2 = \exp\{\alpha_0 + \sum_{j=1}^{p_{\mathrm{D}}} \alpha_j z_{tj}\}$. As mentioned previously, when it is known that σ_t^2 is a constant (or $N_{\mathrm{D},n}$ is an empty set), the problem of estimating $N_{\mathrm{L},n}$ has been tackled in the literature. More specifically, Basu and Michailidis (2015) and Wu and Wu (2016) show that $N_{\mathrm{L},n}$ can be consistently estimated via Lasso if $\{\mathbf{x}_t\}$ is independent of $\{\eta_t\} = \{\epsilon_t\}$ and $\{\epsilon_t\}$ is a short-memory process. In addition, Han and Tsay (2019) show that the same property holds for Lasso in high-dimensional ARX models with GARCH errors. Instead of using Lasso, Hsu et al. (2019) propose identifying $N_{\mathrm{L},n}$ in high-dimensional ARX models using the orthogonal greedy algorithm (OGA, Temlyakov, 2000), together with the high-dimensional information criterion (HDIC, Ing and Lai, 2011) and Trim (a backward elimination method based on HDIC). Their method, referred to as the OGA+HDIC+Trim, was originally introduced by Ing and Lai (2011), who establish OGA+HDIC+Trim's selection consistency in high-dimensional regression models with i.i.d. observations. Hsu et al. (2019) show that OGA+HDIC+Trim's consistency carries over to high-dimensional ARX models. However, when $N_{\mathrm{D},n}$ is non-empty, to the best of our knowledge, no consistent estimate of $N_{\mathrm{D},n}$ or $N_{\mathrm{L},n}$ is available in the literature.

In this paper, a variable selection method intended to fill this gap is proposed. This method, modified from OGA+HDIC+Trim, is called two-stage OGA+HDIC+Trim (Twohit). Twohit contains two parts. In the first part, OGA+HDIC+Trim is used to select $N_{\mathrm{L},n}$ by ignoring the heteroscedasticity of η_t. In the second part, a natural log transformation is first taken for the least squares residuals of the regression function selected in the first part (subject to left truncation). Then, the transformed data is modeled by a linear combination of the dispersion variables (see (17)), and OGA+HDIC+Trim is used again to select $N_{\mathrm{D},n}$. The key contribution of this work is to show that Twohit consistently estimates $N_{\mathrm{D},n}$ and

$N_{\mathrm{L},n}$, regardless of whether Model I or II is assumed. The rest of the paper is organized as follows. The details of Twohit are described in Section 2. The consistency of Twoit in estimating $N_{\mathrm{L},n}$ and $N_{\mathrm{D},n}$ is reported in Section 3.1 and Section 3.2, respectively. In Section 4, simulation results are given to corroborate our theoretical findings. In Section 5, we analyze the aforementioned monthly growth rate data using Twohit. We conclude in Section 6. The proofs of the results in Section 3.1 are provided in the Appendix, whereas the proofs of the results in Section 3.2, along with additional details on our real data analysis, are deferred to the supplemental material.

We end this section with some notation used throughout the paper. For vector $\boldsymbol{a}$, $\|\boldsymbol{a}\|_1$ and $\|\boldsymbol{a}\|_2$ denote its $L1$-norm and $L2$-norm, respectively. For matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_2$ and $\lambda_{\min}(\boldsymbol{A})$ denote its spectral norm and minimum eigenvalue, respectively. The cardinality of set A is denoted by $\sharp(A)$. In addition, for sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, $a_n \asymp b_n$ means $C_1^{-1} \leq b_n/a_n \leq C_1$ for all large n, where C_1 is some positive constant.

2 Methodology

Ing and Lai (2011) consider high-dimensional regression models with i.i.d. observations, and propose using OGA+HDIC+Trim to select input variables. This method consists of (a) OGA: an iterative forward inclusion of input variables in a "greedy" manner, (b) HDIC: a stopping rule to terminate forward inclusion of variables, and (c) Trim: a backward elimination of variables according to HDIC. Ing and Lai (2011) establish the selection consistency of OGA+HDIC+Trim, which is subsequently generalized by Hsu et al. (2019) to high-dimensional ARX models.

By ignoring the heteroscedasticity of η_t, Twohit begins with choosing $N_{\mathrm{L},n}$ through OGA+HDIC+Trim. Define

$$\boldsymbol{y} = (y_1 - \bar{y}, \ldots, y_n - \bar{y})^\top,\ \boldsymbol{X}_j = (x_{1j} - \bar{x}_j, \ldots, x_{nj} - \bar{x}_j)^\top,\ j = 1, \ldots, p_{\mathrm{L}}, \tag{10}$$

and

$$\hat{\sigma}^2_{\mathrm{L},J} = n^{-1}\boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{H}_{\mathrm{L},J})\boldsymbol{y},$$

where $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i), i = 1, \ldots, n$ are data available up to time n, $\bar{y} = n^{-1}\sum_{t=1}^n y_t$, $\bar{x}_j = n^{-1}\sum_{t=1}^n x_{tj}$, $J \subseteq \{1, \ldots, p_{\mathrm{L}}\}$, $\boldsymbol{I}$ is the $n \times n$ identity matrix, and $\boldsymbol{H}_{\mathrm{L},J}$ is the orthogonal

projection matrix onto the subspace spanned by $\{\boldsymbol{X}_j : j \in J\}$. The method is described in detail as follows.

Algorithm 1 : OGA+HDIC+Trim

1. (OGA). Let $K_{\mathrm{L},n}$ be an upper bound for the number of iterations to be specified in Sections 3.1 and 4.1. Initialize the algorithm by setting $\boldsymbol{R}_{\mathrm{L},0} = \boldsymbol{y}$ and $\hat{J}_{\mathrm{L},0} = \emptyset$. For $i = 1$ to $K_{\mathrm{L},n}$, define

$$\hat{j}_{\mathrm{L},i} = \arg \max_{1 \le j \le p_{\mathrm{L}}} \frac{|\boldsymbol{R}_{\mathrm{L},i-1}^{\top} \boldsymbol{X}_j|}{\|\boldsymbol{X}_j\|_2},$$

   and update $\hat{J}_{\mathrm{L},i-1}$ and $\boldsymbol{R}_{\mathrm{L},i-1}$ by $\hat{J}_{\mathrm{L},i} = \hat{J}_{\mathrm{L},i-1} \bigcup \{\hat{j}_{\mathrm{L},i}\}$ and $\boldsymbol{R}_{\mathrm{L},i} = (\boldsymbol{I} - \boldsymbol{H}_{\mathrm{L},\hat{J}_{\mathrm{L},i}})\boldsymbol{y}$, respectively.

2. (HDIC). Set

$$\hat{k}_{\mathrm{L},n} = \mathrm{argmin}_{1 \le k \le K_{\mathrm{L},n}} \mathrm{HDIC}(\hat{J}_{\mathrm{L},k}),$$

   where for $J \subset \{1, \ldots, p_{\mathrm{L}}\}$,

$$\mathrm{HDIC}(J) = n \log \hat{\sigma}^2_{\mathrm{L},J} + \sharp(J) G_{\mathrm{L}}(p_{\mathrm{L}}, n), \tag{11}$$

   and $G_{\mathrm{L}}(p_{\mathrm{L}}, n)$ is a penalty term to be specified in Sections 3.1 and 4.1.

3. (Trim). Output

$$\hat{N}_{\mathrm{L},n} = \begin{cases} \{\hat{j}_{\mathrm{L},l} : \mathrm{HDIC}(\hat{J}_{\mathrm{L},\hat{k}_{\mathrm{L},n}} - \{\hat{j}_{\mathrm{L},l}\}) > \mathrm{HDIC}(\hat{J}_{\mathrm{L},\hat{k}_{\mathrm{L},n}}), 1 \le l \le \hat{k}_{\mathrm{L},n}\}, & \text{if } \hat{k}_{\mathrm{L},n} > 1; \\ \{\hat{j}_{\mathrm{L},1}\}, & \text{if } \hat{k}_{\mathrm{L},n} = 1. \end{cases}$$

It is worth mentioning that the major difference between HDIC and conventional consistent information criteria such BIC and HQ is that the penalty terms in the latter criteria depend only on the sample size n, whereas the penalty term $G_{\mathrm{L}}(p_{\mathrm{L}}, n)$ in the former depends not only on n but also on the number of candidate variables p_{L}, and hence can be much larger than those in BIC and HQ when $p_{\mathrm{L}} \gg n$; see Section 3.1 for details. The larger penalty term in HDIC is used to adjust for potential spuriousness of the variables greedily chosen by OGA from among p_{L} candidate variables. Although OGA+HDIC+Trim selects input variables without taking the heteroscedasticity of $\eta_t = \sigma_t \epsilon_t$ into account, we show in Section 3.1 that

$$\lim_{n \to \infty} P(\hat{N}_{\mathrm{L},n} = N_{\mathrm{L},n}) = 1 \tag{12}$$

still follows if, among other assumptions, σ_t has finite higher-order moments (see Assumption (A1)) and $G_{\mathrm{L}}(p_{\mathrm{L}}, n)$ diverges to ∞ at a suitable rate.

The second part of Twohit is to select dispersion variables using $\mathrm{OGA}_{\mathcal{D}}$+$\mathrm{HDIC}_{\mathcal{D}}$+$\mathrm{Trim}_{\mathcal{D}}$, where $\mathrm{OGA}_{\mathcal{D}}$, $\mathrm{HDIC}_{\mathcal{D}}$, and $\mathrm{Trim}_{\mathcal{D}}$, respectively, are counterparts to OGA, HDIC, and Trim. Define

$$\tilde{\eta}_t^2 = \max\{\hat{\eta}_t^2, \underline{c}_n\}, \;\; r_t = \log \tilde{\eta}_t^2, \tag{13}$$

$$\boldsymbol{r} = (r_1 - \bar{r}, \ldots, r_n - \bar{r})^\top, \boldsymbol{Z}_j = (z_{1j} - \bar{z}_j, \ldots, z_{nj} - \bar{z}_j)^\top, \; j = 1, \ldots, p_{\mathrm{D}}, \tag{14}$$

and

$$\hat{\sigma}^2_{\mathrm{D},J} = n^{-1} \boldsymbol{r}^\top (\boldsymbol{I} - \boldsymbol{H}_{\mathrm{D},J}) \boldsymbol{r},$$

where $(\hat{\eta}_1, \ldots, \hat{\eta}_n)^\top = (\boldsymbol{I} - \boldsymbol{H}_{\mathrm{L},\hat{N}_{\mathrm{L},n}})\boldsymbol{y}$ are the ordinary least squares residuals of the model selected by OGA+HDIC+Trim, $\underline{c}_n$ is a small positive constant, $\bar{r} = n^{-1} \sum_{t=1}^n r_t$, $\bar{z}_j = n^{-1} \sum_{t=1}^n z_{tj}$, $J \subseteq \{1, \ldots, p_{\mathrm{D}}\}$, and $\boldsymbol{H}_{\mathrm{D},J}$ is the orthogonal projection matrix onto the subspace spanned by $\{\boldsymbol{Z}_j : j \in J\}$. $\mathrm{OGA}_{\mathcal{D}}$+$\mathrm{HDIC}_{\mathcal{D}}$+$\mathrm{Trim}_{\mathcal{D}}$ is described as follows.

Algorithm 2 : $\text{OGA}_{\mathcal{D}}$+$\text{HDIC}_{\mathcal{D}}$+$\text{Trim}_{\mathcal{D}}$

1. ($\text{OGA}_{\mathcal{D}}$). Let $K_{\text{D},n}$ be an upper bound for the number of iterations to be specified in Sections 3.2 and 4.1. Initialize the algorithm by setting $\boldsymbol{R}_{\text{D},0} = \boldsymbol{r}$ and $\hat{J}_{\text{D},0} = \emptyset$. For $i = 1$ to $K_{\text{D},n}$, define

$$\hat{j}_{\text{D},i} = \arg \max_{1 \le j \le p_{\text{D}}} \frac{|\boldsymbol{R}_{\text{D},i-1}^{\top} \boldsymbol{Z}_j|}{\|\boldsymbol{Z}_j\|_2},$$

and update $\hat{J}_{\text{D},i-1}$ and $\boldsymbol{R}_{\text{D},i-1}$ by $\hat{J}_{\text{D},i} = \hat{J}_{\text{D},i-1} \bigcup \{\hat{j}_{\text{D},i}\}$ and $\boldsymbol{R}_{\text{D},i} = (\boldsymbol{I} - \boldsymbol{H}_{\text{D},\hat{J}_{\text{D},i}})\boldsymbol{r}$, respectively. (Note that the choice of $\underline{c}_n$ in $\tilde{\eta}_t^2$ will be discussed in Sections 3.2 and 4.1).

2. ($\text{HDIC}_{\mathcal{D}}$). Set

$$\hat{k}_{\text{D},n} = \text{argmin}_{1 \le k \le K_{\text{D},n}} \text{HDIC}_{\text{D}}(\hat{J}_{\text{D},k}),$$

where

$$\text{HDIC}_{\text{D}}(J) = n \log \hat{\sigma}_{\text{D},J}^2 + \sharp(J) G_{\text{D}}(p_{\text{D}}, n), \tag{15}$$

and $G_{\text{D}}(p_{\text{D}}, n)$ is a penalty term to be specified in Sections 3.2 and 4.1.

3. ($\text{Trim}_{\mathcal{D}}$). Output

$$\hat{N}_{\text{D},n} = \begin{cases} \{\hat{j}_{\text{D},l} : \text{HDIC}_{\text{D}}(\hat{J}_{\text{D},\hat{k}_{\text{D},n}} - \{\hat{j}_{\text{D},l}\}) > \text{HDIC}_{\text{D}}(\hat{J}_{\text{D},\hat{k}_{\text{D},n}}), 1 \le l \le \hat{k}_{\text{D},n}\}, & \text{if } \hat{k}_{\text{D},n} > 1; \\ \{\hat{j}_{\text{D},1}\}, & \text{if } \hat{k}_{\text{D},n} = 1. \end{cases}$$

We now briefly explain why $\boldsymbol{r}$ is used in OGA$_{\mathcal{D}}$+HDIC$_{\mathcal{D}}$+Trim$_{\mathcal{D}}$. In view of (4), it is clear that

$$\log(\eta_t^2) = \tilde{\alpha}_0 + \sum_{j=1}^{p_{\mathrm{D}}} \alpha_j z_{tj} + \varepsilon_t, \quad t = 1, 2, \ldots, n, \tag{16}$$

where $\tilde{\alpha}_0 = \alpha_0 + E(\log \epsilon_t^2)$ and $\varepsilon_t = \log \epsilon_t^2 - E(\log \epsilon_t^2)$. While η_t on the left-hand side of (16) is unobservable, it can be estimated by $\hat{\eta}_t$. However, $\log \hat{\eta}_t^2$ may face numerical and statistical difficulties when $\hat{\eta}_t$ is very close to 0. We therefore adopt a left-truncated version, $\tilde{\eta}_t^2$, of $\hat{\eta}_t^2$ to estimate η_t^2, and reexpress (16) as

$$r_t = \tilde{\alpha}_0 + \sum_{j=1}^{p_{\mathrm{D}}} \alpha_j z_{tj} + \varepsilon_t + \Theta_{t,n}, \quad t = 1, 2, \ldots, n, \tag{17}$$

where $\Theta_{t,n} = \log(\tilde{\eta}_t^2) - \log(\eta_t^2)$. When $\underline{c}_n$ in (13) ($G_{\mathrm{D}}(p_{\mathrm{D}}, n)$ in (15)) converges (diverges) to 0 (∞) at a suitable rate and the distributions of η_t^2 satisfy some smoothness conditions at the origin (see (30)), it is argued in Section 3.2 that the impact of $\Theta_{t,n}$ vanishes asymptotically, and

$$\lim_{n \to \infty} P(\hat{N}_{\mathrm{D},n} = N_{\mathrm{D},n}) = 1. \tag{18}$$

3 Theoretical Properties

This section aims to develop the selection consistency of Twohit when p_{L} and p_{D} are allowed to be much larger than n. In particular, the selection consistency of OGA+HDIC+Trim in selecting $N_{\mathrm{L},n}$ and that of OGA$_{\mathcal{D}}$+HDIC$_{\mathcal{D}}$+Trim$_{\mathcal{D}}$ in selecting $N_{\mathrm{D},n}$ is established in Sections 3.1 and 3.2, respectively.

3.1 Consistency of OGA+HDIC+Trim in Selecting $N_{\mathrm{L},n}$

In this section, we assume that $\beta_0 = 0$ and $\{\boldsymbol{x}_t\}$ is a covariance stationary time series satisfying $E(\boldsymbol{x}_t) = 0$ and $E(x_{tj}^2) = 1$ for all j. We only show that OGA+HDIC+Trim is consistent when $\bar{y}$ and $\bar{x}_j$ in (10) are set to 0. However, our argument can be easily generalized to prove the consistency of OGA+HDIC+Trim in situations where β_0, $E(\boldsymbol{x}_t)$, or $E(x_{tj}^2)$ is unknown.

Let $\boldsymbol{x} = (x_1, \ldots, x_{p_{\rm L}})^\top$ be independent of and have the same covariance structure as $\{\boldsymbol{x}_t\}$, $y(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}$, and $\boldsymbol{\Gamma}_{\rm L}(J) = E(\boldsymbol{x}_t(J)\boldsymbol{x}_t^\top(J))$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_{\rm L}})^\top$ and $\boldsymbol{x}_t(J) = (x_{tj}, j \in J)^\top$. Recall that $\hat{J}_{{\rm L},m}$ is the index set determined by OGA at the m-th iteration. We start by investigating the convergence rate of

$$E_n[(y(\boldsymbol{x}) - \hat{y}_{\hat{J}_{{\rm L},m}}(\boldsymbol{x}))^2],$$

which plays a crucial role in proving (12). Here, $\hat{y}_J(\boldsymbol{x}) = \boldsymbol{x}^\top(J)\hat{\boldsymbol{\beta}}(J)$, $\boldsymbol{x}(J) = (x_i, i \in J)^\top$, $\hat{\boldsymbol{\beta}}(J) = (\sum_{t=1}^n \boldsymbol{x}_t(J)\boldsymbol{x}_t^\top(J))^{-1} \sum_{t=1}^n \boldsymbol{x}_t(J) y_t$, and

$$E_n(\cdot) = E(\cdot | y_1, \boldsymbol{x}_1, \boldsymbol{z}_1, \ldots, y_n, \boldsymbol{x}_n, \boldsymbol{z}_n).$$

The following assumptions are needed in our analysis.

(A1) (a) For Model I, there exists $q_1 \geq 2$ such that $E|w_1|^{\max\{q_1,4\}} < \infty$, $\sup_{-\infty<t<\infty} E|\sigma_t|^{2q_1} < \infty$, and $\max_{1\leq t\leq n, 1\leq i\leq p_{\rm L}} E|x_{ti}|^{2q_1} = O(1)$.

(b) For Model II, (i) there exists $q_1 \geq 2$ such that $\sup_{-\infty<t<\infty} E|\epsilon_t|^{3q_1} < \infty$, $\sup_{-\infty<t<\infty} E|\sigma_t|^{3q_1} < \infty$, and $\max_{1\leq t\leq n, 1\leq i\leq p_{\rm L}} E|x_{ti}|^{3q_1} = O(1)$; (ii) ϵ_t^2 is a stationary process satisfying $\lim_{k\to\infty} {\rm cov}(\epsilon_1^2, \epsilon_{1+k}^2) = 0$.

(A2) For some $q_2 \geq 2$,

$$\max_{1\leq i,j\leq p_{\rm L}} E\left| n^{-1/2} \sum_{t=1}^n (x_{ti}x_{tj} - E(x_{ti}x_{tj})) \right|^{2q_2} = O(1). \tag{19}$$

(A3) For some $0 < \underline{q} < \min\{q_1, q_2\}$, $p_{\rm L}^{2/\underline{q}}/n^{1-2d}$, where $0 < d < 1/2$ is defined in (7) and $d = 0$ if (6) or (8) follows.

(A4) $\sup_{n\geq 1} \sum_{j=1}^{p_{\rm L}} |\beta_j| < \infty$.

(A5) There are some $\delta_{\rm L} > 0$ and $M_{\rm L} > 0$ such that for all larger n,

$$\begin{aligned} &\min_{1\leq \sharp(J)\leq K_{{\rm L},n}} \lambda_{\min}(\boldsymbol{\Gamma}_{\rm L}(J)) > \delta_{\rm L}, \\ &\max_{1\leq \sharp(J)\leq K_{{\rm L},n}, i\notin J} \|\boldsymbol{\Gamma}_{\rm L}^{-1}(J)\boldsymbol{g}_{{\rm L},i}(J)\|_1 \leq M_{\rm L}, \end{aligned} \tag{20}$$

where $\boldsymbol{g}_{{\rm L},i}(J) = E(\boldsymbol{x}_t(J)x_{ti})$.

Some comments are in order.

(1). Assumptions (A1)(a) (or (A1)(b)(i)) and (A2)–(A5) resemble (F1)–(F5) of Hsu et al. (2019), which are made to ensure that OGA has the desired asymptotic property in high-dimensional time series models with homogeneous errors. In particular, (A2), (A4), and (A5) are almost the same as (F1), (F4), and (F5) of Hsu et al. (2019), respectively. Condition (A1)(a) (or (A1)(b)(i)) is the key assumption leading to

$$\max_{1\le i\le p_{\mathrm{L}}}\left|\frac{1}{n}\sum_{t=1}^{n}x_{ti}\eta_t\right| = O_p(n^{-1/2+d}p_{\mathrm{L}}^{1/q_1});$$

see Lemma A.1 in the Appendix for details. Hence, (A1)(a) (or (A1)(b)(i)) is similar in spirit to (F2) of Hsu et al. (2019), which imposes a moment condition on $n^{-1}\sum_{t=1}^{n}x_{ti}\epsilon_t$.

(2). In some cases, assumptions like

$$\max_{1\le t\le n,1\le i\le p_{\mathrm{L}}}E|x_{ti}|^{2q} = O(1), q\ge 2, \tag{21}$$

used in (A1) can imply that (19) in (A2) holds with $q_2 = q/2$. For example, assume that $\{x_{ti}\}$ admits an infinite moving-average representation

$$x_{ti} = \sum_{j=0}^{\infty}b_j(i)\nu_{t-j}(i), \tag{22}$$

where $\{\nu_t(i)\}$ is a martingale difference sequence with respect to an increasing sequence of σ-fields $\{\mathcal{G}_t\}$ and $\max_{1\le i\le p_{\mathrm{L}}}\sum_{j=0}^{\infty}|(j+1)b_j(i)| < C_0$ for some positive constant C_0. Also assume

$$\nu_t^2(i) - E(\nu_t^2(i)) = \sum_{j=0}^{\infty}\theta_j(i)\eta_{t-j}(i), \tag{23}$$

where $\{\eta_t(i),\mathcal{G}_t\}$ is a martingale difference sequence and $\max_{1\le i\le p_{\mathrm{L}}}\sum_{j=0}^{\infty}|\theta_j(i)| + \max_{1\le i\le p_{\mathrm{L}}}n\sum_{j\ge n}\theta_j^2(i) < C_0^*$ for some positive constant C_0^*. Then, by Burkholder's inequality, (21)–(23), and some algebraic manipulations, it can be shown that (A2) holds with $q_2 = q/2$. Note that (22) and (23) are satisfied not only by linear processes with i.i.d. innovations, but also by linear processes with stationary GARCH innovations.

(3). To illustrate the flexibility of (A2), we consider the following high-dimensional ARX model,

$$y_t = \sum_{j=1}^{k} \phi_j y_{t-j} + \sum_{v=1}^{p} \sum_{j=0}^{r_v} \eta_j^{(v)} s_{t-j}^{(v)} + \sigma_t \epsilon_t, \ 1, \ldots, n,$$

where p is a positive integer which can be larger than n, k and $r_v, 1 \le v \le p$, are positive integers bounded from above, $1 - \sum_{j=1}^{k} \phi_j z^j \neq 0$ for $|z| \le 1$, $\sum_{j=1}^{k} |\phi_j| + \sum_{v=1}^{p} \sum_{j=0}^{r_v} |\eta_j^{(v)}| < \infty$, $s_t^{(v)} = \sum_{j=0}^{\infty} \psi_j^{(v)} \delta_{t-j}^{(v)}$ with $\sum_{j=0}^{\infty} (\psi_j^{(v)})^2 < \infty$ and $\boldsymbol{\delta}_t(p) = (\delta_t^{(1)}, \ldots, \delta_t^{(p)})^\top$ being independent random vectors satisfying $\mathrm{E}\left(\boldsymbol{\delta}_t(p)\boldsymbol{\delta}_t^\top(p)\right) = \Sigma_p$ (a p-dimensional positive definite matrix) and $\mathrm{E}(\boldsymbol{\delta}_t(p)) = \mathbf{0}$, $\{\epsilon_t\}$ is a stationary GARCH(r_1, r_2) process independent of $\{\boldsymbol{\delta}_t(p)\}$, with $1 \le r_1 + r_2 < \infty$, and $\{\sigma_t\}$ is a sequence of positive random variables independent of $\{\epsilon_t\}$. It is clear that the input vectors at time t is $\boldsymbol{x}_t = (y_{t-1}, \ldots, y_{t-k}, s_t^{(1)}, \ldots s_{t-r_1}^{(1)}, \ldots, s_t^{(p)}, \ldots, s_{t-r_p}^{(p)})^\top$ and the number of input variables is $p_{\mathrm{L}} = k + \sum_{v=1}^{p}(r_v + 1)$. Let $\gamma_i(v), i = 1, 2, \ldots$, be the autocovariance function of $\{s_t^{(v)}\}$. Then by the First Moment Bound Theorem of Findley and Wei (1993), the Burkholder inequality, and some algebraic manipulations, it can be shown that (A2) holds true, provided $\max_{1 \le v \le p} \sum_{i=-\infty}^{\infty} \gamma_i^2(v) < C_1$, $\sup_{-\infty < t < \infty} \max_{1 \le v \le p} E|\delta_t^{(v)}|^{4q_2} + E|\epsilon_t|^{4q_2} < C_2$, and

$$\sigma_t^2 = \exp\left\{ \alpha_0 + \sum_{j=0}^{\infty} \boldsymbol{w}_j^\top \boldsymbol{\eta}_{t-j} \right\}, \tag{24}$$

where C_1 and C_2 are some positive constants, $\{\boldsymbol{\eta}_t\}$ is a sequence of i.i.d. p_{D}-dimensional sub-Gaussian random vectors in which p_{D} can be larger than n, and $\boldsymbol{w}_j$ are p_{D}-dimensional coefficient vectors satisfying $\sum_{j=0}^{\infty} \|\boldsymbol{w}_j\|_1 < \infty$.

(4). By (A3), p_{L} is allowed to be larger than n if $\min\{q_1, q_2\} > 2/(1-2d)$. This implies that the stronger the dependence of ϵ_t, the more stringent moment assumptions are needed for OGA to handle the case of $p_{\mathrm{L}} > n$. Note that (A3) is more restrictive than (F3) of Hsu et al. (2019), which is (A3) with $d = 0$. However, if we assume that σ_t^2 is a constant and $\{\epsilon_t\}$ is a stationary fractionally integrated process of order d (i.e., $(1 - B)^d \epsilon_t = w_t$, where B is a back-shift operator) with $0 \le d < 1/4$, then (A3) can be weakened to the latter one.

(5). Conditions (A4) and (A5) introduce mild restrictions on the regression coefficients and the correlations among regressors. They are frequently made to analyze the performance of OGA in high-dimensional regression or time series models with homogeneous errors; see Ing and Lai (2011), Hsu et al. (2019), and Ing (2019).

(6). Finally, we remark that (A1)(b)(ii) is satisfied by a broad class conditional heteroscedastic time series. In particular, it is shown in Proposition 3.1 of Giraitis et al. (2000) that (A1)(b)(ii) is fulfilled by the stationary ARCH(∞) process,

$$\epsilon_t^2 = \rho_t \xi_t, \quad \rho_t = b_0 + \sum_{j=1}^{\infty} b_j \epsilon_{t-j}^2, \tag{25}$$

where ξ_t are i.i.d. nonnegative random variables satisfying $E(\xi_0^2) = 1$ and b_j are nonnegative numbers obeying $\sum_{j=1}^{\infty} b_j < 1$.

Theorem 1 *Assume that* Model I (Model II), (A1)(a) ((A1)(b)(i)), *and* (A2)–(A5) *hold. Suppose*

$$K_{\mathrm{L},n} \asymp n^{1/2-d} p_{\mathrm{L}}^{-1/\underline{q}}. \tag{26}$$

Then

$$\max_{1 \le m \le K_{\mathrm{L},n}} \frac{E_n[(y(\boldsymbol{x}) - \hat{y}_{\hat{J}_{\mathrm{L},m}}(\boldsymbol{x}))^2]}{m^{-1} + m p_{\mathrm{L}}^{2/\underline{q}} / n^{1-2d}} = O_p(1). \tag{27}$$

Theorem 1 reveals that $E_n[(y(\boldsymbol{x}) - \hat{y}_{\hat{J}_{\mathrm{L},m}}(\boldsymbol{x}))^2]$ is uniformly bounded by the sum of two terms, m^{-1} and $m p_{\mathrm{L}}^{2/\underline{q}} / n^{1-2d}$. The first term, m^{-1}, caused by approximating the regression function using OGA, decreases as the number of iterations increases. The second term, $m p_{\mathrm{L}}^{2/\underline{q}} / n^{1-2d}$, is associated with the estimation error. While this term decreases with the sample size, it increases with the numbers of OGA iterations as well as the candidate variables. It also becomes larger as d increases or $\underline{q}$ decreases. When m grows to ∞ at a rate not exceeding $n^{1/2-d} p_{\mathrm{L}}^{-1/\underline{q}}$, the aforementioned sum converges to 0, suggesting that $\hat{y}_{\hat{J}_{\mathrm{L},m}}(\boldsymbol{x})$ provides a good approximation of $y(\boldsymbol{x})$. This feature, together with the following 'beta-min' condition (see (A6)) and an assumption on the penalty term, $G_{\mathrm{L}}(p_{\mathrm{L}}, n)$, of HDIC (see (29)), ensures the selection consistency of OGA+HDIC+Trim, as detailed in Theorem 2.

(A6) There exists $\gamma_{\rm L}$ satisfying $0 \le \gamma_{\rm L} < 1/2 - d$ and $n^{\gamma_{\rm L}} = o(n^{1/2-d}/p_{\rm L}^{1/q})$ such that

$$\liminf_{n\to\infty} n^{\gamma_{\rm L}} \min_{j\in N_{{\rm L},n}} \beta_j^2 > 0,$$

where $0 < d < 1/2$ is defined in (7) and $d = 0$ if (6) or (8) follows.

Theorem 2 *Assume that* Model I (Model II), (A1)(a) ((A1)(b)), (A2)–(A6), (26), *and*

$$\frac{1}{n}\sum_{t=1}^{n}(\sigma_t^2 - E(\sigma_t^2)) = o_p(1) \quad \textit{and} \quad \liminf_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} E(\sigma_t^2) > 0 \tag{28}$$

hold. In addition, suppose that $G_{\rm L}(p_{\rm L}, n)$ *in* (11) *satisfies*

$$\frac{G_{\rm L}(p_{\rm L}, n)}{n^{1-2\gamma_{\rm L}}} = o(1) \textit{ and } \frac{n^{2d} p_{\rm L}^{2/q}}{G_{\rm L}(p_{\rm L}, n)} = o(1). \tag{29}$$

Then (12) *follows.*

We close this section by noting that (28) is a high-level assumption fulfilled by a wide range of time series models. For example, it can be shown that (24) satisfies (28). Moreover, Theorem 1 in the supplement of Gao and Ling (2019) shows that if $\mathcal{Y}_t \equiv (\sigma_t^2 - E(\sigma_t^2))$ is a stationary process satisfying $\sup_{-\infty<t<\infty} E|\mathcal{Y}_t|^\iota < \infty$ for some $1 < \iota < 2$ and a strong mixing condition, then for some $\delta \in (0, 1)$,

$$n^{-1+\delta}\sum_{t=1}^{n} \mathcal{Y}_t \to 0 \text{ a.s.},$$

yielding (28).

3.2 Consistency of OGA$_{\mathcal{D}}$+HDIC$_{\mathcal{D}}$+Trim$_{\mathcal{D}}$ in Selecting $N_{{\rm D},n}$

This section is concerned with the performance of OGA$_{\mathcal{D}}$+HDIC$_{\mathcal{D}}$+Trim$_{\mathcal{D}}$. In order to simplify the exposition, we assume in (17) that $\tilde{\alpha}_0$ is known, $E(\boldsymbol{z}_t) = 0$, and $E(z_{tj}^2) = 1$. Our goal is to establish the selection consistency of OGA$_{\mathcal{D}}$+HDIC$_{\mathcal{D}}$+Trim$_{\mathcal{D}}$ when $\bar{r}$ and $\bar{z}_j$ in (14) are replaced by $\tilde{\alpha}_0$ and 0, respectively. Our argument can be easily generalized to show the consistency of the method in situations where $\tilde{\alpha}_0$ is unknown, $E(\boldsymbol{z}_t) \neq 0$,

or $E(z_{tj}^2) \neq 1$. Note that the major difference between (17) and the conventional high-dimensional regression model is the presence of $\Theta_{t,n} = \log(\tilde{\eta}_t^2) - \log(\eta_t^2)$. Therefore, the issue here is to control the local behavior of $\Theta_{t,n}$ when η_t^2 is near the origin. To this end, we impose the following condition.

(A0′) There exist positive constants δ_1 (which can be arbitrarily small) and τ such that for all $0 < s < \delta_1$,

$$\sup_{t \geq 1} P(\eta_t^2 < s) \leq c_\tau s^\tau, \tag{30}$$

where c_τ is a positive constant that may depend on τ.

If ϵ_t satisfies (25) with $\sup_{t\geq 1} P(\xi_t < s) \leq cs^\tau$ for some $c > 0$, and $\sup_{t\geq 1} E(|\sigma_t^{-2\tau}|) < \infty$, then (30) follows. Moreover, when ϵ_t is a stationary Gaussian process and $\sup_{t\geq 1} E(|\sigma_t^{-1}|) < \infty$, it can be shown that (30) holds with $\tau = 1/2$. Although it is possible to verify (30) under more general conditions, we leave this issue for future research. In addition to (A0′), we also need a series assumptions parallel to (A1)–(A6). Throughout this section, we assume that (A1) is true for a sufficiently large q_1 and (A6) holds for $\gamma_{\mathrm{L}} = 0$ in order to avoid excessive technicalities. Define $\boldsymbol{z}_t(J) = (z_{tj}, j \in J)^\top$.

(A1′) There is $q_3 \geq 4$ such that

$$\sup_{t \geq 1} E|\log \epsilon_t^2|^{q_3} < \infty, \tag{31}$$

and

$$\max_{1 \leq i \leq p_D} E\left|\sum_{t=1}^n \xi_t z_{ti}\right|^{q_3} \leq C_{q_3}\left(\sum_{t=1}^n \xi_t^2\right)^{q_3/2}, \tag{32}$$

where $\xi_1, \ldots, \xi_n$ are any real numbers.

(A2′) $\{\boldsymbol{z}_t\}$ is a covariance stationary time series satisfying

$$\max_{1 \leq i,j \leq p_{\mathrm{D}}} E\left|n^{-1/2}\sum_{t=1}^n (z_{ti}z_{tj} - E(z_{ti}z_{tj}))\right|^{2q_4} = O(1),$$

where $q_4 \geq 2$.

(A3′) $p_{\mathrm{D}}^{2/q}/n^{2\kappa} = o(1)$, where $0 < d < 1/2$ is defined in (7), and $d = 0$ if (6) or (8) follows.

(A4′) $\sup_{n\geq 1}\sum_{j=1}^{p_{\mathrm{D}}} |\alpha_j| < \infty$.

(A5′) There are some $\delta_{\mathrm{D}} > 0$ and $M_{\mathrm{D}} > 0$ such that for all larger n,

$$\min_{1\leq \sharp(J)\leq K_{\mathrm{D},n}} \lambda_{\min}(\mathbf{\Gamma}_{\mathrm{D}}(J)) > \delta_{\mathrm{D}}, \tag{33}$$
$$\max_{1\leq \sharp(J)\leq K_{\mathrm{D},n}, i\notin J} \|\mathbf{\Gamma}_{\mathrm{D}}^{-1}(J)\boldsymbol{g}_{\mathrm{D},i}(J)\|_1 \leq M_{\mathrm{D}}.$$

where $\mathbf{\Gamma}_{\mathrm{D}}(J) = E(\boldsymbol{z}_t(J)\boldsymbol{z}_t^{\top}(J))$ and $\boldsymbol{g}_{\mathrm{D},i}(J) = E(\boldsymbol{z}_t(J)z_i)$.

(A6′) There exists $0 \leq \gamma_{\mathrm{D}} < \kappa$ such that $n^{\gamma_{\mathrm{D}}} = o(n^{\kappa}/p_{\mathrm{D}}^{1/q})$ and

$$\liminf_{n\to\infty} n^{\gamma_{\mathrm{D}}} \min_{j\in N_{\mathrm{D},n}} \alpha_j^2 > 0,$$

where κ is defined in (A3′).

It can be shown that (31) holds when ϵ_t is a stationary Gaussian process or when ϵ_t is a stationary ARCH (∞) process (see (25)) with a finite second moment and with ξ_t obeying $\sup_{t\geq 1} E|\log \xi_t|^{q_3} < \infty$. Equations (31) and (32) imply

$$\max_{1\leq i\leq p_{\mathrm{D}}} \left|\frac{1}{n}\sum_{t=1}^{n} z_{ti}\varepsilon_t\right| = O_p(n^{-1/2}p_{\mathrm{D}}^{1/q_3}),$$

which is crucial for proving the consistency of OGA$_{\mathcal{D}}$+HDIC$_{\mathcal{D}}$+Trim$_{\mathcal{D}}$; see the supplementary material for details. By using Lemma 2 of Wei (1987) and Theorem 2.1 of Ing and Wei (2006), (32) holds true, provided $\{z_{ti}\}$ is a linear process generated by an i.i.d. sequence $\{\delta_{ti}\}$ satisfying $\max_{1\leq i\leq p_{\mathrm{D}}} E|\delta_{ti}|^{q_3} = O(1)$, and has a square summable autocovariance function. Assumptions (A2′), (A4′), and (A5′) play the same roles as those played by (A2), (A4), and (A5) in the analysis of OGA+HDIC+Trim. Assumptions (A3′) and (A6′) appear to be more stringent than (A3) and (A6) because of $\kappa < 1/2 - d$. However, (A3′) does not preclude $p_{\mathrm{D}} \gg n$ if q is sufficiently large, and (A6′) still allows that $\min_{j\in N_{\mathrm{D},n}} |\alpha_j|$ converges to 0 slowly. The main result of this section is given in the next theorem.

Theorem 3 *Assume that the same assumptions as in Theorem 2 and* (A0′)–(A6′) *hold. Also assume that* $K_{\mathrm{D},n} \asymp n^{\kappa}/p_{\mathrm{D}}^{1/q}$,

$$\frac{1}{n}\sum_{t=1}^{n}(\varepsilon_t^2 - E(\varepsilon_t^2)) = o_p(1), \quad \liminf_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}E(\varepsilon_t^2) > 0, \tag{34}$$

$\underline{c}_n \asymp n^{-1/2+d}$ *in* (13)*, and* $G_{\mathrm{D}}(p_{\mathrm{D}}, n)$ *in* (15) *obeys*

$$\frac{G_{\mathrm{D}}(p_{\mathrm{D}}, n)}{n^{1-2\gamma_{\mathrm{D}}}} = o(1), \quad \frac{n^{1-2\kappa}p_D^{2/q}}{G_{\mathrm{D}}(p_{\mathrm{D}}, n)} = o(1). \tag{35}$$

Then (18) *holds true.*

Condition (34), playing a role similar to (28) in Theorem 2, is also easily satisfied in practice. For example, under (5) and (6) (or (7)), (34) holds, provided w_t obeys mild moment conditions and its distribution follows some smoothness conditions at the origin. In addition, (34) follows when ϵ_t is a stationary GARCH process satisfying mild moment conditions and the logarithms of the absolute values of its corresponding innovations are i.i.d. random variables having a finite second moment.

4 Simulation Studies

We carry out simulation studies to evaluate the finite sample performance of Twohit. Section 4.1 provides a guideline for choosing tuning parameters in Twohit. In Section 4.2, we compare Twohit with the doubly regularized approach of Daye et al. (2012), Lasso, and adaptive Lasso using simulated data from Models I and II.

4.1 Selection of Tuning Parameters in Twohit

Consider $K_{\mathrm{L},n}$ and $G_{\mathrm{L}}(p_{\mathrm{L}}, n)$, $\underline{c}_n$, $K_{\mathrm{D},n}$, and $G_{\mathrm{D}}(p_{\mathrm{D}}, n)$ in Twohit. In our simulation study, we set

$$\begin{aligned}
K_{\mathrm{L},n} &= \min\{\lfloor C_{\mathrm{L},1} n^{\iota} p_{\mathrm{L}}^{-1/q_{\mathrm{L}}} \rfloor, p_{\mathrm{L}}\}, \\
G_{\mathrm{L}}(p_{\mathrm{L}}, n) &= C_{\mathrm{L},2} n^{1-2\iota} p_{\mathrm{L}}^{2/q_{\mathrm{L}}} \log\log n, \\
\underline{c}_n &= 10^{-8} n^{-\iota}, \\
K_{\mathrm{D},n} &= \min\{\lfloor C_{\mathrm{D},1} n^{\nu} p_{\mathrm{D}}^{-1/q_{\mathrm{D}}} \rfloor, p_{\mathrm{D}}\}, \\
G_{\mathrm{D}}(p_{\mathrm{D}}, n) &= C_{\mathrm{D},2} n^{1-2\nu} p_{\mathrm{D}}^{2/q_{\mathrm{D}}} \log\log n,
\end{aligned}$$

where $C_{\mathrm{L},1}, C_{\mathrm{L},2}, q_{\mathrm{L}}, q_{\mathrm{D}}, \iota, \nu, C_{\mathrm{D},1}$, and $C_{\mathrm{D},2}$ are tuning parameters. For a given

$$\boldsymbol{\vartheta} = (C_{\mathrm{L},1}, C_{\mathrm{L},2}, q_{\mathrm{L}}, q_{\mathrm{D}}, \iota, \nu, C_{\mathrm{D},1}, C_{\mathrm{D},2})^{\top},$$

we use Twohit in Section 2 to determine $\hat{N}_{\mathrm{L},n} = \hat{N}_{\mathrm{L},n}(\boldsymbol{\vartheta})$ and $\hat{N}_{\mathrm{D},n} = \hat{N}_{\mathrm{D},n}(\boldsymbol{\vartheta})$. We then minimize the loss function

$$\sum_{t=1}^{n} \left(\alpha_0 + \sum_{i \in \hat{N}_{\mathrm{D},n}(\boldsymbol{\vartheta})} \alpha_i z_{ti} \right) + \sum_{t=1}^{n} \frac{\left(y_t - \beta_0 - \sum_{j \in \hat{N}_{\mathrm{L},n}(\boldsymbol{\vartheta})} \beta_j x_{tj} \right)^2}{\exp\left\{ \alpha_0 + \sum_{i \in \hat{N}_{\mathrm{D},n}(\boldsymbol{\vartheta})} \alpha_i z_{ti} \right\}},$$

with respect to α_i and β_j, and denote the corresponding minimum value by $L(\boldsymbol{\vartheta})$. Let $\boldsymbol{\Theta}$ be a range of $\boldsymbol{\vartheta}$ depending on the user's choice. Define

$$\boldsymbol{\vartheta}^* = \operatorname*{argmin}_{\boldsymbol{\vartheta} \in \boldsymbol{\Theta}} L(\boldsymbol{\vartheta}) + (\sharp(\hat{N}_{\mathrm{L},n}(\boldsymbol{\vartheta})) + \sharp(\hat{N}_{\mathrm{D},n}(\boldsymbol{\vartheta})))P_n,$$

where P_n is a prescribed positive number that may vary with n. Then $\hat{N}_{\mathrm{L},n}(\boldsymbol{\vartheta}^*)$ and $\hat{N}_{\mathrm{D},n}(\boldsymbol{\vartheta}^*)$ are the final outputs of Twohit.

Note that $L(\boldsymbol{\vartheta}) + (\sharp(\hat{N}_{\mathrm{L},n}(\boldsymbol{\vartheta})) + \sharp(\hat{N}_{\mathrm{D},n}(\boldsymbol{\vartheta})))P_n$ is nothing but a conventional information criterion (with penalty P_n) when $\{\epsilon_t\}$ is assumed to be a sequence of i.i.d. $\mathcal{N}(0,1)$ random variables. In the next section, we set $P_n = 2\log\log n$ and $\boldsymbol{\Theta} = \boldsymbol{\vartheta}_1 \times \boldsymbol{\vartheta}_4 \times \boldsymbol{\vartheta}_2 \times \boldsymbol{\vartheta}_2 \times \boldsymbol{\vartheta}_3 \times \boldsymbol{\vartheta}_3 \times \boldsymbol{\vartheta}_1 \times \boldsymbol{\vartheta}_4$, where $\boldsymbol{\vartheta}_1 = \{5, 10\}$, $\boldsymbol{\vartheta}_2 = \{4, 5\}$, $\boldsymbol{\vartheta}_3 = \{0.05, 0.2, 0.35, 0.5\}$, and $\boldsymbol{\vartheta}_4 = \{0.25, 0.5\}$. We have tried to use different P_n and $\boldsymbol{\Theta}$, but those given here usually lead to better performance.

4.2 Performance Evaluation

As mentioned previously, in addition to Twohit, this section also considers the performance of the doubly regularized approach of Daye et al. (2012) (denoted by HHR), Lasso, and adaptive Lasso. To implement HHR, we use the source code from `https://sites.google.com/site/zhongyindaye/software`, and follow the suggestion of Daye et al. (2012) to select tuning parameters therein by the Akaike information criterion (AIC). To perform Lasso (adaptive Lasso) under Model I or Model II, whose corresponding negative log-likelihood function is not jointly convex in α_i and β_j even when $\{\epsilon_t\}$ is a sequence of i.i.d. standard normal random variables, we use TwLasso (TwAdaLasso), which is Twohit with OGA+HDIC+Trim and $\text{OGA}_{\mathcal{D}}$+$\text{HDIC}_{\mathcal{D}}$+$\text{Trim}_{\mathcal{D}}$ replaced by Lasso (adaptive Lasso). Moreover, the Lasso and adaptive Lasso in TwLasso and TwAdaLasso are implemented through the `glmnet` and `parcor` packages in `R`, respectively.

The performance of the aforementioned methods is evaluated on Examples 1–3, each containing several data sets generated by Model I (Examples 1 and 2) or Model II (Example 3). The performance measures are given by the average (over 100 replications) true positive rates (ATPR), the average (over 100 replications) false positive rates (AFPR), `E` (the frequency, in 100 replications, of selecting exactly the relevant variables), and $\texttt{E}^*$ (the frequency, in 100 replications, of including all relevant variables). All sample sizes in these examples are 400.

Example 1

Let $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n$ be i.i.d. p_{L}-dimensional normal random vectors with zero mean and covariance matrix $\boldsymbol{\Sigma} = (h^{|i-j|})_{1 \le i,j \le p_{\text{L}}}$, where $h \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $p_{\text{L}} = 4000$. We also set $\beta_0 = 0.5$, $(\beta_1, \ldots, \beta_{15}) = (7.5, 7.5, 7.5, 0, 0, 0, 10, 10, 10, 0, 0, 0, 12.5, 12.5, 12.5)$, $\beta_{16} = \cdots = \beta_{4000} = 0$, $\boldsymbol{z}_t = \boldsymbol{x}_t$ (yielding $p_{\text{D}} = p_{\text{L}} = 4000$), $\alpha_0 = 0.1$, $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (0, 2.5, 0, 0, -2.5)$, and $\alpha_6 = \cdots = \alpha_{4000} = 0$. Moreover, $\{\epsilon_t\}$ is generated by one of the following time series models (TSM):

TSM 1: $\epsilon_t = w_t$, TSM 2: $\epsilon_t = 0.6\epsilon_{t-1} + w_t$,

TSM 3: $\epsilon_t = w_t + 0.6w_{t-1}$, TSM 4: $(1 - B)^{0.3}\epsilon_t = w_t$,

where B is the backshift operator and $\{w_t\}$ is a sequence of i.i.d. standard normal random variables. Under these specifications, the ATPR, AFPR, E, and E* of Twohit, HHR, TwLasso, and TwAdaLasso are summarized in Table 1.

Note that $\mathrm{Var}(\boldsymbol{\beta}^\top \boldsymbol{x}_t)/E(\sigma_t^2)$ =1.88, 2.84, 6.85, 38.8, and 827.7 when $h = 0.1, 0.3, 0.5, 0.7$, and 0.9, respectively. Therefore, a larger h leads to a larger signal-to-noise ratio, which may enhance the chance of identifying the relevant variables and excluding the irrelevant ones. As shown in Table 1, for each TSM $i, 1 \leq i \leq 4$, the ATPR (AFPR) of Twohit tends to increase (decrease) as h increases from 0.1 to 0.9. On the other hand, because $E(x_{ti}x_{tj}) = h^{|i-j|}$, a large h also introduces high correlations between the relevant variables and a few irrelevant ones. Once these irrelevant variables are included by OGA or OGA$_{\mathcal{D}}$, they are often not easily eliminated by Trim or Trim$_{\mathcal{D}}$, thereby worsening the performance of Twohit on E. Indeed, Table 1 reveals that although the E* of Twohit appears to increase with h, the lowest E of Twohit also occurs at the highest value of h.

When $h < 0.9$, TwAdaLasso is comparable with Twohit in terms of ATPR and E*, but is worse than the latter in terms of AFPR and E. When $h = 0.9$, TwAdaLasso is inferior to Twohit on all performance measures. For all values of h, TwLasso suffers from a small ATPR and a large AFPR compared to TwAdaLasso. The performance of TwLasso on E* and E is also much worse than TwAdaLasso. HHR works slightly better than TwLasso on ATPR and E* when $h = 0.1$. The method, however, faces a severe false positive problem. As a result, its AFPR values are extremely large and its E values are zero at $h = 0.1$. When h grows to 0.9, the ATPR of HHR substantially deteriorates and its AFPR remains large, yielding the smallest E* and E among all methods. Finally, we mention that there is no systematic change in the performance of the methods considered when $\{\epsilon_t\}$ varies from an i.i.d. process (TSM 1) to a short-memory process (TSM 2 or 3) to a long-memory process (TSM 4).

Example 2

Let $\{\boldsymbol{x}_t = (x_{t1}, \ldots, x_{t4000})^\top\}$ be a sequence of i.i.d. random vectors, where $x_{t1}, \ldots, x_{t12}$ are i.i.d. standard normal random variables and

$$x_{tj} = d_{tj} + \sum_{l=1}^{12} x_{tl},\ j = 13, \ldots, 4000,$$

Table 1: The ATPR, AFPR, E, and E* of Twohit, HHR, TwLasso, and TwAdaLasso in Example 1.

	$h = 0.1$															
	TSM 1				TSM 2				TSM 3				TSM 4			
Method	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*
Twohit	**0.969**	**1.73**e-4	**72**	**83**	0.941	**2.94**e-4	**64**	**82**	0.924	**1.99**e-4	**67**	**79**	**0.937**	**1.78**e-4	**60**	**78**
HHR	0.925	5.14e-2	0	53	0.925	5.13e-2	0	56	**0.940**	5.18e-2	0	63	0.934	5.16e-2	0	60
TwLasso	0.925	9.20e-3	0	52	0.922	8.84e-3	0	52	0.900	9.33e-3	0	55	0.911	9.02e-3	0	45
TwAdaLasso	0.965	4.09e-4	20	80	**0.950**	4.02e-4	17	79	0.925	3.35e-4	17	78	**0.937**	4.06e-4	14	72
	$h = 0.3$															
	TSM 1				TSM 2				TSM 3				TSM 4			
Method	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*
Twohit	0.956	**1.70**e-4	**71**	**85**	0.948	**2.19**e-4	**65**	**83**	0.949	**2.49**e-4	**55**	**79**	0.940	**1.58**e-4	**65**	79
HHR	0.890	4.97e-2	0	32	0.917	5.05e-2	0	50	0.913	5.05e-2	0	45	0.907	5.04e-2	0	39
TwLasso	0.941	7.86e-3	0	65	0.945	7.76e-3	0	61	0.940	7.69e-3	0	61	0.939	8.07e-3	0	62
TwAdaLasso	**0.963**	2.39e-4	29	**85**	**0.958**	3.18e-4	30	79	**0.961**	3.14e-4	21	**79**	**0.955**	4.12e-4	19	**82**
	$h = 0.5$															
	TSM 1				TSM 2				TSM 3				TSM 4			
Method	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*
Twohit	0.969	**9.01**e-5	**72**	**87**	0.960	**2.04**e-4	**64**	82	0.964	**8.89**e-5	**79**	92	0.959	**1.97**e-4	**59**	85
HHR	0.837	4.08e-2	0	6	0.845	4.39e-2	0	9	0.830	4.24e-2	0	6	0.836	4.32e-2	0	6
TwLasso	0.950	7.15e-3	0	69	0.965	7.85e-3	0	77	0.948	7.19e-3	0	72	0.952	6.86e-3	0	62
TwAdaLasso	**0.976**	2.72e-4	29	**87**	**0.975**	2.90e-4	35	**87**	**0.975**	2.69e-4	40	**93**	**0.976**	2.72e-4	28	**86**
	$h = 0.7$															
	TSM 1				TSM 2				TSM 3				TSM 4			
Method	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*
Twohit	0.992	**4.76**e-5	**87**	**98**	0.971	**1.29**e-4	**70**	89	0.992	**7.64**e-5	**73**	**96**	0.983	**1.18**e-4	**73**	92
HHR	0.820	1.42e-2	0	0	0.819	2.06e-2	0	0	0.819	1.80e-2	0	0	0.818	1.71e-2	0	0
TwLasso	0.983	6.54e-3	0	88	0.978	6.93e-3	0	84	0.975	5.87e-3	0	82	0.978	6.54e-3	0	83
TwAdaLasso	**0.993**	1.39e-4	60	96	**0.986**	1.84e-4	53	**94**	**0.995**	1.84e-4	38	**96**	**0.990**	1.97e-4	37	**93**
	$h = 0.9$															
	TSM 1				TSM 2				TSM 3				TSM 4			
Method	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*
Twohit	**0.993**	**1.04**e-4	**64**	**92**	**0.984**	**1.59**e-4	**46**	**90**	**0.980**	**1.69**e-4	**42**	**88**	**0.990**	**1.54**e-4	**45**	**90**
HHR	0.818	5.19e-3	0	0	0.818	4.98e-3	0	0	0.818	5.64e-3	0	0	0.818	5.21e-3	0	0
TwLasso	0.929	3.59e-3	0	51	0.961	4.55e-3	0	69	0.949	4.25e-3	0	65	0.944	4.88e-3	0	61
TwAdaLasso	0.955	3.53e-4	17	73	0.982	2.87e-4	21	84	0.967	3.15e-4	21	75	0.963	3.25e-4	16	75

with $(0.25)^{-1/2}(d_{t13}, \ldots, d_{t4000})^\top$ following a 3988-dimensional standard normal distribution. Set $\boldsymbol{z}_t = \boldsymbol{x}_t$, $\beta_0 = 0.2$, $(\beta_1, \ldots, \beta_{10}) = (3, 3.75, 4.5, 5.25, 6, 6.75, 7.5, 8.25, 9, 9.75)$, $\beta_{11} = \cdots = \beta_{4000} = 0$, $\alpha_0 = 0.1$, $\alpha_{11} = -1$, $\alpha_{12} = 1$, and $\alpha_j = 0$ if $j \notin \{0, 11, 12\}$. Moreover, $\{\epsilon_t\}$ is generated by TSM 1, 2, 4, or

TSM 5: $(1 - B)^{0.45}\epsilon_t = w_t$,

where w_t is defined as in Example 1. Note that the serial dependence of $\{\epsilon_t\}$ in TSM 5 is much stronger than in TSM 4, although both are long-memory processes. The performance of Twohit, HHR, TwLasso, and TwAdaLasso on ATPR, AFPR, E, and E* are presented in Table 2.

In this example, Lasso-type methods, HHR, TwLasso, and TwAdaLasso, may encounter the intrinsic difficulty that the irrepresentable condition (Zhao and Yu, 2006) fails to hold; see Example 3 of Ing and Lai (2011) for further discussion. As observed in Table 2, the ATPR of TwLasso is around 0.92, which is much larger than that of HHR and TwAdaLasso, but much smaller than that of Twohit. The performance of TwAdaLasso in terms of AFPR is much better than TwLasso and HHR, but is obviously inferior to Twohit. Moreover, the E* and E values of the Lasso-type methods are all (close to) zero. In contrast, Twohit works perfectly on E* when $\{\epsilon_t\}$ is generated by TSM 1, 2 or 4, and the corresponding E value, ranging from 71 to 86, is also reasonably large. However, when $\{\epsilon_t\}$ is generated by TSM 5, Twohit becomes relatively unsatisfactory. This, together with the simulation results obtained in Example 1, suggests that the performance of Twohit deteriorates under a very strong serial dependence.

Table 2: The ATPR, AFPR, E, and E* of Twohit, HHR, TwLasso, and TwAdaLasso in Example 2.

	TSM 1				TSM 2				TSM 4				TSM 5			
Method	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*
Twohit	**1.000**	**0.91**e-4	**86**	**100**	**1.000**	**1.19**e-4	**75**	**100**	**1.000**	**1.38**e-4	**71**	**100**	**0.983**	**1.93**e-4	**43**	**81**
HHR	0.805	8.31e-3	0	1	0.732	6.62e-3	0	0	0.804	7.89e-3	0	1	0.635	7.38e-3	0	1
TwLasso	0.916	1.05e-2	0	0	0.917	8.56e-3	0	0	0.917	9.79e-3	0	0	0.916	8.46e-3	0	0
TwAdaLasso	0.793	1.06e-3	0	0	0.844	1.07e-3	0	0	0.815	1.10e-3	0	0	0.824	1.05e-3	0	0

Example 3

Let $\boldsymbol{z}_t$ be the same as that in Example 2 and $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{t4050})^\top$, where $x_{tj} = z_{tj}$ for $j = 1, \ldots, 4000$ and $x_{t,4000+j} = y_{t-j}$ for $j = 1, \ldots, 50$. In addition, y_t is set to zero for $t \leq 0$. We generate $\{\epsilon_t\}$ according to the following GARCH$(1,1)$ model,

$$\epsilon_t = \nu_t w_t,$$

where $\nu_t^2 = 1 + 0.5\epsilon_{t-1}^2 + 0.3\nu_{t-1}^2$ and $\{w_t\}$ is a sequence of i.i.d. standard normal random variables. The coefficients $\beta_0, \ldots, \beta_{4000}$ and $\alpha_0, \ldots, \alpha_{4000}$ are also the same as those in Example 2. On the other hand, we consider three different cases for $(\beta_{4001}, \ldots, \beta_{4050})$ (the AR coefficients corresponding to $(y_{t-1}, \ldots, y_{t-50})$):

Case 1: $(\beta_{4001}, \beta_{4002}) = (0.5, 0.25)$, and $\beta_j = 0$ if $j = 4003, \ldots, 4050$.

Case 2: $(\beta_{4001}, \ldots, \beta_{4003}) = (0.5, 0.25, 0.125)$, and $\beta_j = 0$ if $j = 4004, \ldots, 4050$.

Case 3: $(\beta_{4001}, \ldots, \beta_{4004}) = (0.5, 0.25, 0.125, 0.0625)$, and $\beta_j = 0$ if $j = 4005, \ldots, 4050$.

The simulation results are reported in Table 3. Like Example 2, the irrepresentable condition does not hold in this example; hence Lasso-type methods (HHR, TwLasso, and TwAdaLasso) do not work well on all performance measures. However, Twohit performs quite satisfactorily, in particular in terms of E. We also notice that the larger the smallest non-zero AR coefficient, the higher the value of Twohit on E.

Table 3: The ATPR, AFPR, E, and E* of Twohit, HHR, TwLasso, and TwAdaLasso in Example 3.

	Case 1				Case 2				Case 3			
Method	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*	ATPR	AFPR	E	E*
Twohit	**0.985**	**2.12**e-5	**94**	**95**	**0.983**	**3.48**e-5	**89**	**93**	**0.971**	**3.73**e-5	**88**	**90**
HHR	0.389	1.53e-2	0	5	0.209	2.82e-3	0	0	0.205	2.54e-3	0	0
TwLasso	0.924	7.80e-3	0	2	0.923	7.47e-3	0	0	0.909	7.09e-3	0	0
TwAdaLasso	0.873	1.06e-3	0	0	0.841	1.06e-3	0	0	0.801	1.15e-3	0	0

5 Real Data Analysis

In this section, we use Twohit to choose important explanatory variables for the monthly growth rate of the semiconductor world market billings (WMB) from among a large number of candidate variables. We collect the monthly semiconductor WMB, denoted by WMB_t, and other 66 monthly variables (suggested by Liu and Weng (2018) as candidate predictor variables) from the World Semiconductor Trade Statistics; see Tables S2.1 and S2.2 in the supplement for the description of these variables. Following Liu and Weng (2018), we define the monthly growth rate of the semiconductor WMB by

$$y_t = \log(\text{WMB}_t) - \log(\text{WMB}_{t-12}), \tag{36}$$

and take a two-step transformation procedure for the candidate variables (except for CLI, CS, UTL, ISR, Bill, PPI, and PPI3) in order to obtain (possibly) stationary series.

We begin by analyzing y_t based on Model II, in which the dispersion variables, z_{tj}, are the aforementioned 66 variables (after the two-step transformation) and their lagged values (up to 24 months) and the regression variables, x_{tj}, contain all z_{tj} and $y_{t-1},\ldots,$ y_{t-24}, yielding $n = 240$, $p_{\text{L}} = 1608$, and $p_{\text{D}} = 1584$. We use Twohit to select regression and dispersion variables and obtain the estimates of ϵ_t, denoted by $\hat{\epsilon}_t(\text{II})$, based on the selected model. The time plot, ACF plot, and partial ACF (PACF) plot of $\hat{\epsilon}_t(\text{II})$, along with the ACF and PACF plots of $\hat{\epsilon}_t^2(\text{II})$, are given in Fig. 2. These plots suggest that there is no serial correlation and ARCH/GARCH effect in $\hat{\epsilon}_t(\text{II})$. The estimated coefficients and their standard errors are presented in Table S2.3.

We next analyze $\{y_t\}$ based on Model I, in which z_{tj} are the same as those of Model II and $x_{tj} = z_{tj}$ for all j. Therefore, $p_{\text{L}} = p_{\text{D}} = 1584 > n = 240$. We select regression and dispersion coefficients using Twohit, and obtain the estimates of ϵ_t, $\hat{\epsilon}_t(\text{I})$. According to the ACF and PACF plots of $\hat{\epsilon}_t(\text{I})$, we postulate an AR(2) model for $\{\epsilon_t\}$,

$$(1 - \phi_1 B - \phi_2 B^2)\epsilon_t = w_t,$$

and then simultaneously estimate ϕ_1, ϕ_2, and the selected regression and dispersion coefficients. These estimates and their standard errors are reported in Table S2.4. Moreover, the time plot, ACF plot, and PACF plot of estimates of w_t, $\hat{w}_t$, are given in Fig. 3, showing that there is no serial correlation in $\hat{w}_t$.

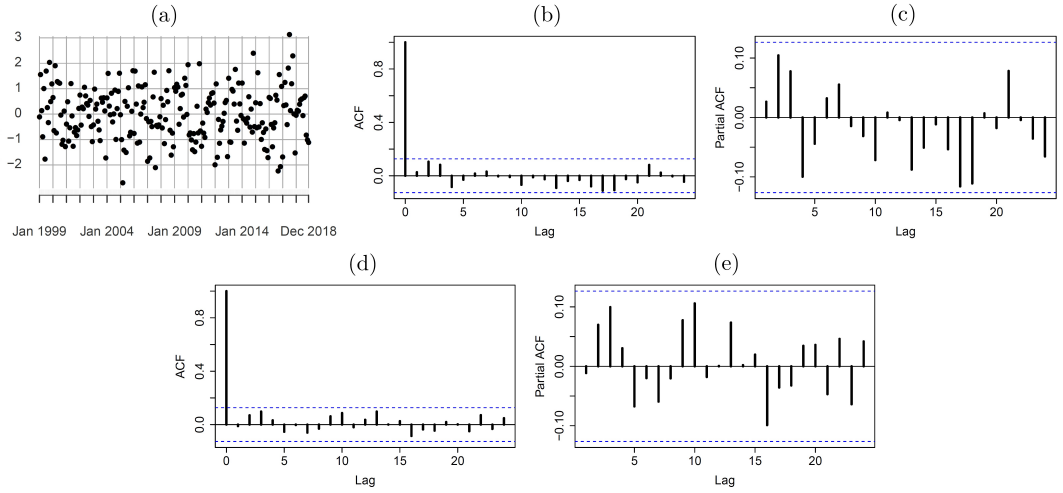

Fig. 2. Residual analysis for Model II: (a) the time plot of $\hat{\epsilon}_t$(II), (b) the ACF plot of $\hat{\epsilon}_t$(II), (c) the PACF plot of $\hat{\epsilon}_t$(II), (d) the ACF plot of $\hat{\epsilon}_t^2$(II), (e) the ACF plot of $\hat{\epsilon}_t^2$(II), and the blue dashed lines in (b)–(e) are two-standard-deviation limits.

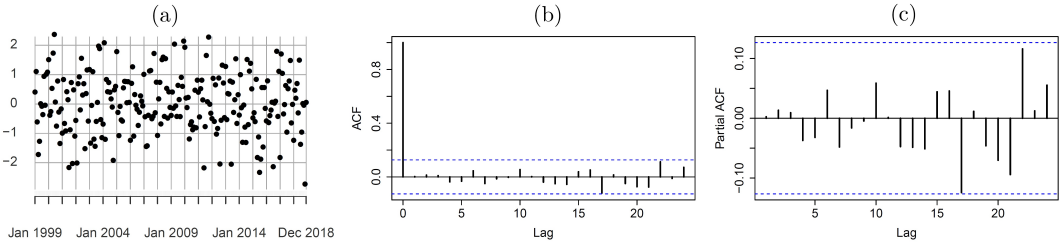

Fig. 3. Residual analysis for Model I: (a) the time plot of $\hat{w}_t$, (b) the ACF plot of $\hat{w}_t$, (c) the PACF plot of $\hat{w}_t$, and the blue dashed lines in (b) and (c) are two-standard-deviation limits.

Although Fig. 2 and Fig. 3 indicate that the two models built upon Models I and II are adequate, it is shown in Tables S2.3 and S2.4 that the model built upon Model II has the smaller AIC and BIC values, and hence seems to be more recommendable than the model built upon Model I. It is worth pointing out that the former model does not include the three most crucial determinants (NO, TI, and UTL) of the semiconductor industry cycles suggested by industry practitioners (see Table 2 in Liu (2005)), whereas its competitor does. One possible explanation of this phenomenon is that y_{t-1} is included by the former model at the first OGA iteration owing to its high correlation with the response variable y_t. Once y_{t-1} is chosen by OGA, it is difficult for the aforementioned crucial determinants to enter the regression equation because they are statistically confounded with y_{t-1}.

6 Concluding Remarks

This paper has addressed the important problem of selecting high-dimensional regression models with heteroscedastic and serially correlated errors. When the serial correlation or heteroscedasticity does not exist in the error terms, this type of problem has been undertaken in the past; see, e.g., Belloni et al. (2014), Basu and Michailidis (2015), Wu and Wu (2016), Gu and Zou (2016), Han and Tsay (2019), and Hsu et al. (2019). Their results, however, are not directly applicable to situations where both heteroscedasticity and serial correlation occur. We fill this gap by proposing a two-part selection procedure, Twohit, and proving its consistency in selecting regression and dispersion variables in situations where the model error contains a short-memory, long-memory, or conditionally heteroscedastic component. We also show that Twohit works well in finite samples compared to other competing methods.

While Twohit focuses on the dispersion function $\sigma_t^2 = \exp\{\alpha_0 + \sum_{j=1}^{p_{\mathrm{D}}} \alpha_j z_{tj}\}$, it is possible to extend the consistency result of Twohit to another popular dispersion function $\sigma_t^{*^2} = (\alpha_0 + \sum_{j=1}^{p_{\mathrm{D}}} \alpha_j z_{tj})^2$, which has been considered by many authors either with fixed p_{D} or with $p_{\mathrm{D}} \gg n$; see, e.g., Efron (1991), Koenker and Bassett (1982), Koenker and Zhao (1994), and Gu and Zou (2016). Let $\eta_t^* = \sigma_t^* \epsilon_t$. Then, it holds that

$$\eta_t^{*^2} = \left(\alpha_0 + \sum_{j=1}^{p_{\mathrm{D}}} \alpha_j z_{tj} \right)^2 + \sigma_t^{*^2} (\epsilon_t^2 - 1), \tag{37}$$

where, without loss of generality, we assume $E(\epsilon_t^2) = 1$. When $p_{\rm D} \gg n$, (37) is a high-dimensional interaction model with heteroscedastic and time series error $\sigma_t^{*^2}(\epsilon_t^2 - 1)$. Since the interaction model obeys the strong heredity condition (see Hao and Zhang, 2014), it is expected that Twohit (with a suitable modification for OGA$_{\mathcal{D}}$+HDIC$_{\mathcal{D}}$+Trim$_{\mathcal{D}}$) can still achieve selection consistency when σ_t^2 (or (16)) is replaced by $\sigma_t^{*^2}$ (or (37)). However, this extension requires a detailed study and is left for future work.

Acknowledgments

We thank comments by Ruey Tsay and other participants at 2018 Workshop on High Dimensional Statistical Analysis. We also thank two anonymous referees for insightful and constructive comments. Chiou's and Ing's research was supported in part by the Science Vanguard Research Program of the Ministry of Science and Technology (MOST) of Taiwan, and Guo's research was supported by the MOST of Taiwan under Grant MOST 106-2118-M-110-003-MY2.

Appendix A Proofs of Theorems 1 and 2

To prove Theorems 1 and 2, we need some supporting lemmas. Recall that we assume that $\beta_0 = 0$, $E(\boldsymbol{x}_t) = 0$ and $E(x_{tj}^2) = 1$, and set $\bar{y}$ and $\bar{x}_j$ in (10) to 0. Throughout this appendix, C stands for a generic positive constant independent of n.

Lemma A.1 *Assume that either*

(i) (5), (6) *or* (7), *and* (A1)(a), *or*

(ii) (8), (9), *and* (A1)(b)(i)

holds. Then

$$\max_{1\le i\le p_{\rm L}} \left| n^{-1} \sum_{t=1}^{n} \eta_t x_{ti} \right| = O_p(n^{-1/2+d} p_{\rm L}^{1/q_1}),$$

recalling that $0 < d < 1/2$ *is defined in* (7) *and* $d = 0$ *if* (6) *or* (8) *is assumed.*

PROOF. It suffices to show that

$$\max_{1\le i\le p_{\rm L}} E\left(\left| n^{-1/2-d} \sum_{t=1}^{n} \eta_t x_{ti} \right| \right) = O(1). \tag{A.1}$$

We first prove (A.1) under (i). Let $\mathbf{\Gamma}_{\epsilon,n} = (\gamma_\epsilon(i-j))_{1\le i,j\le n}$, where $\gamma_\epsilon(s) = E(\epsilon_1\epsilon_{1+s})$. In view of (7), $\gamma_\epsilon(s) = \sum_{j=0}^{\infty}\psi_j\psi_{j+|s|} = O(|s+1|^{-1+2d})$, which, together with (6), yields $\sum_{j=0}^{n}|\gamma_\epsilon(j)| = O(n^{2d})$. Since for any n-dimensional unit vector $\boldsymbol{a}$, $\boldsymbol{a}^\top\mathbf{\Gamma}_{\epsilon,n}\boldsymbol{a} \le C\sum_{j=0}^{n}|\gamma_\epsilon(j)|$, it holds that

$$\|\mathbf{\Gamma}_{\epsilon,n}\|_2 \le \sup_{\|\boldsymbol{a}\|_2=1}\boldsymbol{a}^\top\mathbf{\Gamma}_{\epsilon,n}\boldsymbol{a} = O(n^{2d}). \tag{A.2}$$

By (A.2), (A1)(a), independence of $\{w_s\}$ and $\{(\boldsymbol{x}_s,\boldsymbol{z}_s)\}$, Lemma 2 of Wei (1987), Jensen's inequality, and Cauchy-Schwarz inequality, one has for all $1\le i\le p_{\rm L}$,

$$\begin{aligned}
&E\left|n^{-1/2-d}\sum_{t=1}^{n}\eta_t x_{ti}\right|^{q_1}\\
&= n^{-q_1(1/2+d)}E\left\{E\left(\Big|\sum_{s=-\infty}^{n}\Big(\sum_{t=\max\{1,s\}}^{n}\sigma_t x_{ti}\psi_{t-s}\Big)w_s\Big|^{q_1}\,\Big|\,\boldsymbol{x}_s,\boldsymbol{z}_s,-\infty<s<\infty\right)\right\}\\
&\le Cn^{-q_1(1/2+d)}E\left(\sum_{s=-\infty}^{n}\left(\sum_{t=\max\{1,s\}}^{n}\sigma_t x_{ti}\psi_{t-s}\right)^2\right)^{q_1/2}\\
&= Cn^{-q_1(1/2+d)}E\left|\sum_{k=1}^{n}\sum_{l=1}^{n}\sigma_k x_{ki}\sigma_l x_{li}\sum_{s=-\infty}^{\min\{k,l\}}\psi_{k-s}\psi_{l-s}\right|^{q_1/2}\\
&\le Cn^{-q_1 d}\|\mathbf{\Gamma}_{\epsilon,n}\|_2^{q_1/2}E\left|n^{-1}\sum_{k=1}^{n}\sigma_k^2x_{ki}^2\right|^{q_1/2}\\
&\le Cn^{-1}\sum_{k=1}^{n}E|\sigma_k x_{ki}|^{q_1} \le C\left(\max_{1\le t\le n}E|\sigma_t|^{2q_1}\right)^{1/2}\left(\max_{1\le t\le n,1\le i\le p_{\rm L}}E|x_{ti}|^{2q_1}\right)^{1/2},
\end{aligned}$$

yielding (A.1). We next show that (A.1) holds true under (ii). Note first that for each $i=1,\ldots,p_{\rm L}$, $\{\sigma_t\epsilon_t x_{ti},\mathcal{F}_t\}$ is a martingale difference sequence. By (A1)(b)(i), Burkholder's inequality, Jensen's inequality, and Hölder's inequality, one has for all $1\le i\le p_{\rm L}$,

$$\begin{aligned}
E\left|n^{-1/2}\sum_{t=1}^{n}\eta_t x_{ti}\right|^{q_1} &= n^{-q_1/2}E\left|\sum_{t=1}^{n}\sigma_t\epsilon_t x_{ti}\right|^{q_1}\\
&\le CE\left(n^{-1}\sum_{t=1}^{n}(\sigma_t\epsilon_t x_{ti})^2\right)^{q_1/2}\\
&\le Cn^{-1}\sum_{t=1}^{n}E|\sigma_t\epsilon_t x_{ti}|^{q_1}\\
&\le C\left(\max_{1\le t\le n}E|\epsilon_t|^{3q_1}\right)^{1/3}\left(\max_{1\le t\le n}E|\sigma_t|^{3q_1}\right)^{1/3}\left(\max_{1\le t\le n,1\le i\le p_{\rm L}}E|x_{ti}|^{3q_1}\right)^{1/3},
\end{aligned}$$

and hence (A.1) follows. □

Lemma A.2 *Assume that* (A2) *holds. Then*

$$\max_{1\le i,j\le p_{\mathrm{L}}}\left|n^{-1}\sum_{t=1}^{n}x_{ti}x_{tj}-E(x_{ti}x_{tj})\right|=O_p(n^{-1/2}p_{\mathrm{L}}^{1/q_2}).$$

PROOF. The proof of this lemma is elementary and is therefore omitted. □

Lemma A.3 *Under* (A2) *and* (20)*,*

$$\max_{1\le\sharp(J)\le K_{\mathrm{L},n}}\|\hat{\boldsymbol{\Gamma}}_{\mathrm{L},n}(J)-\boldsymbol{\Gamma}_{\mathrm{L}}(J)\|_2=O_p(K_{\mathrm{L},n}n^{-1/2}p_{\mathrm{L}}^{1/q_2}),\tag{A.3}$$

where $\hat{\boldsymbol{\Gamma}}_{\mathrm{L},n}(J)=n^{-1}\sum_{t=1}^{n}\boldsymbol{x}_t(J)\boldsymbol{x}_t^{\top}(J)$*. Moreover, if* $K_{\mathrm{L},n}=o(n^{1/2}p_{\mathrm{L}}^{-1/q_2})$*, then*

$$\max_{1\le\sharp(J)\le K_{\mathrm{L},,n}}\|\hat{\boldsymbol{\Gamma}}_{\mathrm{L},n}^{-1}(J)-\boldsymbol{\Gamma}_{\mathrm{L}}^{-1}(J)\|_2=o_p(1).\tag{A.4}$$

PROOF. Since

$$\max_{1\le\sharp(J)\le K_{\mathrm{L},n}}\|\hat{\boldsymbol{\Gamma}}_{\mathrm{L},n}(J)-\boldsymbol{\Gamma}_{\mathrm{L}}(J)\|_2\le K_{\mathrm{L},n}\max_{1\le i,j\le p_{\mathrm{L}}}\left|n^{-1}\sum_{t=1}^{n}x_{ti}x_{tj}-E(x_{ti}x_{tj})\right|,$$

(A.3) follows directly from Lemma A.2. Moreover, (A.4) is ensured by (20), (A.3), $K_{\mathrm{L},n}=o(n^{1/2}p_{\mathrm{L}}^{-1/q_2})$,

$$\|\hat{\boldsymbol{\Gamma}}_{\mathrm{L},n}^{-1}(J)-\boldsymbol{\Gamma}_{\mathrm{L}}^{-1}(J)\|_2\le(\|\hat{\boldsymbol{\Gamma}}_{\mathrm{L},n}^{-1}(J)-\boldsymbol{\Gamma}_{\mathrm{L}}^{-1}(J)\|_2+\|\boldsymbol{\Gamma}_{\mathrm{L}}^{-1}(J)\|_2)\|\hat{\boldsymbol{\Gamma}}_{\mathrm{L},n}(J)-\boldsymbol{\Gamma}_{\mathrm{L}}(J)\|_2\|\boldsymbol{\Gamma}_{\mathrm{L}}^{-1}(J)\|_2,$$

and some algebraic manipulations. □

Proof of Theorem 1. Note first that

$$E_n[(y(\boldsymbol{x})-\hat{y}_{\hat{J}_{\mathrm{L},m}}(\boldsymbol{x}))^2]=E_n[(y(\boldsymbol{x})-y_{\hat{J}_{\mathrm{L},m}}(\boldsymbol{x}))^2]+E_n[(\hat{y}_{\hat{J}_{\mathrm{L},m}}(\boldsymbol{x})-y_{\hat{J}_{\mathrm{L},m}}(\boldsymbol{x}))^2].$$

Therefore, (27) is ensured by

$$\max_{1\le m\le K_{\mathrm{L},n}}\frac{E_n[(y(\boldsymbol{x})-y_{\hat{J}_{\mathrm{L},m}}(\boldsymbol{x}))^2]}{m^{-1}}=O_p(1),\tag{A.5}$$

and

$$\max_{1\le m\le K_{\mathrm{L},n}}\frac{E_n[(\hat{y}_{\hat{J}_{\mathrm{L},m}}(\boldsymbol{x})-y_{\hat{J}_{\mathrm{L},m}}(\boldsymbol{x}))^2]}{mp_{\mathrm{L}}^{2/q}/n^{1-2d}}=O_p(1),\tag{A.6}$$

where $y_J(\boldsymbol{x}) = \boldsymbol{x}^\top(J)\boldsymbol{\Gamma}_{\mathrm{L}}^{-1}(J)E(\boldsymbol{x}(J)y(\boldsymbol{x}))$.

Let $(\hat{y}_{1;J},\ldots,\hat{y}_{n;J})^\top = \boldsymbol{H}_{\mathrm{L},J}\boldsymbol{y}$ and $(\hat{x}_{1i;J},\ldots,\hat{x}_{ni;J})^\top = \boldsymbol{H}_{\mathrm{L},J}\boldsymbol{X}_i$, $\hat{x}_{ti;J}^{\perp} = x_{ti} - \hat{x}_{ti;J}$, and $x_{i;J}^{\perp} = x_i - x_{i;J}$, where $\boldsymbol{H}_{\mathrm{L},J}$ is defined after (10) and $x_{i;J} = \boldsymbol{x}^\top(J)\boldsymbol{\Gamma}_{\mathrm{L}}^{-1}(J)\boldsymbol{g}_{\mathrm{L},i}(J)$. By (A1)–(A3), (A5), Lemmas A.1–A.3, $K_{\mathrm{L},n} = O(n^{1/2-d}p_{\mathrm{L}}^{-1/\underline{q}})$, and an argument similar to that used to prove of (3.8) of Ing and Lai (2011), we obtain

$$\max_{1\le i\le p_{\mathrm{L},n}}\left|n^{-1}\sum_{t=1}^{n}x_{ti}^2 - 1\right| = o_p(1),$$

$$\max_{1\le \sharp(J)\le K_{\mathrm{L},n}}\|\hat{\boldsymbol{\Gamma}}_{\mathrm{L},n}^{-1}(J)\|_2 = O_p(1), \tag{A.7}$$

$$\max_{\sharp(J)\le K_{\mathrm{L},n}-1, i\notin J}\left|n^{-1}\sum_{t=1}^{n}\eta_t\hat{x}_{ti;J}^{\perp}\right| = O_p(n^{-1/2+d}p_{\mathrm{L}}^{1/q_1}), \tag{A.8}$$

$$\max_{\sharp(J)\le K_{\mathrm{L},n}-1, i,j\notin J}\left|n^{-1}\sum_{t=1}^{n}x_{tj}\hat{x}_{ti;J}^{\perp} - E(x_j x_{i;J}^{\perp})\right| = O_p(n^{-1/2}p_{\mathrm{L}}^{1/q_2}), \tag{A.9}$$

which further imply

$$\max_{\{(J,i):\sharp(J)\le K_{\mathrm{L},n}-1, i\notin J\}}|\hat{\mu}_{\mathrm{L},J,i} - \mu_{\mathrm{L},J,i}| = O_p(n^{-1/2+d}p_{\mathrm{L}}^{1/\underline{q}}), \tag{A.10}$$

where $\mu_{\mathrm{L},J,i} = E[(y(\boldsymbol{x}) - y_J(\boldsymbol{x}))x_i]$ and

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

Equation (A.10) ensures that for any small $\varrho > 0$, there exists a large constant V_ϱ for which

$$P\left(\max_{\{(J,i):\sharp(J)\le K_{\mathrm{L},n}-1, i\notin J\}}|\hat{\mu}_{\mathrm{L},J,i} - \mu_{\mathrm{L},J,i}| > V_\varrho n^{-1/2+d}p_{\mathrm{L}}^{1/\underline{q}}\right) < \varrho. \tag{A.11}$$

For $1 \le m \le K_{\mathrm{L},n}$, define

$$A_{\mathrm{L},n}(m) = \left\{\max_{\{(J,i):\sharp(J)\le m-1, i\notin J\}}|\hat{\mu}_{\mathrm{L},J,i} - \mu_{\mathrm{L},J,i}| \le V_\varrho n^{-1/2+d}p_{\mathrm{L}}^{1/\underline{q}}\right\},$$

and

$$B_{\mathrm{L},n}(m) = \left\{\min_{0\le i\le m-1}\max_{1\le j\le p_{\mathrm{L}}}|\mu_{\mathrm{L},\hat{J}_{\mathrm{L},i},j}| > \tilde{\xi}V_\varrho n^{-1/2+d}p_{\mathrm{L}}^{1/\underline{q}}\right\},$$

where $2 < \tilde{\xi} < \infty$. By an argument similar to that used to prove (3.11) and (3.12) of Ing and Lai (2011), it holds that

$$E_n[\{y(\boldsymbol{x}) - y_{\hat{J}_{\mathrm{L},m}}(\boldsymbol{x})\}^2] \le C_{\tilde{\xi},V_\varrho}\left(\sum_{j=1}^{p_{\mathrm{L}}}|\beta_j|\right)^2 m^{-1} \quad \text{on} \quad A_{\mathrm{L},n}(m)\cap B_{\mathrm{L},n}(m), \tag{A.12}$$

and

$$E_n[\{y(\boldsymbol{x}) - y_{\hat{J}_{\mathrm{L},m}}(\boldsymbol{x})\}^2] \leq \left(\sum_{j=1}^{p_{\mathrm{L}}} |\beta_j|\right) \tilde{\xi} V_\varrho n^{-1/2+d} p_{\mathrm{L}}^{1/\underline{q}} \quad \text{on} \quad B_{\mathrm{L},n}^c(m), \tag{A.13}$$

where $C_{\tilde{\xi},V_\varrho}$ is some positive constant depending on $\tilde{\xi}$ and V_ϱ. Consequently, (A.5) follows from $A_{\mathrm{L},n}(m) \subseteq A_{\mathrm{L},n}(K_{\mathrm{L},n})$, (A.11)–(A.13), $K_{\mathrm{L},n} = O(n^{1/2-d} p_{\mathrm{L}}^{-1/\underline{q}})$, and (A4). Moreover, by (A4), (A5), (A.7), and Lemmas A.1–A.3, (A.6) also holds true. Thus, the proof is complete. □

To prove Theorem 2, we also need the following lemma.

Lemma A.4 *Assume that* Model I (Model II), (A1)(a) ((A1)(b)), (A2), (A3), (A5), (20), *and the first relation of* (28) *hold. Suppose* $K_{\mathrm{L},n} \asymp n^{1/2-d} p_{\mathrm{L}}^{-1/\underline{q}}$. *Then*

$$\max_{1 \leq \sharp(J) \leq K_{\mathrm{L},n}} \left| n^{-1} \boldsymbol{\eta}^\top (\boldsymbol{I} - \boldsymbol{H}_{\mathrm{L},J}) \boldsymbol{\eta} - n^{-1} \sum_{t=1}^n E(\eta_t^2) \right| = o_p(1), \tag{A.14}$$

where $\boldsymbol{\eta}^\top = (\eta_1, \ldots, \eta_n)$.

PROOF. The left-hand side of (A.14) is bounded above by

$$\left| n^{-1} \sum_{t=1}^n (\eta_t^2 - E(\eta_t^2)) \right| + \max_{1 \leq \sharp(J) \leq K_{\mathrm{L},n}} \|\hat{\boldsymbol{\Gamma}}_{\mathrm{L},n}^{-1}(J)\|_2 \left\| n^{-1} \sum_{t=1}^n \boldsymbol{x}_t(J) \eta_t \right\|_2^2. \tag{A.15}$$

By (A2), (A3), (A5), (A.7), Lemma A.1, and $K_{\mathrm{L},n} = O(n^{1/2-d} p_{\mathrm{L}}^{-1/\underline{q}})$, it follows that

$$\max_{1 \leq \sharp(J) \leq K_{\mathrm{L},n}} \|\hat{\boldsymbol{\Gamma}}_{\mathrm{L},n}^{-1}(J)\|_2 \left\| n^{-1} \sum_{t=1}^n \boldsymbol{x}_t(J) \eta_t \right\|_2^2 = o_p(1). \tag{A.16}$$

The first term of (A.15) is bounded above by

$$\left| n^{-1} \sum_{t=1}^n \sigma_t^2 (\epsilon_t^2 - E(\epsilon_t^2)) \right| + \left| n^{-1} \sum_{t=1}^n (\sigma_t^2 - E(\sigma_t^2)) E(\epsilon_t^2) \right|. \tag{A.17}$$

Since $\{\epsilon_t\}$ is independent of $\{\boldsymbol{z}_t\}$, it follows from (A1)(a) ((A1)(b)(i)) and Cauchy-Schwarz inequality that

$$\begin{aligned} E\left(n^{-1} \sum_{t=1}^n \sigma_t^2 (\epsilon_t^2 - E(\epsilon_t^2)) \right)^2 &= n^{-2} \sum_{t=1}^n \sum_{s=1}^n E(\sigma_t^2 \sigma_s^2) E[(\epsilon_t^2 - E(\epsilon_t^2))(\epsilon_s^2 - E(\epsilon_s^2))] \\ &\leq n^{-2} \sum_{t=1}^n \sum_{s=1}^n [E(\sigma_t^4) E(\sigma_s^4)]^{1/2} E[(\epsilon_t^2 - E(\epsilon_t^2))(\epsilon_s^2 - E(\epsilon_s^2))] \\ &\leq CE\left(n^{-1} \sum_{t=1}^n (\epsilon_t^2 - E(\epsilon_t^2)) \right)^2. \end{aligned} \tag{A.18}$$

For Model I, by (A1)(a) and the moment bounds for quadratic forms of linear processes (see Findley and Wei, 1993), one obtains

$$E\left(n^{-1}\sum_{t=1}^{n}(\epsilon_t^2 - E(\epsilon_t^2))\right)^2 \le Cn^{-2}E\left(\sum_{t=1}^{n}\sum_{s=1}^{n}\gamma_\epsilon^2(t-s)\right) = o(1). \tag{A.19}$$

For Model II, by (A1)(b)(i) and (A1)(b)(ii), we have

$$E\left(n^{-1}\sum_{t=1}^{n}(\epsilon_t^2 - E(\epsilon_t^2))\right)^2 = n^{-2}\sum_{t=-n+1}^{n-1}(n-|t|)\mathrm{Cov}(\epsilon_1^2, \epsilon_{1+|t|}^2) = o(1). \tag{A.20}$$

Moreover, the first relation of (28) implies

$$\left|\frac{1}{n}\sum_{t=1}^{n}(\sigma_t^2 - E(\sigma_t^2))E(\epsilon_t^2)\right| = o_p(1). \tag{A.21}$$

Consequently, (A.14) is ensured by (A.15)–(A.21). □

Proof of Theorem 2. It follows from Theorem 1, (A6), and an argument used in Theorem 3 of Ing and Lai (2011) that

$$\lim_{n\to\infty} P(\mathcal{D}_{\mathrm{L},n}) = 1, \tag{A.22}$$

where $\mathcal{D}_{\mathrm{L},n} = \{N_{\mathrm{L},n} \subseteq \hat{J}_{\mathrm{L},\lfloor an^{\gamma_{\mathrm{L}}}\rfloor}\}$ and a is a large constant. Define $\tilde{k}_{\mathrm{L},n} = \min\{k : 1 \le k \le K_{\mathrm{L},n}, N_{\mathrm{L},n} \subseteq \hat{J}_{\mathrm{L},k}\}$ and $K_{\mathrm{L},n}+1$ if $N_{\mathrm{L},n} - \hat{J}_{\mathrm{L},K_{\mathrm{L},n}} \ne \emptyset$. We first show that

$$\lim_{n\to\infty} P(\hat{k}_{\mathrm{L},n} = \tilde{k}_{\mathrm{L},n}) = 1, \tag{A.23}$$

which is guaranteed by

$$P(\hat{k}_{\mathrm{L},n} < \tilde{k}_{\mathrm{L},n}) = o(1), \tag{A.24}$$

and

$$P(\hat{k}_{\mathrm{L},n} > \tilde{k}_{\mathrm{L},n}) = o(1). \tag{A.25}$$

Straightforward calculations yield that

$$\begin{aligned}&\{\hat{k}_{\mathrm{L},n} < \tilde{k}_{\mathrm{L},n},\ \mathcal{D}_{\mathrm{L},n}\} \subseteq \mathcal{M}_n\\ &\equiv \left\{2\beta_{\hat{j}_{\mathrm{L},\tilde{k}_{\mathrm{L},n}}}\hat{U}_{\mathrm{L},2,n} - \lambda\varpi_{\mathrm{L},n}|\hat{U}_{\mathrm{L},3,n}| \le -\beta^2_{\hat{j}_{\mathrm{L},\tilde{k}_{\mathrm{L},n}}}\hat{U}_{\mathrm{L},1,n} + \lambda\varpi_{\mathrm{L},n}\left(n^{-1}\sum_{t=1}^{n}E(\eta_t^2)\right)\right\},\end{aligned} \tag{A.26}$$

where λ is some positive constant,

$$\begin{aligned}
\hat{U}_{\mathrm{L},1,n} &= n^{-1}\boldsymbol{X}^{\top}_{\hat{j}_{\mathrm{L},\tilde{k}_{\mathrm{L},n}}}(\boldsymbol{I}-\boldsymbol{H}_{\mathrm{L},\hat{J}_{\mathrm{L},\tilde{k}_{\mathrm{L},n}-1}})\boldsymbol{X}_{\hat{j}_{\mathrm{L},\tilde{k}_{\mathrm{L},n}}},\\
\hat{U}_{\mathrm{L},2,n} &= n^{-1}\boldsymbol{X}^{\top}_{\hat{j}_{\mathrm{L},\tilde{k}_{\mathrm{L},n}}}(\boldsymbol{I}-\boldsymbol{H}_{\mathrm{L},\hat{J}_{\mathrm{L},\tilde{k}_{\mathrm{L},n}-1}})\boldsymbol{\eta},\\
\hat{U}_{\mathrm{L},3,n} &= \hat{\sigma}^2_{\mathrm{L},\hat{J}_{\mathrm{L},\tilde{k}_{\mathrm{L},n}}}-n^{-1}\sum_{t=1}^{n}E(\eta_t^2),
\end{aligned}$$

and $\varpi_{\mathrm{L},n}=\lfloor an^{\gamma_{\mathrm{L}}}\rfloor G_{\mathrm{L}}(p_{\mathrm{L}},n)/n$. By (20) and (A.3),

$$\begin{aligned}
\hat{U}_{\mathrm{L},1,n}I_{\mathcal{D}_{\mathrm{L},n}} &\geq \lambda_{\min}(\hat{\boldsymbol{\Gamma}}_{\mathrm{L},n}(\hat{J}_{\mathrm{L},\lfloor an^{\gamma_{\mathrm{L}}}\rfloor}))\\
&\geq \lambda_{\min}(\boldsymbol{\Gamma}_{\mathrm{L}}(\hat{J}_{\mathrm{L},\lfloor an^{\gamma_{\mathrm{L}}}\rfloor}))-\|\hat{\boldsymbol{\Gamma}}_{\mathrm{L},n}(\hat{J}_{\mathrm{L},\lfloor an^{\gamma_{\mathrm{L}}}\rfloor})-\boldsymbol{\Gamma}_{\mathrm{L}}(\hat{J}_{\mathrm{L},\lfloor an^{\gamma_{\mathrm{L}}}\rfloor})\|_2\\
&\geq v_{\mathrm{L},n}+o_p(1),
\end{aligned}\tag{A.27}$$

where $v_{\mathrm{L},n}=\min_{1\leq\sharp(J)\leq\lfloor an^{\gamma_{\mathrm{L}}}\rfloor}\lambda_{\min}(\boldsymbol{\Gamma}_{\mathrm{L}}(J))>\delta_{\mathrm{L}}$ for all large n. By (A.8) and $K_{\mathrm{L},n}/n^{\gamma_{\mathrm{L}}}\to\infty$,

$$|\hat{U}_{\mathrm{L},2,n}|I_{\mathcal{D}_{\mathrm{L},n}}\leq\max_{\sharp(J)\leq\lfloor an^{\gamma_{\mathrm{L}}}\rfloor-1,i\notin J}\left|\frac{1}{n}\sum_{t=1}^{n}\eta_t\hat{x}^{\perp}_{ti;J}\right|=O_p(n^{-1/2+d}p_{\mathrm{L}}^{1/q_1})=o_p(n^{-\gamma_{\mathrm{L}}/2}).\tag{A.28}$$

Moreover, it follows from Lemma A.4 that

$$|\hat{U}_{\mathrm{L},3,n}|I_{\mathcal{D}_{\mathrm{L},n}}=o_p(1).\tag{A.29}$$

By making use of (28), (29), (A.22), (A.27)–(A.29), one obtained $P(\mathcal{M}_n)=o(1)$, which, together with (A.26), leads to (A.24).

To show (A.25), we note that by (28) and some algebraic manipulations,

$$\begin{aligned}
&\{\hat{k}_{\mathrm{L},n}>\tilde{k}_{\mathrm{L},n}\}\subseteq\mathcal{Q}_n\\
&\equiv\big\{2(\hat{k}_{\mathrm{L},n}-\tilde{k}_{\mathrm{L},n})(\hat{a}_{\mathrm{L},n}+\hat{b}_{\mathrm{L},n})+n\tilde{\varpi}_{\mathrm{L},n}|\hat{U}_{\mathrm{L},3,n}|\geq\delta_0 n\tilde{\varpi}_{\mathrm{L},n}\big\},
\end{aligned}\tag{A.30}$$

where δ_0 is some positive constant, $\tilde{\varpi}_{\mathrm{L},n}=1-\exp\{-n^{-1}(\hat{k}_{\mathrm{L},n}-\tilde{k}_{\mathrm{L},n})G_{\mathrm{L}}(p_{\mathrm{L}},n)\}$, and

$$\begin{aligned}
\hat{a}_{\mathrm{L},n} &= \|\hat{\boldsymbol{\Gamma}}^{-1}_{\mathrm{L},n}(\hat{J}_{\mathrm{L},K_{\mathrm{L},n}})\|_2\max_{1\leq j\leq p_{\mathrm{L}}}\left(n^{-1/2}\sum_{t=1}^{n}\eta_t x_{tj}\right)^2,\\
\hat{b}_{\mathrm{L},n} &= \|\hat{\boldsymbol{\Gamma}}^{-1}_{\mathrm{L},n}(\hat{J}_{\mathrm{L},K_{\mathrm{L},n}})\|_2\max_{1\leq\sharp(J)\leq\tilde{k}_{\mathrm{L},n},i\notin J}\left(n^{-1/2}\sum_{t=1}^{n}\eta_t\hat{x}_{ti;J}\right)^2.
\end{aligned}$$

Since $\hat{k}_{\mathrm{L},n}\leq K_{\mathrm{L},n}=O(n^{1/2-d}p_{\mathrm{L}}^{-1/\underline{q}})$, there exists $\lambda>0$ such that

$$(n\tilde{\varpi}_{\mathrm{L},n})/(\hat{k}_{\mathrm{L},n}-\tilde{k}_{\mathrm{L},n})\geq\lambda\min\{n^{1/2+d}p_{\mathrm{L}}^{1/\underline{q}},G_{\mathrm{L}}(p_{\mathrm{L}},n)\}\quad\text{on }\{\hat{k}_{\mathrm{L},n}>\tilde{k}_{\mathrm{L},n}\}.\tag{A.31}$$

Moreover, straightforward calculations give $\hat{a}_{\mathrm{L},n} + \hat{b}_{\mathrm{L},n} = O_p(n^{2d} p_{\mathrm{L}}^{2/q})$, which, in conjunction with (A3), (A.22), and (A.29)–(A.31), implies $P(\mathcal{Q}_n) = o(1)$. In view of this and (A.30), (A.25) follows. Thus, the proof of (A.23) is complete.

For $\tilde{k}_{\mathrm{L},n} > 1$, define

$$\tilde{\delta}_{\mathrm{L},l} = \begin{cases} 1, & \text{if } \mathrm{HDIC}(\hat{J}_{\mathrm{L},\tilde{k}_{\mathrm{L},n}} - \{\hat{j}_{\mathrm{L},l}\}) > \mathrm{HDIC}(\hat{J}_{\mathrm{L},\tilde{k}_{\mathrm{L},n}}); \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} P(\hat{N}_{\mathrm{L},n} \neq N_{\mathrm{L},n}) \quad \leq \quad & P(\tilde{\mathcal{G}}_{\mathrm{L},n}) + P(\tilde{\mathcal{H}}_{\mathrm{L},n}) + P(\hat{k}_{\mathrm{L},n} \neq \tilde{k}_{\mathrm{L},n}) \\ & + P(\hat{N}_{\mathrm{L},n} \neq N_{\mathrm{L},n}, \hat{k}_{\mathrm{L},n} = 1) + P(N_{\mathrm{L},n} \nsubseteq \hat{J}_{\mathrm{L},\hat{k}_{\mathrm{L},n}}), \end{aligned} \tag{A.32}$$

where $\tilde{\mathcal{G}}_{\mathrm{L},n} = \{\tilde{\delta}_{\mathrm{L},l} = 0 \text{ and } \beta_{\hat{j}_{\mathrm{L},l}} \neq 0 \text{ for some } 1 \leq l \leq \tilde{k}_{\mathrm{L},n}, \tilde{k}_{\mathrm{L},n} = \hat{k}_{\mathrm{L},n} > 1, N_{\mathrm{L},n} \subseteq \hat{J}_{\mathrm{L},\hat{k}_{\mathrm{L},n}}\}$ and $\tilde{\mathcal{H}}_{\mathrm{L},n} = \{\tilde{\delta}_{\mathrm{L},l} = 1 \text{ and } \beta_{\hat{j}_{\mathrm{L},l}} = 0 \text{ for some } 1 \leq l \leq \tilde{k}_{\mathrm{L},n}, \tilde{k}_{\mathrm{L},n} = \hat{k}_{\mathrm{L},n} > 1, N_{\mathrm{L},n} \subseteq \hat{J}_{\mathrm{L},\hat{k}_{\mathrm{L},n}}\}$. By an argument similar to that used to prove (A.23), it holds that

$$P(\tilde{\mathcal{G}}_{\mathrm{L},n}) \leq P(\mathcal{G}_{\mathrm{L},n}) = o(1) \quad \text{and} \quad P(\tilde{\mathcal{H}}_{\mathrm{L},n}) \leq P(\mathcal{H}_{\mathrm{L},n}) = o(1), \tag{A.33}$$

where with ζ denoting any positive constant and $\boldsymbol{X}_J = (\boldsymbol{X}_i, i \in J)$,

$$\begin{aligned} \mathcal{G}_{\mathrm{L},n} &= \{\tilde{U}_{\mathrm{L},1,n} < v_{\mathrm{L},n}/2\} \cup \{|\tilde{U}_{\mathrm{L},2,n}| > \zeta n^{-\gamma_{\mathrm{L}}/2}\} \cup \{|\hat{U}_{\mathrm{L},3,n}| > \zeta\}, \\ \mathcal{H}_{\mathrm{L},n} &= \{|\hat{U}_{\mathrm{L},3,n}| > \zeta\} \cup \{(\tilde{a}_{\mathrm{L},n} + \tilde{b}_{\mathrm{L},n}) \geq \zeta G_{\mathrm{L}}(p_{\mathrm{L}}, n)\}, \\ \tilde{U}_{\mathrm{L},1,n} &= n^{-1} \boldsymbol{X}_{\hat{j}_{\mathrm{L},l}}^{\top} (\boldsymbol{I} - \boldsymbol{H}_{\mathrm{L},\hat{J}_{\mathrm{L},\tilde{k}_{\mathrm{L},n}} - \{\hat{j}_{\mathrm{L},l}\}}) \boldsymbol{X}_{\hat{j}_{\mathrm{L},l}}, \\ \tilde{U}_{\mathrm{L},2,n} &= n^{-1} \boldsymbol{X}_{\hat{j}_{\mathrm{L},l}}^{\top} (\boldsymbol{I} - \boldsymbol{H}_{\mathrm{L},\hat{J}_{\mathrm{L},\tilde{k}_{\mathrm{L},n}} - \{\hat{j}_{\mathrm{L},l}\}}) \boldsymbol{\eta}, \\ \tilde{a}_{\mathrm{L},n} &= \|\hat{\boldsymbol{\Gamma}}_{\mathrm{L},n}^{-1}(\hat{J}_{\mathrm{L},\tilde{k}_{\mathrm{L},n}})\|_2 \max_{1 \leq j \leq p_{\mathrm{L}}} \left(n^{-1/2} \sum_{t=1}^{n} \eta_t x_{tj} \right)^2, \\ \tilde{b}_{\mathrm{L},n} &= \|\hat{\boldsymbol{\Gamma}}_{\mathrm{L},n}^{-1}(\hat{J}_{\mathrm{L},\tilde{k}_{\mathrm{L},n}})\|_2 \max_{1 \leq \sharp(J) \leq \tilde{k}_{\mathrm{L},n} - 1, i \notin J} \left(n^{-1/2} \sum_{t=1}^{n} \eta_t \hat{x}_{ti;J} \right)^2. \end{aligned}$$

In addition, (A.23) gives

$$P(\hat{k}_{\mathrm{L},n} \neq \tilde{k}_{\mathrm{L},n}) + P(\hat{N}_{\mathrm{L},n} \neq N_{\mathrm{L},n}, \hat{k}_{\mathrm{L},n} = 1) + P(N_{\mathrm{L},n} \nsubseteq \hat{J}_{\mathrm{L},\hat{k}_{\mathrm{L},n}}) = o(1). \tag{A.34}$$

By (A.32)–(A.34), (12) follows. □

Appendix B Supplementary Appendix

The supplementary material contains the proof of Theorem 3 and some auxiliary lemmas used to prove it.

References

Basu, S., Michailidis, G., 2015. Regularized estimation in sparse high-dimensional time series models. The Annals of Statistics 43, 1535–1567.

Belloni, A., Chernozhukov, V., Wang, L., 2014. Pivotal estimation via square-root Lasso in nonparametric regression. The Annals of Statistics 42, 757–788.

Box, G.E.P., Tiao, G.C., 1975. Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association 70, 70–79.

Chien, C.-F., Chen, Y.-J., Wu, J.-Z., 2016. Big data analytics for modeling WAT parameter variation induced by process tool in semiconductor manufacturing and empirical study. In *Proceedings of the 2016 Winter Simulation Conference*, Piscataway, NJ, USA: IEEE Press: 2512–2522.

Daye, Z. J., Chen, J., Li, H., 2012. High-dimensional heteroscedastic regression with an application to eQTL data analysis. Biometrics 68, 316–326.

Efron, B., 1991. Regression percentiles using asymmetric squared error loss. Statistica Sinica 1, 93–125.

Findley, D.F., Wei, C.-Z., 1993. Moment bounds for deriving time series CLT's and model selection procedures. Statistica Sinica 3, 453–480.

Gao, Z., Ling, S., 2019. Statistical inference for structurally changed threshold autoregressive models. Statistica Sinica, to appear.

Giraitis, L., Kokoszka, P., Leipus, R., 2000. Stationary ARCH models: dependence structure and central limit theorem. Econometric Theory 16, 3–22.

Gu, Y., Zou, H., 2016. High-dimensional generalizations of asymmetric least squares regression and their applications. The Annals of Statistics 44, 2661–2694.

Han, Y., Tsay, R.S., 2019. High-dimensional Linear Regression for Dependent Data with Applications to Nowcasting. Statistica Sinica, to appear.

Hao, N., Zhang, H.H., 2014. Interaction screening for ultra-high dimensional data. Journal of the American Statistical Association 109, 1285–1301.

Hsu, H.-L., Ing, C.-K., Tong, H., 2019. On model selection from a finite family of possibly misspecified time series models. The Annals of Statistics 47, 1061–1087.

Ing, C.-K., 2019. Model selection for high-dimensional linear regression with dependent observations. The Annals of Statistics, to appear.

Ing, C.-K, Lai, T.L., 2011. A stepwise regression method and consistent model selection for high-dimensional sparse linear models. Statistica Sinica 21, 1473–1513.

Ing, C.-K., Wei, C.-Z., 2006. A maximal moment inequality for long range dependent time series with applications to estimation and model selection. Statistica Sinica 16, 721–740.

Koenker, R., Bassett, G., 1982. Robust tests for heteroscedasticity based on regression quantiles. Econometrica 50, 43–61.

Koenker, R., Zhao, Q., 1994. L-estimation for linear heteroscedastic models. Journal of Nonparametric Statistics 3, 223–235.

Liu, W.-H., 2005. Determinants of the semiconductor industry cycles. Journal of Policy Modeling 27, 853–866.

Liu, W.-H., Weng, S.-S., 2018. On predicting the semiconductor industry cycle: a Bayesian model averaging approach. Empirical Economics 54, 673–703.

Temlyakov, V.N., 2000. Weak greedy algorithms. Advanced in Computational Mathematics 12, 213–227.

Tiao, G.C., 1985. Autoregressive moving average models, intervention problems and outlier detection in time series. In: Hannan, E.J., Krishnaiah, P.R., Rao, M.M. (Eds.), Handbook of Statistics, vol. 5. North-Holland, Amsterdam, pp. 85–118.

Tsay, R.S., 1984. Regression models with time series errors. Journal of the American Statistical Association 79, 118–124.

Wei, C.-Z., 1987. Adaptive prediction by least squares predictors in stochastic regression models with applications to time series. The Annals of Statistics 15, 1667–1682.

Wu, W.-B., Wu, Y.N., 2016. Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. Electronic Journal of Statistics 10, 352–379.

Zhao, P., Yu, B., 2006. On model selection consistency of Lasso. Journal of Machine Learning Research 7, 2541–2563.

Supplement to "Variable Selection for High-Dimensional Regression Models with Time Series and Heteroscedastic Errors"

Hai-Tang Chiou[a], Meihui Guo[b], and Ching-Kang Ing[a]
[a]Institute of Statistics, National Tsing Hua University
[b]Department of Applied Mathematics, National Sun Yat-sen University

S1 Proofs of Theorem 3

Recall that we assume in (17) that $\tilde{\alpha}_0$ is known, $E(\boldsymbol{z}_t) = 0$, and $E(z_{tj}^2) = 1$. In addition, $\bar{r}$ and $\bar{z}_j$ in (14) are set to $\tilde{\alpha}_0$ and 0, respectively. For the sake of convenience, we may assume, without loss of generality, that $\tilde{\alpha}_0 = 0$. Throughout this supplement, C stands for a generic positive constant independent of n. We start by several useful lemmas.

Lemma S1.1 *Assume* (A2′). *Then*

$$\max_{1\le i,j\le p_{\rm D}} \left| n^{-1}\sum_{t=1}^{n} z_{ti}z_{tj} - E(z_{ti}z_{tj}) \right| = O_p(n^{-1/2}p_{\rm D}^{1/q_4}).$$

PROOF. The proof of this lemma is elementary and is therefore omitted. □

Lemma S1.2 *Assume* (A1′). *Then*

$$\max_{1\le i\le p_{\rm D}} \left| n^{-1}\sum_{t=1}^{n} \varepsilon_t z_{ti} \right| = O_p(n^{-1/2}p_{\rm D}^{1/q_3}).$$

PROOF. Since $\{\varepsilon_t\}$ is independent of $\{\boldsymbol{z}_t\}$, by (A1′) and Jensen's inequality,

$$E\left| n^{-1/2}\sum_{t=1}^{n} \varepsilon_t z_{ti} \right|^{q_3} = E\left[E\left(\left| n^{-1/2}\sum_{t=1}^{n} \varepsilon_t z_{ti} \right|^{q_3} \middle| \varepsilon_t, 1\le t\le n \right)\right] \le C_{q_3} E\left(n^{-1}\sum_{t=1}^{n}\varepsilon_t^2 \right)^{q_3/2} \le C,$$

leading to the desired conclusion. □

Lemma S1.3 *Assume* (A2′) *and* (33). *Then*

$$\max_{1\le \sharp(J)\le K_{{\rm D},n}} \|\hat{\boldsymbol{\Gamma}}_{{\rm D},n}(J) - \boldsymbol{\Gamma}_{\rm D}(J)\|_2 = O_p(K_{{\rm D},n}n^{-1/2}p_{\rm D}^{1/q_4}),$$

where $\hat{\boldsymbol{\Gamma}}_{{\rm D},n}(J) = n^{-1}\sum_{t=1}^{n}\boldsymbol{z}_t(J)\boldsymbol{z}_t^{\top}(J)$. *Moreover, if* $K_{{\rm D},n} = o(n^{1/2}p_{\rm D}^{-1/q_4})$, *then*

$$\max_{1\le \sharp(J)\le K_{{\rm D},n}} \|\hat{\boldsymbol{\Gamma}}_{{\rm D},n}^{-1}(J) - \boldsymbol{\Gamma}_{\rm D}^{-1}(J)\|_2 = o_p(1).$$

PROOF. The proof of this lemma is similar to that of Lemma A.3. The details are omitted. □

Lemma S1.4 *Assume that the same assumptions as in Theorem 2,* (30)*, and* (A1′)–(A4′) *hold. Then, for* $\underline{c}_n \asymp n^{-1/2+d}$*,*

$$\begin{aligned} n^{-1}\sum_{t=1}^{n}\Theta_{t,n}^2 &= O_p(n^{-2\kappa}), \\ \max_{1\le i\le p_{\mathrm{D}}}\left|n^{-1}\sum_{t=1}^{n}\Theta_{t,n}z_{ti}\right| &= O_p(n^{-\kappa}). \end{aligned} \tag{S1.1}$$

PROOF. Note first that (A1) and Lemma S1.1 yield

$$\max_{1\le i\le p_{\mathrm{D}}}\left|n^{-1}\sum_{t=1}^{n}\Theta_{t,n}z_{ti}\right| \le \left\{n^{-1}\sum_{t=1}^{n}(\log\tilde{\eta}_t^2-\log\eta_t^2)^2\right\}^{1/2}\max_{1\le i\le p_{\mathrm{D}}}\left\{n^{-1}\sum_{t=1}^{n}z_{ti}^2\right\}^{1/2}, \tag{S1.2}$$

and

$$\max_{1\le i\le p_{\mathrm{D}}} n^{-1}\sum_{t=1}^{n}z_{ti}^2 \le \max_{1\le i\le p_{\mathrm{D}}}\left|n^{-1}\sum_{t=1}^{n}(z_{ti}^2-E(z_{ti}^2))\right| + \max_{1\le t\le n,1\le i\le p_{\mathrm{D}}}E(z_{ti}^2) = O_p(1). \tag{S1.3}$$

Since (A1) holds for any $q_1>0$ and since $0<\kappa<(1-2d)/(2+8/\tau)$, we may assume, without loss of generality, that $\kappa=(1-2d-2/q_1)/(2+8/\tau)$. By Theorem 2, (A1), and some algebraic manipulations, we obtain

$$\max_{1\le t\le n}|\tilde{\eta}_t^2-\eta_t^2| = O_p(n^{-1/2+d+1/q_1}). \tag{S1.4}$$

Let $\varsigma_1>0$ be arbitrarily small. By (S1.4), there is a large s (depending on ς_1) such that for all large n,

$$P(A_n)<\varsigma_1, \tag{S1.5}$$

where $A_n=\{\max_{1\le t\le n}|\tilde{\eta}_t^2-\eta_t^2|\le sn^{-1/2+d+1/q_1}\}$. Moreover,

$$n^{-1}\sum_{t=1}^{n}\left(\log\tilde{\eta}_t^2-\log\eta_t^2\right)^2 1_{A_n} \le (S1)+(S2)+(S3), \tag{S1.6}$$

with $\delta=2\kappa/\tau=(1-2d-2/q_1)/(4+\tau)$,

$$\begin{aligned} (S1) &= 2n^{-1}\sum_{t=1}^{n}(\log\tilde{\eta}_t^2)^2 1_{\{\eta_t^2<n^{-2\delta}\}}1_{A_n}, \\ (S2) &= 2n^{-1}\sum_{t=1}^{n}(\log\eta_t^2)^2 1_{\{\eta_t^2<n^{-2\delta}\}}1_{A_n}, \\ (S3) &= n^{-1}\sum_{t=1}^{n}(\log(1+(\tilde{\eta}_t^2-\eta_t^2)/\eta_t^2))^2 1_{\{\eta_t^2\ge n^{-2\delta}\}}1_{A_n}. \end{aligned}$$

By (A1), (A1$'$), (A3$'$), (A4$'$), (30), (32), (S1.4), Cauchy-Schwarz inequality, and Taylor's theorem, it holds that

$$\max_{1\le t\le n} P(\eta_t^2 < n^{-2\delta}) = O(n^{-4\kappa}), \tag{S1.7}$$

$$E(S1) \le \frac{C}{n}\sum_{t=1}^{n}(\log n)^2 P(\eta_t^2 < n^{-2\delta}) = O\big((\log n)^2 n^{-4\kappa}\big), \tag{S1.8}$$

$$E(S2) \le \frac{C}{n}\sum_{t=1}^{n}\big\{E[(\log \eta_t^2)^4]\big\}^{1/2}\big\{P(\eta_t^2 < n^{-2\delta})\big\}^{1/2} = O\big(n^{-2\kappa}\big), \tag{S1.9}$$

$$(S3) = O_p(n^{-2\kappa}). \tag{S1.10}$$

By (S1.5)–(S1.10) and Markov's inequality, we obtain the first equation of (S1.1), which, in conjunction with (S1.2) and (S1.3), leads to the second one. □

Let $\boldsymbol{z} = (z_1, \ldots, z_{p_{\mathrm{L}}})^\top$ be independent of and have the same covariance structure as $\{\boldsymbol{z}_t\}$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{p_{\mathrm{D}}})^\top$, and $r(\boldsymbol{z}) = \boldsymbol{z}^\top\boldsymbol{\alpha}$. Recall that $\hat{J}_{\mathrm{D},m}$ is the index set determined by $\mathrm{OGA}_{\mathcal{D}}$ at the m-th iteration. The next lemma investigates the convergence rate of

$$E_n[(r(\boldsymbol{z}) - \hat{r}_{\hat{J}_{\mathrm{D},m}}(\boldsymbol{z}))^2].$$

Here, for $J \subset \{1, \ldots, p_{\mathrm{D}}\}$, $\hat{r}_J(\boldsymbol{z}) = \boldsymbol{z}^\top(J)\hat{\boldsymbol{\alpha}}(J)$, $\hat{\boldsymbol{\alpha}}(J) = (\sum_{t=1}^n \boldsymbol{z}_t(J)\boldsymbol{z}_t^\top(J))^{-1}\sum_{t=1}^n \boldsymbol{z}_t(J)r_t$, and $\boldsymbol{z}(J) = (z_i, i \in J)^\top$.

Lemma S1.5 *Assume that the same assumptions as in Lemma S1.4 hold and $K_{\mathrm{D},n} \asymp n^\kappa p_{\mathrm{D}}^{-1/q}$. Then*

$$\max_{1\le m\le K_{\mathrm{D},n}} \frac{E_n[(r(\boldsymbol{z}) - \hat{r}_{\hat{J}_{\mathrm{D},m}}(\boldsymbol{z}))^2]}{m^{-1} + n^{-2\kappa} m p_{\mathrm{D}}^{2/q}} = O_p(1).$$

PROOF. Let $(\hat{r}_{1;J}, \ldots, \hat{r}_{n;J})^\top = \boldsymbol{H}_{\mathrm{D},J}\boldsymbol{r}$, $(\hat{z}_{1i;J}, \ldots, \hat{z}_{ni;J})^\top = \boldsymbol{H}_{\mathrm{D},J}\boldsymbol{Z}_i$, $\hat{z}_{ti;J}^{\perp} = z_{ti} - \hat{z}_{ti;J}$, and $z_{i;J}^{\perp} = z_i - z_{i;J}$, where $\boldsymbol{H}_{\mathrm{D},J}$ is defined after (14) and $z_{i;J} = \boldsymbol{z}^\top(J)\boldsymbol{\Gamma}_{\mathrm{D}}^{-1}(J)\boldsymbol{g}_{\mathrm{D},i}(J)$. Define

$$\mu_{\mathrm{D},J,i} = E[(r(\boldsymbol{z}) - r_J(\boldsymbol{z}))z_i] \quad \text{and} \quad \hat{\mu}_{\mathrm{D},J,i} = \frac{n^{-1}\sum_{t=1}^n (r_t - \hat{r}_{t;J})z_{ti}}{(n^{-1}\sum_{t=1}^n z_{ti}^2)^{1/2}},$$

where $r_J(\boldsymbol{z}) = \boldsymbol{z}^\top(J)\boldsymbol{\Gamma}_{\mathrm{D}}^{-1}(J)E(\boldsymbol{z}(J)r(\boldsymbol{z}))$. By Lemmas S1.1–S1.4 and an argument similar to that used to prove (A.7)–(A.9), we obtain

$$\max_{1\le\sharp(J)\le K_{\mathrm{D},n}} \|\hat{\boldsymbol{\Gamma}}_{\mathrm{D},n}^{-1}(J)\|_2 = O_p(1), \tag{S1.11}$$

$$\max_{\sharp(J)\le K_{\mathrm{D},n}-1, i\notin J} \left|\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t\hat{z}_{ti;J}^{\perp}\right| = O_p(n^{-1/2}p_{\mathrm{D}}^{1/q_3}),$$

$$\max_{\sharp(J)\le K_{\mathrm{D},n}-1, i\notin J} \left|\frac{1}{n}\sum_{t=1}^{n}\Theta_{t,n}\hat{z}_{ti;J}^{\perp}\right| = O_p(n^{-\kappa}),$$

$$\max_{\sharp(J)\le K_{\mathrm{D},n}-1, i,j\notin J} \left|\frac{1}{n}\sum_{t=1}^{n}z_{tj}\hat{z}_{ti;J}^{\perp} - E(z_j z_{i;J}^{\perp})\right| = O_p(n^{-1/2}p_{\mathrm{D}}^{1/q_4}),$$

which yield

$$\max_{\{(J,i):\sharp(J)\le K_{\mathrm{D},n}-1, i\notin J\}} |\hat{\mu}_{\mathrm{D},J,i} - \mu_{\mathrm{D},J,i}| = O_p(n^{-\kappa}p_{\mathrm{D}}^{1/q}). \tag{S1.12}$$

With the help of (S1.12), the rest of the proof can be carried out in the same fashion as that of Theorem 1. The details are skipped. □

Lemma S1.6 *Assume that the same assumptions as in Lemma S1.5 and* (A6′) *hold. Then*

$$\lim_{n\to\infty} P(N_{\mathrm{D},n} \subset \hat{J}_{K_{\mathrm{D},n}}) = 1.$$

Moreover, there is a sufficiently large a such that

$$\lim_{n\to\infty} P(\mathcal{D}_{\mathrm{D},n}) = 1,$$

where $\mathcal{D}_{\mathrm{D},n} = \{N_{\mathrm{D},n} \subseteq \hat{J}_{\mathrm{D},\lfloor an^{\gamma_{\mathrm{D}}}\rfloor}\}$.

PROOF. These conclusion follow directly from Lemma S1.5, (A6′), and an argument similar to that used to prove (A.22). □

Define $\tilde{k}_{\mathrm{D},n} = \min\{k : 1 \le k \le K_{\mathrm{D},n}, N_{\mathrm{D},n} \subseteq \hat{J}_{\mathrm{D},k}\}$, and $K_{\mathrm{D},n}+1$ if $N_{\mathrm{D},n} - \hat{J}_{\mathrm{D},K_{\mathrm{D},n}} \neq \emptyset$.

Lemma S1.7 *Assume that the same assumptions as in Lemma S1.6 and* (34) *hold. Then*

$$\hat{\sigma}^2_{\mathrm{D},\hat{J}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}}} - n^{-1}\sum_{t=1}^{n}E(\varepsilon_t^2) = o_p(1).$$

PROOF. Note that $\hat{\sigma}^2_{\mathrm{D},\hat{J}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}}} - n^{-1}\sum_{t=1}^n E(\varepsilon_t^2)$ is equal to

$$n^{-1}\|\boldsymbol{\varepsilon}+\boldsymbol{\Theta}_n\|_2^2 - n^{-1}\sum_{t=1}^n E(\varepsilon_t^2) - n^{-2}(\boldsymbol{\varepsilon}+\boldsymbol{\Theta}_n)^\top \boldsymbol{Z}_{\hat{J}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}}} \hat{\boldsymbol{\Gamma}}^{-1}_{\mathrm{D},n}(\hat{J}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}}) \boldsymbol{Z}^\top_{\hat{J}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}}} (\boldsymbol{\varepsilon}+\boldsymbol{\Theta}_n),$$

where $\boldsymbol{Z}_J = (\boldsymbol{Z}_j, j \in J)$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, and $\boldsymbol{\Theta}_n = (\Theta_{1,n}, \ldots, \Theta_{n,n})^\top$. Thus, $\hat{\sigma}^2_{\mathrm{D},\hat{J}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}}} - n^{-1}\sum_{t=1}^n E(\varepsilon_t^2)$ is bounded above by

$$\begin{aligned}
&\left| n^{-1}\sum_{t=1}^n \varepsilon_t^2 - E(\varepsilon_t^2) \right| + \left(n^{-1}\sum_{t=1}^n \Theta_{t,n}^2 \right) \\
&+2\left\{ \left(\left| n^{-1}\sum_{t=1}^n (\varepsilon_t^2 - E(\varepsilon_t^2)) \right| + n^{-1}\sum_{t=1}^n E(\varepsilon_t^2) \right) \left(n^{-1}\sum_{t=1}^n \Theta_{t,n}^2 \right) \right\}^{1/2} \\
&+\|n^{-1}(\boldsymbol{\varepsilon}+\boldsymbol{\Theta}_n)^\top \boldsymbol{Z}_{\hat{J}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}}}\|_2^2 \|\hat{\boldsymbol{\Gamma}}^{-1}_{\mathrm{D},n}(\hat{J}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}})\|_2. \qquad \text{(S1.13)}
\end{aligned}$$

By (A3′), (S1.11), Lemmas S1.2 and S1.4, $K_{\mathrm{D},n} = O(n^\kappa p_{\mathrm{D}}^{-1/q})$, and $K_{\mathrm{D},n}/n^{\gamma_{\mathrm{D}}} \to \infty$,

$$\begin{aligned}
&\|n^{-1}(\boldsymbol{\varepsilon}+\boldsymbol{\Theta}_n)^\top \boldsymbol{Z}_{\hat{J}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}}}\|_2^2 \|\hat{\boldsymbol{\Gamma}}^{-1}_{\mathrm{D},n}(\hat{J}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}})\|_2 I_{\mathcal{D}_{\mathrm{D},n}} \\
&\le \max_{1\le \sharp(J)\le \lfloor an^{\gamma_{\mathrm{D}}} \rfloor} \|\hat{\boldsymbol{\Gamma}}^{-1}_{\mathrm{D},n}(J)\|_2 \times an^{\gamma_{\mathrm{D}}} \max_{1\le j\le p_{\mathrm{D}}} \left(n^{-1}\sum_{t=1}^n (\varepsilon_t + \Theta_{t,n}) z_{tj} \right)^2 \\
&= o_p(1). \qquad \text{(S1.14)}
\end{aligned}$$

Hence, the desired conclusion follows from Lemmas S1.4 and S1.6, (A1′), (A3′), (34), (S1.11), (S1.13), and (S1.14). □

Proof of Theorem 3. We first show that

$$\lim_{n\to\infty} P(\hat{k}_{\mathrm{D},n} = \tilde{k}_{\mathrm{D},n}) = 1, \qquad \text{(S1.15)}$$

which is ensured by

$$P(\hat{k}_{\mathrm{D},n} < \tilde{k}_{\mathrm{D},n}) = o(1), \qquad \text{(S1.16)}$$

and

$$P(\hat{k}_{\mathrm{D},n} > \tilde{k}_{\mathrm{D},n}) = o(1). \qquad \text{(S1.17)}$$

To show (S1.16), note first that

$$\begin{aligned}
&\{\hat{k}_{\mathrm{D},n} < \tilde{k}_{\mathrm{D},n}, \mathcal{D}_{\mathrm{D},n}\} \subseteq \mathcal{M}_{1,n} \\
&\equiv \left\{ 2\alpha_{\hat{j}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}}} \hat{U}_{\mathrm{D},2,n} - \lambda\varpi_{\mathrm{D},n}|\hat{U}_{\mathrm{D},3,n}| \le -\alpha^2_{\hat{j}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}}} \hat{U}_{\mathrm{D},1,n} + \lambda\varpi_{\mathrm{D},n} \left(n^{-1}\sum_{t=1}^n E(\varepsilon_t^2) \right) \right\}, \qquad \text{(S1.18)}
\end{aligned}$$

where λ is some positive constant,

$$\begin{aligned}
\hat{U}_{\mathrm{D},1,n} &= n^{-1}\boldsymbol{Z}_{\hat{j}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}}}^{\top}(\boldsymbol{I}-\boldsymbol{H}_{\mathrm{D},\hat{j}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}-1}})\boldsymbol{Z}_{\hat{j}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}}},\\
\hat{U}_{\mathrm{D},2,n} &= n^{-1}\boldsymbol{Z}_{\hat{j}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}}}^{\top}(\boldsymbol{I}-\boldsymbol{H}_{\mathrm{D},\hat{j}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}-1}})(\boldsymbol{\varepsilon}+\boldsymbol{\Theta}_n),\\
\hat{U}_{\mathrm{D},3,n} &= \hat{\sigma}^2_{\mathrm{D},\hat{j}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}}}-n^{-1}\sum_{t=1}^{n}E(\varepsilon_t^2),
\end{aligned}$$

and $\varpi_{\mathrm{D},n}=n^{-1}\lfloor an^{\gamma_{\mathrm{D}}}\rfloor G_{\mathrm{D}}(p_{\mathrm{D}},n)$. Using Lemmas S1.1–S1.4 and S1.7 and an argument similar to that used to prove (A.27) and (A.28), one obtains

$$\hat{U}_{\mathrm{D},1,n}I_{\mathcal{D}_{\mathrm{D},n}}\geq v_{\mathrm{D},n}+o_p(1),\ |\hat{U}_{\mathrm{D},2,n}|I_{\mathcal{D}_{\mathrm{D},n}}=o_p(n^{-\gamma_{\mathrm{D}}/2}),\ |\hat{U}_{\mathrm{D},3,n}|I_{\mathcal{D}_{\mathrm{D},n}}=o_p(1), \tag{S1.19}$$

where $v_{\mathrm{D},n}=\min_{1\leq\sharp(J)\leq\lfloor an^{\gamma_{\mathrm{D}}}\rfloor}\lambda_{\min}(\boldsymbol{\Gamma}_{\mathrm{D}}(J))>\delta_{\mathrm{D}}$ for all large n. Now, by making use of (34), (35), Lemma S1.6, and (S1.19), $P(\mathcal{M}_{1,n})=o(1)$, which, together with (S1.18), leads to (S1.16).

To prove (S1.17), we obtain after straightforward calculations that

$$\{\hat{k}_{\mathrm{D},n}>\tilde{k}_{\mathrm{D},n}\}\subseteq\big\{2(\hat{k}_{\mathrm{D},n}-\tilde{k}_{\mathrm{D},n})(\hat{a}_{\mathrm{D},n}+\hat{b}_{\mathrm{D},n})+n\tilde{\varpi}_{\mathrm{D},n}|\hat{U}_{\mathrm{D},3,n}|\geq\delta_2 n\tilde{\varpi}_{\mathrm{D},n}\big\}, \tag{S1.20}$$

where δ_2 is some positive constant, $\tilde{\varpi}_{\mathrm{D},n}=1-\exp\{-n^{-1}(\hat{k}_{\mathrm{D},n}-\tilde{k}_{\mathrm{D},n})G_{\mathrm{D}}(p_{\mathrm{D}},n)\}$ and

$$\begin{aligned}
\hat{a}_{\mathrm{D},n} &= \|\hat{\boldsymbol{\Gamma}}^{-1}_{\mathrm{D},n}(\hat{J}_{\mathrm{D},K_{\mathrm{D},n}})\|_2\max_{1\leq j\leq p_{\mathrm{D}}}\left(n^{-1/2}\sum_{t=1}^{n}(\varepsilon_t+\Theta_{t,n})z_{tj}\right)^2,\\
\hat{b}_{\mathrm{D},n} &= \|\hat{\boldsymbol{\Gamma}}^{-1}_{\mathrm{D},n}(\hat{J}_{\mathrm{D},K_{\mathrm{D},n}})\|_2\max_{1\leq\sharp(J)\leq\tilde{k}_{\mathrm{D},n},i\notin J}\left(n^{-1/2}\sum_{t=1}^{n}(\varepsilon_t+\Theta_{t,n})\hat{z}_{ti;J}\right)^2.
\end{aligned}$$

By $\hat{k}_{\mathrm{D},n}\leq K_{\mathrm{D},n}=O(n^{\kappa}p_{\mathrm{D}}^{-1/q})$, there exists $\lambda>0$ such that

$$(n\tilde{\varpi}_{\mathrm{D},n})/(\hat{k}_{\mathrm{D},n}-\tilde{k}_{\mathrm{D},n})\geq\lambda\min\{n^{1-\kappa}p_{\mathrm{D}}^{1/q},G_{\mathrm{D}}(p_{\mathrm{D}},n)\}\quad\text{on }\{\hat{k}_{\mathrm{D},n}>\tilde{k}_{\mathrm{D},n}\}. \tag{S1.21}$$

After some algebraic manipulations, we have

$$\hat{a}_{\mathrm{D},n}+\hat{b}_{\mathrm{D},n}=O_p(p_{\mathrm{D}}^{2/q}+n^{1-2\kappa}),$$

which, together with (A3′), (S1.20), (S1.21), and Lemmas S1.6 and S1.7, gives (S1.17). Thus, (S1.15) is proved.

For $\tilde{k}_{\mathrm{D},n}>1$, define $\tilde{\delta}_{\mathrm{D},l}=1$ if $\mathrm{HDIC}_{\mathrm{D}}(\hat{J}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}}-\{\hat{j}_{\mathrm{D},l}\})>\mathrm{HDIC}_{\mathrm{D}}(\hat{J}_{\mathrm{D},\tilde{k}_{\mathrm{D},n}})$ and $\tilde{\delta}_{\mathrm{D},l}=0$ otherwise. Then,

$$\begin{aligned}
P(\hat{N}_{\mathrm{D},n}\neq N_{\mathrm{D},n})\leq{}&P(\tilde{\mathcal{G}}_{\mathrm{D},n})+P(\tilde{\mathcal{H}}_{\mathrm{D},n})\\
&+P(\hat{k}_{\mathrm{D},n}\neq\tilde{k}_{\mathrm{D},n})+P(\hat{N}_{\mathrm{D},n}\neq N_{\mathrm{D},n},\hat{k}_{\mathrm{D},n}=1)+P(N_{\mathrm{D},n}\not\subseteq\hat{J}_{\hat{k}_{\mathrm{D},n}}),
\end{aligned} \tag{S1.22}$$

where $\tilde{\mathcal{G}}_{\mathrm{D},n} = \{\tilde{\delta}_{\mathrm{D},l} = 0 \text{ and } \alpha_{\hat{j}_{\mathrm{D},l}} \neq 0 \text{ for some } 1 \leq l \leq \tilde{k}_{\mathrm{D},n}, \tilde{k}_{\mathrm{D},n} = \hat{k}_{\mathrm{D},n} > 1, N_{\mathrm{D},n} \subseteq \hat{J}_{\hat{k}_{\mathrm{D},n}}\}$ and $\tilde{\mathcal{H}}_{\mathrm{D},n} = \{\tilde{\delta}_{\mathrm{D},l} = 1 \text{ and } \alpha_{\hat{j}_{\mathrm{D},l}} = 0 \text{ for some } 1 \leq l \leq \tilde{k}_{\mathrm{D},n}, \tilde{k}_{\mathrm{D},n} = \hat{k}_{\mathrm{D},n} > 1, N_{\mathrm{D},n} \subseteq \hat{J}_{\mathrm{D},\hat{k}_{\mathrm{D},n}}\}$. By an argument similar to that used to prove (S1.15), it holds that $P(\tilde{\mathcal{G}}_{\mathrm{D},n}) = o(1)$ and $P(\tilde{\mathcal{H}}_{\mathrm{D},n}) = o(1)$. In addition, (S1.15) and Lemma S1.6 imply $P(\hat{k}_{\mathrm{D},n} \neq \tilde{k}_{\mathrm{D},n}) = o(1)$, $P(\hat{N}_{\mathrm{D},n} \neq N_{\mathrm{D},n}, \hat{k}_{\mathrm{D},n} = 1) = o(1)$, and $P(N_{\mathrm{D},n} \nsubseteq \hat{J}_{\mathrm{D},\hat{k}_{\mathrm{D},n}}) = o(1)$. These equations and (S1.22) lead immediately to (18). □

S2 Tables for Real Data Analysis

Table S2.1: Variable descriptions. "OECD" represents Organisation for Economic Co-operation and Development; "SEMI" stands for Semiconductor Equipment and Materials International.

	Variable	Description	Source
	Macroeconomic Variables		
1	FF	Federal Funds Rate	Federal Reserve
2	CLI	Composite Leading Index	OECD
3	IP	US Industrial Production Index	Federal Reserve
4	CS	Consumer Sentiment Index	University of Michigan
	Financial variables		
5	SOX	Philadelphia Semiconductor Index	Yahoo! Finance
6	NDQ	NASDAQ Composite Index	Yahoo! Finance
7	DJ	Dow Jones Industrial Average Index	Yahoo! Finance
	Semiconductor Variables		
8	CAP	Capacity	Federal Reserve
9	SIP	Industrial Production Index	Federal Reserve
10	UTL	Capacity Utilization Ratio	Federal Reserve
11	ISR	Inventories to Shipments Ratios (Computer and Electronic Products)	Bureau of Census
12	NO	New Orders (Computer and Electronic Products)	Bureau of Census
13	FGI	Finished Goods Inventories (Computer and Electronic Products)	Bureau of Census
14	MSI	Materials and Supplies Inventories (Computer and Electronic Products)	Bureau of Census
15	VS	Value of Shipments (Computer and Electronic Products)	Bureau of Census
16	TI	Total Inventories (Computer and Electronic Products)	Bureau of Census
17	Bill	Billings for Semiconductor Manufacturing Equipment	SEMI
18	PPI	Producer Price Index (Electronic Components and Accessories)	Bureau of Labor Statistics
19	ES	Retail Sales for Electronics and Appliance Stores	Bureau of Census
20	ESA	Wholesale Sales for Electrical and Electronic Goods	Bureau of Census
21	EIN	Wholesale Inventories for Electrical and Electronic Goods	Bureau of Census
	Industrial Production Index		
22	IP1	Computer and Electronic Product	Federal Reserve
23	IP2	Computer and Peripheral Equipment	Federal Reserve
24	IP3	Communications Equipment	Federal Reserve
25	IP4	Audio and Video Equipment	Federal Reserve
26	IP6	Electrical Equipment, Appliance, and Component	Federal Reserve
27	IP7	Battery	Federal Reserve
28	IP8	Communication and Energy Wire and Cable	Federal Reserve

Table S2.2: Variable descriptions (continued).

	Variable	Description	Source
	Producer Price Index		
29	PPI3	Other Semiconductor Devices (Parts such as Chips, Wafers, and Heat Sinks)	Bureau of Labor Statistics
	New Orders		
30	NO2	Construction Machinery Manufacturing	Bureau of Census
31	NO7	Other Electronic Component Manufacturing	Bureau of Census
32	NO9	Household Appliance Manufacturing	Bureau of Census
33	NO11	Computers and Related Products	Bureau of Census
34	NO12	Communication Equipment	Bureau of Census
35	NO14	Electrical Equipment Manufacturing	Bureau of Census
36	NO15	Search and Navigation Equipment (Nondefense)	Bureau of Census
	Total Inventories		
37	TI1	Farm Machinery and Equipment Manufacturing	Bureau of Census
38	TI2	Construction Machinery Manufacturing	Bureau of Census
39	TI3	Computer Storage Device Manufacturing	Bureau of Census
40	TI4	Other Computer Peripheral Equipment Manufacturing	Bureau of Census
41	TI5	Communications Equipment Manufacturing (Nondefense)	Bureau of Census
42	TI6	Audio and Video Equipment	Bureau of Census
43	TI7	Other Electronic Component Manufacturing	Bureau of Census
44	TI8	Electrical Equipment, Appliances, and Components	Bureau of Census
45	TI9	Household Appliance Manufacturing	Bureau of Census
46	TI10	Battery Manufacturing	Bureau of Census
47	TI11	Computers and Related Products	Bureau of Census
48	TI12	Communication Equipment	Bureau of Census
49	TI13	Information Technology Industries	Bureau of Census
50	TI14	Electrical Equipment Manufacturing	Bureau of Census
51	TI15	Search and Navigation Equipment (Nondefense)	Bureau of Census
	Value of Shipments		
52	VS1	Farm Machinery and Equipment Manufacturing	Bureau of Census
53	VS2	Construction Machinery Manufacturing	Bureau of Census
54	VS3	Computer Storage Device Manufacturing	Bureau of Census
55	VS4	Other Computer Peripheral Equipment Manufacturing	Bureau of Census
56	VS5	Communications Equipment Manufacturing (Nondefense)	Bureau of Census
57	VS6	Audio and Video Equipment	Bureau of Census
58	VS7	Other Electronic Component Manufacturing	Bureau of Census
59	VS8	Electrical Equipment, Appliances, and Components	Bureau of Census
60	VS9	Household Appliance Manufacturing	Bureau of Census
61	VS10	Battery Manufacturing	Bureau of Census
62	VS11	Computers and Related Products	Bureau of Census
63	VS12	Communication Equipment	Bureau of Census
64	VS13	Information Technology Industries	Bureau of Census
65	VS14	Electrical Equipment Manufacturing	Bureau of Census
66	VS15	Search and Navigation Equipment (Nondefense)	Bureau of Census

Table S2.3: Variables selected by Twohit under Model II and their coefficient estimates

Regression model (β)					
Variable(lag)	Est.	S.E.	Variable(lag)	Est.	S.E.
intercept	**0.00502**	0.00223	NO14(4)	**0.15633**	0.02118
$y_t(1)$	**0.82070**	0.01146	NO14(20)	**0.19847**	0.02091
SOX(2)	**0.06035**	0.01581	NO15(21)	**-0.03884**	0.00643
SOX(6)	**0.07574**	0.01732	TI3(4)	**-0.22313**	0.03989
SIP(12)	**-0.89537**	0.11948	TI4(2)	**0.29334**	0.06827
FGI(2)	**-0.32394**	0.07048	TI9(21)	**0.71267**	0.08566
FGI(16)	**0.24811**	0.07172	TI10(12)	**0.29017**	0.05388
MSI(3)	**0.67855**	0.09244	TI12(10)	**-0.62192**	0.06310
MSI(14)	**-0.89731**	0.09346	VS3(10)	**0.08423**	0.00913
ES(12)	**1.02190**	0.09301	VS4(7)	**-0.15338**	0.01769
EIN(10)	**-0.63792**	0.14271	VS6(16)	**-0.09810**	0.01688
IP1(1)	**2.01883**	0.16786	VS7(12)	**-0.40061**	0.04432
IP3(6)	**0.44903**	0.06870	VS9(22)	**0.37633**	0.04783
IP7(21)	**-0.30576**	0.04480	VS12(5)	**0.18091**	0.02841
IP8(15)	**0.26290**	0.06454	VS14(12)	**-0.42075**	0.04919
NO7(18)	**0.11655**	0.02069	IP(18)	**-2.08027**	0.27617
NO9(21)	**0.11959**	0.02215			
Dispersion model (α)					
Variable(lag)	Est.	S.E.	Variable(lag)	Est.	S.E.
intercept	**-7.38764**	0.93013	TI6(19)	**-13.65783**	1.77437
SOX(14)	**2.98043**	0.82968	VS1(21)	1.28667	0.88379
ISR(17)	0.09709	0.60577	VS4(19)	**2.31033**	0.95088
NO7(13)	**-7.20692**	1.16834	VS6(10)	**1.86396**	0.82516
NO7(23)	**5.16120**	1.14999			
Information criterion					
AIC	-971.3615				
BIC	-832.1359				

Note: Coefficients with absolute values larger than 1.96 standard errors are boldfaced.

Table S2.4: Variables selected by Twohit under Model I and their coefficient estimates

Regression model (β)					
Variable(lag)	Est.	S.E.	Variable(lag)	Est.	S.E.
intercept	0.63236	0.59236	ES(1)	**-0.55477**	0.12094
FF(5)	**-0.05186**	0.01721	ES(10)	**0.49835**	0.12103
FF(14)	**-0.10049**	0.02018	IP1(4)	**2.59567**	0.26452
CLI(16)	**-0.03054**	0.00561	IP3(5)	**0.44699**	0.08640
CLI(24)	**0.02719**	0.00380	IP4(24)	**-0.08947**	0.02444
IP(10)	**-2.05307**	0.37445	IP7(14)	**0.33045**	0.05875
SOX(6)	**0.12761**	0.02462	TI1(8)	**0.30947**	0.11051
NDQ(2)	**0.23045**	0.03694	TI4(21)	**-0.39709**	0.08245
NDQ(4)	**0.21216**	0.03746	TI6(15)	**-0.25269**	0.04181
NDQ(7)	**0.15578**	0.03606	TI6(18)	**-0.43609**	0.04561
NDQ(9)	**0.25454**	0.03385	TI9(14)	**0.72934**	0.11655
SIP(1)	**2.61711**	0.16375	TI11(2)	**0.62424**	0.09295
UTL(12)	**-0.00561**	0.00058	TI13(11)	**-1.11378**	0.21753
NO(2)	**0.28929**	0.04636	TI13(14)	**-1.66225**	0.24722
FGI(2)	**-0.42104**	0.09951	VS3(10)	**0.08848**	0.01365
TI(1)	**1.20487**	0.24303	VS8(1)	**0.65022**	0.10446
Bill(1)	**0.00020**	0.00001	VS14(8)	**0.34628**	0.06484
Bill(10)	**-0.00011**	0.00001			
Dispersion model (α)					
Variable(lag)	Est.	S.E.	Variable(lag)	Est.	S.E.
intercept	**-6.67482**	0.14447	NO2(23)	**0.92285**	0.40358
CAP(11)	**19.67328**	9.98880	TI1(9)	-6.00693	4.11759
ESA(1)	**-20.31691**	4.13612	TI14(14)	**-22.06525**	6.16805
IP2(23)	-3.97144	3.57243	VS3(3)	**2.03680**	0.69172
IP7(9)	**-12.25308**	2.83380	VS5(3)	2.78433	1.77836
The model for ϵ_t: $\epsilon_t = \phi_1\epsilon_{t-1} + \phi_2\epsilon_{t-2} + w_t$					
	Est.	S.E.		Est.	S.E.
ϕ_1	**0.34949**	0.06296	ϕ_2	**0.22457**	0.06386
Information criterion					
AIC	-819.4105				
BIC	-662.7818				

Note: Coefficients with absolute values larger than 1.96 standard errors are boldfaced.

References

Chiou, H.-T., Guo, M., Ing, C.-K., 2019. Variable selection for high-dimensional regression models with time series and heteroscedastic errors.

Ing, C.-K, Lai, T.L., 2011. A stepwise regression method and consistent model selection for high-dimensional sparse linear models. Statistica Sinica 21, 1473–1513.