國立臺灣大學理學院應用數學科學研究所

碩士論文

Institute of Applied Mathematical Sciences

College of Science

National Taiwan University

Master Thesis

高維度時間序列並帶有測量誤差模型之模型選擇

Model Selection for High-Dimensional Time Series Models with Measurement Errors

黃學涵

Hsueh-Han Huang

指導教授：銀慶剛 博士

Advisor: Ching-Kang Ing, Ph.D.

中華民國 106 年 6 月

June 2017

doi:10.6342/NTU201700925

摘要

我們使用一個叫做正交化貪婪演算法的快速逐步迴歸對高維度時間序列並帶有測量誤差的模型做模型選擇。在一個弱稀疏的條件下，我們推導出了正交化貪婪演算法預測誤差的收斂速度。在一個強稀疏的條件下，發展出一套擁有一致性的選模準則。

關鍵詞: 高維度、測量誤差、正交化貪婪演算法、稀疏性、時間序列

doi:10.6342/NTU201700925

Abstract

We use a fast stepwise regression method, called orthogonal greedy algorithm (OGA) to select variables for high-dimensional time series model with measurement errors. Under a weak sparsity condition, we derive a convergence rate of OGA, which is expressed in terms of the number of iterations, the sample size and the order of the moment imposed on the error process. Under a strong sparsity condition, we develop a consistent model selection procedure using OGA and a high-dimensional information criterion.

Keywords: High-dimensional, measurement error, OGA, sparsity, time series.

doi:10.6342/NTU201700925

目　錄

doi:10.6342/NTU201700925

表　目　錄

doi:10.6342/NTU201700925

1 Introduction

Consider the simple linear regression without intercept

$$y = \underset{\sim}{\beta}^T \underset{\sim}{x} + \xi,$$

where $\underset{\sim}{x} = (x_1, x_2, ..., x_p)^T$ is a p-dimensional random vector satisfying $E(\underset{\sim}{x}) = \underset{\sim}{0} = (0, 0, ..., 0)^T$ and $E(\underset{\sim}{x}\underset{\sim}{x}^T) = \Sigma_x$. $\underset{\sim}{\beta} = (\beta_1, \beta_2, ..., \beta_p)^T$ is a p-dimensional constant vector. $E(\xi) = 0$, $E(\xi^2) = \sigma_\xi^2 > 0$, and $\underset{\sim}{x}$ and ξ are independent.

Assume that $\underset{\sim}{w} = \underset{\sim}{x} + \underset{\sim}{\eta}$ is observed instead of $\underset{\sim}{x}$, where $\underset{\sim}{w} = (w_1, w_2, ..., w_p)^T$, $\underset{\sim}{\eta} = (\eta_1, \eta_2, ..., \eta_p)^T$ is a vector of measurement errors, and $y^\star = y + \eta_y$ is observed instead of y, η_y is a measurement error, where $E(\underset{\sim}{\eta}) = \underset{\sim}{0}$, $E(\underset{\sim}{\eta}\underset{\sim}{\eta}^T) = \Sigma_\eta$ and $\underset{\sim}{\eta}$ is independent of $(\underset{\sim}{x}, \xi)$, $E(\eta_y) = 0$, $E(\eta_y^2) = \sigma_{\eta_y}^2$ and η_y is independent of $(\underset{\sim}{x}, \underset{\sim}{\eta}, \xi)$. To make complicated things simple, we assume that $\Sigma_\eta = diag(\sigma_{\eta_1}^2, \sigma_{\eta_2}^2, ..., \sigma_{\eta_p}^2)$ is a diagonal matrix. Note that since η_y can be absorb into ξ, we still denote y as $y^\star$ for simplicity and view ξ as the random errors after absorbing η_y.

If we regress y on $\underset{\sim}{w}$, then it follows that

$$y = \underset{\sim}{\beta}^{\star T} \underset{\sim}{w} + \xi^\star,$$

where $\underset{\sim}{\beta}^\star = \underset{\sim}{\beta} - \underset{\sim}{U}$ and $\underset{\sim}{U} = (\Sigma_x + \Sigma_\eta)^{-1} \Sigma_\eta \underset{\sim}{\beta}$, noting that $\underset{\sim}{\beta}^\star = \underset{\sim}{\beta}$ if $\forall i = 1, 2, ..., p,\ \sigma_{\eta_i} = 0$.

Let $(y_t, \underset{\sim}{w}_t^T), t = 1, 2, ..., n$, be observations, where $\underset{\sim}{w}_t = (w_{t1}, w_{t2}, ..., w_{tp})^T$. We allow $(y_t, \underset{\sim}{x}_t^T, \underset{\sim}{\eta}_t^T, \xi_t)$ be a stationary time series and $p >> n$. When p is larger than n, there are computational difficulties in estimating the regression coefficients by standard regression methods. Ing and Lai (2011) propose the orthogonal greedy algorithm (OGA) to circumvent the computation in high dimensional inversion matrix. They derive the convergence rate of OGA and provide a consistent model selection procedure for high-dimensional time independent models. Ing and Huang (2016) generalize the results to multivariate time series model setting and relax the moment bound assumptions from exponential moment bounds to polynomial moment bounds. However, none of them consider measurement errors in their models. Since we often face data with measurement errors that cannot be ignored in many applications, recently, high-dimensional models with measurement errors has been widely studied. Loh and Wainwright (2012) propose a non-convex modification of Lasso for doing high-dimensional models with measurement errors and missing data, they also consider a time series model setting but only the cases within class of VAR(1) models with an upper restricted eigenvalue condition for sample covariance matrix. Datta and Zou (2016) propose a modification which is called Convex Conditioned Lasso (CoCoLasso) to circumvent the problem of non-convexity and the method can handle with a general class of corrupted data, but they only develop theories of the case that the true regressors are fixed design and there is no theory of model selection for a time series model setting. Belloni, Rosenbaum and Tsybakov (2014), (2016) use a Dantzig Selector type method named matrix uncertainty (MU) selector for doing high-dimensional model with measurement errors, but they do not consider a time series model setting. To our best knowledge, the existing papers that are related with high-dimensional model with measurement errors seldom consider a time series setting additionally, and none of them use greedy algorithm to do model selection with the previous model settings. This paper focuses on the OGA method and generalizes the results in Ing and Lai (2011) and Ing and Huang (2016) to a new dimension: high-dimensional time series

doi:10.6342/NTU201700925

models with measurement errors under polynomial moment bound assumptions.

In this paper, We provide an upper bound of the number of iterations and derive the uniform convergence rate of empirical prediction error of OGA under a weak sparsity condition. We prove the sure screening property of OGA under a strong sparsity condition. We propose an information criterion to do model selection, together with a trimming method, the whole procedure is shown to achieve the oracle property. We also provide simulation studies to show that with proper order of moment bounds, OGA+HDBIC+Trim successfully identifies the smallest correct model with high ratios in some general model settings. Although some additional conditions are needed, the necessary conditions for OGA to do consistent model selection for models with measurement errors remain simple.

The rest of this paper is organized as follows: in Section 2, we introduce OGA and noiseless OGA. In Section 3, we derive the convergence rate of OGA. In Section 4, we prove the sure screening property of OGA, and introduce our model selection criterion along the OGA path which is called high-dimensional information criterion (HDIC). We also proposed a trimming method to exclude redundant variables and prove that OGA+HDIC+Trim achieves model selection consistency. In Section 5, we present simulation studies to illustrate the performance of OGA+HDIC+Trim.

2 OGA and Noiseless OGA

In this secition, we briefly introduce OGA and noiseless OGA that are proposed by Ing and Lai (2011).

doi:10.6342/NTU201700925

Denote $\hat{y}_k(\underset{\sim}{w})$ as a sequence of linear approximations of the regression function $y(\underset{\sim}{w}) = \underset{\sim}{\beta}^{\star T}\underset{\sim}{w}$. Initializing with $\hat{y}_0(\cdot) = 0$, it computes the residuals $U_t^{(k)} := y_t - \hat{y}_k(\underset{\sim}{w}_t),\ 1 \le t \le n$, at the end of the kth iterations and chooses $w_{t,\hat{j}_{k+1}}$ on which $U_t^{(k)}$ is regressed, such that

$$\hat{j}_{k+1} = \arg\min_{1\le j\le p} \sum_{t=1}^{n} (U_t^{(k)} - \tilde{\beta}_j^{(k)} w_{tj})^2,$$

where $\tilde{\beta}_j^{(k)} = \frac{\sum_{t=1}^{n} U_t^{(k)} w_{tj}}{\sum_{t=1}^{n} w_{tj}^2}$. We update

$$\hat{y}_{k+1}(\underset{\sim}{w}_t) = \hat{y}_k(\underset{\sim}{w}_t) + \hat{\beta}^{(k)}_{\hat{j}_{k+1}} w^{\perp}_{t,\hat{j}_{k+1}}$$

where $\hat{\beta}^{(k)}_{\hat{j}_{k+1}} = \frac{\sum_{t=1}^{n} U_t^{(k)} w^{\perp}_{t,\hat{j}_{k+1}}}{\sum_{t=1}^{n} w^{\perp 2}_{t,\hat{j}_{k+1}}}$, $w^{\perp}_{t,\hat{j}_{k+1}}$ is the tth component of vector $\mathbf{w}^{\perp}_{\hat{j}_{k+1}} = \mathbf{w}_{\hat{j}_{k+1}} - \hat{\mathbf{w}}_{\hat{j}_{k+1}}$, $\hat{\mathbf{w}}_{\hat{j}_{k+1}}$ is the projection of $\mathbf{w}_{\hat{j}_{k+1}}$ into the linear space spanned by $(\mathbf{w}_{\hat{j}_1}, \mathbf{w}_{\hat{j}_2}, ..., \mathbf{w}_{\hat{j}_k})$, where $\mathbf{w}_j = (w_{1j}, w_{2j}, ..., w_{nj})^T$. The orthogonalization of the predictor variables allows us to use componentwise linear regression to compute OLS, thereby circumventing the difficulties with computing the inverse of high-dimensional matrix.

Noiseless OGA is similar to OGA but replaces y_t by its mean $y(\underset{\sim}{w}_t)$. In the next section, we'll use noiseless OGA to derive the convergence rate of the empirical prediction error of OGA. More details of OGA and noiseless OGA can be found in Ing and Lai (2011).

3 Uniform Convergence Rate of Empirical Prediction Error

In this section, we derive the convergence rate for OGA in linear regression time series models with measurement errors in which the number of regressors is allowed to be much larger than the number of observations.

According to OGA, $\hat{y}_m(\underset{\sim}{w}_t) = \underset{\sim}{w}_t^T(\hat{J}_m)\hat{\underset{\sim}{\beta}}(\hat{J}_m)$, where $\hat{J}_m$ is the index set of the variable selected by OGA after m iterations, $\underset{\sim}{w}_t(J) = (w_{ti}, i \in J)^T$ and $\hat{\underset{\sim}{\beta}}(J) = (\sum_{t=1}^{n} \underset{\sim}{w}_t(J)\underset{\sim}{w}_t^T(J))^{-1} \sum_{t=1}^{n} \underset{\sim}{w}_t(J)y_t$ is the LSE based on model J. Let K_n denote a prescribed upper bound on the number m of OGA iterations. To provide the uniform convergence rate of the empirical norm $\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_m(\underset{\sim}{w}_t) - \underset{\sim}{\beta}^{\star T}\underset{\sim}{w}_t)^2$, $1 \leq m \leq K_n$, we make the following assumptions below.

Assume $\{\xi_t\}_{t=1}^{n}$ is a martingale difference sequence with respect to an increasing sequence of σ-fields $\{\mathcal{F}_t\}$, $\{\eta_{ti}\}_{t=1}^{n}, i = 1, 2, ..., n$ are martingale difference sequences with respect to an increasing sequence of σ-fields $\left\{\tilde{\mathcal{F}}_t\right\}$, $w_{ti}, i = 1, 2, ..., n$, are $\mathcal{F}_{t-1}$-measurable, $x_{ti}, i = 1, 2, ..., n$, are $\tilde{\mathcal{F}}_{t-1}$-measurable and there exist q_1, q_2 with $q_2 > q_1 \geq 2$ s.t.

(C1) $\max\limits_{1\leq t\leq n, 1\leq i\leq p} E|x_{ti}|^{2q_1} = O(1)$,

$\sup\limits_{1\leq t<\infty, 1\leq i\leq p} E[|\eta_{ti}|^{2q_1}|\tilde{\mathcal{F}}_{t-1}] \leq C_1 < \infty$ a.s., for some $C_1 > 0$,

$\sup\limits_{1\leq t<\infty} E[|\xi_t|^{q_1}|\mathcal{F}_{t-1}] \leq C_2 < \infty$ a.s., for some $C_2 > 0$,

(C2) $\max\limits_{1\leq i,j\leq p} E|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(w_{ti}w_{tj} - \sigma_{ij})|^{2q_2} = O(1)$,

$\max\limits_{1\leq i,j\leq p} E|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(x_{ti}x_{tj} - \sigma_{xij})|^{2q_2} = O(1)$,

where $\sigma_{ij} = E(w_{ti}w_{tj}), \sigma_{xij} = E(x_{ti}x_{tj})$.

Remark. If w_{tj} has a linear representation

$$w_{tj} = \sum_{k=-\infty}^{\infty} a(k)\alpha_j(t-k)$$

where $(\alpha_j(t), \mathcal{F}_t)$ are martingale difference sequences with $-\infty < t < \infty$ and

$$E[\alpha_j(t)^2 | \mathcal{F}_{t-1}] = 1$$

and there exists a positive constant C_{q_2} s.t.

$$\sup_{-\infty<t<\infty} E[\alpha_j(t)^{4q_2} | \mathcal{F}_{t-1}] \leq C_{q_2}$$

with the spectral density function of w_{tj}, denoted f_j is square integrable,

$$\max_{1\leq j\leq p} \sum_{k=-\infty}^{\infty} [E(w_{tj} w_{t+k,j})]^2 = \max_{1\leq j\leq p} \frac{1}{2\pi} \int_{-\pi}^{\pi} f_j^2(\lambda) d\lambda = O(1).$$

Then, by (2.10) in Findley and Wei (1993), the first condition in (C2) holds.

(C3) $||\underset{\sim}{\beta}||_1 < \infty$.

This assumption is the weak sparsity condition on the uncontaminated regression coefficients.

(C4) $||\underset{\sim}{U}||_1 = ||(\Sigma_x + \Sigma_\eta)^{-1} \Sigma_\eta \underset{\sim}{\beta}||_1 < \infty$.

This assumption and (C4) assure the weak sparsity condition on the regression coefficients contaminated by measurement errors.

Remark. There are many ways to achieve (C4), for example, if the values of measurement errors are restricted by the number of regressors, say $\max\limits_{1\leq i\leq p} \sigma^2_{\eta_i} = O(\frac{1}{\sqrt{p}})$, then (C4) holds, since $||\underset{\sim}{\beta}^\star||_1 = ||\underset{\sim}{\beta} - (\Sigma_x + \Sigma_\eta)^{-1} \Sigma_\eta \underset{\sim}{\beta}||_1 \leq ||\underset{\sim}{\beta}||_1 (1 + \sqrt{p} \frac{\max\limits_{1\leq i\leq p} \sigma^2_{\eta_i}}{\lambda_{min}(\Sigma_x) + \min\limits_{1\leq i\leq p} \sigma^2_{\eta_i}})$. Another important example is the case that $(x_{t1}, x_{t2}..., x_{tp})$ has special covariance structure, for example, uncorrelated structure. But, in general, (C4) does not hold without further conditions.

(C5) $\frac{n}{p^{\frac{2}{q_1}}} \to \infty$ as $n \to \infty$.

doi:10.6342/NTU201700925

The following theorem gives the rate of convergence, which holds uniformly over $1 \leq m \leq K_n$, for the empirical prediction error of OGA. The uniform convergence rate varies with the prescribed order of moments q_1 in (C1). When the order of moments q_1 is smaller, the uniform convergence rate becomes larger due to the weaker moment assumptions. Define $R(J) = E(\underset{\sim}{w}_1(J)\underset{\sim}{w}_1(J)^T)$ and $\underset{\sim}{\gamma}_i(J) = E(w_{1i}\underset{\sim}{w}_1(J))$.

Theorem 1. *Assume* (C1)-(C5). *Suppose* $K_n = O(\sqrt{\frac{n}{p^{\frac{2}{q_1}}}})$, *and*

$$\min_{1\leq \#(J) \leq K_n} \lambda_{min}(R(J)) > \delta, \max_{1 \leq \#(J) \leq K_n, i \notin J} ||R^{-1}(J)\underset{\sim}{\gamma}_i(J)||_1 < C^\star < \infty, \quad (3.1)$$

for some $\delta, C^\star > 0$. *Then*

$$\max_{1 \leq m \leq K_n} \left(\frac{n^{-1}\sum_{t=1}^{n}(\hat{y}_m(\underset{\sim}{w}_t) - \underset{\sim}{\beta}^{\star T}\underset{\sim}{w}_t)^2}{m^{-1} + mn^{-1}p^{\frac{2}{q_1}}} \right) = O_p(1).$$

Proof.

$$\begin{aligned} &\tfrac{1}{n}\textstyle\sum_{t=1}^{n}(\hat{y}_m(\underset{\sim}{w}_t) - \underset{\sim}{\beta}^{\star T}\underset{\sim}{w}_t)^2 \\ &= \tfrac{1}{n}(Y(\mathbf{w}) - H_{\hat{J}_m}Y)^T(Y(\mathbf{w}) - H_{\hat{J}_m}Y) \\ &= \tfrac{1}{n}Y^T(\mathbf{w})(I - H_{\hat{J}_m})Y(\mathbf{w}) + \tfrac{1}{n}(Y - Y(\mathbf{w}))^T H_{\hat{J}_m}(Y - Y(\mathbf{w})), \end{aligned}$$

where $Y(\mathbf{w}) = (\underset{\sim}{w}_1^T\underset{\sim}{\beta}^\star, \underset{\sim}{w}_2^T\underset{\sim}{\beta}^\star, ..., \underset{\sim}{w}_n^T\underset{\sim}{\beta}^\star)^T$, $Y = (y_1, y_2, ..., y_n)^T$, and H_J is a projection matrix project vectors into the linear space spanned by $(\mathbf{w}_i, i \in J)$, where $\mathbf{w}_i = (w_{1i}, w_{2i}, ..., w_{ni})^T$. Let

$$\mu_{J,i} = \frac{Y^T(\mathbf{w})(I-H_J)\mathbf{w}_i}{n^{\frac{1}{2}}||\mathbf{w}_i||}, \hat{\mu}_{J,i} = \frac{Y^T(I-H_J)\mathbf{w}_i}{n^{\frac{1}{2}}||\mathbf{w}_i||},$$

where $||\cdot|| = ||\cdot||_2$ denotes the L_2-norm in this paper. Consider two events

$$A_n(k) = \left\{ \max_{(J,i):\#(J)\leq k-1, i \notin J} |\hat{\mu}_{J,i} - \mu_{J,i}| \leq s(\sqrt{\frac{p^{\frac{2}{q_1}}}{n}}) \right\},$$

$$B_n(k) = \left\{ \min_{0 \leq i \leq k-1} \max_{1 \leq j \leq p} |\mu_{\hat{J}_i,j}| > \tilde{\xi}_0 s(\sqrt{\frac{p^{\frac{2}{q_1}}}{n}}) \right\},$$

doi:10.6342/NTU201700925

where s is a positive constant independent of n and k, $\tilde{\xi}_0 = 2/(1-\xi_0)$, for $0 < \xi_0 < 1$.

On $A_n(m) \cap B_n(m)$, for $1 \leq q \leq m$,

$$\begin{aligned}|\mu_{\hat{J}_{q-1},\hat{j}_q}| &\geq -|\hat{\mu}_{\hat{J}_{q-1},\hat{j}_q} - \mu_{\hat{J}_{q-1},\hat{j}_q}| + |\hat{\mu}_{\hat{J}_{q-1},\hat{j}_q}| \\ &\geq -2s\sqrt{\frac{p^{\frac{2}{q_1}}}{n}} + \max_{1\leq j\leq p} |\mu_{\hat{J}_{q-1},j}| \\ &\geq \xi_0 \max_{1\leq j\leq p} |\mu_{\hat{J}_{q-1},j}|.\end{aligned}$$

This is the generalization of noiseless OGA in the Appendix B in Ing and Lai (2011). So, by Lemma B1 in Ing and Lai (2011), (C3), and (C4),

$$\frac{1}{n}Y^T(\mathbf{w})(I - H_{\hat{J}_m})Y(\mathbf{w}) = O_p(\frac{1}{1+m\xi_0^2}). \tag{3.2}$$

On $B_n^c(m)$, by (C3), (C4) and Lemma 1 in Appendix,

$$\begin{aligned}\tfrac{1}{n}Y^T(\mathbf{w})(I - H_{\hat{J}_m})Y(\mathbf{w}) &\leq \min_{1\leq i\leq m-1} \tfrac{1}{n}Y^T(\mathbf{w})(I - H_{\hat{J}_i})Y(\mathbf{w}) \\ &\leq \max_{1\leq j\leq p} ||\underset{\sim}{\beta}^\star||_1 \tfrac{||w_j||}{n^{1/2}} \tilde{\xi}_0 s\sqrt{\frac{p^{\frac{2}{q_1}}}{n}} \\ &= O_p(\frac{1}{m}). \end{aligned} \tag{3.3}$$

It remains to prove that $\forall\epsilon > 0, \exists s > 0$ s.t.

$$P(A_n^c(m)) \leq \epsilon, \tag{3.4}$$

and the proof is shown in the Appendix.

So, by (3.2)-(3.4), we have

$$\frac{1}{n}Y^T(\mathbf{w})(I - H_{\hat{J}_m})Y(\mathbf{w}) = O_p(\frac{1}{m}). \tag{3.5}$$

On the other hand,

$$\begin{aligned}
&\tfrac{1}{n}(Y - Y(\mathbf{w}))^T H_{\hat{J}_m}(Y - Y(\mathbf{w})) \\
&= \tfrac{1}{n}\boldsymbol{\xi}^{\star T} H_{\hat{J}_m}\boldsymbol{\xi}^\star \\
&\leq ||\hat{R}^{-1}(\hat{J}_m)||m \max_{1\leq i\leq p}(\tfrac{1}{n}\textstyle\sum_{t=1}^n \xi_t^\star w_{ti})^2 \\
&= O_p(\frac{mp^{\frac{2}{q_1}}}{n}), \qquad (3.6)
\end{aligned}$$

where $\boldsymbol{\xi}^\star = (\xi_1^\star, \xi_2^\star, ..., \xi_n^\star)^T$. Theorem 1 follows form (3.5) and (3.6).

4 Sure Screening Property and Model Selection Consistency

In the first part of this section, we prove the sure screening property of OGA under a strong sparsity condition:

(C6) $\exists L_n$ satisfies $L_n \to 0$ and $\sqrt{\frac{n}{p^{\frac{2}{q_1}}}}L_n^2 \to \infty$ as $n \to \infty$ s.t. for any $\beta_j \neq 0$, $|\beta_j| \geq (\frac{\max_{1\leq i\leq p}\sigma^2_{\eta_i}||\underset{\sim}{\beta}||_1}{\lambda_{min}(\Sigma_x)+\min_{1\leq i\leq p}\sigma^2_{\eta_i}}) + L_n$.

Theorem 2. *Assume* (C1)-(C6), (3.1) *and* $K_n = O(\sqrt{\frac{n}{p^{\frac{2}{q_1}}}})$. *Then* $\lim_{n\to\infty} P(N \subseteq \hat{J}_{K_n}) = 1$, *where* $N = \{1 \leq j \leq p : \beta_j \neq 0\}$ *denote the set of relevant input variables.*

Proof. Let $m_0 = \lfloor aL_n^{-2} \rfloor = o(K_n)$, for some positive constant a. Consider a event

$$A_n^\star(k) = \left\{ \max_{(J,i):\#(J)\leq k-1, i\notin J} |\hat{\mu}_{J,i} - \mu_{J,i}| \leq sL_n^2 \right\},$$

for some positive constant s independent of n and k. By (3.4), we have $\forall s > 0$, $\lim_{n\to\infty} P(A_n^{\star c}(K_n)) = 0$, which implies $\lim_{n\to\infty} P(A_n^{\star c}(m_0)) = 0$. So, by similar arguments in the proof of Theorem 1, $\lim_{n\to\infty} P(F_n) = 0$, where

$$F_n = \{\tfrac{1}{n}Y^T(\mathbf{w})(I - H_{\hat{J}_{m_0}})Y(\mathbf{w}) > Cm_0^{-1}\},$$

for some $C > 0$. By (C3), (C6), it follows that $\#(N) = O(1)$, yielding $\#(N \cup \hat{J}_{m_0}) = o(K_n)$. So, on $\{N \cap \hat{J}^c_{m_0} \neq \emptyset\}$, when n is large,

$$\begin{aligned}
&\tfrac{1}{n}Y^T(\mathbf{w})(I - H_{\hat{J}_{m_0}})Y(\mathbf{w}) \\
&= \tfrac{1}{n}\beta^{\star T}_{N\cap \hat{J}^c_{m_0}} \mathbf{w}^T_{N\cap \hat{J}^c_{m_0}} (I - H_{\hat{J}_{m_0}})\mathbf{w}_{N\cap \hat{J}^c_{m_0}} \beta^{\star}_{N\cap \hat{J}^c_{m_0}} \\
&\geq (\min_{j\in N} \beta_j^{\star 2}) \min_{1\leq \#(J)\leq K_n} \lambda_{min}(\hat{R}(J)) \\
&\geq bL_n^2,
\end{aligned}$$

for some $b > 0$, where $\mathbf{w}_{N\cap \hat{J}^c_{m_0}} = (\mathbf{w}_i, i \in N \cap \hat{J}^c_{m_0})$, $\beta^{\star}_{N\cap \hat{J}^c_{m_0}} = (\beta^{\star}_i, i \in N \cap \hat{J}^c_{m_0})^T$. The last inequality above follows from Lemma 3, (C6) and (3.1). By choosing a in $m_0 = \lfloor aL_n^{-2} \rfloor$ large enough, we have $bL_n^2 > Cm_0^{-1}$, and the proof of Theorem 2 is complete.

To choose the smallest number of iterations that include all relevant variables, we propose a high-dimensional information criterion (HDIC). Define $\hat{\sigma}^2_J = n^{-1}\sum_{t=1}^n (y_t - \hat{y}_{t;J})^2$, where $\hat{y}_{t;J}$ denotes the fitted value of y_t when $Y = (y_1, y_2, ..., y_n)^T$ is projected into the linear space spanned by $\mathbf{w}_j$, $j \in J \neq \emptyset$, setting $\hat{y}_{t;J} = 0$ if $J = \emptyset$. Let

$$\begin{aligned}
&\text{HDIC}(J) = n\log \hat{\sigma}^2_J + \#(J) w_n p^{\frac{2}{q_1}}, \\
&\hat{k}_n = \arg\min_{1\leq k\leq K_n} \text{HDIC}(\hat{J}_k), \\
&w_n \to \infty, w_n p^{\frac{2}{q_1}} = o(n{L_n}^4), \qquad (4.1)
\end{aligned}$$

$$\tilde{k}_n = \min\{k : 1 \leq k \leq K_n, N \subseteq \hat{J}_k\}(\min \emptyset = K_n).$$

Note that $\hat{k}_n$ is the number of OGA iterations we choose according to HDIC, and $\tilde{k}_n$ is the minimal number of iterations that includes all relevant regressors along an OGA path.

To achieve consistency of model selection under (C6), the strong sparsity condition, we need to assume the contaminated regression coefficients converges to the uncontaminated regression coefficients in an appropriate rate, which means the measurement errors must converges to 0 in probability with some rate:

(C7) $||\underset{\sim}{U}||_1 = O(\sqrt{\frac{p^{\frac{2}{q_1}}}{n}})$,

note that $\max\limits_{1\leq i\leq p} \sigma^2_{\eta_i} = O(\frac{1}{\sqrt{p}}\sqrt{\frac{p^{\frac{2}{q_1}}}{n}})$ assures (C7). If the regressors are uncorrelated, then $\max\limits_{1\leq i\leq p} \sigma^2_{\eta_i} = O(\sqrt{\frac{p^{\frac{2}{q_1}}}{n}})$ assures (C7), which is weaker than general conditions.

In addition, we assume a weak dependency on the square of regression errors:

(C8) $\max\limits_{1\leq t\leq n} E(\xi_t^4) = O(1)$ and $E(\xi_t^2\xi_{t+h}^2) - \sigma_\xi^4 = o(1)$ as $h \to \infty$,
where $\sigma_\xi^2 = E(\xi_t^2),\ \forall t = 1, 2, ..., n.$

This assumption is used to derive weak law of large numbers of ξ_t^2. The following theorem proves that $\hat{k}$ approaches $\tilde{k}$ when n grows in probability sense.

Theorem 3. With the same notation and assumptions as in Theorem 2, suppose (3.1), (C7) and (C8) holds, $K_n = O(\sqrt{\frac{n}{p^{\frac{2}{q_1}}}})$. Then $\lim\limits_{n\to\infty} P(\hat{k}_n = \tilde{k}_n) = 1$.

Proof. For notational simplicity, dropping the subscript n in $\tilde{k}_n$ and $\hat{k}_n$. Let $D_n = \{N \subseteq \hat{J}_{m_0}\} = \{\tilde{k} \leq m_0\}$, by Theorem 2, $\lim\limits_{n\to\infty} P(D_n) = 1$. On $\{\hat{k} < \tilde{k}\}$, by definition of $\hat{k}$, it follows that

$$\exp(\text{HDIC}(\hat{J}_{\hat{k}})/n) \leq \exp(\text{HDIC}(\hat{J}_{\tilde{k}})/n),$$

so,

$$\hat{\sigma}^2_{\hat{J}_{\tilde{k}-1}} - \hat{\sigma}^2_{\hat{J}_{\tilde{k}}} \leq \hat{\sigma}^2_{\hat{J}_{\hat{k}}} - \hat{\sigma}^2_{\hat{J}_{\tilde{k}}} \leq \hat{\sigma}^2_{\hat{J}_{\tilde{k}}} \{\exp(n^{-1} w_n \tilde{k} p^{\frac{2}{q_1}}) - \exp(n^{-1} w_n \hat{k} p^{\frac{2}{q_1}})\}. \quad (4.2)$$

Note that

$$\begin{aligned}
& n^{-1}\{\textstyle\sum_{t=1}^n (y_t - \hat{y}_{t;\hat{J}_{\tilde{k}-1}})^2 - \sum_{t=1}^n (y_t - \hat{y}_{t;\hat{J}_{\tilde{k}}})^2\} \\
&= n^{-1}(\beta^\star_{\hat{j}_{\tilde{k}}} \mathbf{w}_{\hat{j}_{\tilde{k}}} + \textstyle\sum_{l \notin \hat{J}_{\tilde{k}}} \beta^\star_l \mathbf{w}_l + \boldsymbol{\xi}^\star)^T (H_{\hat{J}_{\tilde{k}}} - H_{\hat{J}_{\tilde{k}-1}})(\beta^\star_{\hat{j}_{\tilde{k}}} \mathbf{w}_{\hat{j}_{\tilde{k}}} + \sum_{l \notin \hat{J}_{\tilde{k}}} \beta^\star_l \mathbf{w}_l + \boldsymbol{\xi}^\star) \\
&= \tfrac{\{\beta^\star_{\hat{j}_{\tilde{k}}} \mathbf{w}^T_{\hat{j}_{\tilde{k}}} (I - H_{\hat{J}_{\tilde{k}-1}}) \mathbf{w}_{\hat{j}_{\tilde{k}}} + \mathbf{w}^T_{\hat{j}_{\tilde{k}}} (I - H_{\hat{J}_{\tilde{k}-1}}) \boldsymbol{\xi}^\star\}^2}{n \mathbf{w}^T_{\hat{j}_{\tilde{k}}} (I - H_{\hat{J}_{\tilde{k}-1}}) \mathbf{w}_{\hat{j}_{\tilde{k}}}} \\
& + n^{-1}(\textstyle\sum_{l \notin \hat{J}_{\tilde{k}}} \beta^\star_l \mathbf{w}_l)^T (H_{\hat{J}_{\tilde{k}}} - H_{\hat{J}_{\tilde{k}-1}})(\sum_{l \notin \hat{J}_{\tilde{k}}} \beta^\star_l \mathbf{w}_l) \\
& + 2n^{-1}(\textstyle\sum_{l \notin \hat{J}_{\tilde{k}}} \beta^\star_l \mathbf{w}_l)^T (H_{\hat{J}_{\tilde{k}}} - H_{\hat{J}_{\tilde{k}-1}})(\beta^\star_{\hat{j}_{\tilde{k}}} \mathbf{w}_{\hat{j}_{\tilde{k}}} + \boldsymbol{\xi}^\star).
\end{aligned}$$

By (4.2),

$$\begin{aligned}
& \beta^{\star 2}_{\hat{j}_{\tilde{k}}} \hat{A}_n + 2\beta^\star_{\hat{j}_{\tilde{k}}} \hat{B}_n + \hat{A}_n^{-1} \hat{B}_n^2 + \hat{D}_n + 2\hat{E}_n \\
& \leq \lambda n^{-1} w_n p^{\frac{2}{q_1}} m_0 (\hat{C}_n + \sigma^2_{\xi^\star}) \text{ on } \{\hat{k} < \tilde{k}\} \bigcap D_n, \qquad (4.3)
\end{aligned}$$

for some $\lambda > 0$, where

$$\begin{aligned}
\hat{A}_n &= n^{-1} \mathbf{w}^T_{\hat{j}_{\tilde{k}}} (I - H_{\hat{J}_{\tilde{k}-1}}) \mathbf{w}_{\hat{j}_{\tilde{k}}} \\
\hat{B}_n &= n^{-1} \mathbf{w}^T_{\hat{j}_{\tilde{k}}} (I - H_{\hat{J}_{\tilde{k}-1}}) \boldsymbol{\xi}^\star \\
\hat{C}_n &= \hat{\sigma}^2_{\hat{J}_{\tilde{k}}} - \sigma^2_{\xi^\star} \\
\hat{D}_n &= n^{-1}(\textstyle\sum_{l \notin \hat{J}_{\tilde{k}}} \beta^\star_l \mathbf{w}_l)^T (H_{\hat{J}_{\tilde{k}}} - H_{\hat{J}_{\tilde{k}-1}})(\sum_{l \notin \hat{J}_{\tilde{k}}} \beta^\star_l \mathbf{w}_l) \\
\hat{E}_n &= n^{-1}(\textstyle\sum_{l \notin \hat{J}_{\tilde{k}}} \beta^\star_l \mathbf{w}_l)^T (H_{\hat{J}_{\tilde{k}}} - H_{\hat{J}_{\tilde{k}-1}})(\beta^\star_{\hat{j}_{\tilde{k}}} \mathbf{w}_{\hat{j}_{\tilde{k}}} + \boldsymbol{\xi}^\star).
\end{aligned}$$

In the Appendix, it is shown that $\forall \theta > 0$,

$$P(\hat{A}_n < \frac{v_n}{2}, D_n) + P(|\hat{B}_n| \geq \theta L_n, D_n) + P(|\hat{C}_n| \geq \theta, D_n) + P(|\hat{E}_n| \geq \theta L_n^2, D_n) = o(1), \qquad (4.4)$$

where $v_n = \min_{1 \leq \#(J) \leq m_0} \lambda_{min}(R(J))$. From (4.3), (4.4), (C6), $\lim\limits_{n \to \infty} P(D_n) = 1$, $\hat{D}_n$, $\hat{A}_n^{-1} \hat{B}_n^2 \geq 0$, and θ is arbitrary, it follows that $P(\hat{k} < \tilde{k}) = o(1)$.

On $\{\hat{k} > \tilde{k}\}$, by definition of $\hat{k}$, it follows that

doi:10.6342/NTU201700925

$$\hat{\sigma}^2_{\hat{J}_{\hat{k}}} \exp(n^{-1} w_n \hat{k} p^{\frac{2}{q_1}}) \leq \hat{\sigma}^2_{\hat{J}_{\tilde{k}}} \exp(n^{-1} w_n \tilde{k} p^{\frac{2}{q_1}}),$$

so, it can be derived that

$$\begin{aligned}
&\boldsymbol{\xi}^{\star T}(H_{\hat{J}_{\hat{k}}} - H_{\hat{J}_{\tilde{k}}})\boldsymbol{\xi}^{\star} \\
&+(\textstyle\sum_{l \notin \hat{J}_{\tilde{k}}} \beta_l^{\star}\mathbf{w}_l)^T (H_{\hat{J}_{\hat{k}}} - H_{\hat{J}_{\tilde{k}}})(\textstyle\sum_{l \notin \hat{J}_{\tilde{k}}} \beta_l^{\star}\mathbf{w}_l) \\
&+2(\textstyle\sum_{l \notin \hat{J}_{\tilde{k}}} \beta_l^{\star}\mathbf{w}_l)^T (H_{\hat{J}_{\hat{k}}} - H_{\hat{J}_{\tilde{k}}})\boldsymbol{\xi}^{\star} \\
&\geq \{\boldsymbol{\xi}^{\star T}(I - H_{\hat{J}_{\tilde{k}}})\boldsymbol{\xi}^{\star} \\
&+(\textstyle\sum_{l \notin \hat{J}_{\tilde{k}}} \beta_l^{\star}\mathbf{w}_l)^T (I - H_{\hat{J}_{\tilde{k}}})(\textstyle\sum_{l \notin \hat{J}_{\tilde{k}}} \beta_l^{\star}\mathbf{w}_l) \\
&+2(\textstyle\sum_{l \notin \hat{J}_{\tilde{k}}} \beta_l^{\star}\mathbf{w}_l)^T (I - H_{\hat{J}_{\tilde{k}}})\boldsymbol{\xi}^{\star}\} \\
&\times(1 - \exp(-n^{-1} w_n (\hat{k} - \tilde{k}) p^{\frac{2}{q_1}})). \qquad (4.5)
\end{aligned}$$

Let $F_{\hat{k},\tilde{k}}$ denote the $n \times (\hat{k} - \tilde{k})$ matrix whose column vectors are $\mathbf{w}_j$, $j \in \hat{J}_{\hat{k}} - \hat{J}_{\tilde{k}}$. Since

$$\begin{aligned}
&\boldsymbol{\xi}^{\star T}(H_{\hat{J}_{\hat{k}}} - H_{\hat{J}_{\tilde{k}}})\boldsymbol{\xi}^{\star} \\
&= \boldsymbol{\xi}^{\star T}(I - H_{\hat{J}_{\tilde{k}}}) F_{\hat{k},\tilde{k}} \{F^T_{\hat{k},\tilde{k}}(I - H_{\hat{J}_{\tilde{k}}}) F_{\hat{k},\tilde{k}}\}^{-1} F^T_{\hat{k},\tilde{k}}(I - H_{\hat{J}_{\tilde{k}}})\boldsymbol{\xi}^{\star} \\
&\leq ||\hat{R}^{-1}(\hat{J}_{K_n})|| || n^{-\frac{1}{2}} F^T_{\hat{k},\tilde{k}}(I - H_{\hat{J}_{\tilde{k}}})\boldsymbol{\xi}^{\star}||^2 \\
&\leq ||\hat{R}^{-1}(\hat{J}_{K_n})||(2||n^{-\frac{1}{2}} F^T_{\hat{k},\tilde{k}}\boldsymbol{\xi}^{\star}||^2 + 2||n^{-\frac{1}{2}} F^T_{\hat{k},\tilde{k}} H_{\hat{J}_{\tilde{k}}}\boldsymbol{\xi}^{\star}||^2) \\
&\leq 2(\hat{k} - \tilde{k})(\hat{a}_n + \hat{b}_n),
\end{aligned}$$

where

$$\begin{aligned}
\hat{a}_n &= ||\hat{R}^{-1}(\hat{J}_{K_n})|| \max_{1 \leq i \leq p} (n^{-\frac{1}{2}} \textstyle\sum_{t=1}^n w_{ti}\xi_t^{\star})^2, \\
& \qquad (4.6) \\
\hat{b}_n &= ||\hat{R}^{-1}(\hat{J}_{K_n})|| \max_{1 \leq \#(J) \leq \tilde{k}, i \notin J} (n^{-\frac{1}{2}} \textstyle\sum_{t=1}^n w_{\widehat{ti};J}\xi_t^{\star})^2,
\end{aligned}$$

and it is shown in the Appendix that

$$
\begin{aligned}
&P((\hat{k}-\tilde{k})(\hat{a}_n+\hat{b}_n) \geq \theta n(1-\exp(-n^{-1}w_n(\hat{k}-\tilde{k})p^{\frac{2}{q_1}})), \hat{k}>\tilde{k}) \\
&+P((\textstyle\sum_{l\notin \hat{J}_{\tilde{k}}} \beta_l^\star \mathbf{w}_l)^T(H_{\hat{J}_{\hat{k}}}-H_{\hat{J}_{\tilde{k}}})(\textstyle\sum_{l\notin \hat{J}_{\hat{k}}} \beta_l^\star \mathbf{w}_l) \geq \theta n(1-\exp(-n^{-1}w_n(\hat{k}-\tilde{k})p^{\frac{2}{q_1}})), \hat{k}> \\
&\quad \tilde{k}) \\
&+P(|(\textstyle\sum_{l\notin \hat{J}_{\tilde{k}}} \beta_l^\star \mathbf{w}_l)^T(H_{\hat{J}_{\hat{k}}}-H_{\hat{J}_{\tilde{k}}})\boldsymbol{\xi}^\star| \geq \theta n(1-\exp(-n^{-1}w_n(\hat{k}-\tilde{k})p^{\frac{2}{q_1}})), \hat{k}>\tilde{k}) \\
&+P((\textstyle\sum_{l\notin \hat{j}_{\tilde{k}}} \beta_l^\star \mathbf{w}_l)^T(I-H_{\hat{J}_{\tilde{k}}})(\textstyle\sum_{l\notin \hat{j}_{\tilde{k}}} \beta_l^\star \mathbf{w}_l) \geq \theta n, \hat{k}>\tilde{k}) \\
&+P(|(\textstyle\sum_{l\notin \hat{j}_{\tilde{k}}} \beta_l^\star \mathbf{w}_l)^T(I-H_{\hat{J}_{\tilde{k}}})\boldsymbol{\xi}^\star| \geq \theta n, \hat{k}>\tilde{k}) \\
&= o(1). \qquad (4.7)
\end{aligned}
$$

So, by (A.9), (4.5), (4.7), it follows that $P(\hat{k}>\tilde{k})=o(1)$, and the proof of Theorem 3 is complete.

Even the true model will be included by OGA+HDIC, some redundant variables could be contained. So, we provide a trimming method to trim out redundant variables, Let

$$\hat{N}=\{\hat{j}_l : \mathrm{HDIC}(\hat{J}_{\hat{k}}-\{\hat{j}_l\}) > \mathrm{HDIC}(\hat{J}_{\hat{k}}), 1\leq l\leq \hat{k}\} \text{ if } \hat{k}>1,$$

and $\hat{N}=\{\hat{j}_1\}$ if $\hat{k}=1$. $\hat{N}$ is the subset of $\hat{J}_{\hat{k}}$ after trimming. The following theorem shows that OGA+HDIC+Trim will achieve the oracle property.

Theorem 4. *Under the same assumption as in Theorem* 3, $\lim\limits_{n\to\infty} P(\hat{N}=N)=1$.

Proof. For $\hat{k}>1$, define $\delta_l=1$ if HDIC$(\hat{J}_{\hat{k}}-\{\hat{j}_l\})$ >HDIC$(\hat{J}_{\hat{k}})$ and $\delta_l=0$ otherwise. Then

$$\begin{aligned} P(\hat{N} \neq N) &\leq P(\hat{N} \neq N, \hat{k} > 1, N \subseteq \hat{J}_{\hat{k}}) + P(N \nsubseteq \hat{J}_{\hat{k}}) + P(\hat{N} \neq N, \hat{k} = 1) \\ &\leq P(\delta_l = 1 \text{ and } \beta_{\hat{j}_l} = 0 \text{ for some } 1 \leq l \leq \tilde{k}, N \subseteq \hat{J}_{\tilde{k}}, \tilde{k} > 1) \\ &+P(\delta_l = 0 \text{ and } \beta_{\hat{j}_l} \neq 0 \text{ for some } 1 \leq l \leq \tilde{k}, N \subseteq \hat{J}_{\tilde{k}}, \tilde{k} > 1) \\ &+P(\hat{k} \neq \tilde{k}) + P(N \nsubseteq \hat{J}_{\hat{k}}) + P(\hat{N} \neq N, \hat{k} = 1). \end{aligned} \tag{4.8}$$

Let $\hat{J}_{\tilde{k}} - \{\hat{j}_l\} = Q_l$. On $\{\hat{k} = \tilde{k}\}$, Since by similar arguments in the proof of Theorem 3, it can be derived that $\forall \theta > 0,\ 1 \leq l \leq \hat{k}$,

$$P((\tilde{a}_n + \tilde{b}_n) \geq \theta n(1 - \exp(-n^{-1} w_n(\hat{k} - \tilde{k}) p^{\frac{2}{q_1}}))) = o(1), \tag{4.9}$$

in which $\tilde{a}_n$ and $\tilde{b}_n$ are the same as $\hat{a}_n$ and $\hat{b}_n$ in (4.6) but with K_n replaced by $\tilde{k}$, and $\tilde{k}$ replaced by $\tilde{k} - 1$, and

$$P((\sum_{r \notin Q_l} \beta_r^\star \mathbf{w}_r)^T (H_{\hat{J}_{\tilde{k}}} - H_{Q_l})(\sum_{r \notin Q_l} \beta_r^\star \mathbf{w}_r) \geq \theta n(1 - \exp(-n^{-1} w_n(\hat{k} - \tilde{k}) p^{\frac{2}{q_1}}))) = o(1), \tag{4.10}$$

$$P(|(\sum_{r \notin Q_l} \beta_r^\star \mathbf{w}_r)^T (H_{\hat{J}_{\tilde{k}}} - H_{Q_l}) \boldsymbol{\xi}^\star| \geq \theta n(1 - \exp(-n^{-1} w_n(\hat{k} - \tilde{k}) p^{\frac{2}{q_1}}))) = o(1), \tag{4.11}$$

$$P(|n^{-1} \boldsymbol{\xi}^{\star T} (I - H_{Q_l}) \boldsymbol{\xi}^\star - \sigma^2_{\xi^\star}| \geq \theta) = o(1), \tag{4.12}$$

$$P((\sum_{r \notin Q_l} \beta_r^\star \mathbf{w}_r)^T (I - H_{Q_l})(\sum_{r \notin Q_l} \beta_r^\star \mathbf{w}_r) \geq \theta n) = o(1), \tag{4.13}$$

$$P(|(\sum_{r \notin Q_l} \beta_r^\star \mathbf{w}_r)^T (I - H_{Q_l}) \boldsymbol{\xi}^\star| \geq \theta n) = o(1). \tag{4.14}$$

So, by (4.9)-(4.14), it follows that

$$P(\delta_l = 1 \text{ and } \beta_{\hat{j}_l} = 0 \text{ for some } 1 \leq l \leq \tilde{k}, N \subseteq \hat{J}_{\tilde{k}}, \tilde{k} > 1) = o(1). \tag{4.15}$$

On the other hand,

$$P(|n^{-1} \mathbf{w}_{\hat{j}_l}^T (I - H_{Q_l}) \boldsymbol{\xi}^\star| \geq \theta L_n, D_n) = o(1), \tag{4.16}$$

$$P(n^{-1}\mathbf{w}_{\hat{j}_l}^T(I-H_{Q_l})\mathbf{w}_{\hat{j}_l} \le \frac{v_n}{2}) = o(1), \tag{4.17}$$

$$P(|n^{-1}(\sum_{l\notin \hat{J}_{\tilde{k}}} \beta_l^\star \mathbf{w}_l)^T(H_{\hat{J}_{\tilde{k}}} - H_{Q_l})(\beta_{\hat{j}_l}\mathbf{w}_{\hat{j}_l} + \boldsymbol{\xi}^\star)| \ge \theta L_n^2) = o(1). \tag{4.18}$$

So, by (A.9), (4.16)-(4.18) and similar arguments in the proof of Theorem 3, it follows that

$$P(\delta_l = 0 \text{ and } \beta_{\hat{j}_l} \neq 0 \text{ for some } 1 \le l \le \tilde{k}, N \subseteq \hat{J}_{\tilde{k}}, \tilde{k} > 1) = o(1). \tag{4.19}$$

Finally, by (4.8), (4.15), (4.19) and Theorem 2 and 3, we have the desired conclusion.

5 Simulation Studies

In this section, we report simulation studies of the performance of OGA+ HDBIC+Trim. These simulations consider the regression model

$$y_t^\star = \sum_{j=1}^{p'} \beta_j^\star w_{tj} + \sum_{j=p'+1}^{p} \beta_j^\star w_{tj} + \xi_t^\star,\ t = 1, 2, ..., n, \tag{5.1}$$

where $\beta_{p'+1}, \beta_{p'+2}, ..., \beta_p = 0,\ p \gg n,\ \eta_{tj}$ are i.i.d. $N(0, \sigma_\eta^2),\ \forall t = 1, 2, ..., n,\ j = 1, 2, ..., p$, and are independent of x_{tj}. ξ_t are i.i.d. $N(0, \sigma_\xi^2)$ and are independent of x_{tj}, η_{tj}. η_{yt} are i.i.d. $N(0, \sigma_{\eta_y}^2)$ and are independent of x_{tj}, η_{tj}, ξ_t

Examples 1 and 2 consider the case

$$x_{tj} = d_{tj} + \tilde{\eta}\tilde{x}_t, \tag{5.2}$$

in which $\tilde{\eta} \ge 0$ and $(d_{t1}, d_{t2}, ..., d_{tj}, \tilde{x}_t)^T,\ t = 1, 2, ..., n$ are i.i.d. normal with mean $(1, 1, ..., 1, 0)^T$ and covariance matrix $\mathbf{I}$. We standardize the variance of x_{tj} by replacing x_{tj} with $\frac{x_{tj}}{\sqrt{1+\tilde{\eta}^2}}$. Since for any $J \subset \{1, 2, ..., p\}$ and $1 \le i \le p$ with $i \notin J$,

$$\lambda_{min}(R(J)) = \frac{1}{1+\tilde{\eta}^2} + \sigma_\eta^2 > 0 \text{ and } ||R^{-1}(J)\gamma_i(J)||_1 < 1,$$

(3.1) is satisfied. Moreover, $\mathrm{Corr}(w_{ti}, w_{tj}) = \frac{\tilde{\eta}^2}{1+\tilde{\eta}^2}$ increases when $\tilde{\eta}$ grows.

Example 1. Consider (5.1) with $p' = 5$, $(\beta_1, \beta_2, ..., \beta_5) = (3, -3.5, 4, -2.8, 3.2)$, $\sigma_\xi^2 = 1$, $\sigma_{\eta_y}^2 = 0.01$ and assume that (5.2) holds. The cases $\tilde{\eta} = 0$, which means the regressors are uncorrelated, $\sigma_\eta^2 = 0.01, 0.5, 0.1$, and $(n, p) = (50, 1000), (100, 2000), (200, 4000)$ are considered here. We choose $K_n = \lfloor 5(n/p^{\frac{2}{q_1}})^{\frac{1}{2}} \rfloor$ and allow q_1 to vary between 4 and 15. We have also allowed D in $K_n = \lfloor D(n/p^{\frac{2}{q_1}})^{\frac{1}{2}} \rfloor$ to vary between 3 and 10, and the results are similar to those for $D = 5$. We perform 1000 simulations on each case. Define the mean squared prediction errors

$$\mathrm{MSPE} = \frac{1}{1000} \sum_{l=1}^{1000} (\sum_{j=1}^{p} \beta_j^\star w_{n+1}^{(l)} - \hat{y}_{n+1}^{(l)})^2$$

in which $x_{n+1,1}^{(l)}, x_{n+1,2}^{(l)}, ..., x_{n+1,p}^{(l)}$ are the regressors associated with $y_{n+1}^{(l)}$, the new outcome in the lth simulation run, and $\hat{y}_{n+1}^{(l)}$ denotes the predictor of $y_{n+1}^{(l)}$. Table 1 shows that OGA+HDBIC+Trim is very sensitive to the order of moment bounds q_1, it performs well with proper q_1, but performs poorly with improper q_1. If q_1 is too small, the penalty for the number of predictor variables in HDBIC is too large, so, OGA+HDBIC tends to be underfitting; if q_1 is too large, the penalty for the number of predictor variables in HDBIC is too small, so, OGA+HDBIC tends to be overfitting. With moderate order of moment bounds ($q_1 = 8, 10$), in the simulations for $n \geq 100$, OGA includes the 5 relevant regressors within K_n iterations for 99.9% or more of the simulations, and HDBIC+Trim identify the smallest correct model for 98% or more of the simulations.

Table1. Frequency, in 1000 simulations, of including all five relevant variables (Correct), of selecting exactly the relevant variables (E), of selecting all relevant variables and i irrelevant variables (E+i).

σ_η^2	q_1	n	p	E	E+1	E+2	E+3	E+4	E+5	Correct	MSPE
0.01	4	50	1000	0	0	0	0	0	0	0	64.02502
		100	2000	0	0	0	0	0	0	0	53.08281
		200	4000	0	0	0	0	0	0	0	55.59686
	6	50	1000	623	0	0	0	0	0	623	24.54740
		100	2000	1000	0	0	0	0	0	1000	0.15931
		200	4000	1000	0	0	0	0	0	1000	0.08096
	8	50	1000	911	18	0	0	0	0	929	4.34789
		100	2000	1000	0	0	0	0	0	1000	0.17550
		200	4000	1000	0	0	0	0	0	1000	0.08053
	10	50	1000	571	129	43	17	17	7	922	10.29011
		100	2000	983	16	1	0	0	0	1000	0.17837
		200	4000	999	1	0	0	0	0	1000	0.16207
	15	50	1000	0	0	0	0	0	0	914	14.44628
		100	2000	21	12	10	7	3	2	1000	4.70902
		200	4000	677	225	75	14	5	2	1000	0.19443
0.05	5	50	1000	0	0	0	0	0	0	0	65.54043
		100	2000	0	0	0	0	0	0	0	53.91476
		200	4000	0	0	0	0	0	0	0	47.64495
	6	50	1000	2	0	0	0	0	0	2	59.51543
		100	2000	689	0	0	0	0	0	689	16.94148
		200	4000	1000	0	0	0	0	0	1000	0.18862
	8	50	1000	816	16	2	0	0	0	834	13.14926
		100	2000	1000	0	0	0	0	0	1000	0.39365
		200	4000	1000	0	0	0	0	0	1000	0.17408
	10	50	1000	522	118	36	21	8	14	861	13.67555
		100	2000	983	16	1	0	0	0	1000	0.44005
		200	4000	998	2	0	0	0	0	1000	0.17630
	15	50	1000	0	0	0	0	0	0	854	26.64408
		100	2000	11	17	12	10	3	0	1000	10.66257
		200	4000	683	218	75	19	1	2	1000	0.43310

σ_η^2	q_1	n	p	E	E+1	E+2	E+3	E+4	E+5	Correct	MSPE
0.1	5.5	50	1000	0	0	0	0	0	0	0	59.78768
		100	2000	0	0	0	0	0	0	0	50.27432
		200	4000	28	0	0	0	0	0	28	45.76851
	6	50	1000	0	0	0	0	0	0	0	59.57889
		100	2000	20	0	0	0	0	0	20	46.85445
		200	4000	973	0	0	0	0	0	973	1.57907
	8	50	1000	507	9	0	0	0	0	516	28.21477
		100	2000	999	0	0	0	0	0	999	0.65213
		200	4000	1000	0	0	0	0	0	1000	0.31026
	10	50	1000	493	94	36	12	8	9	744	22.35131
		100	2000	987	13	0	0	0	0	1000	0.63598
		200	4000	999	1	0	0	0	0	1000	0.29356
	15	50	1000	0	0	0	0	0	0	773	43.94780
		100	2000	16	13	8	4	5	0	1000	16.68930
		200	4000	684	222	66	20	3	2	1000	0.83435

Example 2. The settings of this example are the same with Example 1, but we allow σ_η^2 to have a rate of convergence such that

$$||\underset{\sim}{U}||_1 \leq C(\sqrt{\frac{p^{\frac{2}{q_1}}}{n}}), \tag{5.3}$$

for some C varies between 0.01 and 45. Two cases are considered here:

Case 1: Consider $\tilde{\eta} = 0$, which means the regressors are uncorrelated, and let $\sigma_\eta^2 = C\sqrt{\frac{p^{\frac{2}{q_1}}}{n}}$. In this case, the inequality in (5.3) becomes an equality. Table 2 shows that OGA+HDBIC+Trim agrees with the asymptotic theory of Theorem 4. In the cases of $n = 50$, $p = 1000$, OGA can include all relevant variables at least 90% of the time if $C \leq 1$ ($\sigma_\eta^2 \leq 0.02$), furthermore, with proper order of moment bounds ($q_1 = 8$), OGA+HDBIC+Trim can identify the smallest correct model over 88% of the time. In the cases of $n \geq 100$, OGA always include all relevant variables when $C \leq 5$ ($\sigma_\eta^2 \leq 0.085$), furthermore, when $q_1 = 8, 10$, HDBIC+Trim identifies the smallest correct model at least 98% of the time. In the case of $n = 200$, $p = 4000$, $q_1 = 12$, even if $\sigma_\eta^2 = 0.625$, which is 62.5% of the variance of the real input variables, OGA+HDBIC+Trim could still identify the smallest correct model for 91.5% of the time. Since the penalty term of each number of predictor variables in HDBIC is $\log(n)p^{\frac{2}{q_1}}$, when n is small, a small q_1 is appropriate to prevent from being overfitting; When n is large, a larger q_1 can be tolerated without being seriously overfitting.

Case 2: Consider $\tilde{\eta} = 2$, which means the regressors are highly correlated (80%), and let $\sigma_\eta^2 = C\frac{1}{\sqrt{p}}\sqrt{\frac{p^{\frac{2}{q_1}}}{n}}$, which implies (5.3). Table 3 shows that in the cases of $n = 50,\ p = 1000$, the performance of OGA is getting worse with the ratio of including all relevant variables decreases to about $50 \sim 60\%$ of the time when $C = 0.01$ (σ_η^2 is about 0.0001) due to the highly correlatedness of the regressors. However, when $n \geq 100,\ q_1 = 10, 12,\ C \leq 5$ ($\sigma_\eta^2 \leq 0.024$), OGA can include all relevant variables for 98% or more of the time, and HDBIC+Trim identifies the smallest correct model for 80% or more of the time. In the case of $n = 200,\ p = 4000,\ q_1 = 10,\ C = 35$ (σ_η^2 is about 0.09), HDBIC+Trim can identify the smallest correct model for 85% of the time.

doi:10.6342/NTU201700925

Table2. Case 1 in Example 2, with $\underset{\sim}{\eta} = 0$ and $\sigma_{\eta}^{2} = C\sqrt{\frac{p^{\frac{2}{q_1}}}{n}}$. The other notations are the same in Table 1.

q_1	C	n	p	E	E+1	E+2	E+3	E+4	E+5	Correct	MSPE	σ_{η}^{2}
8	0.01	50	1000	913	13	0	0	0	0	926	6.28107	0.00020
		100	2000	999	1	0	0	0	0	1000	0.11205	0.00016
		200	4000	1000	0	0	0	0	0	1000	0.05414	0.00012
	1	50	1000	887	13	0	0	0	0	900	7.44817	0.02075
		100	2000	1000	0	0	0	0	0	1000	0.18092	0.01592
		200	4000	1000	0	0	0	0	0	1000	0.08328	0.01223
	5	50	1000	459	8	1	0	0	0	468	31.75064	0.11312
		100	2000	999	1	0	0	0	0	1000	0.53822	0.08503
		200	4000	1000	0	0	0	0	0	1000	0.20888	0.06431
	20	50	1000	0	0	0	0	0	0	0	45.31566	0.68492
		100	2000	6	0	0	0	0	0	6	34.84350	0.45657
		200	4000	963	0	0	0	0	0	963	1.03303	0.31875
10	0.01	50	1000	569	125	46	24	20	16	920	9.10107	0.00017
		100	2000	984	16	0	0	0	0	1000	0.10932	0.00013
		200	4000	1000	0	0	0	0	0	1000	0.05647	0.00010
	1	50	1000	559	130	59	26	11	9	911	8.49444	0.01740
		100	2000	980	19	1	0	0	0	1000	0.19531	0.01313
		200	4000	997	3	0	0	0	0	1000	0.07708	0.00992
	5	50	1000	493	101	35	18	10	8	763	18.06490	0.09350
		100	2000	983	16	1	0	0	0	1000	0.50227	0.06929
		200	4000	999	1	0	0	0	0	1000	0.19478	0.05165
	35	50	1000	0	0	0	0	0	0	0	42.26501	1.49096
		100	2000	13	2	0	0	0	0	15	28.75040	0.83021
		200	4000	901	0	0	0	0	0	901	1.74922	0.52387
12	0.01	50	1000	6	2	1	0	1	1	932	14.44258	0.00015
		100	2000	789	151	37	16	4	1	1000	0.21749	0.00011
		200	4000	970	28	2	0	0	0	1000	0.05420	0.00009
	1	50	1000	0	4	1	1	0	0	908	17.25881	0.01548
		100	2000	781	170	35	5	4	1	1000	0.39322	0.01155
		200	4000	973	25	1	1	0	0	1000	0.08424	0.00863
	5	50	1000	3	3	1	0	2	8	777	41.15342	0.08249
		100	2000	798	147	39	7	3	2	1000	0.81430	0.06055
		200	4000	999	1	0	0	0	0	1000	0.19478	0.05165
	45	50	1000	0	0	0	0	0	0	0	119.84890	2.18342
		100	2000	14	3	1	0	0	0	18	26.23490	1.05687
		200	4000	915	25	2	0	0	0	942	1.53381	0.62584

Table3. Case 2 in Example 2, with $\underset{\sim}{\eta} = 2$ and $\sigma_\eta^2 = C\frac{1}{\sqrt{p}}\sqrt{\frac{p^{\frac{2}{q_1}}}{n}}$. The other notations are the same in Table 1.

q_1	C	n	p	E	E+1	E+2	E+3	E+4	E+5	Correct	MSPE	σ_η^2
8	0.01	50	1000	507	3	3	0	0	0	513	4.75287	0.00011
		100	2000	1000	0	0	0	0	0	1000	0.04985	0.00006
		200	4000	1000	0	0	0	0	0	1000	0.02725	0.00003
	1	50	1000	214	3	0	0	0	0	217	7.69615	0.01061
		100	2000	994	0	0	0	0	0	994	0.09892	0.00578
		200	4000	1000	0	0	0	0	0	1000	0.03679	0.00315
	5	50	1000	2	0	0	0	0	0	2	11.83424	0.05303
		100	2000	711	0	0	0	0	0	711	0.45313	0.02891
		200	4000	1000	0	0	0	0	0	1000	0.08263	0.01576
	15	50	1000	0	0	0	0	0	0	0	15.98425	0.15908
		100	2000	2	0	0	0	0	0	2	9.95848	0.08674
		200	4000	971	0	0	0	0	0	971	0.48686	0.04729
10	0.01	50	1000	466	77	21	21	3	4	610	3.65043	0.00010
		100	2000	990	10	0	0	0	0	1000	0.05153	0.00005
		200	4000	1000	0	0	0	0	0	1000	0.02699	0.00003
	1	50	1000	343	43	19	6	2	3	429	5.37950	0.00892
		100	2000	982	17	0	0	0	0	999	0.07926	0.00478
		200	4000	997	3	0	0	0	0	1000	0.02730	0.00256
	5	50	1000	52	8	2	1	0	0	67	9.36974	0.04462
		100	2000	958	21	1	0	0	0	980	0.33916	0.02391
		200	4000	1000	1	0	0	0	0	1000	0.07646	0.01281
	35	50	1000	0	0	0	0	0	0	0	20.92983	0.31231
		100	2000	1	0	0	0	0	0	1	13.52782	0.16736
		200	4000	851	8	0	0	0	0	859	1.35893	0.08969
12	0.01	50	1000	64	20	13	7	10	14	573	5.72614	0.00008
		100	2000	834	122	32	6	5	0	999	0.09410	0.00004
		200	4000	980	19	1	0	0	0	1000	0.02973	0.00002
	1	50	1000	59	17	14	9	8	1	466	7.44604	0.00795
		100	2000	855	105	28	7	4	0	1000	0.10421	0.00421
		200	4000	978	22	0	0	0	0	1000	0.03700	0.00223
	5	50	1000	12	10	2	2	3	0	97	12.88040	0.03976
		100	2000	803	137	35	10	4	0	993	0.31848	0.02106
		200	4000	969	30	1	0	0	0	1000	0.07507	0.01116
	40	50	1000	0	0	0	0	0	0	0	26.24179	0.31811
		100	2000	7	7	0	0	0	0	14	11.19198	0.16851
		200	4000	816	144	2	0	0	0	962	1.10595	0.08927

doi:10.6342/NTU201700925

References

Abhirup Datta and Hui Zou. (2016). CoCoLasso for High-dimensional Error-in-variables Regression. https://arxiv.org/abs/1510.07123.

Alexandre Belloni, Mathieu Rosenbaum and Alexandre B. Tsybakov. (2014). An $\{\ell_1,\ \ell_2,\ \ell_\infty\}$-Regularization Approach to High-Dimensional Errors-in-variables Models. https://arxiv.org/abs/1412.7216.

Alexandre Belloni, Mathieu Rosenbaum and Alexandre B. Tsybakov. (2016). Linear and Conic Programming Estimators in High-Dimensional Errors-in-variables Models. https://arxiv.org/abs/1408.0241.

Ching-Kang Ing and Tze Leung Lai. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. Statist. Sinica, 1473-1513.

Ching-Kang Ing and Kunling Huang. (2016). Model Selection for High-Dimensional Multivariate Time Dependent Models (Unpublished master's thesis). National Taiwan University, Taipei City.

C. Z. Wei. (1987). Adaptive prediction by least squares predictors in stochastic regression models with applications to time series. Ann. Statist. 15(4):1667-1682.

David F. Findley and Ching-Zong Wei. (1993). Moment bounds for deriving time series CLTs and model selection procedures. Statist. Sinica, 453-480.

Po-Ling Loh and Martin J. Wainwright. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. Ann. Statist. 40(3):1637 -1664.

Temlyakov, V. N. (2000). Weak greedy algorithms. Adv. Comput. Math. 12, 213-227.

doi:10.6342/NTU201700925

Appendix

Before we prove (3.4), three lemmas are needed first:

Lemma 1. *Assume* (C2), $\max\limits_{1\le i,j\le p}\frac{1}{n}|\sum_{t=1}^{n}(w_{ti}w_{tj}-\sigma_{ij})| = o_p(\sqrt{\frac{p^{\frac{2}{q_1}}}{n}})$ *and*

$$\max_{1\le i,j\le p}\frac{1}{n}|\sum_{t=1}^{n}(x_{ti}x_{tj}-\sigma_{xij})| = o_p(\sqrt{\frac{p^{\frac{2}{q_1}}}{n}}).$$

Lemma 2. *Assume* (C1)-(C4), $\max\limits_{1\le i\le p}\frac{1}{n}|\sum_{t=1}^{n}\xi_t^{\star}w_{ti}| = O_p(\sqrt{\frac{p^{\frac{2}{q_1}}}{n}})$.

Lemma 3. *Assume* (C1)-(C4), (3.1) *and* $K_n = O(\sqrt{\frac{n}{p^{\frac{2}{q_1}}}})$,

$$\max_{1\le\#(J)\le K_n}||\hat{R}(J)-R(J)|| = o_p(1), \qquad (A.1)$$

$$\max_{1\le\#(J)\le K_n}||\hat{R}^{-1}(J)-R^{-1}(J)|| = o_p(1), \qquad (A.2)$$

$$\max_{1\le\#(J)\le K_n}||\hat{R}^{-1}(J)||-\delta^{-1} = o_p(1). \qquad (A.3)$$

Proof of Lemma 1. Given $M>0$,

$$\begin{aligned}
&P(\max_{1\le i,j\le p}\frac{1}{n}|\sum_{t=1}^{n}(w_{ti}w_{tj}-\sigma_{ij})| > M\sqrt{\frac{p^{\frac{2}{q_2}}}{n}})\\
&\le M^{-2q_2}n^{q_2}p^{-2}E(\max_{1\le i,j\le p}\frac{1}{n}|\sum_{t=1}^{n}(w_{ti}w_{tj}-\sigma_{ij})|)^{2q_2}\\
&\le M^{-2q_2}\max_{1\le i,j\le p}E(\frac{1}{\sqrt{n}}|\sum_{t=1}^{n}(w_{ti}w_{tj}-\sigma_{ij})|)^{2q_2},
\end{aligned}$$

it follows from (C2) that

$$\begin{aligned}
\max_{1\le i,j\le p}\frac{1}{n}|\sum_{t=1}^{n}(w_{ti}w_{tj}-\sigma_{ij})| &= O_p(\sqrt{\frac{p^{\frac{2}{q_2}}}{n}})\\
&= o_p(\sqrt{\frac{p^{\frac{2}{q_1}}}{n}}).
\end{aligned}$$

doi:10.6342/NTU201700925

Similarly,

$$\max_{1\le i,j\le p}\frac{1}{n}\Big|\sum_{t=1}^{n}(x_{ti}x_{tj}-\sigma_{xij})\Big| = o_p\Big(\sqrt{\frac{p^{\frac{2}{q_1}}}{n}}\Big).$$

Proof of Lemma 2. Note that

$$\begin{aligned}
&E(\max_{1\le i\le p}\tfrac{1}{n}|\textstyle\sum_{t=1}^{n}\xi_t^{\star}w_{ti}|)\\
&= E(\max_{1\le i\le p}\tfrac{1}{n}|\textstyle\sum_{t=1}^{n}[\xi_t-\underset{\sim}{\beta}^T(\underset{\sim}{\eta}_t-\Sigma_\eta(\Sigma_x+\Sigma_\eta)^{-1}(\underset{\sim}{x}_t+\underset{\sim}{\eta}_t))](x_{ti}+\eta_{ti})|)\\
&\le E(\max_{1\le i\le p}\tfrac{1}{n}|\textstyle\sum_{t=1}^{n}\xi_t w_{ti}|)\\
&\ + E(\max_{1\le i\le p}\tfrac{1}{n}|\textstyle\sum_{t=1}^{n}\underset{\sim}{\beta}^T\Sigma_x(\Sigma_x+\Sigma_\eta)^{-1}\underset{\sim}{\eta}_t x_{ti}|)\\
&\ + E(\max_{1\le i\le p}\tfrac{1}{n}|\textstyle\sum_{t=1}^{n}\underset{\sim}{\beta}^T\Sigma_\eta(\Sigma_x+\Sigma_\eta)^{-1}\underset{\sim}{x}_t\eta_{ti}|)\\
&\ + E(\max_{1\le i\le p}\tfrac{1}{n}|\textstyle\sum_{t=1}^{n}\underset{\sim}{\beta}^T\Sigma_\eta(\Sigma_x+\Sigma_\eta)^{-1}\underset{\sim}{x}_t x_{ti}-\underset{\sim}{\beta}^T\Sigma_x(\Sigma_x+\Sigma_\eta)^{-1}\underset{\sim}{\eta}_t\eta_{ti}|)\\
&:= E_1+E_2+E_3+E_4.
\end{aligned}$$

Since

$$\begin{aligned}
&E(\max_{1\le i\le p}\tfrac{1}{n}|\textstyle\sum_{t=1}^{n}\xi_t w_{ti}|)^{q_1}\\
&\le p\max_{1\le i\le p}E(\tfrac{1}{n}|\textstyle\sum_{t=1}^{n}\xi_t w_{ti}|)^{q_1}\\
&\le p\max_{1\le i\le p}n^{-q_1}E(\sup_{1\le k\le n}|\textstyle\sum_{t=1}^{k}\xi_t w_{ti}|)^{q_1}\\
&\le p\max_{1\le i\le p}n^{-q_1}KE(\textstyle\sum_{t=1}^{n}w_{ti}^2)^{\frac{q_1}{2}}\\
&= p\max_{1\le i\le p}n^{-\frac{q_1}{2}}KE(\tfrac{1}{n}\textstyle\sum_{t=1}^{n}w_{ti}^2)^{\frac{q_1}{2}}\\
&\le pn^{-\frac{q_1}{2}}K\max_{1\le i\le p,1\le t\le n}E|w_{ti}|^{q_1},
\end{aligned}$$

where the third inequality comes from Lemma 2 in Wei (1987), K is a positive constant depends only on q_1 and C_2 in (C1); the last inequality comes from Jensen's inequality. From (C1), it follows that

$$E_1 = O(\sqrt{\frac{p^{\frac{2}{q_1}}}{n}}). \qquad (A.4)$$

Since

$$\begin{aligned} E_2 &= E(\max_{1\le i\le p} |\tfrac{1}{n} \textstyle\sum_{t=1}^{n} \underset{\sim}{\beta}^T [I - \Sigma_\eta(\Sigma_x + \Sigma_\eta)^{-1}] \underset{\sim}{\eta}_t x_{ti}|) \\ &\le ||\underset{\sim}{\beta}^\star||_1 E(\max_{1\le i,j\le p} |\tfrac{1}{n} \textstyle\sum_{t=1}^{n} \eta_{tj} x_{ti}|), \end{aligned}$$

by (C1) and similar arguments in the derivation of (A.4), it follows that $E(\max_{1\le i,j\le p} |\frac{1}{n} \sum_{t=1}^{n} \eta_{tj} x_{ti}|) = O(\sqrt{\frac{p^{\frac{2}{q_1}}}{n}})$, combined it with (C3) and (C4), it follows that

$$E_2 = O(\sqrt{\frac{p^{\frac{2}{q_1}}}{n}}). \qquad (A.5)$$

Similarly,

$$E_3 = O(\sqrt{\frac{p^{\frac{2}{q_1}}}{n}}). \qquad (A.6)$$

Finally,

$$\begin{aligned} E_4 &= E\left\{\max_{1\le i\le p} |\tfrac{1}{n} \textstyle\sum_{t=1}^{n} \underset{\sim}{\beta}^T [\underset{\sim}{\eta}_t \eta_{ti} - \Sigma_\eta(\Sigma_x + \Sigma_\eta)^{-1} (\underset{\sim}{x}_t x_{ti} + \underset{\sim}{\eta}_t \eta_{ti})]|\right\} \\ &= E\left\{\max_{1\le i\le p} |\tfrac{1}{n} \textstyle\sum_{t=1}^{n} \underset{\sim}{\beta}^T [\underset{\sim}{\eta}_t \eta_{ti} - E(\underset{\sim}{\eta}_t \eta_{ti}) - \Sigma_\eta(\Sigma_x + \Sigma_\eta)^{-1}\right. \\ &\quad \left. \times (\underset{\sim}{x}_t x_{ti} - E(\underset{\sim}{x}_t x_{ti}) + \underset{\sim}{\eta}_t \eta_{ti} - E(\underset{\sim}{\eta}_t \eta_{ti}))]|\right\} \\ &= E\left\{\max_{1\le i\le p} |\tfrac{1}{n} \textstyle\sum_{t=1}^{n} \underset{\sim}{\beta}^T [\underset{\sim}{w}_t w_{ti} - E(\underset{\sim}{w}_t w_{ti}) - \underset{\sim}{x}_t x_{ti} + E(\underset{\sim}{x}_t x_{ti}) - \underset{\sim}{x}_t \eta_{ti} - \underset{\sim}{\eta}_t x_{ti}\right. \\ &\quad \left. -\Sigma_x(\Sigma_0 + \Sigma_x)^{-1} (\underset{\sim}{w}_t w_{ti} - E(\underset{\sim}{w}_t w_{ti}) - \underset{\sim}{x}_t \eta_{ti} - \underset{\sim}{\eta}_t x_{ti})]|\right\} \\ &\le ||\underset{\sim}{\beta}^\star||_1 [E(\max_{1\le i,j\le p} |\tfrac{1}{n} \textstyle\sum_{t=1}^{n} (w_{ti} w_{tj} - \sigma_{ij})|) \\ &\quad +2E(\max_{1\le i,j\le p} |\tfrac{1}{n} \textstyle\sum_{t=1}^{n} \eta_{tj} x_{ti}|)] + ||\underset{\sim}{\beta}||_1 E(\max_{1\le i,j\le p} |\tfrac{1}{n} \textstyle\sum_{t=1}^{n} (x_{ti} x_{tj} - \sigma_{xij})|) \\ &= O(\sqrt{\frac{p^{\frac{2}{q_1}}}{n}}). \qquad (A.7) \end{aligned}$$

The last equality comes from Lemma 1, (A.5) , (C3) and (C4). By (A.4)-(A.7) and Markov's inequality, the proof of Lemma 2 is complete.

Proof of Lemma 3. For (A.1), note that

$$\max_{1\le\#(J)\le K_n} ||\hat{R}(J)-R(J)|| \le K_n \max_{1\le i,j\le p} \tfrac{1}{n}|\textstyle\sum_{t=1}^{n}(w_{ti}w_{tj}-\sigma_{ij})|.$$

Since $K_n = O(\sqrt{\frac{n}{p^{\frac{2}{q_1}}}})$ and $\max_{1\le i,j\le p} \frac{1}{n}|\sum_{t=1}^{n}(w_{ti}w_{tj}-\sigma_{ij})| = o_p(\sqrt{\frac{p^{\frac{2}{q_1}}}{n}})$ by Lemma 1, we have the desired conclusion. By (A7) and (A8) in Ing and Lai (2011) and (A.1), we have (A.2). Furthermore, since $\max_{1\le\#(J)\le K_n} ||R^{-1}(J)|| < \delta^{-1}$ by (3.1),

$$\max_{1\le\#(J)\le K_n} ||\hat{R}^{-1}(J)|| - \delta^{-1} \le \max_{1\le\#(J)\le K_n} ||\hat{R}^{-1}(J)-R^{-1}(J)||,$$

so, we have (A.3), and the tools for proving (3.4) are ready.

Proof of (3.4). Since by Lemma 1, $n^{\frac{1}{2}}||\mathbf{w}_i|| \xrightarrow{p} \sigma_{ii} \ge \min_{1\le i\le p}|\sigma_{ii}|$, as $n\to\infty$, $\forall i = 1,2,...,n$, it suffices to prove that $\forall\epsilon>0, \exists s>0$ s.t.

$$P(\max_{(J,i):\#(J)\le m-1, i\notin J} \tfrac{1}{n}|\textstyle\sum_{t=1}^{n}\xi_t^\star \hat{w}_{ti;J}^{\perp}| > s\sigma^\star\sqrt{\frac{p^{\frac{2}{q_1}}}{n}}) \le \epsilon,$$

where $\sigma^\star = \min_{1\le i\le p}|\sigma_{ii}|$. Notice that

$$\begin{aligned}
&\max_{(J,i):\#(J)\le m-1, i\notin J} \tfrac{1}{n}|\textstyle\sum_{t=1}^{n}\xi_t^\star \hat{w}_{ti;J}^{\perp}| \\
&\le \max_{1\le i\le p}\tfrac{1}{n}|\textstyle\sum_{t=1}^{n}\xi_t^\star w_{ti}| \\
&+ \max_{(J,i):\#(J)\le K_n-1, i\notin J} |(\tfrac{1}{n}\textstyle\sum_{t=1}^{n}\xi_t^\star w_t(J))^T\hat{R}^{-1}(J)(\tfrac{1}{n}\sum_{t=1}^{n}w_{ti;J}^{\perp}w_t(J))| \\
&+ \max_{(J,i):\#(J)\le K_n-1, i\notin J} |\gamma_i{}^T(J)R^{-1}(J)\tfrac{1}{n}\textstyle\sum_{t=1}^{n}\xi_t^\star w_t(J))| \\
&:= S_{1,n}+S_{2,n}+S_{3,n},
\end{aligned}$$

where $w^{\perp}_{ti;J} = w_{ti} - \gamma_i^T(J)R^{-1}(J)w_t(J)$. Since

$$S_{3,n} \leq \max_{1\leq i\leq p} |\frac{1}{n}\sum_{t=1}^{n} \xi_t^{\star} w_{ti}| \max_{1\leq \#(J)\leq K_n, i\notin J} ||R^{-1}(J)\gamma_i(J)||_1,$$

by Lemma 2 and (3.1), $S_{1,n}, S_{3,n}$ are $O_p(\sqrt{\frac{p^{\frac{2}{q_1}}}{n}})$. Notice that

$$S_{2,n} \leq (\max_{1\leq \#(J)\leq K_n} ||\hat{R}^{-1}(J)||)K_n(1 + \max_{1\leq \#(J)\leq K_n, i\notin J} ||R^{-1}(J)\gamma_i(J)||_1)$$
$$\times(\max_{1\leq i,j\leq p} |\tfrac{1}{n}\textstyle\sum_{t=1}^{n} w_{ti}w_{tj} - \sigma_{ij}|)(\max_{1\leq i\leq p} |\tfrac{1}{n}\textstyle\sum_{t=1}^{n} \xi_t^{\star} w_{ti}|),$$

so by Lemma 1-3, (3.1) and $K_n = O(\sqrt{\frac{n}{p^{\frac{2}{q_1}}}})$, it follows that $S_{2,n} = o_p(\sqrt{\frac{p^{\frac{2}{q_1}}}{n}})$, and the proof of (3.4) is complete.

Proof of (4.4). By the proof of (3.4), it follows that

$$\begin{aligned} P(|\hat{B}_n| \geq \theta L_n, D_n) &\leq P(\max_{1\leq \#(J)\leq m_0-1, i\notin J} |\tfrac{1}{n}\textstyle\sum_{t=1}^{n} \xi_t^{\star}\hat{w}^{\perp}_{ti;J}| \geq \theta L_n) \\ &= o(1). \qquad (A.8) \end{aligned}$$

Since

$$\begin{aligned} \hat{C}_n = {} & (n^{-1}\boldsymbol{\xi}^{\star T}\boldsymbol{\xi}^{\star} - \sigma^2_{\xi^{\star}}) + n^{-1}(\textstyle\sum_{l\notin \hat{J}_{\tilde{k}}} \beta_l^{\star}\mathbf{w}_l)^T(I - H_{\hat{J}_{\tilde{k}}})(\textstyle\sum_{l\notin \hat{J}_{\tilde{k}}} \beta_l^{\star}\mathbf{w}_l) \\ & +2n^{-1}(\textstyle\sum_{l\notin \hat{J}_{\tilde{k}}} \beta_l^{\star}\mathbf{w}_l)^T(I - H_{\hat{J}_{\tilde{k}}})\boldsymbol{\xi}^{\star} - n^{-1}\boldsymbol{\xi}^{\star T}H_{\hat{J}_{\tilde{k}}}\boldsymbol{\xi}^{\star}, \end{aligned}$$

it follows that

$$\begin{aligned} P(|\hat{C}_n| \geq \theta, D_n) \leq {} & P(|\tfrac{1}{n}\boldsymbol{\xi}^{\star T}\boldsymbol{\xi}^{\star} - \sigma^2_{\xi^{\star}}| \geq \tfrac{\theta}{4}) \\ & +P(||\underset{\sim}{U}||_1^2(\max_{1\leq i,j\leq p} \tfrac{1}{n}|\textstyle\sum_{t=1}^{n} w_{ti}w_{tj} - \sigma_{ij}| + \max_{1\leq i,j\leq p} |\sigma_{ij}|) \geq \tfrac{\theta}{4}) \\ & +P(2||\underset{\sim}{U}||_1\tfrac{1}{n} \max_{1\leq \#(J)\leq m_0-1, i\notin J} |\tfrac{1}{n}\textstyle\sum_{t=1}^{n} \xi_t^{\star}\hat{w}^{\perp}_{ti;J}| \geq \tfrac{\theta}{4}) \\ & +P(\max_{1\leq \#(J)\leq m_0} ||\hat{R}^{-1}(J)||m_0 \max_{1\leq i\leq p}(\tfrac{1}{n}\textstyle\sum_{t=1}^{n} \xi_t^{\star} w_{ti})^2 \geq \tfrac{\theta}{4}) \\ = {} & o(1), \qquad (A.9) \end{aligned}$$

the equality comes from (C1), (C7), (C8), (A.8), Lemma 1-3. Note that

$$\begin{aligned} P(\hat{A}_n < \tfrac{v_n}{2}, D_n) &\leq P(\lambda_{min}(\hat{R}(\hat{J}_{\tilde{k}})) < \tfrac{v_n}{2}, D_n) \\ &\leq P(\lambda_{min}(\hat{R}(\hat{J}_{m_0})) < \tfrac{v_n}{2}) \\ &\leq P(\max_{1 \leq \#(J) \leq m_0} ||\hat{R}(J) - R(J)|| > \tfrac{\delta}{2}) \\ &= o(1), \qquad (A.10) \end{aligned}$$

the equality comes from Lemma 3, and

$$\begin{aligned} &P(|\hat{E}_n| \geq \theta L_n^2, D_n) \\ &\leq P(||\underset{\sim}{U}||_1 \max_{1 \leq j \leq p} |\beta_j^\star|(\max_{1 \leq i,j \leq p} \tfrac{1}{n}|\textstyle\sum_{t=1}^n w_{ti} w_{tj} - \sigma_{ij}| + \max_{1 \leq i,j \leq p} |\sigma_{ij}|) \geq \tfrac{\theta}{2} L_n^2) \\ &+P(||\underset{\sim}{U}||_1 2(S_{2,n} + S_{3,n}) \geq \tfrac{\theta}{2} L_n^2) \\ &= o(1). \qquad (A.11) \end{aligned}$$

where $S_{2,n},\ S_{3,n}$ are the same as those in the proof of (3.4), the equality comes from (C1), (C3), (C7), Lemma 1 and the proof of (3.4). By (A.8)-(A.11), the proof of (4.4) is complete.

Proof of (4.7). Since ∃ a constant $\tilde{\lambda} > 0$ and $\zeta_n \to \infty$ s.t.

$$\begin{aligned} \frac{n(1 - \exp(-n^{-1} w_n (\hat{k} - \tilde{k}) p^{\frac{2}{q_1}}))}{\hat{k} - \tilde{k}} &\geq \tilde{\lambda} \min\{(np^{\frac{2}{q_1}})^{\frac{1}{2}}, w_n p^{\frac{2}{q_1}}\} \\ &= \zeta_n p^{\frac{2}{q_1}}, \qquad (A.12) \end{aligned}$$

it follows from (A.12) that

$$\begin{aligned} &P((\hat{k} - \tilde{k})(\hat{a}_n + \hat{b}_n) \geq \theta n(1 - \exp(-n^{-1} w_n (\hat{k} - \tilde{k}) p^{\frac{2}{q_1}})), \hat{k} > \tilde{k}) \\ &\leq P(||\hat{R}^{-1}(\hat{J}_{K_n})|| \max_{1 \leq i \leq p} (n^{-\frac{1}{2}} \textstyle\sum_{t=1}^n w_{ti} \xi_t^\star)^2 \geq \tfrac{\theta}{2} \zeta_n p^{\frac{2}{q_1}}) \\ &+P(||\hat{R}^{-1}(\hat{J}_{K_n})|| || n(S_{2,n} + S_{3,n})|| \geq \tfrac{\theta}{2} \zeta_n p^{\frac{2}{q_1}}) \\ &= o(1), \qquad (A.13) \end{aligned}$$

the equality comes from Lemma 2, 3 and proof of (3.4). Note that

$P((\sum_{l \notin \hat{J}_{\tilde{k}}} \beta_l^\star \mathbf{w}_l)^T (H_{\hat{J}_{\hat{k}}} - H_{\hat{J}_{\tilde{k}}})(\sum_{l \notin \hat{J}_{\tilde{k}}} \beta_l^\star \mathbf{w}_l)$

$$\geq \theta n(1-\exp(-n^{-1}w_n(\hat{k}-\tilde{k})p^{\frac{2}{q_1}})), \hat{k} > \tilde{k})$$

$$\leq P(||\underset{\sim}{U}||_1^2(\max_{1\leq i,j\leq p} \tfrac{1}{n}|\textstyle\sum_{t=1}^n w_{ti}w_{tj} - \sigma_{ij}| + \max_{1\leq i,j\leq p} |\sigma_{ij}|)$$

$$\geq \theta \min\{\sqrt{\tfrac{p^{\frac{2}{q_1}}}{n}}, n^{-1}w_n p^{\frac{2}{q_1}}\}(\hat{k}-\tilde{k}), \hat{k} > \tilde{k})$$

$$\leq P(||\underset{\sim}{U}||_1^2(\max_{1\leq i,j\leq p} \tfrac{1}{n}|\textstyle\sum_{t=1}^n w_{ti}w_{tj} - \sigma_{ij}| + \max_{1\leq i,j\leq p} |\sigma_{ij}|) \geq \theta \min\{\sqrt{\tfrac{p^{\frac{2}{q_1}}}{n}}, L_n^4\})$$

$$= o(1), \qquad (A.14)$$

and

$$P(|(\textstyle\sum_{l\notin \hat{J}_{\tilde{k}}} \beta_l^\star \mathbf{w}_l)^T(H_{\hat{J}_{\hat{k}}} - H_{\hat{J}_{\tilde{k}}})\underset{\sim}{\xi}^\star| \geq \theta n(1-\exp(-n^{-1}w_n(\hat{k}-\tilde{k})p^{\frac{2}{q_1}})), \hat{k} > \tilde{k})$$
$$\leq P(||\underset{\sim}{U}||_1 2(S_{2,n} + S_{3,n}) \geq \theta \min\{\sqrt{\tfrac{p^{\frac{2}{q_1}}}{n}}, L_n^4\})$$
$$= o(1). \qquad (A.15)$$

and similar to (A.14),(A.15), it follows that

$$P((\sum_{l\notin \hat{J}_{\tilde{k}}} \beta_l^\star \mathbf{w}_l)^T(I - H_{\hat{J}_{\tilde{k}}})(\sum_{l\notin \hat{J}_{\tilde{k}}} \beta_l^\star \mathbf{w}_l) \geq \theta n, \hat{k} > \tilde{k}) = o(1), \qquad (A.16)$$

$$P(|(\sum_{l\notin \hat{J}_{\tilde{k}}} \beta_l^\star \mathbf{w}_l)^T(I - H_{\hat{J}_{\tilde{k}}})\underset{\sim}{\xi}^\star| \geq \theta n, \hat{k} > \tilde{k}) = o(1). \qquad (A.17)$$

So, by (A.13)-(A.17), the proof of (4.7) is complete.

doi:10.6342/NTU201700925