Statistica Sinica **21** (2011), 1473-1513
doi:http://dx.doi.org/10.5705/ss.2010.081

A STEPWISE REGRESSION METHOD AND CONSISTENT MODEL SELECTION FOR HIGH-DIMENSIONAL SPARSE LINEAR MODELS

Ching-Kang Ing and Tze Leung Lai

Academia Sinica and Stanford University

Abstract: We introduce a fast stepwise regression method, called the orthogonal greedy algorithm (OGA), that selects input variables to enter a p-dimensional linear regression model (with $p \gg n$, the sample size) sequentially so that the selected variable at each step minimizes the residual sum squares. We derive the convergence rate of OGA and develop a consistent model selection procedure along the OGA path that can adjust for potential spuriousness of the greedily chosen regressors among a large number of candidate variables. The resultant regression estimate is shown to have the oracle property of being equivalent to least squares regression on an asymptotically minimal set of relevant regressors under a strong sparsity condition.

Key words and phrases: Componentwise linear regression, greedy algorithm, high-dimensional information criterion, Lasso, oracle property and inequalities, sparsity.

1. Introduction

Consider the linear regression model

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

with p predictor variables $x_{t1}, x_{t2}, \ldots, x_{tp}$ that are uncorrelated with the mean-zero random disturbances ε_t. When p is larger than n, there are computational and statistical difficulties in estimating the regression coefficients by standard regression methods. Major advances to resolve these difficulties have been made in the past decade with the introduction of L_2-boosting (Bühlmann and Yu (2003)), LARS (Efron et al. (2004)), and Lasso (Tibshirani (1996)) which has an extensive literature because much recent attention has focused on its underlying principle, namely, l_1-penalized least squares. It has also been shown that consistent estimation of the regression function

$$y(\mathbf{x}) = \alpha + \boldsymbol{\beta}^\top \mathbf{x},\ \text{where } \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top, \mathbf{x} = (x_1, \ldots, x_p)^\top, \tag{1.2}$$

is still possible under a sparsity condition on the regression coefficients. In particular, by assuming the "weak sparsity" condition that the regression coefficients are absolutely summable, Bühlmann (2006) has shown that for $p = \exp(O(n^{\xi}))$ with $0 < \xi < 1$, the conditional mean squared prediction error

$$\text{CPE} := E\{(y(\mathbf{x}) - \hat{y}_m(\mathbf{x}))^2 | y_1, \mathbf{x}_1, \ldots, y_n, \mathbf{x}_n\} \tag{1.3}$$

of the L_2-boosting predictor $\hat{y}_m(\mathbf{x})$ (defined in Section 2.1) can converge in probability to 0 if $m = m_n \to \infty$ sufficiently slowly, but there are no results on the convergence rate. The most comprehensive theory to date for high-dimensional regression methods has been developed for Lasso, and Section 3 gives a brief overview of this theory, including recent work on "oracle inequalities" for Lasso and a closely related method, the Dantzig selector.

A method that is widely used in applied regression analysis to handle a large number of input variables, albeit without Lasso's strong theoretical justification, is stepwise least squares regression which consists of (a) forward selection of input variables in a "greedy" manner so that the selected variable at each step minimizes the residual sum of squares after least squares regression is performed on it together with previously selected variables, (b) a stopping rule to terminate forward inclusion of variables, and (c) stepwise backward elimination of variables according to some criterion. In this paper we develop an asymptotic theory for a version of stepwise regression in the context of high-dimensional regression ($p \gg n$) under certain sparsity assumptions, and demonstrate its advantages in simulation studies of its finite-sample performance.

The forward stepwise component of this procedure is called the *orthogonal greedy algorithm* (OGA) or *orthogonal matching pursuit* in information theory, compressed sensing and approximation theory, which focuses on approximations in noiseless models (i.e., $\varepsilon_t = 0$ in (1.1)); see Temlyakov (2000), Tropp (2004), and Tropp and Gilbert (2007). We also develop a fast iterative procedure for updating OGA that uses componentwise linear regression similar to the L_2-boosting procedure of Bühlmann and Yu (2003) and does not require matrix inversion. Section 3 gives an oracle inequality for OGA and the rate of convergence of the squared prediction error (1.3) in which $\hat{y}_m(\cdot)$ is the OGA predictor, under the *weak sparsity* condition that $\sum_{j=1}^{p} |\beta_j|$ remains bounded as $n \to \infty$.

In Section 4, we develop a consistent model selection procedure along an OGA path under a "strong sparsity" condition that the nonzero regression coefficients satisfying the weak sparsity condition are not too small. Applying the convergence rate of OGA established in Theorem 1, we prove that, with probability approaching 1 as $n \to \infty$, the OGA path includes all relevant regressors when the number of iterations is large enough. The sharp convergence rate in

Theorem 1 also suggests the possibility of developing high-dimensional modifications of penalized model selection criteria like BIC and proving their consistency by an extension of the arguments of Hannan and Quinn (1979). We call such modification a *high-dimensional information criterion* (HDIC). This combined estimation and variable selection scheme, which we denote by OGA+HDIC, is shown in Theorem 4 to select the smallest set of all relevant variables along the OGA path with probability approaching 1 (and is therefore variable-selection consistent). We then further trim this set by making use of HDIC to come up with the minimal set of regressors under the strong sparsity condition; the oracle property of this approach is established in Theorem 5. In this connection, Section 4 also reviews recent work on variable selection in high-dimensional sparse linear models, and, in particular, the one proposed by Chen and Chen (2008) and developments in Lasso and adaptive Lasso. Section 5 presents simulation studies to illustrate the performance of OGA+HDIC and some issues raised in the review. Concluding remarks and further discussion are given in Section 6.

2. L_2-Boosting, Forward Stepwise Regression and Temlyakov's Greedy Algorithms

We begin this section by reviewing Bühlmann and Yu's (2003) L_2-boosting and then represent forward stepwise regression as an alternative L_2-boosting method. The "population versions" of these two methods are Temlyakov's (2000) pure greedy and orthogonal greedy algorithms (PGA and OGA). Replacing y_t by $y_t - \bar{y}$ and x_{tj} by $x_{tj} - \bar{x}_j$, where $\bar{x}_j = n^{-1}\sum_{t=1}^n x_{tj}$ and $\bar{y} = n^{-1}\sum_{t=1}^n y_t$, it will be assumed that $\alpha = 0$. Let $\mathbf{x}_t = (x_{t1}, \ldots, x_{tp})^\top$.

2.1. PGA iterations

Bühlmann and Yu's (2003) L_2-boosting is an iterative procedure that generates a sequence of linear approximations $\hat{y}_k(\mathbf{x})$ of the regression function (1.2) (with $\alpha = 0$), by applying componentwise linear least squares to the residuals obtained at each iteration. Initializing with $\hat{y}_0(\cdot) = 0$, it computes the residuals $U_t^{(k)} := y_t - \hat{y}_k(\mathbf{x}_t), 1 \le t \le n$, at the end of the kth iteration and chooses $x_{t,\hat{j}_{k+1}}$ on which the pseudo-responses $U_t^{(k)}$ are regressed, such that

$$\hat{j}_{k+1} = \arg\min_{1\le j\le p} \sum_{t=1}^n (U_t^{(k)} - \tilde{\beta}_j^{(k)} x_{tj})^2, \tag{2.1}$$

where $\tilde{\beta}_j^{(k)} = \sum_{t=1}^n U_t^{(k)} x_{tj} / \sum_{t=1}^n x_{tj}^2$. This yields the update

$$\hat{y}_{k+1}(\mathbf{x}) = \hat{y}_k(\mathbf{x}) + \tilde{\beta}_{\hat{j}_{k+1}}^{(k)} x_{\hat{j}_{k+1}}. \tag{2.2}$$

The procedure is then repeated until a pre-specified upper bound m on the number of iterations is reached. When the procedure stops at the mth iteration, $y(\mathbf{x})$ in (1.2) is approximated by $\hat{y}_m(\mathbf{x})$. Note that the same predictor variable can be entered at several iterations, and one can also use smaller step sizes to modify the increments as $\hat{y}_{k+1}(\mathbf{x}_t) = \hat{y}_k(\mathbf{x}_t) + \delta\tilde{\beta}^{(k)}_{\hat{j}_{k+1}} x_{t,\hat{j}_{k+1}}$, $0 < \delta \le 1$, during the iterations; see Bühlmann (2006, p.562).

2.2. Forward stepwise regression via OGA iterations

Like PGA, OGA uses the variable selector (2.1). Since $\sum_{t=1}^n (U_t^{(k)} - \tilde{\beta}_j^{(k)} x_{tj})^2 / \sum_{t=1}^n (U_t^{(k)})^2 = 1 - r_j^2$, where r_j is the correlation coefficient between x_{tj} and $U_t^{(k)}$, (2.1) chooses the predictor that is most correlated with $U_t^{(k)}$ at the kth stage. However, our implementation of OGA updates (2.2) in another way and also carries out an additional linear transformation of the vector $\mathbf{X}_{\hat{j}_{k+1}}$ to form $\mathbf{X}^{\perp}_{\hat{j}_{k+1}}$, where $\mathbf{X}_j = (x_{1j}, \ldots, x_{nj})^\top$. Our idea is to orthogonalize the predictor variables sequentially so that OLS can be computed by *componentwise* linear regression, thereby *circumventing difficulties with inverting high-dimensional matrices in the usual implementation of OLS*. With the orthogonal vectors $\mathbf{X}_{\hat{j}_1}, \mathbf{X}^{\perp}_{\hat{j}_2}, \ldots, \mathbf{X}^{\perp}_{\hat{j}_k}$ already computed in the previous stages, we can compute the projection $\hat{\mathbf{X}}_{\hat{j}_{k+1}}$ of $\mathbf{X}_{\hat{j}_{k+1}}$ into the linear space spanned by $\mathbf{X}_{\hat{j}_1}, \mathbf{X}^{\perp}_{\hat{j}_2}, \ldots, \mathbf{X}^{\perp}_{\hat{j}_k}$ by adding the k projections into the respective one-dimensional linear spaces (with each projection being componentwise linear regression of $x_{t,\hat{j}_{k+1}}$ on $x^{\perp}_{t,\hat{j}_i}$ for some $i \le k$). This also yields the residual vector $\mathbf{X}^{\perp}_{\hat{j}_{k+1}} = \mathbf{X}_{\hat{j}_{k+1}} - \hat{\mathbf{X}}_{\hat{j}_{k+1}}$. With $\mathbf{X}^{\perp}_{\hat{j}_{k+1}} = (x^{\perp}_{1,\hat{j}_{k+1}}, \ldots, x^{\perp}_{n,\hat{j}_{k+1}})^\top$ thus computed, OGA uses the following update in lieu of (2.2):

$$\hat{y}_{k+1}(\mathbf{x}_t) = \hat{y}_k(\mathbf{x}_t) + \hat{\beta}^{(k)}_{\hat{j}_{k+1}} x^{\perp}_{t,\hat{j}_{k+1}}, \tag{2.3}$$

where $\hat{\beta}^{(k)}_{\hat{j}_{k+1}} = (\sum_{t=1}^n U_t^{(k)} x^{\perp}_{t,\hat{j}_{k+1}}) / \sum_{t=1}^n (x^{\perp}_{t,\hat{j}_{k+1}})^2$.

Note that OGA is equivalent to the least squares regression of y_t on $(x_{t,\hat{j}_1}, \ldots, x_{t,\hat{j}_{k+1}})^\top$ at stage $k+1$ when it chooses the predictor $x_{t,\hat{j}_{k+1}}$ that is most correlated with $U_t^{(k)}$. By sequentially orthogonalizing the input variables, OGA preserves the attractive computational features of componentwise linear regression in PGA while replacing (2.2) by a considerably more efficient OLS update. Since $\sum_{t=1}^n U_t^{(k)} x^{\perp}_{tj} = \sum_{t=1}^n U_t^{(k)} x_{tj}$, $\tilde{\beta}_j^{(k)}$ and $\hat{\beta}_j^{(k)}$ only differ in their denominators, $\sum_{t=1}^n x_{tj}^2$ and $\sum_{t=1}^n (x^{\perp}_{tj})^2$. Note that OGA still uses $\tilde{\beta}_j^{(k)}$, which does not require computation of vector $\mathbf{X}_j^{\perp}$, for variable selection. However, because $U_t^{(k)}$ are the residuals in regressing y_t on $(x_{t,\hat{j}_1}, \ldots, x_{t,\hat{j}_k})^\top$ for OGA, the corresponding

variable selector for $\hat{j}_{k+1}$ in (2.1) can be restricted to $j \notin \{\hat{j}_1, \ldots, \hat{j}_k\}$. Therefore, unlike PGA for which the same predictor variable can be entered repeatedly, OGA excludes variables that are already precluded from further consideration in (2.1).

2.3. Population version of OGA

Let $y, z_1, \ldots, z_p$ be square integrable random variables having zero means and such that $E(z_i^2) = 1$. Let $\mathbf{z} = (z_1, \ldots, z_p)^\top$. The population version of OGA, which is a special case of Temlyakov's (2000) greedy algorithms, is an iterative scheme which chooses $j_1, j_2, \ldots$ sequentially by

$$j_{k+1} = \arg \max_{1 \le j \le p} |E(u_k z_j)|, \text{ where } u_k = y - \tilde{y}_k(\mathbf{z}), \tag{2.4}$$

and which updates $\tilde{y}_k(\mathbf{z})$ by the best linear predictor $\sum_{j \in J_{k+1}} \lambda_j z_j$ of y that minimizes $E(y - \sum_{j \in J_{k+1}} \lambda_j z_j)^2$, where $J_{k+1} = \{j_1, \ldots, j_{k+1}\}$ and $\tilde{y}_0(\mathbf{z}) = 0$.

3. An Oracle-Type Inequality and Convergence Rates under Weak Sparsity

In the first part of this section, we prove convergence rates for OGA in linear regression models in which the number of regressors is allowed to be much larger than the number of observations. Specifically, we assume that $p = p_n \to \infty$ and

(C1) $\log p_n = o(n)$,

which is weaker than Bühlmann's (2006) assumption (A1) for PGA. Moreover, similar to Bühlmann's assumptions (A2)−(A4), we assume that the $(\varepsilon_t, \mathbf{x}_t)$ in (1.1) are i.i.d., such that ε_t is independent of $\mathbf{x}_t$, and

(C2) $E\{\exp(s\varepsilon)\} < \infty$ for $|s| \le s_0$,

where $(\varepsilon, \mathbf{x})$ denotes an independent replicate of $(\varepsilon_t, \mathbf{x}_t)$. As in Section 2, we assume that $\alpha = 0$ and $E(\mathbf{x}) = \mathbf{0}$. Letting $\sigma_j^2 = E(x_j^2)$, $z_j = x_j/\sigma_j$, and $z_{tj} = x_{tj}/\sigma_j$, we assume that there exists $s_1 > 0$ such that

(C3) $\limsup_{n\to\infty} \max_{1 \le j \le p_n} E\{\exp(s_1 z_j^2)\} < \infty$.

This assumption is used to derive exponential bounds for moderate deviation probabilities of the sample correlation matrix of $\mathbf{x}_t$. In addition, we assume the weak sparsity condition

(C4) $\sup_{n \ge 1} \sum_{j=1}^{p_n} |\beta_j \sigma_j| < \infty$,

which is somewhat weaker than Bühlmann's assumption (A2). While Bühlmann's moment condition (A4) is weaker than (C2), his (A3) requires $x_j (1 \le j \le p_n)$ to be uniformly bounded random variables and (C3) is considerably weaker. The second part of this section gives an inequality for OGA similar to Bickel, Ritov, and Tsybakov's (2009) oracle inequality for Lasso. In this connection we also review related inequalities in the recent literature.

3.1. Uniform convergence rates

Let K_n denote a prescribed upper bound on the number m of OGA iterations. Let

$$\mathbf{\Gamma}(J) = E\{\mathbf{z}(J)\mathbf{z}^\top(J)\}, \quad \mathbf{g}_i(J) = E(z_i \mathbf{z}(J)), \tag{3.1}$$

where $\mathbf{z}(J)$ is a subvector of $(z_1, \ldots, z_p)^\top$ and J denotes the associated subset of indices $1, \ldots, p$. We assume that for some $\delta > 0$, $M > 0$, and all large n,

$$\min_{1 \le \sharp(J) \le K_n} \lambda_{\min}(\mathbf{\Gamma}(J)) > \delta, \quad \max_{1 \le \sharp(J) \le K_n, i \notin J} \|\mathbf{\Gamma}^{-1}(J)\mathbf{g}_i(J)\|_1 < M, \tag{3.2}$$

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

$$\|\boldsymbol{\nu}\|_1 = \sum_{j=1}^{k} |\nu_j| \text{ for } \boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)^\top. \tag{3.3}$$

The following theorem gives the rate of convergence, which holds uniformly over $1 \le m \le K_n$, for the CPE (defined in (1.3)) of OGA provided the correlation matrix of the regressors satisfies (3.2), whose meaning will be discussed in Sections 3.2 and Example 3 of Section 5.

Theorem 1. *Assume* (C1)-(C4) *and* (3.2). *Suppose* $K_n \to \infty$ *such that* $K_n = O((n/\log p_n)^{1/2})$. *Then for* OGA,

$$\max_{1 \le m \le K_n} \left(\frac{E[\{y(\mathbf{x}) - \hat{y}_m(\mathbf{x})\}^2 | y_1, \mathbf{x}_1, \ldots, y_n, \mathbf{x}_n]}{m^{-1} + n^{-1} m \log p_n} \right) = O_p(1).$$

Let $y(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x}$ and let $y_J(\mathbf{x})$ denote the best linear predictor of $y(\mathbf{x})$ based on $\{x_j, j \in J\}$, where J is a subset of $\{1, \ldots, p_n\}$. Let J_k be the set of input variables selected by the population version of OGA at the end of stage k. Then by Theorem 3 of Temlyakov (2000), the squared bias in approximating $y(\mathbf{x})$ by $y_{J_m}(\mathbf{x})$ is $E(y(\mathbf{x}) - y_{J_m}(\mathbf{x}))^2 = O(m^{-1})$. Since OGA uses $\hat{y}_m(\cdot)$ instead of $y_{J_m}(\cdot)$, it has not only larger squared bias but also variance in the least squares estimates $\hat{\beta}_{\hat{j}_i}, i = 1, \ldots, m$. The variance is of order $O(n^{-1} m \log p_n)$, noting that m is the number of estimated regression coefficients, $O(n^{-1})$ is the variance per coefficient, and $O(\log p_n)$ is the variance inflation factor due to data-dependent

selection of $\hat{j}_i$ from $\{1,\ldots,p_n\}$. Combining the squared bias with the variance suggests that $O(m^{-1}+n^{-1}m\log p_n)$ is the smallest order one can expect for $E_n(\{y(\mathbf{x})-\hat{y}_m(\mathbf{x})\}^2)$, and standard bias-variance tradeoff suggests that m should not be chosen to be larger than $O((n/\log p_n)^{1/2})$. Here and in the sequel, we use $E_n(\cdot)$ to denote $E[\cdot|y_1,\mathbf{x}_1,\ldots,y_n,\mathbf{x}_n]$. Theorem 1 says that, uniformly in $m=O((n/\log p_n)^{1/2})$, OGA can indeed attain this heuristically best order of $m^{-1}+n^{-1}m\log p_n$ for $E_n(\{y(\mathbf{x})-\hat{y}_m(\mathbf{x})\}^2)$. Section 3.2 gives further discussion of these bias-variance considerations and the restriction on K_n.

Proof of Theorem 1. Let $\hat{J}_k=\{\hat{j}_1,\ldots,\hat{j}_k\}$ and note that $\hat{J}_k$ is independent of $(y,\mathbf{x},\varepsilon)$. Replacing x_{tj} by x_{tj}/σ_j and x_j by x_j/σ_j in the OGA and its population version, we can assume without loss of generality that $\sigma_j=1$ for $1\le j\le p_n$, and therefore $z_j=x_j$; recall that (C4) actually involves $\sum_{j=1}^{p_n}|\beta_j|\sigma_j$. For $i\notin J$, let

$$\mu_{J,i}=E[\{y(\mathbf{x})-y_J(\mathbf{x})\}x_i],\quad \hat{\mu}_{J,i}=\frac{n^{-1}\sum_{t=1}^n(y_t-\hat{y}_{t;J})x_{ti}}{\left(n^{-1}\sum_{t=1}^n x_{ti}^2\right)^{1/2}},\tag{3.4}$$

where $\hat{y}_{t;J}$ denotes the fitted value of y_t when $\mathbf{Y}=(y_1,\ldots,y_n)^\top$ is projected into the linear space spanned by $\mathbf{X}_j, j\in J\neq\emptyset$, setting $\hat{y}_{t;J}=0$ if $J=\emptyset$. Note that $\hat{\mu}_{J,i}$ is the method-of-moments estimate of $\mu_{J,i}$; the denominator $(n^{-1}\sum_{t=1}^n x_{ti}^2)^{1/2}$ in (3.4) is used to estimate σ_j (which is assumed to be 1), recalling that $E(x_{ti})=0$. In view of (1.2) with $\alpha=0$, for $i\notin J$,

$$\mu_{J,i}=\sum_{j\notin J}\beta_j E[(x_j-x_j^{(J)})x_i]=\sum_{j\notin J}\beta_j E[x_j(x_i-x_i^{(J)})]=\sum_{j\notin J}\beta_j E(x_j x_{i;J}^{\perp}),\tag{3.5}$$

where $x_{i;J}^{\perp}=x_i-x_i^{(J)}$ and $x_i^{(J)}$ is the projection (in L_2) of x_i into the linear space spanned by $\{x_j, j\in J\}$, i.e.,

$$x_i^{(J)}=\mathbf{x}_J^\top\mathbf{\Gamma}^{-1}(J)\mathbf{g}_i(J),\ \text{with } \mathbf{x}_J=(x_l,l\in J).\tag{3.6}$$

Since $y_t=\sum_{j=1}^{p_n}\beta_j x_{tj}+\varepsilon_t$ and since $\sum_{t=1}^n(\varepsilon_t-\hat{\varepsilon}_{t;J})x_{ti}=\sum_{t=1}^n\varepsilon_t\hat{x}_{ti;J}^{\perp}$, where $\hat{x}_{ti;J}^{\perp}=x_{ti}-\hat{x}_{ti;J}$, and $\hat{\varepsilon}_{t;J}$ and $\hat{x}_{ti;J}$ are the fitted values of ε_t and x_{ti} when $\boldsymbol{\varepsilon}=(\varepsilon_1,\ldots,\varepsilon_n)^\top$ and $\mathbf{X}_i$ are projected into the linear space spanned by $\mathbf{X}_j, j\in J$, it follows from (3.4) and (3.5) that

$$\hat{\mu}_{J,i}-\mu_{J,i}=\frac{\sum_{t=1}^n\varepsilon_t\hat{x}_{ti;J}^{\perp}}{\sqrt{n}(\sum_{t=1}^n x_{ti}^2)^{1/2}}+\sum_{j\notin J}\beta_j\left\{\frac{n^{-1}\sum_{t=1}^n x_{tj}\hat{x}_{ti;J}^{\perp}}{(n^{-1}\sum_{t=1}^n x_{ti}^2)^{1/2}}-E(x_j x_{i;J}^{\perp})\right\}.\tag{3.7}$$

In Appendix A, we make use of (C2) and (C3), together with (3.2) and (3.6), to derive exponential bounds for the right-hand side of (3.7), and combine these

exponential bounds with (C1) and (C4) to show that there exists a positive constant s, independent of m and n, such that

$$\lim_{n\to\infty} P(A_n^c(K_n)) = 0, \text{ where}$$
$$A_n(m) = \left\{ \max_{(J,i):\sharp(J)\le m-1, i\notin J} |\hat{\mu}_{J,i} - \mu_{J,i}| \le s(\frac{\log p_n}{n})^{1/2} \right\}. \tag{3.8}$$

For any $0 < \xi < 1$, let $\tilde{\xi} = 2/(1-\xi)$ and take

$$B_n(m) = \left\{ \min_{0\le i\le m-1} \max_{1\le j\le p_n} |\mu_{\hat{\jmath}_i,j}| > \tilde{\xi} s(\log \frac{p_n}{n})^{1/2} \right\}, \tag{3.9}$$

in which we set $\mu_{J,j} = 0$ if $j \in J$, and $\mu_{\hat{\jmath}_0,j} = \mu_{\emptyset,j}$. We now show that for all $1 \le q \le m$,

$$|\mu_{\hat{\jmath}_{q-1},\hat{\jmath}_q}| \ge \xi \max_{1\le i\le p_n} |\mu_{\hat{\jmath}_{q-1},i}| \text{ on } A_n(m) \bigcap B_n(m), \tag{3.10}$$

by noting that on $A_n(m) \bigcap B_n(m)$,

$$\begin{aligned}
|\mu_{\hat{\jmath}_{q-1},\hat{\jmath}_q}| &\ge -|\hat{\mu}_{\hat{\jmath}_{q-1},\hat{\jmath}_q} - \mu_{\hat{\jmath}_{q-1},\hat{\jmath}_q}| + |\hat{\mu}_{\hat{\jmath}_{q-1},\hat{\jmath}_q}| \\
&\ge - \max_{(J,i):\sharp(J)\le m-1, i\notin J} |\hat{\mu}_{J,i} - \mu_{J,i}| + |\hat{\mu}_{\hat{\jmath}_{q-1},\hat{\jmath}_q}| \\
&\ge -s(\log \frac{p_n}{n})^{1/2} + \max_{1\le j\le p_n} |\hat{\mu}_{\hat{\jmath}_{q-1},j}| \ \ (\text{since } |\hat{\mu}_{\hat{\jmath}_{q-1},\hat{\jmath}_q}| = \max_{1\le j\le p_n} |\hat{\mu}_{\hat{\jmath}_{q-1},j}|) \\
&\ge -2s(\log \frac{p_n}{n})^{1/2} + \max_{1\le j\le p_n} |\mu_{\hat{\jmath}_{q-1},j}| \ge \xi \max_{1\le j\le p_n} |\mu_{\hat{\jmath}_{q-1},j}|,
\end{aligned}$$

since $2s(n^{-1}\log p_n)^{1/2} < (2/\tilde{\xi}) \max_{1\le j\le p_n} |\mu_{\hat{\jmath}_{q-1},j}|$ on $B_n(m)$ and $1 - \xi = 2/\tilde{\xi}$.

Consider the "semi-population version" of OGA that uses the variable selector $(\hat{\jmath}_1, \hat{\jmath}_2, \cdots)$ but still approximates $y(\mathbf{x})$ by $\sum_{j\in\hat{\jmath}_{k+1}} \lambda_j x_j$, where the λ_j are the same as those for the population version of OGA. In view of (3.10), this semi-population version is the "weak orthogonal greedy algorithm" introduced by Temlyakov (2000, pp.216-217), whose Theorem 3 can be applied to conclude that

$$E_n[\{y(\mathbf{x}) - y_{\hat{\jmath}_m}(\mathbf{x})\}^2] \le (\sum_{j=1}^{p_n} |\beta_j|)^2 (1 + m\xi^2)^{-1} \text{on } A_n(m) \bigcap B_n(m). \tag{3.11}$$

For $0 \le i \le m-1$, $E_n[\{y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})\}^2] \le E_n[\{y(\mathbf{x}) - y_{\hat{J}_i}(\mathbf{x})\}^2]$, and therefore

$$\begin{aligned} E_n[\{y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})\}^2] &\le \min_{0\le i\le m-1} E_n\{(y(\mathbf{x}) - y_{\hat{J}_i}(\mathbf{x}))(\sum_{j=1}^{p_n} \beta_j x_j)\} \\ &\le \min_{0\le i\le m-1} \max_{1\le j\le p_n} |\mu_{\hat{J}_i,j}| \sum_{j=1}^{p_n} |\beta_j| \\ &\le \tilde{\xi} s(n^{-1}\log p_n)^{1/2} \sum_{j=1}^{p_n} |\beta_j| \text{ on } B_n^c(m). \end{aligned}$$

Combining this with (C4), (3.11), and the assumption that $m \le K_n = O((n / \log p_n)^{1/2})$ yields

$$E_n[\{y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})\}^2] I_{A_n(m)} \le C^* m^{-1} \tag{3.12}$$

for some constant $C^* > 0$. Moreover, since $A_n(K_n) \subseteq A_n(m)$, it follows from (3.8) and (3.12) that $\max_{1\le m\le K_n} mE_n[\{y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})\}^2] = O_p(1)$. Theorem 1 follows from this and

$$\max_{1\le m\le K_n} \frac{nE_n[\{\hat{y}_m(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})\}^2]}{m\log p_n} = O_p(1), \tag{3.13}$$

whose proof is given in Appendix A, noting that

$$E_n[\{y(\mathbf{x}) - \hat{y}_m(\mathbf{x})\}^2] = E_n[\{y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})\}^2] + E_n[\{\hat{y}_m(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})\}^2].$$

3.2. A bias-variance bound

In this section, we assume that the x_{tj} in (1.1) are nonrandom constants and develop an upper bound for the empirical norm

$$\|\hat{y}_m(\cdot) - y(\cdot)\|_n^2 = n^{-1}\sum_{t=1}^{n} (\hat{y}_m(\mathbf{x}_t) - y(\mathbf{x}_t))^2 \tag{3.14}$$

of OGA, providing an analog of the oracle inequalities of Candes and Tao (2007), Bunea, Tsybakov, and Wegkamp (2007) Bickel, Ritov, and Tsybakov (2009) and Candes and Plan (2009) for Lasso and Dantzig selector that will be reviewed below. In the approximation theory literature, the ε_t in (1.1) are usually assumed to be either zero or nonrandom. In the case $\varepsilon_t = 0$ for all t, an upper bound for (3.14) has been obtained by Tropp (2004). When the ε_t are nonzero but nonrandom, a bound for the bias of the OGA estimate has also been given by Donoho, Elad, and Temlyakov (2006). When the ε_t in (1.1) are zero-mean

random variables, an upper bound for (3.14) should involve the variance besides the bias of the regression estimate and should also provide insights into the bias-variance tradeoff, as is the case with the following theorem for which p can be much larger than n. Noting that the regression function in (1.1) has infinitely many representations when $p > n$, we introduce the representation set

$$\mathbf{B} = \{\mathbf{b} : \mathbf{X}\mathbf{b} = (y(\mathbf{x}_1), \ldots, y(\mathbf{x}_n))^\top\}, \tag{3.15}$$

where $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ is $n \times p$. In addition, for $J \subseteq \{1, \ldots, p\}$ and $1 \le i \le p$ with $i \notin J$, let $\mathbf{B}_{J,i} = \{\boldsymbol{\theta}_{J,i} : \mathbf{X}_J^\top \mathbf{X}_i = \mathbf{X}_J^\top \mathbf{X}_J \boldsymbol{\theta}_{J,i}\}$. Moreover, take

$$r_p = \arg \min_{0<r<1/2} \frac{1}{r}\Big\{1 + \Big(\frac{\log \sqrt{1/(1-2r)}}{\log p}\Big)\Big\},\ \tilde{r}_p = \frac{1}{1-2r_p}. \tag{3.16}$$

Note that as $p \to \infty$, $r_p \to 1/2$ and $\tilde{r}_p = o(p^\eta)$ for any $\eta > 0$.

Theorem 2. *Suppose ε_t are i.i.d. normal random variables with $E(\varepsilon_t) = 0$ and $E(\varepsilon_t^2) = \sigma^2$. Assume that x_{tj} are nonrandom constants, normalized so that $n^{-1}\sum_{t=1}^n x_{tj}^2 = 1$, and satisfying*

$$\max_{1 \le \sharp(J) \le \lfloor n/\log p \rfloor, i \notin J} \inf_{\boldsymbol{\theta}_{J,i} \in \mathbf{B}_{J,i}} \|\boldsymbol{\theta}_{J,i}\|_1 < M \quad \text{for some } M > 0. \tag{3.17}$$

Let $0 < \xi < 1$, $C > \sqrt{2}(1+M)$, $s > \{1 + (2\log p)^{-1} \log \tilde{r}_p\}/r_p$, where r_p and $\tilde{r}_p$ are defined by (3.16)*, and*

$$\omega_{m,n} = (\inf_{\mathbf{b} \in \mathbf{B}} \|\mathbf{b}\|_1) \max\left\{\frac{\inf_{\mathbf{b}\in\mathbf{B}} \|\mathbf{b}\|_1}{1 + m\xi^2}, \frac{2C\sigma}{1-\xi}\Big(\frac{\log p}{n}\Big)^{1/2}\right\}. \tag{3.18}$$

Then for all $p \ge 3$, $n \ge \log p$, and $1 \le m \le \lfloor n/\log p \rfloor$,

$$\|\hat{y}_m(\cdot) - y(\cdot)\|_n^2 \le \omega_{m,n} + s\sigma^2 m \frac{(\log p)}{n} \tag{3.19}$$

with probability at least

$$1 - p\exp\left\{-\frac{C^2 \log p}{2(1+M)^2}\right\} - \frac{\tilde{r}_p^{1/2} p^{-(sr_p - 1)}}{1 - \tilde{r}_p^{1/2} p^{-(sr_p-1)}}.$$

The upper bound (3.19) for the prediction risk of OGA is a sum of a variance term, $s\sigma^2 m(\log p)/n$, and a squared bias term, $\omega_{m,n}$. The variance term is the usual "least squares" risk $m\sigma^2/n$ multiplied by a risk inflation factor $s\log p$; see Foster and George (1994) for a detailed discussion of the idea of risk inflation. The squared bias term is the maximum of $(\inf_{\mathbf{b}\in\mathbf{B}} \|\mathbf{b}\|_1)^2/(1 + m\xi^2)$, which is the approximation error of the "noiseless" OGA, and $2C\sigma(1 -$

$\xi)^{-1}\inf_{\mathbf{b}\in\mathbf{B}}\|\mathbf{b}\|_1(n^{-1}\log p)^{1/2}$, which is the error caused by the discrepancy between the noiseless OGA and the sample OGA; see (B.5), (B.6), (B.7), and (B.8) in Appendix B.

The $\|\boldsymbol{\theta}_{J,i}\|_1$ in (3.17) is closely related to the "cumulative coherence function" introduced by Tropp (2004). Since Theorem 2 does not put any restriction on M and $\inf_{\mathbf{b}\in\mathbf{B}}\|\mathbf{b}\|_1$, the theorem can be applied to any design matrix although a large value of M or $\inf_{\mathbf{b}\in\mathbf{B}}\|\mathbf{b}\|_1$ will result in a large bound on the right-hand side of (3.19). Note that the population analog of $\|\boldsymbol{\theta}_{J,i}\|_1$ for random regressors is $\|\boldsymbol{\Gamma}^{-1}(J)\mathbf{g}_i(J)\|_1$, which appears in the second part of (3.2), the first part of which assumes that $\boldsymbol{\Gamma}(J)$ is uniformly positive definite for $1\le\sharp(J)\le K_n$. Note also the similarity of (3.17) and the second part of (3.2). Although (3.2) makes an assumption on $\lambda_{\min}(\boldsymbol{\Gamma}(J))$, it does not make assumptions on $\lambda_{\max}(\boldsymbol{\Gamma}(J))$; this is similar to the "restricted eigenvalue assumption" introduced by Bickel, Ritov, and Tsybakov (2009) but differs from the "sparse Riesz condition" that will be discussed in the second paragraph of Section 5.

When M and $\inf_{\mathbf{b}\in\mathbf{B}}\|\mathbf{b}\|_1$ are bounded by a positive constant independent of n and p, the upper bound in (3.19) suggests that choosing $m=D(n/\log p)^{1/2}$ for some $D>0$ can provide the best bias-variance tradeoff, for which (3.19) reduces to

$$\|\hat{y}_m(\cdot)-y(\cdot)\|_n^2\le d(n^{-1}\log p)^{1/2}, \tag{3.20}$$

where d does not depend on n and p. Note that $D(n/\log p)^{1/2}$ can be used to explain why there is no loss in efficiency for the choice $K_n=O((n/\log p)^{1/2})$ in Theorem 1. We can regard (3.20) as an analog of the oracle inequality of Bickel, Ritov, and Tsybakov (2009, Thm. 6.2) for the Lasso predictor $\hat{y}_{\mathrm{Lasso}(r)}(\mathbf{x}_t)=\mathbf{x}_t^\top\hat{\boldsymbol{\beta}}_{\mathrm{Lasso}(r)}$, where

$$\hat{\boldsymbol{\beta}}_{\mathrm{Lasso}(r)}=\arg\min_{\mathbf{c}\in R^p}\Big\{n^{-1}\sum_{t=1}^n(y_t-\mathbf{x}_t^\top\mathbf{c})^2+2r\|\mathbf{c}\|_1\Big\}, \tag{3.21}$$

with $r>0$. Letting $M(\mathbf{b})$ denote the number of nonzero components of $\mathbf{b}\in\mathbf{B}$ and defining $\bar{Q}=\inf_{\mathbf{b}\in\mathbf{B}}M(\mathbf{b})$, they assume instead of (3.17) that $\mathbf{X}^\top\mathbf{X}$ satisfies a restricted eigenvalue assumption $\mathrm{RE}(\bar{Q},3)$, and show under the same assumptions of Theorem 2 (except for (3.17)) that for $r=A\sigma(n^{-1}\log p)^{1/2}$ with $A>2\sqrt{2}$,

$$\|\hat{y}_{\mathrm{Lasso}(r)}(\cdot)-y(\cdot)\|_n^2\le F\frac{\bar{Q}\log p}{n} \tag{3.22}$$

with probability at least $1-p^{1-A^2/8}$, where F is a positive constant depending only on A,σ, and $1/\kappa$, in which $\kappa=\kappa(\bar{Q},3)$ is the defining restricted eigenvalue of $\mathrm{RE}(\bar{Q},3)$. Suppose that F in (3.22) is bounded by a constant independent of n and p and that $\log p$ is small relative to n. Then (3.20) and (3.22) suggest that the

risk bound for Lasso is smaller (or larger) than that of OGA if $\bar{Q} \ll (n/\log p)^{1/2}$ (or $\bar{Q} \gg (n/\log p)^{1/2}$).

4. Consistent Model Selection under Strong Sparsity

The convergence rate theory of OGA in Theorems 1 and 2 suggests terminating OGA iterations after $K_n = O((n/\log p)^{1/2})$ steps or, equivalently, after K_n input variables have been included in the regression model. We propose to choose along the OGA path the model that has the smallest value of a suitably chosen criterion, which we call a "high-dimensional information criterion" (HDIC). Specifically, for a non-empty subset J of $\{1, \ldots, p\}$, let $\hat{\sigma}_J^2 = n^{-1} \sum_{t=1}^n (y_t - \hat{y}_{t;J})^2$, where $\hat{y}_{t;J}$ is defined below (3.4). Let

$$\mathrm{HDIC}(J) = n \log \hat{\sigma}_J^2 + \sharp(J) w_n \log p, \tag{4.1}$$

$$\hat{k}_n = \arg \min_{1 \le k \le K_n} \mathrm{HDIC}(\hat{J}_k), \tag{4.2}$$

in which different criteria correspond to different choices of w_n and $\hat{J}_k = \{\hat{j}_1, \ldots, \hat{j}_k\}$. Note that $\hat{\sigma}^2_{\hat{j}_k}$, and therefore $\mathrm{HDIC}(\hat{J}_k)$ also, can be readily computed at the kth OGA iteration, and therefore this model selection method along the OGA path involves little additional computational cost. In particular, $w_n = \log n$ corresponds to HDBIC; without the $\log p$ factor, (4.1) reduces to the usual BIC. The case $w_n = c \log\log n$ with $c > 2$ corresponds to the high-dimensional Hannan-Quinn criterion (HDHQ), and the case $w_n = c$ corresponds to HDAIC, recalling that $\mathrm{AIC}(J) = n \log \hat{\sigma}_J^2 + 2\sharp(J)$. For fixed p, Zhang, Li, and Tsai (2010) also consider general penalities w_n in their genelized information criterion (GIC) that "makes a connection between the classical variable selection criterion and the regularization parameter selections for the nonconcave penalized likelihood approaches."

A standard method to select the number m of input variables to enter the regression model is cross-validation (CV) or its variants such as C_p and AIC that aim at striking a suitable balance between squared bias and variance. However, these variable selection methods do not work well when $p \gg n$, as shown by Chen, Ing and Lai (2011) who propose to modify these criteria by including a factor $\log p$, as in HDAIC, for weakly sparse models. Whereas that paper considers the general weakly sparse setting in which the β_j may all be nonzero, we focus here on the case in which a substantial fraction of the β_j is 0. We call an input variable "relevant" if its associated β_j is nonzero, and "irrelevant" otherwise. In Section 4.1, we review the literature on variable selection consistency (i.e., selecting all relevant variables and no irrelevant variables, with probability approaching 1) and on the sure screening property (i.e., including all relevant variables with probability approaching 1). To achieve consistency of variable

selection, some lower bound (which may approach 0 as $n \to \infty$) on the absolute values of nonzero regression coefficients needs to be imposed. We therefore assume a "strong sparsity" condition:

(C5) There exists $0 \le \gamma < 1$ such that $n^\gamma = o((n/\log p_n)^{1/2})$ and

$$\liminf_{n\to\infty} n^\gamma \min_{1\le j\le p_n:\beta_j\neq 0} \beta_j^2\sigma_j^2 > 0.$$

Note that (C5) imposes a lower bound on $\beta_j^2\sigma_j^2$ for nonzero β_j. This is more natural than a lower bound on $|\beta_j|$ since the predictor of y_i involves $\beta_j x_{ij}$. Instead of imposing an upper bound on the number of nonzero regression coefficients, as in the oracle inequality (3.22) for Lasso, we assume strong sparsity in Section 4.2 and first show that OGA has the sure screening property if the number K_n of iterations is large enough. We then show that the best fitting model can be chosen along an OGA path by minimizing a high-dimensional information criterion (4.1) with $\lim_{n\to\infty} w_n = \infty$. In Section 4.3, we further use HDIC to remove irrelevant variables included along the OGA path so that the resultant procedure has the oracle property that it is equivalent to performing ordinary least squares regression on the set of relevant regressors.

4.1. Sure screening and consistent variable selection

When $p = p_n \to \infty$ but the number of nonzero regression parameters remains bounded, Chen and Chen (2008) propose to modify the usual BIC by

$$\mathrm{BIC}_\gamma(J) = n\log\hat\sigma_J^2 + \sharp(J)\log n + 2\gamma\log\tau_{\sharp(J)}, \text{ where } \tau_j = \binom{p}{j}, \tag{4.3}$$

in which $J \subset \{1,\dots,p\}$ is non-empty and $0 \le \gamma \le 1$ is related to a prior distribution on the parameter space partitioned as $\bigcup_{j=1}^p \boldsymbol{\Theta}_j$, with $\boldsymbol{\Theta}_j$ consisting of all $\boldsymbol{\beta} = (\beta_1,\dots,\beta_p)^\top$ such that exactly j of the β_i's are nonzero. The prior distribution puts equal probability to each of τ_j choices of the j relevant regressors for $\boldsymbol{\beta} \in \boldsymbol{\Theta}_j$, and assigns to $\boldsymbol{\Theta}_j$ probability proportional to $1/\tau_j^{1-\gamma}$. Assuming the ε_t to be normal, extension of Schwarz's (1978) argument yields the "extended BIC" (4.3). Chen and Chen (2008) propose to choose the J that has the smallest $\mathrm{BIC}_\gamma(J)$. Since there are $2^p - 1$ non-empty subsets of $\{1,\dots,p\}$, this approach is "computationally infeasible" for large p. They therefore propose to apply BIC_γ to a manageable subset of models selected by other methods, e.g., Lasso(r) over a range of r. Wang (2009) recently proposed to use forward stepwise regression to select this manageable subset; his method and results are discussed in Sections 4.2 and 5.

As pointed out by Zhao and Yu (2006), "obtaining (sparse) models through classical model selection methods usually involve heavy combinatorial search," and Lasso "provides a computationally feasible way for model selection." On the other hand, Leng, Lin, and Wahba (2006) have shown that Lasso is not variable-selection consistent when prediction accuracy is used as the criterion for choosing the penalty r in (3.21). However, if $p_n = O(n^\kappa)$ for some $\kappa > 0$ and $r = r_n$ is chosen to grow at an appropriate rate with n, Zhao and Yu (2006) proved that Lasso is variable-selection consistent under a "strong irrepresentable condition" on the design matrix, and additional regularity conditions. A closely related result is model selection consistency of Lasso with suitably chosen penalty in Gaussian graphical models under a similar condition, which is called the "neighborhood stability condition." As noted by Zhao and Yu (2006), Lasso can fail to distinguish irrelevant predictors that are highly correlated with relevant predictors, and the strong irrepresentable condition is used to rule out such cases. It is closely related to the "coherence condition" of Donoho, Elad, and Temlyakov (2006) and is "almost necessary and sufficient" for Lasso to be sign-consistent for some choice of penalty. Under a sparse Riesz condition, Zhang and Huang (2008) have studied sparsity and bias properties of Lasso-based model selection methods.

Fan and Li (2001) pointed out that the l_1-penalty used by Lasso may lead to severe bias for large regression coefficients and proposed a "smoothly chipped absolute deviation" (SCAD) penalty to address this problem. Because the associated minimization problem is non-convex and direct computation is infeasible for large p, multi-step procedures in which each step involves convex optimization have been introduced, as in the local quadratic approximation of Fan and Li (2001) and the local linear approximation (LLA) of Zou and Li (2008), who also show that the one-step LLA estimator has certain oracle properties if the initial estimator is suitably chosen. Zhou, van de Geer, and Bühlmann (2009) have pointed out that one such procedure is Zou's (2006) adaptive Lasso, which uses the Lasso as an initial estimator to determine the weights for a second-stage weighted Lasso (that replaces $\|\mathbf{c}\|_1 = \sum_{i=1}^p |c_i|$ in (3.21) by a weighted sum $\sum_{i=1}^p \omega_i |c_i|$). They have also substantially weakened the conditions of Huang, Ma, and Zhang (2008) on the variable selection consistency of adaptive Lasso, which Zou (2006) established earlier for the case of fixed p.

The concept of sure screening was introduced by Fan and Lv (2008), who also proposed a method called "sure independence screening" (SIS) that has the sure screening property in sparse high-dimensional regression models satisfying certain conditions. For given positive integer d, SIS selects d regressors whose sample correlation coefficients with y_t have the largest d absolute values. Although SIS with suitably chosen $d = d_n$ has been shown by Fan and Lv (2008, Sec. 5) to

have the sure screening property without the irrepresentable (or neighborhood stability) condition mentioned earlier for Lasso, it requires an assumption on the maximum eigenvalue of the covariance matrix of the candidate regressors that can fail to hold when all regressors are equally correlated, as will be shown in Section 5. Fan and Lv (2010) give a comprehensive overview of SIS and a modification called "iterative sure independence screening" (ISIS), their applications to feature selection for classification, and penalized likelihood methods for variable selection in sparse linear and generalized linear models.

4.2. OGA+HDIC in strongly sparse linear models

The procedure proposed in the first paragraph of Section 4, which we call OGA+HDIC, consists of (i) carrying out K_n OGA iterations, (ii) computing $\mathrm{HDIC}(\hat{J}_k)$ defined by (4.1) at the end of the kth iteration, i.e., after k regressors are selected for the linear regression model, and (iii) choosing the k that minimizes $\mathrm{HDIC}(\hat{J}_k)$ over $1 \le k \le K_n$ after the OGA iterations terminate. Concerning the ingredient (i) of the procedure, we make use of Theorem 1 to show that it has the sure screening property in strongly sparse linear models.

Theorem 3. *Assume* (C1)−(C5) *and* (3.2). *Suppose* $K_n/n^\gamma \to \infty$ *and* $K_n = O((n/\log p_n)^{1/2})$. *Then* $\lim_{n\to\infty} P(N_n \subset \hat{J}_{K_n}) = 1$, *where* $N_n = \{1 \le j \le p_n : \beta_j \ne 0\}$ *denotes the set of relevant input variables.*

Proof. Without loss of generality, assume that $\sigma_j = 1$ so that $z_j = x_j$ for $1 \le j \le p_n$. Let $a > 0$ and define $A_n(m)$ by (3.8), in which $m = \lfloor an^\gamma \rfloor = o(K_n)$. By (3.8) and (3.12),

$$\begin{aligned} &\lim_{n\to\infty} P(A_n^c(m)) \le \lim_{n\to\infty} P(A_n^c(K_n)) = 0, \\ &E_n\{[y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})]^2\} I_{A_n(m)} \le C^* m^{-1}. \end{aligned} \tag{4.4}$$

From (4.4), it follows that

$$\lim_{n\to\infty} P(F_n) = 0, \text{where } F_n = \{E_n[y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})]^2 > C^* m^{-1}\}. \tag{4.5}$$

For $J \subseteq \{1, \dots, p_n\}$ and $j \in J$, let $\tilde{\beta}_j(J)$ be the coefficient of x_j in the best linear predictor $\sum_{i\in J} \tilde{\beta}_i(J) x_i$ of y that minimizes $E(y - \sum_{i\in J} \lambda_i x_i)^2$. Define $\tilde{\beta}_j(J) = 0$ if $j \notin J$. Note that

$$E_n[y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})]^2 = E_n\Big\{ \sum_{j\in \hat{J}_m \cup N_n} (\beta_j - \tilde{\beta}_j(\hat{J}_m)) x_j \Big\}^2. \tag{4.6}$$

From (C4) and (C5), it follows that $\sharp(N_n) = o(n^{\gamma/2})$, yielding $\sharp(\hat{J}_m \cup N_n) = o(K_n)$, and it then follows from (4.6) that for all large n,

$$E_n[\{y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})\}^2] \ge (\min_{j\in N_n} \beta_j^2) \min_{1\le \sharp(J)\le K_n} \lambda_{\min}(\mathbf{\Gamma}(J)) \text{ on } \{N_n \cap \hat{J}_m^c \ne \emptyset\}. \tag{4.7}$$

Combining (4.7) with (C5) and (3.2) then yields $E_n[\{y(\mathbf{x})-y_{\hat{J}_m}(\mathbf{x})\}^2] \geq bn^{-\gamma}$ on $\{N_n \cap \hat{J}_m^c \neq \emptyset\}$, for some $b > 0$ and all large n. By choosing the a in $m = \lfloor an^\gamma \rfloor$ large enough, we have $bn^{-\gamma} > C^* m^{-1}$, implying that $\{N_n \cap \hat{J}_m^c \neq \emptyset\} \subseteq F_n$, where F_n is defined in (4.5). Hence by (4.5), $\lim_{n\to\infty} P(N_n \subseteq \hat{J}_m) = 1$. Therefore, the OGA path that terminates after $m = \lfloor an^\gamma \rfloor$ iterations contains all relevant regressors with probability approaching 1. This is also true for the OGA path that terminates after K_n iterations if $K_n/m \to \infty$.

To explain the importance of the factor $\log p_n$ in the definition (4.1) of HDIC when $p_n \gg n$, suppose $x_1, \ldots, x_{p_n}$ are uncorrelated, i.e., $\mathbf{\Gamma}(J) = \mathbf{I}$, for which "hard thresholding" (Donoho and Johnstone (1994)) can be used for variable selection. Assuming for simplicity that σ^2 and σ_j^2 are known, note that $(\hat{\beta}_j - \beta_j)/(\sigma^2/\sum_{t=1}^n x_{tj}^2)^{1/2}$, $1 \leq j \leq p_n$, are asymptotically independent standard normal random variables in this case. Since $\max_{1\leq j\leq p_n} |n^{-1}\sum_{t=1}^n x_{tj}^2 - \sigma_j^2|$ converges to 0 in probability (see Lemma A.2 in Appendix A), it follows that

$$\max_{1\leq j\leq p_n} n(\hat{\beta}_j - \beta_j)^2 \frac{\sigma_j^2}{\sigma^2} - (2\log p_n - \log\log p_n)$$

has a limiting Gumbel distribution. In view of (C5) that assumes $\beta_j^2\sigma_j^2 \geq cn^{-\gamma}$ for nonzero β_j and some positive constant c, screening out the regressors with $\hat{\beta}_j^2\sigma_j^2 < (\sigma^2 w_n \log p_n)/n$ yields consistent variable selection if $w_n \log p_n = o(n^{1-\gamma})$ and $\liminf_{n\to\infty} w_n > 2$. Such w_n can indeed be chosen if $n^\gamma = o(n/\log p_n)$, recalling that $\log p_n = o(n)$. In the more general case where $x_1, \ldots, x_{p_n}$ are correlated and therefore so are the $\hat{\beta}_j$, we make use of (3.2) and regard the threshold $(\sigma^2 w_n \log p_n)/n$ as a penalty for including an input variable in the regression model. The preceding argument then leads to the criterion (4.1) and suggests selecting the regressor set $\hat{J}_k$ that minimizes $\mathrm{HDIC}(\hat{J}_k)$. We next establish, under strong sparsity, consistency of variable selection along OGA paths by HDIC with w_n in (4.1) satisfying

$$w_n \to \infty,\ w_n \log p_n = o(n^{1-2\gamma}). \tag{4.8}$$

Define the minimal number of relevant regressors along an OGA path by

$$\widetilde{k}_n = \min\{k : 1 \leq k \leq K_n,\ N_n \subseteq \hat{J}_k\}\ (\min\emptyset = K_n). \tag{4.9}$$

Theorem 4. *With the same notation and assumptions as in Theorem* 3*, suppose* (4.8) *holds,* $K_n/n^\gamma \to \infty$*, and* $K_n = O((n/\log p_n)^{1/2})$*. Then* $\lim_{n\to\infty} P(\hat{k}_n = \tilde{k}_n) = 1$.

Proof. Assume $\sigma_j^2 = 1$ as in the proof of Theorem 3, and drop the subscript n in $\tilde{k}_n$ and $\hat{k}_n$ for notational simplicity. We first show that $P(\hat{k} < \tilde{k}) = o(1)$. As shown in the proof of Theorem 3, for sufficiently large a,

$$\lim_{n\to\infty} P(\mathcal{D}_n) = 1, \text{ where } \mathcal{D}_n = \{N_n \subseteq \hat{J}_{\lfloor an^\gamma \rfloor}\} = \{\tilde{k} \le an^\gamma\}. \tag{4.10}$$

On $\{\hat{k} < \tilde{k}\}$, $\exp\{\mathrm{HDIC}(\hat{J}_{\hat{k}})/n\} \le \exp\{\mathrm{HDIC}(\hat{J}_{\tilde{k}})/n\}$ and $\hat{\sigma}^2_{\hat{J}_{\hat{k}}} \ge \hat{\sigma}^2_{\hat{J}_{\tilde{k}-1}}$, so

$$\hat{\sigma}^2_{\hat{J}_{\tilde{k}-1}} - \hat{\sigma}^2_{\hat{J}_{\tilde{k}}} \le \hat{\sigma}^2_{\hat{J}_{\tilde{k}}}\{\exp(n^{-1}w_n\tilde{k}\log p_n) - \exp(n^{-1}w_n\hat{k}\log p_n)\}. \tag{4.11}$$

Let $\mathbf{H}_J$ denote the projection matrix associated with projections into the linear space spanned by $\mathbf{X}_j, j \in J \subseteq \{1,\dots,p\}$. Then, on the set $\mathcal{D}_n$,

$$\begin{aligned} &n^{-1}\{\sum_{t=1}^n (y_t - \hat{y}_{t;\hat{J}_{\tilde{k}-1}})^2 - \sum_{t=1}^n (y_t - \hat{y}_{t;\hat{J}_{\tilde{k}}})^2\} \\ &\quad = n^{-1}(\beta_{\hat{J}_{\tilde{k}}}\mathbf{X}_{\hat{J}_{\tilde{k}}} + \boldsymbol{\varepsilon})^\top(\mathbf{H}_{\hat{J}_{\tilde{k}}} - \mathbf{H}_{\hat{J}_{\tilde{k}-1}})(\beta_{\hat{J}_{\tilde{k}}}\mathbf{X}_{\hat{J}_{\tilde{k}}} + \boldsymbol{\varepsilon}) \\ &\quad = \frac{\left\{\beta_{\hat{J}_{\tilde{k}}}\mathbf{X}^\top_{\hat{J}_{\tilde{k}}}(\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}-1}})\mathbf{X}_{\hat{J}_{\tilde{k}}} + \mathbf{X}^\top_{\hat{J}_{\tilde{k}}}(\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}-1}})\boldsymbol{\varepsilon}\right\}^2}{n\mathbf{X}^\top_{\hat{J}_{\tilde{k}}}(\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}-1}})\mathbf{X}_{\hat{J}_{\tilde{k}}}}. \end{aligned} \tag{4.12}$$

Simple algebra shows that the last expression in (4.12) can be written as $\beta^2_{\hat{J}_{\tilde{k}}}\hat{A}_n + 2\beta_{\hat{J}_{\tilde{k}}}\hat{B}_n + \hat{A}_n^{-1}\hat{B}_n^2$, where

$$\hat{A}_n = n^{-1}\mathbf{X}^\top_{\hat{J}_{\tilde{k}}}(\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}-1}})\mathbf{X}_{\hat{J}_{\tilde{k}}} \text{ and } \hat{B}_n = n^{-1}\mathbf{X}^\top_{\hat{J}_{\tilde{k}}}(\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}-1}})\boldsymbol{\varepsilon},\ \hat{C}_n = \hat{\sigma}^2_{\hat{J}_{\tilde{k}}} - \sigma^2. \tag{4.13}$$

Hence it follows from (4.8), (4.11), and (4.12) that there exists $\lambda > 0$ such that $\beta^2_{\hat{J}_{\tilde{k}}}\hat{A}_n + 2\beta_{\hat{J}_{\tilde{k}}}\hat{B}_n + \hat{A}_n^{-1}\hat{B}_n^2 \le \lambda n^{-1}w_n(\log p_n)\lfloor an^\gamma\rfloor(\hat{C}_n + \sigma^2)$ on $\{\hat{k} < \tilde{k}\}\bigcap\mathcal{D}_n$, which implies that

$$\begin{aligned} &2\beta_{\hat{J}_{\tilde{k}}}\hat{B}_n - \lambda n^{-1}w_n(\log p_n)\lfloor an^\gamma\rfloor|\hat{C}_n| \\ &\quad \le -\beta^2_{\hat{J}_{\tilde{k}}}\hat{A}_n + \lambda n^{-1}w_n(\log p_n)\lfloor an^\gamma\rfloor\sigma^2 \text{ on } \{\hat{k} < \tilde{k}\}\bigcap\mathcal{D}_n. \end{aligned} \tag{4.14}$$

Define $v_n = \min_{1\le \sharp(J)\le\lfloor an^\gamma\rfloor}\lambda_{\min}(\boldsymbol{\Gamma}(J))$. By (3.2), $v_n > \delta$ for all large n. In Appendix A, it is shown that for any $\theta > 0$,

$$P(\hat{A}_n \le \frac{v_n}{2}, \mathcal{D}_n) + P(|\hat{B}_n| \ge \theta n^{-\gamma/2}, \mathcal{D}_n) + P(w_n(\log p_n)|\hat{C}_n| \ge \theta n^{1-2\gamma}, \mathcal{D}_n) = o(1). \tag{4.15}$$

From (C5), (4.10), (4.14), and (4.15), it follows that $P(\hat{k} < \tilde{k}) = o(1)$.

It remains to prove $P(\hat{k} > \tilde{k}) = o(1)$. Note that on $\{\hat{k} > \tilde{k}\}$,

$$\hat{\sigma}^2_{\hat{J}_{\hat{k}}} \exp(n^{-1} w_n \hat{k} \log p_n) \le \hat{\sigma}^2_{\hat{J}_{\tilde{k}}} \exp(n^{-1} w_n \tilde{k} \log p_n).$$

Since $\beta_j = 0$ for $j \notin \hat{J}_{\tilde{k}}$, this implies the following counterpart of (4.11) and (4.12) on $\{\hat{k} > \tilde{k}\}$:

$$\boldsymbol{\varepsilon}^\top (\mathbf{H}_{\hat{J}_{\hat{k}}} - \mathbf{H}_{\hat{J}_{\tilde{k}}}) \boldsymbol{\varepsilon} \ge \boldsymbol{\varepsilon}^\top (\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}}}) \boldsymbol{\varepsilon} \{1 - \exp(-n^{-1} w_n (\hat{k} - \tilde{k}) \log p_n)\}. \tag{4.16}$$

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

$$\begin{aligned}
\boldsymbol{\varepsilon}^\top (\mathbf{H}_{\hat{J}_{\hat{k}}} - \mathbf{H}_{\hat{J}_{\tilde{k}}}) \boldsymbol{\varepsilon} &= \boldsymbol{\varepsilon}^\top (\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}}}) \mathbf{F}_{\hat{k},\tilde{k}} \{\mathbf{F}^\top_{\hat{k},\tilde{k}} (\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}}}) \mathbf{F}_{\hat{k},\tilde{k}}\}^{-1} \mathbf{F}^\top_{\hat{k},\tilde{k}} (\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}}}) \boldsymbol{\varepsilon} \\
&\le \|\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_{K_n})\| \|n^{-1/2} \mathbf{F}^\top_{\hat{k},\tilde{k}} (\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}}}) \boldsymbol{\varepsilon}\|^2 \\
&\le 2\|\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_{K_n})\| \|n^{-1/2} \mathbf{F}^\top_{\hat{k},\tilde{k}} \boldsymbol{\varepsilon}\|^2 + 2\|\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_{K_n})\| \|n^{-1/2} \mathbf{F}^\top_{\hat{k},\tilde{k}} \mathbf{H}_{\hat{J}_{\tilde{k}}} \boldsymbol{\varepsilon}\|^2 \\
&\le 2(\hat{k} - \tilde{k})(\hat{a}_n + \hat{b}_n),
\end{aligned} \tag{4.17}$$

where $\hat{\boldsymbol{\Gamma}}(J)$ denotes the sample covariance matrix that estimates $\boldsymbol{\Gamma}(J)$ for $J \subseteq \{1, \ldots, p_n\}$ (recalling that $\sigma_j^2 = 1$) and $\|\mathbf{L}\| = \sup_{\|\mathbf{x}\|=1} \|\mathbf{L}\mathbf{x}\|$ for a nonnegative definite matrix $\mathbf{L}$,

$$\begin{aligned}
\hat{a}_n &= \|\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_{K_n})\| \max_{1 \le j \le p_n} \Big(n^{-1/2} \sum_{t=1}^n x_{tj} \varepsilon_t \Big)^2, \\
\hat{b}_n &= \|\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_{K_n})\| \max_{1 \le \sharp(J) \le \tilde{k}, i \notin J} \Big(n^{-1/2} \sum_{t=1}^n \varepsilon_t \hat{x}_{ti;J} \Big)^2.
\end{aligned} \tag{4.18}$$

Since $n^{-1} \boldsymbol{\varepsilon}^\top (\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}}}) \boldsymbol{\varepsilon} - \sigma^2 = \hat{C}_n$, combining (4.17) with (4.16) yields

$$\begin{aligned}
&2(\hat{k} - \tilde{k})(\hat{a}_n + \hat{b}_n) + |\hat{C}_n| n [1 - \exp(-n^{-1} w_n (\hat{k} - \tilde{k}) \log p_n)] \\
&\quad \ge n\sigma^2 [1 - \exp(-n^{-1} w_n (\hat{k} - \tilde{k}) \log p_n)] \text{ on } \{\hat{k} > \tilde{k}\}.
\end{aligned} \tag{4.19}$$

In Appendix A it is shown that for any $\theta > 0$,

$$\begin{aligned}
&P\{(\hat{a}_n + \hat{b}_n)(\hat{k} - \tilde{k}) \ge \theta n [1 - \exp(-n^{-1} w_n (\hat{k} - \tilde{k}) \log p_n)], \hat{k} > \tilde{k}\} = o(1), \\
&P\{|\hat{C}_n| \ge \theta\} = o(1).
\end{aligned} \tag{4.20}$$

From (4.19) and (4.20), $P(\hat{k} > \tilde{k}) = o(1)$ follows.

Theorem 4 is a much stronger result than Theorem 2 of Wang (2009) that only establishes the sure screening property $\lim_{n\to\infty} P(N_n \subseteq \hat{J}_{\hat{m}_n}) = 1$ of using the extended BIC (EBIC) to choose variables along an OGA path, where

$\hat{m}_n = \arg\min_{1\le m\le n} \mathrm{EBIC}(\hat{J}_m)$. In addition, Wang proves this result under much stronger assumptions than those of Theorem 4, such as ε_t and $\mathbf{x}_t$ having normal distributions and $a \le \lambda_{\min}(\mathbf{\Sigma}_n) \le \lambda_{\max}(\mathbf{\Sigma}_n) \le b$ for some positive constants a and b and all $n \ge 1$, where $\mathbf{\Sigma}_n$ is the covariance matrix of the p_n-dimensional random vector $\mathbf{x}_t$.

4.3. Further trimming by HDIC to achieve oracle property

Although Theorem 4 shows that $\tilde{k}_n$ can be consistently estimated by $\hat{k}_n$, $\hat{J}_{\hat{k}_n}$ may still contain irrelevant variables that are included along the OGA path, as will be shown in Example 3 of Section 5. To exclude irrelevant variables, we make use of HDIC to define a subset $\hat{N}_n$ of $\hat{J}_{\hat{k}_n}$ by

$$\hat{N}_n = \{\hat{j}_l : \mathrm{HDIC}(\hat{J}_{\hat{k}_n} - \{\hat{j}_l\}) > \mathrm{HDIC}(\hat{J}_{\hat{k}_n}), 1 \le l \le \hat{k}_n\} \text{ if } \hat{k}_n > 1, \tag{4.21}$$

and $\hat{N}_n = \{\hat{j}_1\}$ if $\hat{k}_n = 1$. Note that (4.21) only requires the computation of $\hat{k}_n - 1$ additional least squares estimates and their associated residual sum of squares $\sum_{t=1}^n (y_t - \hat{y}_{t;\hat{J}_{\hat{k}_n} - \{\hat{j}_l\}})^2, 1 \le l < \hat{k}_n$, in contrast to the intractable combinatorial optimization problem of choosing the subset with the smallest extended BIC among all non-empty subsets of $\{1, \ldots, p_n\}$, for which Chen and Chen (2008, Thm. 1) established variable selection consistency under an "asymptotic identifiability" condition and $p_n = O(n^\kappa)$ for some $\kappa > 0$. The following theorem establishes the oracle property of $\hat{N}_n$ and shows that this much simpler procedure, which is denoted by OGA+HDIC+Trim, achieves variable selection consistency.

Theorem 5. *Under the same assumptions as in Theorem* 4, $\lim_{n\to\infty} P(\hat{N}_n = N_n) = 1$.

Proof. As in the proof of Theorem 4, assume $\sigma_j^2 = 1$ and drop the subscript n in $\tilde{k}_n$ and $\hat{k}_n$. For $\tilde{k} > 1$, define $\delta_l = 1$ if $\mathrm{HDIC}(\hat{J}_{\tilde{k}} - \{\hat{j}_l\}) > \mathrm{HDIC}(\hat{J}_{\tilde{k}})$ and $\delta_l = 0$ otherwise. Then

$$\begin{aligned} P(\hat{N}_n \ne N_n) &\le P(\hat{N}_n \ne N_n, \hat{k} > 1, N_n \subseteq \hat{J}_{\hat{k}}) + P(N_n \not\subseteq \hat{J}_{\hat{k}}) + P(\hat{N}_n \ne N_n, \hat{k} = 1) \\ &\le P(\delta_l = 1 \text{ and } \beta_{\hat{j}_l} = 0 \text{ for some } 1 \le l \le \tilde{k},\ N_n \subseteq \hat{J}_{\tilde{k}}, \tilde{k} > 1) \\ &\quad + P(\delta_l = 0 \text{ and } \beta_{\hat{j}_l} \ne 0 \text{ for some } 1 \le l \le \tilde{k},\ N_n \subseteq \hat{J}_{\tilde{k}}, \tilde{k} > 1) \\ &\quad + P(\hat{k} \ne \tilde{k}) + P(N_n \not\subseteq \hat{J}_{\hat{k}}) + P(\hat{N}_n \ne N_n, \hat{k} = 1). \end{aligned} \tag{4.22}$$

With $\hat{C}_n$ given in (4.13) and v_n given below (4.14), let

$$\begin{aligned} \mathcal{G}_n = &\Big\{ \max_{\sharp(J)\le \tilde{k}-1, i\notin J} |n^{-1} \sum_{t=1}^n \varepsilon_t \hat{x}_{ti;J}^{\perp}| \ge \theta n^{-\gamma/2} \Big\} \\ &\bigcup \Big\{ |\hat{C}_n| \ge \frac{\theta n^{1-\gamma}}{w_n \log p_n} \Big\} \bigcup \Big\{ \lambda_{\min}(\hat{\Gamma}(\hat{J}_{\tilde{k}})) \le \frac{v_n}{2} \Big\}. \end{aligned}$$

By (C5) and arguments similar to those in (4.11)-(4.15), there exists $\theta > 0$ such that for all large n,

$$P(\delta_l = 0 \text{ and } \beta_{\hat{j}_l} \neq 0 \text{ for some } 1 \le l \le \tilde{k}, N_n \subseteq \hat{J}_{\tilde{k}}, \tilde{k} > 1) \le P(\mathcal{G}_n) = o(1). \tag{4.23}$$

Moreover, by arguments similar to those used in (4.16)-(4.20), there exists $\theta > 0$ such that for all large n,

$$P(\delta_l = 1 \text{ and } \beta_{\hat{j}_l} = 0 \text{ for some } 1 \le l \le \tilde{k}, N_n \subseteq \hat{J}_{\tilde{k}}, \tilde{k} > 1) \le P(\mathcal{H}_n) = o(1), \tag{4.24}$$

where $\mathcal{H}_n = \{\hat{a}_{1,n} + \hat{b}_{1,n} \ge \theta w_n \log p_n\} \bigcup \{|\hat{C}_n| \ge \theta\}$, in which and $\hat{a}_{1,n}$ and $\hat{b}_{1,n}$ are the same as $\hat{a}_n$ and $\hat{b}_n$ in (4.18) but with K_n replaced by $\tilde{k}$, and $\tilde{k}$ replaced by $\tilde{k} - 1$. By Theorem 4, $P(\hat{k} \neq \tilde{k}) + P(N_n \not\subseteq \hat{J}_{\hat{k}}) + P(\hat{N}_n \neq N_n, \hat{k} = 1) = o(1)$, which can be combined with (4.22)−(4.24) to yield the desired conclusion.

5. Simulation Studies

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

$$y_t = \sum_{j=1}^{q} \beta_j x_{tj} + \sum_{j=q+1}^{p} \beta_j x_{tj} + \varepsilon_t, \ t = 1, \ldots, n, \tag{5.1}$$

where $\beta_{q+1} = \cdots = \beta_p = 0$, $p \gg n$, ε_t are i.i.d. $\mathrm{N}(0, \sigma^2)$ and are independent of the x_{tj}. Although (5.1) is a special case of (1.1) with $\alpha = 0$, we do not assume prior knowledge of the value of α and estimate α by $\bar{y} + \sum_{j=1}^{p} \hat{\beta}_j \bar{x}_j$, which is equivalent to centering the y_t and x_{tj} by their sample means, as noted in the first paragraph of Section 2.

Examples 1 and 2 consider the case

$$x_{tj} = d_{tj} + \eta w_t, \tag{5.2}$$

in which $\eta \ge 0$ and $(d_{t1}, \ldots, d_{tp}, w_t)^\top, 1 \le t \le n$, are i.i.d. normal with mean $(1, \ldots, 1, 0)^\top$ and covariance matrix $\mathbf{I}$. Since for any $J \subseteq \{1, \ldots, p\}$ and $1 \le i \le p$ with $i \notin J$,

$$\lambda_{\min}(\mathbf{\Gamma}(J)) = \frac{1}{1 + \eta^2} > 0 \text{ and } \|\mathbf{\Gamma}^{-1}(J)\mathbf{g}_i(J)\|_1 \le 1, \tag{5.3}$$

(3.2) is satisfied; moreover, $\mathrm{Corr}(x_{tj}, x_{tk}) = \eta^2/(1 + \eta^2)$ increases with $\eta > 0$. On the other hand,

$$\max_{1 \le \sharp(J) \le \nu} \lambda_{\max}(\mathbf{\Gamma}(J)) = \frac{1 + \nu\eta^2}{1 + \eta^2}. \tag{5.4}$$

As noted in Section 4.1, Fan and Lv (2008) require $\lambda_{\max}(\mathbf{\Gamma}(\{1,\ldots,p\})) \le cn^r$ for some $c > 0$ and $0 \le r < 1$ in their theory for the sure independence screening method, but this fails to hold for the equi-correlated regressors (5.2) when $\eta > 0$ and $p \gg n$, in view of (5.4). Although Fan and Song (2010) have recently made use of empirical process techniques to remove this condition, they require additional conditions in their Section 5.3 for "controlling false selection rates" by SIS or ISIS. As will be shown in Example 2, ISIS can include all relevant regressors but still have high false selection rates (or, equivalently, serious overfitting) when $\lambda_{\max}(\mathbf{\Gamma}(\{1,\ldots,p\}))$ is larger than n. For nonrandom regressors for which there is no population correlation matrix $\mathbf{\Gamma}(J)$ and the sample version $\hat{\mathbf{\Gamma}}(J)$ is nonrandom, Zhang and Huang (2008) have shown that under the sparse Riesz condition $c_* \le \min_{1\le \sharp(J)\le q^*} \lambda_{\min}(\hat{\mathbf{\Gamma}}(J)) \le \max_{1\le\sharp(J)\le q^*} \lambda_{\max}(\hat{\mathbf{\Gamma}}(J)) \le c^*$ for some $c^* \ge c_* > 0$ and $q^* \ge \{2 + (4c^*/c_*)\}q + 1$, the set of regressors selected by Lasso includes all relevant regressors, with probability approaching 1. If these fixed regressors are actually a realization of (5.2), then in view of (5.3) and (5.4), it is difficult to meet the requirement $q^* \ge \{2+(4c^*/c_*)\}q+1$ in the sparse Riesz condition when $q \ge (2\eta)^{-2}$.

Example 1. Consider (5.1) with $q = 5$, $(\beta_1,\ldots,\beta_5) = (3, -3.5, 4, -2.8, 3.2)$, and assume that $\sigma = 1$ and (5.2) holds. The special case $\eta = 0$, $\sigma = 1$ or 0.1, and $(n,p) = (50, 200)$ or $(100, 400)$ was used by Shao and Chow (2007) to illustrate the performance of their variable screening method. The cases $\eta = 0, 2$ and (n,p) =(50, 1,000), (100, 2,000), (200, 4,000) are considered here to accommodate a much larger number of candidate variables and allow substantial correlations among them. In light of Theorem 4 which requires the number K_n of iterations to satisfy $K_n = O((n/\log p_n)^{1/2})$, we choose $K_n = \lfloor 5(n/\log p_n)^{1/2}\rfloor$. We have also allowed D in $K_n = \lfloor D(n/\log p_n)^{1/2}\rfloor$ to vary between 3 and 10, and the results are similar to those for $D = 5$. Table 1 shows that OGA+HDBIC, OGA+HDHQ and OGA+HDBIC+Trim perform well, in agreement with the asymptotic theory of Theorems 4 and 5. Each result is based on 1,000 simulations. Here and in the sequel, we choose $c = 2.01$ for HDHQ. We have allowed c in HDHQ to vary among 2.01, 2.51, 3.01, 3.51 and 4.01, but the results are quite similar for the different choices of c. In the simulations for $n \ge 100$, OGA always includes the 5 relevant regressors within K_n iterations, and HDBIC and HDHQ identify the smallest correct model for 99% or more of the simulations, irrespective of whether the candidate regressors are uncorrelated ($\eta = 0$) or highly correlated ($\eta = 2$). The performance of OGA+HDBIC+Trim is even better because it can choose the smallest correct model in all simulations.

For comparison, we have also included in Table 1 the performance of OGA +BIC and Wang's (2009) forward regression procedure, denoted by FR, that carries out OGA with n iterations and then chooses $\hat{J}_{\hat{m}_n}$ as the final set of

regressors, where $\hat{m}_n = \arg\min_{1\le m\le n}\mathrm{EBIC}(\hat{J}_m)$, in which EBIC is defined by (4.3) with $\gamma = 1$ and $\tau_j = p^j$. Table 1 shows that for $n \ge 100$, FR (or OGA+BIC) always chooses the largest model along the OGA path, which is $\hat{J}_n$ (or $\hat{J}_{K_n}$). Examination of the simulation runs shows that $\mathrm{BIC}(\hat{J}_k)$ is a decreasing function of k and that $\mathrm{EBIC}(\hat{J}_k)$ is initially decreasing, then increasing and eventually decreasing in k. Define the mean squared prediction errors

$$\mathrm{MSPE} = \frac{1}{1,000}\sum_{l=1}^{1,000}\Big(\sum_{j=1}^{p}\beta_j x_{n+1,j}^{(l)} - \hat{y}_{n+1}^{(l)}\Big)^2, \tag{5.5}$$

in which $x_{n+1,1}^{(l)},\ldots,x_{n+1,p}^{(l)}$ 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 $y_{n+1}^{(l)}$. The MSPEs of OGA+BIC are at least 16 (or 23) times those of OGA+HDBIC, OGA+HDBIC+Trim and OGA+HDHQ when $n = 100$ (or $n = 200$). The MSPEs of FR are even larger than those of OGA+BIC due to more serious overfitting when OGA terminates after n (instead of K_n) iterations. In the case of $n = 50$ and $p = 1{,}000$, OGA can include all relevant regressors (within K_n iterations) about 92% of the time when $\eta = 0$; this ratio decreases to 80% when $\eta = 2$. In addition, HDBIC identifies the smallest correct model for 86% when $\eta = 0$ and 63% when $\eta = 2$. These latter two ratios, however, can be increased to 92% and 79%, respectively, by using the trimming method in Section 4.3. As shown in Table 2 of Shao and Chow (2007), for the case of $n = 100$ and $p = 400$, applying their variable screening method in conjunction with AIC or BIC can only identify the smallest correct model about 50% of the time even when $\eta = 0$. On the other hand, BIC used in conjunction with OGA that terminates after K_n iterations can include all relevant variables in this case, but it also includes all irrelevant variables, as shown in Table 1. Note that p is 20 times the value of n, resulting in many spuriously significant regression coefficients if one does not adjust for multiple testing. The factor $w_n \log p_n$ in the definition (4.8) of HDIC can be regarded as such adjustment, as explained in the paragraph preceding Theorem 4.

Example 2. Consider (5.1) with $q = 9$, $n = 400$, $p = 4{,}000$, $(\beta_1,\ldots,\beta_q)$=(3.2, 3.2, 3.2, 3.2, 4.4, 4.4, 3.5, 3.5, 3.5), and assume that $\sigma^2 = 2.25$ and (5.2) holds with $\eta = 1$. This example satisfies Meinshausen and Bühlmann's (2006) neighborhood stability condition that requires that for some $0 < \delta < 1$ and all $i = q+1,\ldots,p$,

$$|\mathbf{c}_{qi}'\mathbf{R}^{-1}(q)\mathrm{sign}(\boldsymbol{\beta}(q))| < \delta, \tag{5.6}$$

where $\mathbf{x}_t(q) = (x_{t1},\ldots,x_{tq})^\top$, $\mathbf{c}_{qi} = E(\mathbf{x}_t(q)x_{ti})$, $\mathbf{R}(q) = E(\mathbf{x}_t(q)\mathbf{x}_t^\top(q))$, and $\mathrm{sign}(\boldsymbol{\beta}(q)) = (\mathrm{sign}(\beta_1),\ldots,\mathrm{sign}(\beta_q))^\top$. To show that (5.6) holds in this example,

Table 1. Frequency, in 1,000 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), and of selecting the largest model, along the OGA path, which includes all relevant variables (E*).

η	n	p	Method	E	E+1	E+2	E+3	E+4	E+5	E*	Correct	MSPE
0	50	1,000	OGA+HDHQ	812	86	19	8	0	0	1	926	6.150
			OGA+HDBIC	862	52	7	1	0	0	0	922	6.550
			OGA+HDBIC+Trim	919	3	0	0	0	0	0	922	6.550
			OGA+BIC	0	0	0	0	0	0	926	926	8.310
			FR	0	0	0	0	0	0	926	926	8.300
	100	2,000	OGA+HDHQ	993	6	0	1	0	0	0	1,000	0.065
			OGA+HDBIC	999	0	0	1	0	0	0	1,000	0.064
			OGA+HDBIC+Trim	1,000	0	0	0	0	0	0	1,000	0.064
			OGA+BIC	0	0	0	0	0	0	1,000	1,000	1.320
			FR	0	0	0	0	0	0	1,000	1,000	1.729
	200	4,000	OGA+HDHQ	999	1	0	0	0	0	0	1,000	0.034
			OGA+HDBIC	1,000	0	0	0	0	0	0	1,000	0.034
			OGA+HDBIC+Trim	1,000	0	0	0	0	0	0	1,000	0.034
			OGA+BIC	0	0	0	0	0	0	1,000	1,000	0.796
			FR	0	0	0	0	0	0	1,000	1,000	1.612
2	50	1,000	OGA+HDHQ	609	140	36	15	5	2	0	807	13.250
			OGA+HDBIC	629	130	29	5	0	0	0	793	14.110
			OGA+HDBIC+Trim	792	1	0	0	0	0	0	793	14.100
			OGA+BIC	0	0	0	0	0	0	807	807	14.660
			FR	0	0	0	0	0	0	807	807	14.990
	100	2,000	OGA+HDHQ	988	9	3	0	0	0	0	1,000	0.070
			OGA+HDBIC	994	3	3	0	0	0	0	1,000	0.069
			OGA+HDBIC+Trim	1,000	0	0	0	0	0	0	1,000	0.069
			OGA+BIC	0	0	0	0	0	0	1,000	1,000	1.152
			FR	0	0	0	0	0	0	1,000	1,000	1.537
	200	4,000	OGA+HDHQ	1,000	0	0	0	0	0	0	1,000	0.033
			OGA+HDBIC	1,000	0	0	0	0	0	0	1,000	0.033
			OGA+HDBIC+Trim	1,000	0	0	0	0	0	0	1,000	0.033
			OGA+BIC	0	0	0	0	0	0	1,000	1,000	0.779
			FR	0	0	0	0	0	0	1,000	1,000	1.688

straightforward calculations give $\mathbf{c}_{qi} = \eta^2 \mathbf{1}_q$, $\mathbf{R}^{-1}(q) = \mathbf{I} - \{\eta^2/(1+\eta^2 q)\}\mathbf{1}_q \mathbf{1}_q^\top$, and $\operatorname{sign}(\boldsymbol{\beta}(q)) = \mathbf{1}_q$, where $\mathbf{1}_q$ is the q-dimensional vector of 1's. Therefore, for all $i = q+1, \ldots, p$, $|\mathbf{c}_{qi}' \mathbf{R}^{-1}(q) \operatorname{sign}(\boldsymbol{\beta}(q))| = \eta^2 q/(1+\eta^2 q) < 1$. Under (5.6) and some other conditions, Meinshausen and Bühlmann (2006, Thms. 1 and 2) have shown that if $r = r_n$ in the Lasso estimate (3.21) converges to 0 at a rate slower than $n^{-1/2}$, then $\lim_{n\to\infty} P(\hat{L}_n = N_n) = 1$, where $\hat{L}_n$ is the set of regressors whose associated regression coefficients estimated by Lasso(r_n) are nonzero.

Table 2 compares the performance of OGA+HDBIC, OGA+HDHQ and

OGA+HDBIC+
Trim, using $K_n = \lfloor 5(n/\log p)^{1/2} \rfloor$ iterations for OGA, with that of SIS-SCAD, ISIS-SCAD, LARS, Lasso, and adaptive Lasso. To implement Lasso, we use the `Glmnet` package in `R` (Friedman, Hastie, and Tibshirani (2010)) that conducts 5-fold cross-validation to select the optimal penalty r, yielding the estimate Lasso in Table 2. To implement LARS, we use the `LARS` package in `R` (Hastie and Efron (2007)) and conduct 5-fold cross-validation to select the optimal tuning parameter, yielding the estimate LARS in Table 2. For adaptive lasso, we use the `parcor` package in `R` (Kraemer and Schaefer (2010)), which uses an initial Lasso estimate to calculate the weights for the final weighted Lasso estimate. Table 2 shows that OGA+HDBIC and OGA+HDHQ can choose the smallest correct model 98% of the time and choose slightly overfitting models 2% of the time. Moreover, OGA+HDBIC+Trim can choose the smallest correct model in all simulations. The MSPEs of these three methods are near the oracle value $q\sigma^2/n = 0.051$. On the other hand, although Lasso and LARS can include the 9 relevant variables in all simulation runs, they encounter overfitting problems. The smallest number of variables selected by Lasso (or LARS) was 31 (or 97) in the 1,000 simulations, while the largest number was 84 (or 300). Although the model chosen by Lasso is more parsimonious than that chosen by LARS, the MSPE of Lasso is larger than that of LARS. Adaptive Lasso can choose the smallest correct model in all 1,000 simulations. However, its MSPE is still over 5 times, while those of LARS and Lasso are over 10 times the value of $q\sigma^2/n$.

To implement SIS and ISIS followed by the SCAD regularization method (instead of Lasso that uses the l_1-penalty), we use the `SIS` package in `R` (Fan et al. (2010)), which provides the estimates SIS-SCAD and ISIS-SCAD in Table 2. As shown in Table 2, although ISIS-SCAD can include the 9 relevant variables in all simulation runs, it encounters overfitting problems and the resulting MSPE is about 29 times the value of $q\sigma^2/n$. Moreover, SIS-SCAD performs much worse than ISIS-SCAD. It includes the 9 relevant variables in 28.9% of the simulations and its MSPE is over 10 times that of ISIS-SCAD. Besides the mean of the squared prediction errors in Table 2, we also give in Table 3 a 5-number summary of $\{(\sum_{j=1}^{p} \beta_j x_{n+1,j}^{(l)} - \hat{y}_{n+1}^{(l)})^2 : 1 \le l \le 1,000\}$.

Example 3. Let $q = 10$, $(\beta_1, \ldots, \beta_q)$=(3, 3.75, 4.5, 5.25, 6, 6.75, 7.5, 8.25, 9, 9.75), $n = 400$, and $p = 4{,}000$ in (5.1). Assume that $\sigma = 1$, that $x_{t1}, \ldots, x_{tq}$ are i.i.d. standard normal, and

$$x_{tj} = d_{tj} + b \sum_{l=1}^{q} x_{tl}, \text{ for } q+1 \le j \le p, \tag{5.7}$$

where $b = (3/4q)^{1/2}$ and $(d_{t(q+1)}, \ldots, d_{tp})^\top$ are i.i.d. multivariate normal with mean $\mathbf{0}$ and covariance matrix $(1/4)\mathbf{I}$ and are independent of x_{tj} for $1 \le j \le q$.

Table 2. Frequency, in 1,000 simulations, of including all nine relevant variables and selecting all relevant variables in Example 2; see notation in Table 1.

Method	E	E+1	E+2	E+3	Correct	MSPE
OGA+HDHQ	980	18	1	1	1,000	0.067
OGA+HDBIC	982	16	1	1	1,000	0.067
OGA+HDBIC+Trim	1,000	0	0	0	1,000	0.066
SIS-SCAD	127	19	10	14	289	15.170
ISIS-SCAD	0	0	0	0	1,000	1.486
Adaptive Lasso	1,000	0	0	0	1,000	0.289
LARS	0	0	0	0	1,000	0.549
Lasso	0	0	0	0	1,000	0.625

Table 3. 5-number summaries of squared prediction errors in 1,000 simulations.

Method	Minimum	1st Quartile	Median	3nd Quartile	Maximum
OGA+HDHQ	0.000	0.006	0.026	0.084	1.124
OGA+HDBIC	0.000	0.006	0.026	0.084	1.124
OGA+HDBIC+Trim	0.000	0.005	0.026	0.084	1.124
SIS-SCAD	0.000	0.100	2.498	17.210	507.200
ISIS-SCAD	0.000	0.153	0.664	1.845	21.340
Adaptive Lasso	0.000	0.030	0.118	0.360	3.275
LARS	0.000	0.047	0.224	0.719	5.454
Lasso	0.000	0.067	0.280	0.823	7.399

Using the same notation as in the first paragraph of Example 2, straightforward calculations show that for $q+1 \le j \le p$, $\mathbf{c}_{qj} = (b, \ldots, b)^\top$, $\mathbf{R}(q) = \mathbf{I}$, and $\text{sign}(\boldsymbol{\beta}(q)) = (1, \ldots, 1)^\top$. Therefore, for $q+1 \le j \le p$, $|\mathbf{c}_{qj}^\top \mathbf{R}^{-1}(q)\text{sign}(\boldsymbol{\beta}(q))| = (3q/4)^{1/2} = (7.5)^{1/2} > 1$, and hence (5.6) is violated. On the other hand, it is not difficult to show that (3.2) is satisfied in this example and that

$$\min_{q+1 \le i \le p} |E(x_i y)| > \max_{1 \le i \le q} |E(x_i y)|. \tag{5.8}$$

In fact, $|E(x_i y)| = 24.69$ for all $q+1 \le i \le p$ and $\max_{1 \le i \le q} |E(x_i y)| = \beta_q = 9.75$. Making use of (5.8) and Lemmas A.2 and A.4 in Appendix A, it can be shown that $\lim_{n \to \infty} P(\hat{J}_1 \subseteq \{1, \ldots, q\}) = 0$, and therefore with probability approaching 1, the first iteration of OGA selects an irrelevant variable, which remains in the OGA path until the last iteration.

Table 4 shows that although OGA+HDHQ and OGA+HDBIC fail to choose the smallest set of relevant regressors in all 1,000 simulations, consistent with the above asymptotic theory, they include only 1−3 irrelevant variables while correctly including all relevant variables. Moreover, by using HDBIC to define the subset (4.21) of $\hat{J}_{\hat{k}_n}$, OGA+HDBIC+Trim is able to choose all relevant variables

Table 4. Frequency, in 1,000 simulations, of including all nine relevant variables and selecting all relevant variables in Example 3; see notation in Table 1.

Method	E	E+1	E+2	E+3	Correct	MSPE
OGA+HDHQ	0	39	945	16	1,000	0.035
OGA+HDBIC	0	39	945	16	1,000	0.035
OGA+HDBIC+Trim	1,000	0	0	0	1,000	0.028
SIS-SCAD	0	0	0	0	0	51.370
ISIS-SCAD	0	0	0	0	1,000	0.734
Adaptive Lasso	0	0	0	0	0	27.270
LARS	0	0	0	0	0	0.729
Lasso	0	0	0	0	0	2.283

without including any irrelevant variables in all 1,000 simulations, and its MSPE is close to the oracle value of $q\sigma^2/n = 0.025$ while those of OGA+HDBIC and OGA+HDHQ are somewhat larger. Similar to Example 2, ISIS-SCAD includes all relevant regressors in all 1,000 simulations, but also includes many irrelevant regressors. Its MSPE is 0.73, which is about 29 times the value of $q\sigma^2/n$. The performance of SIS-SCAD is again much worse than that of ISIS-SCAD. It fails to include all relevant regressors in all 1,000 simulations and its MSPE is about 70 times that of ISIS-SCAD.

Like SIS-SCAD, LARS, Lasso, and adaptive Lasso also fail to include all 10 relevant regressors in all 1,000 simulations, even though they also include many irrelevant variables. The smallest number of selected variables in the 1,000 simulations is 8 for adaptive Lasso, 234 for Lasso, and 372 for LARS. The average and the largest numbers of selected variables are 12.59 and 19 for adaptive Lasso, 272.2 and 308 for Lasso, and 393.7 and 399 for LARS. On the other hand, the MSPE of LARS is 0.73, which is about 1/3 of that of Lasso and 1/40 of that of adaptive Lasso. This example shows that when Lasso fails to have the sure screening property, adaptive Lasso, which relies on an initial estimate based on Lasso to determine the weights for a second-stage weighted Lasso, may not be able to improve Lasso and may actually perform worse. The example also illustrates an inherent difficulty with high-dimensional sparse linear regression when irrelevant input variables have substantial correlations with relevant ones. Assumptions on the design matrix are needed to ensure that this difficulty is surmountable; in particular, (3.17) or the second part of (3.2) can be viewed as a "sparsity" constraint, when a candidate irrelevant input variable is regressed on the set of variables already selected by the OGA path, to overcome this difficulty.

6. Concluding Remarks and Discussion

Forward stepwise regression is a popular regression method that seems to be particularly suitable for high-dimensional sparse regression models but has

encountered computational and statistical difficulties that hinder its use. On the computational side, direct implementation of least squares regression involves inverting high-dimensional covariance matrices and has led to the L_2-boosting method of Bühlmann and Yu (2003) that uses gradient descent instead. The statistical issue with forward stepwise regression is when to stop sequential variable addition and how to trim back after stopping so that a minimal number of variables can be included in the final regression model. The usual model selection criteria, such as Mallows' C_p, AIC and BIC, do not work well in the high-dimensional case, as we have noted in the second paragraph of Section 4. A major contribution of this paper is a high-dimensional information criterion (HDIC) to be used in conjunction with forward stepwise regression, implemented by OGA, and backward trimming, implemented by (4.21), so that OGA+HDIC+Trim has the oracle property of being equivalent to least squares regression on an asymptotically minimal set of relevant regressors under a strong sparsity assumption. The novel probabilistic arguments in the Appendix used in conjunction with Temlyakov's (2000) bounds for weak orthogonal greedy algorithms show how OGA+HDIC+Trim resolves the issue of spurious variable selection when the number of variables is much larger than the sample size. There are no comparable results in the literature except for Lasso and its variants. In Section 3.2, we have compared the rates of OGA with those of Lasso obtained by Bickel, Ritov, and Tsybakov (2009). In particular, they show that OGA can have substantially better rates than Lasso in some situations, and that the reverse is also true in some other situations. High-dimensional sparse regression is a difficult but important problem and needs an arsenal of methods to address different scenarios. Our results in Sections 3-5 have shown OGA+HDIC or OGA+HDIC+Trim to be worthy of inclusion in this arsenal, which now already includes Lasso and its variants.

Whereas OGA+HDIC+Trim is straightforward to implement, the convex program in Lasso requires numerical optimization and we have relied on open-source software to perform repeated simulations in reasonable time. As noted by Wainwright (2009, p.2183), although a "natural optimization-theoretic formulation" of the problem of estimating a high-dimensional linear regression vector $\boldsymbol{\beta}$ with mostly zero components is via "l_0-minimization, where the l_0-norm of a vector corresponds to the number of nonzero elements," l_0-minimization is "known to be NP-hard" and has motivated the use of "computationally tractable approximations or relaxations," among which is Lasso. Using the l_1-norm as a surrogate for the l_0-norm, Lasso has become very popular because it can be solved by convex programming in polynomial time with standard optimization software. The software packages in `R` used in Examples 2 and 3 have further facilitated the use of Lasso and adaptive Lasso in high-dimensional regression problems despite the

inherent complexity of these methods. In particular, in the simulations in Examples 2 and 3, `Glmnet` is relatively fast and its computation time is comparable to that of OGA+HDBIC+Trim, while the computation time of adaptive Lasso (or LARS) is about 8 (or 100) times that of `Glmnet`. It should be noted that these software packages have made many computational short cuts and simplifying approximations to the original convex optimization problem defining Lasso. On the other hand, the implementation of OGA+HDBIC+Trim is straightforward and does not require any approximation to speed it up.

Barron et al. (2008) have recently made use of empirical process theory to extend the convergence rates of OGA in noiseless models (i.e., $\varepsilon_t = 0$ in (1.1)) to regression models in which ε_t and $\mathbf{x}_t$ are bounded so that $|y_t| \le B$ for some known bound B. They need this bound to apply empirical process theory to the sequence of estimates $\hat{y}_m^{(B)}(\mathbf{x}) = \operatorname{sgn}(\hat{y}_m(\mathbf{x})) \min\{B, |\hat{y}_m(\mathbf{x})|\}$. They propose to terminate OGA after $\lfloor n^a \rfloor$ iterations for some $a \ge 1$ and to select m^* that minimizes $\sum_{i=1}^n \{y_i - \hat{y}_m^{(B)}(\mathbf{x}_i)\}^2 + \kappa m \log n$ over $1 \le m \le \lfloor n^a \rfloor$, for which they show choosing $\kappa \ge 2568B^4(a+5)$ yields their convergence result for $\hat{y}_{m^*}^{(B)}$. In comparison, the convergence rate and oracle inequality in Section 3 are sharper and are directly applicable to $\hat{y}_m$, while the model selection criterion in Section 4 has definitive oracle properties. Wang (2009) terminates OGA after n iterations and selects $\hat{m}_n$ that minimizes $\mathrm{EBIC}(\hat{J}_m)$ over $1 \le m \le n$. Although this method has the sure screening property under conditions that are much stronger than those of Theorem 4, Example 1 has shown that it has serious overfitting problems. Wang actually uses it to screen variables for a second-stage regression analysis using Lasso or adaptive Lasso. Forward stepwise regression followed by cross-validation as a screening method in high-dimensional sparse linear models has also been considered by Wasserman and Roeder (2009), who propose to use out-of-sample least squares estimates for the selected model after partitioning the data into a screening group and a remaining group for out-of-sample final estimation. By using OGA+HDIC+Trim instead, we can already achieve the oracle property without any further refinement.

Acknowledgement

Ing's research was supported by the National Science Council, Taiwan, R.O.C. Lai's research was supported by the National Science Foundation. The authors are grateful to Ray-Bing Chen and Feng Zhang for their help and stimulating discussions, and to Emmanuel Candes, Jerome Friedman, and Robert Tibshirani for valuable suggestions.

Appendix A: Proofs of (3.8), (3.13), (4.15) and (4.20)

The proof of (3.8) relies on the representation (3.7), whose right-hand side involves (i) a weighted sum of the i.i.d. random variables ε_t that satisfy (C2), and (ii) the difference between a nonlinear function of the sample covariance matrix of the x_{tj} that satisfy (C3) and its expected value, recalling that we have assumed $\sigma_j = 1$ in the proof of Theorem 1. The proof of (3.13) also makes use of a similar representation. The following four lemmas give exponential bounds for moderate deviation probabilities of (i) and (ii).

Lemma A.1. *Let $\varepsilon, \varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. random variables such that $E(\varepsilon) = 0$, $E(\varepsilon^2) = \sigma^2$ and* (C2) *holds. Then, for any constants $a_{ni} (1 \le i \le n)$ and $u_n > 0$ such that*

$$u_n \max_{1 \le i \le n} \frac{|a_{ni}|}{A_n} \to 0 \text{ and } \frac{u_n^2}{A_n} \to \infty \text{ as } n \to \infty, \tag{A.1}$$

where $A_n = \sum_{i=1}^n a_{ni}^2$, we have

$$P\{\sum_{i=1}^n a_{ni}\varepsilon_i > u_n\} \le \exp\Big\{\frac{-(1+o(1))u_n^2}{2\sigma^2 A_n}\Big\}. \tag{A.2}$$

Proof. Let $e^{\psi(\theta)} = E(e^{\theta\varepsilon})$, which is finite for $|\theta| < t_0$ by (C2). By the Markov inequality, if $\theta > 0$ and $\max_{1 \le i \le n} |\theta a_{ni}| < t_0$, then

$$P\Big\{\sum_{i=1}^n a_{ni}\varepsilon_i > u_n\Big\} \le \exp\Big\{-\theta u_n + \sum_{i=1}^n \psi(\theta a_{ni})\Big\}. \tag{A.3}$$

By (A.1) and the Taylor approximation $\psi(t) \sim \sigma^2 t^2/2$ as $t \to 0$, $\theta u_n - \sum_{i=1}^n \psi(\theta a_{ni})$ is minimized at $\theta \sim u_n/(\sigma^2 A_n)$ and has minimum value $u_n^2/(2\sigma^2 A_n)$. Putting this minimum value in (A.3) proves (A.2).

Lemma A.2. *With the same notation and assumptions as in Theorem* 1, *and assuming that $\sigma_j = 1$ for all j so that $z_j = x_j$, there exists $C > 0$ such that*

$$\max_{1 \le i,j \le p_n} P\{|\sum_{t=1}^n (x_{ti}x_{tj} - \sigma_{ij})| > n\delta_n\} \le \exp(-Cn\delta_n^2) \tag{A.4}$$

for all large n, where $\sigma_{ij} = \mathrm{Cov}(x_i, x_j)$ and δ_n are positive constants satisfying $\delta_n \to 0$ and $n\delta_n^2 \to \infty$ as $n \to \infty$. Define $\mathbf{\Gamma}(J)$ by (3.1) *and let $\hat{\mathbf{\Gamma}}_n(J)$ be the corresponding sample covariance matrix. Then, for all large n,*

$$P\{\max_{1 \le \sharp(J) \le K_n} \|\hat{\mathbf{\Gamma}}_n(J) - \mathbf{\Gamma}(J)\| > K_n\delta_n\} \le p_n^2 \exp(-Cn\delta_n^2). \tag{A.5}$$

If furthermore $K_n\delta_n = O(1)$, then there exists $c > 0$ such that

$$P\{\max_{1\le\sharp(J)\le K_n} \|\hat{\mathbf{\Gamma}}_n^{-1}(J) - \mathbf{\Gamma}^{-1}(J)\| > K_n\delta_n\} \le p_n^2 \exp(-cn\delta_n^2) \tag{A.6}$$

for all large n, where $\hat{\mathbf{\Gamma}}_n^{-1}$ denotes a generalized inverse when $\hat{\mathbf{\Gamma}}_n$ is singular.

Proof. Since (x_{ti}, x_{tj}) are i.i.d. and (C3) holds, the same argument as that in the proof of Lemma A.1 can be used to prove (A.4) with $C < 1/(2\mathrm{Var}(x_i x_j))$. Letting $\triangle_{ij} = n^{-1}\sum_{t=1}^n x_{ti}x_{tj} - \sigma_{ij}$, note that $\max_{1\le\sharp(J)\le K_n} \|\hat{\mathbf{\Gamma}}_n(J) - \mathbf{\Gamma}(J)\| \le K_n \max_{1\le i,j\le p_n} |\triangle_{ij}|$ and therefore (A.5) follows from (A.4). Since $\lambda_{\min}(\hat{\mathbf{\Gamma}}_n(J)) \ge \lambda_{\min}(\mathbf{\Gamma}(J)) - \|\hat{\mathbf{\Gamma}}_n(J) - \mathbf{\Gamma}(J)\|$, it follows from (3.2) and (A.5) that the probability of $\hat{\mathbf{\Gamma}}_n(J)$ being singular is negligible in (A.6), for which we can therefore assume $\hat{\mathbf{\Gamma}}_n(J)$ to be nonsingular.

To prove (A.6), denote $\hat{\mathbf{\Gamma}}_n(J)$ and $\mathbf{\Gamma}(J)$ by $\hat{\mathbf{\Gamma}}$ and $\mathbf{\Gamma}$ for simplicity. Making use of $\hat{\mathbf{\Gamma}}^{-1} - \mathbf{\Gamma}^{-1} = \mathbf{\Gamma}^{-1}(\mathbf{\Gamma} - \hat{\mathbf{\Gamma}})\hat{\mathbf{\Gamma}}^{-1}$ and $\hat{\mathbf{\Gamma}} = \mathbf{\Gamma}\{\mathbf{I} + \mathbf{\Gamma}^{-1}(\hat{\mathbf{\Gamma}} - \mathbf{\Gamma})\}$, it can be shown that $\|\hat{\mathbf{\Gamma}}^{-1} - \mathbf{\Gamma}^{-1}\|(1 - \|\mathbf{\Gamma}^{-1}\|\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|) \le \|\mathbf{\Gamma}^{-1}\|^2\|\mathbf{\Gamma} - \hat{\mathbf{\Gamma}}\|$, and hence

$$\begin{aligned}&\max_{1\le\sharp(J)\le K_n} \|\hat{\mathbf{\Gamma}}^{-1} - \mathbf{\Gamma}^{-1}\|(1 - \max_{1\le\sharp(J)\le K_n} \|\mathbf{\Gamma}^{-1}\|\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|)\\ &\quad\le \max_{1\le\sharp(J)\le K_n} \|\mathbf{\Gamma}^{-1}\|^2\|\mathbf{\Gamma} - \hat{\mathbf{\Gamma}}\|.\end{aligned} \tag{A.7}$$

By (3.2), $\max_{1\le\sharp(J)\le K_n} \|\mathbf{\Gamma}^{-1}\| \le \delta^{-1}$ for all large n. Letting $G = \sup_{n\ge1} K_n\delta_n$, we bound $P\{\max_{1\le\sharp(J)\le K_n} \|\hat{\mathbf{\Gamma}}^{-1} - \mathbf{\Gamma}^{-1}\| > K_n\delta_n\}$ by

$$\begin{aligned}&P\left\{\max_{1\le\sharp(J)\le K_n} \|\hat{\mathbf{\Gamma}}^{-1} - \mathbf{\Gamma}^{-1}\| > K_n\delta_n, \max_{1\le\sharp(J)\le K_n} \|\mathbf{\Gamma}^{-1}\|\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}\| \le \frac{K_n\delta_n}{G+1}\right\}\\ &\quad + P\left\{\max_{1\le\sharp(J)\le K_n} \|\mathbf{\Gamma}^{-1}\|\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}\| > \frac{K_n\delta_n}{G+1}\right\}\\ &\le P\left\{\max_{1\le\sharp(J)\le K_n} \|\mathbf{\Gamma}^{-1}\|^2\|\mathbf{\Gamma} - \hat{\mathbf{\Gamma}}\| > \frac{K_n\delta_n}{G+1}\right\}\\ &\quad + P\left\{\max_{1\le\sharp(J)\le K_n} \|\mathbf{\Gamma}^{-1}\|\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}\| > \frac{K_n\delta_n}{G+1}\right\},\end{aligned} \tag{A.8}$$

in view of (A.7) and $1 - (G+1)^{-1}K_n\delta_n \ge (G+1)^{-1}$. Since $\max_{1\le\sharp(J)\le K_n} \|\mathbf{\Gamma}^{-1}\|^2 \le \delta^{-2}$ for all large n, combining (A.8) with (A.5) (in which δ_n is replaced by $\delta^2\delta_n/(G+1)$ for the first summand in (A.8), and by $\delta\delta_n/(G+1)$ for the second) yields (A.6) with $c < C\delta^4/(G+1)^2$.

Lemma A.3. *With the same notation and assumptions as in Theorem 1 and assuming $\sigma_j = 1$ for all j, let $n_1 = \sqrt{n}/(\log n)^2$ and $n_{k+1} = \sqrt{n_k}$ for $k \ge 1$. Let u_n be positive constants such that $u_n/n_1 \to \infty$ and $u_n = O(n)$. Let K be a*

positive integer and $\Omega_n = \{\max_{1\le t\le n}|\varepsilon_t| < (\log n)^2\}$. *Then there exists* $\alpha > 0$ *such that for all large* n,

$$\max_{1\le i\le p_n} P\{\max_{1\le t\le n}|x_{ti}| \ge n_1\} \le \exp(-\alpha n_1^2), \tag{A.9}$$

$$\max_{1\le k\le K, 1\le i\le p_n} P\Big(\sum_{t=1}^{n}|\varepsilon_t x_{ti}|I_{\{n_{k+1}\le|x_{ti}|<n_k\}} \ge u_n(\log n)^2, \Omega_n\Big) \le \exp(-\alpha u_n). \tag{A.10}$$

Proof. (A.9) follows from (C3). To prove (A.10), note that on Ω_n,

$$\sum_{t=1}^{n}|\varepsilon_t x_{ti}|I_{\{n_{k+1}\le|x_{ti}|<n_k\}} \le n_k(\log n)^2\sum_{t=1}^{n}I_{\{|x_{ti}|\ge n_{k+1}\}}.$$

Therefore it suffices to show that for all large n and $1\le i\le p_n$, $1\le k\le K$,

$$\exp(-\alpha u_n) \ge P(\sum_{t=1}^{n}I_{\{|x_{ti}|\ge n_{k+1}\}} \ge \frac{u_n}{n_k}) = P\{\text{Binomial}(n,\pi_{n,k,i}) \ge \frac{u_n}{n_k}\},$$

where $\pi_{n,k,i} = P(|x_i| \ge n_{k+1}) \le \exp(-cn_{k+1}^2) = \exp(-cn_k)$ for some $c > 0$, by (C3). The desired conclusion follows from standard bounds for the tail probability of a binomial distribution, recalling that $u_n = O(n)$ and $u_n/n_1 \to \infty$.

Lemma A.4. *With the same notation and assumptions as in Lemma* A.3, *let* δ_n *be positive numbers such that* $\delta_n = O(n^{-\theta})$ *for some* $0 < \theta < 1/2$ *and* $n\delta_n^2 \to \infty$. *Then there exists* $\beta > 0$ *such that for all large* n,

$$\max_{1\le i\le p_n} P\Big(|\sum_{t=1}^{n}\varepsilon_t x_{ti}| \ge n\delta_n, \Omega_n\Big) \le \exp(-\beta n\delta_n^2). \tag{A.11}$$

Proof. Let $n_i, i\ge 1$ be defined as in Lemma A.3. Let K be a positive integer such that $2^{-K} < \theta$. Then since $\delta_n = O(n^{-\theta})$, $n^{2^{-K}} = o(\delta_n^{-1})$. Letting $A^{(1)} = [n_1,\infty)$, $A^{(k)} = [n_k, n_{k-1})$ for $2\le k\le K$, $A^{(K+1)} = [0, n_K)$, note that

$$\begin{aligned} P\Big(|\sum_{t=1}^{n}\varepsilon_t x_{ti}| \ge n\delta_n, \Omega_n\Big) &\le \sum_{k=1}^{K+1} P\Big(|\sum_{t=1}^{n}\varepsilon_t x_{ti}I_{\{|x_{ti}|\in A^{(k)}\}}| \ge \frac{n\delta_n}{K+1}, \Omega_n\Big) \\ &\le P(\max_{1\le t\le n}|x_{ti}| \ge n_1) + \sum_{k=2}^{K+1} P\Big(|\sum_{t=1}^{n}\varepsilon_t x_{ti}I_{\{|x_{ti}|\in A^{(k)}\}}| \ge \frac{n\delta_n}{K+1}, \Omega_n\Big). \end{aligned} \tag{A.12}$$

From (A.10) (in which u_n is replaced by $n\delta_n/\{(K+1)(\log n)^2\}$), it follows that for $2\le k\le K$ and all large n,

$$\max_{1\le i\le p_n} P\Big(|\sum_{t=1}^{n}\varepsilon_t x_{ti}I_{\{|x_{ti}|\in A^{(k)}\}}| \ge \frac{n\delta_n}{K+1}, \Omega_n\Big) \le \exp\Big\{-\frac{\alpha n\delta_n}{(K+1)(\log n)^2}\Big\}, \tag{A.13}$$

where α is some positive constant. Moreover, by (A.9), $\max_{1\le i\le p_n} P(\max_{1\le t\le n} |x_{ti}| \ge n_1) \le \exp\{-\alpha n/(\log n)^4\}$ for some $\alpha > 0$. Putting this bound and (A.13) into (A.12) and noting that $1/\{(\log n)^2\delta_n\} \to \infty$, it remains to show for some $a_1 > 0$ and all large n,

$$\max_{1\le i\le p_n} P\Big(|\sum_{t=1}^n \varepsilon_t x_{ti} I_{\{|x_{ti}|\in A^{(K+1)}\}}| \ge \frac{n\delta_n}{K+1}, \Omega_n\Big) \le \exp(-a_1 n\delta_n^2). \tag{A.14}$$

Let $0 < \lambda < 1$ and $L_i = \{\lambda \le n^{-1}\sum_{t=1}^n x_{ti}^2 I_{\{|x_{ti}|\in A^{(K+1)}\}} < \lambda^{-1}\}$. By an argument similar to that used in the proof of (A.4), it can be shown that there exists $a_2 > 0$ such that for all large n, $\max_{1\le i\le p_n} P(L_i^c) \le \exp(-a_2 n)$. Application of Lemma A.1 after conditioning on $\mathbf{X}_i$, which is independent of $(\varepsilon_1, \ldots, \varepsilon_n)^\top$, shows that there exists $a_3 > 0$ for which

$$\max_{1\le i\le p_n} P\Big(\Big|\sum_{t=1}^n \varepsilon_t x_{ti} I_{\{|x_{ti}|\in A^{(K+1)}\}}\Big| \ge \frac{n\delta_n}{K+1}, L_i\Big) \le \exp(-a_3 n\delta_n^2),$$

for all large n. This completes the proof of (A.14).

Proof of (3.8)**.** Let $\Omega_n = \{\max_{1\le t\le n} |\varepsilon_t| < (\log n)^2\}$. It follows from (C2) that $\lim_{n\to\infty} P(\Omega_n^c) = 0$. Moreover, by (A.4), there exists $C > 0$ such that for any $a > 1/\sqrt{C}$ and all large n,

$$\begin{aligned} &P\Big(\max_{1\le i\le p_n} |n^{-1}\sum_{t=1}^n x_{ti}^2 - 1| > a(\log\frac{p_n}{n})^{1/2}\Big)\big) \\ &\quad \le p_n \exp(-Ca^2 \log p_n) = \exp\{\log p_n - Ca^2 \log p_n\} = o(1). \end{aligned} \tag{A.15}$$

Combining (A.15) with (3.7), (C4), and $\lim_{n\to\infty} P(\Omega_n^c) = 0$, it suffices for the proof of (3.8) to show that for some $\lambda > 0$,

$$P\Big(\max_{\sharp(J)\le K_n-1, i\notin J} |n^{-1}\sum_{t=1}^n \varepsilon_t \hat{x}_{ti;J}^{\perp}| > \lambda(\log\frac{p_n}{n})^{1/2}, \Omega_n\Big) = o(1), \tag{A.16}$$

$$P\Big(\max_{i,j\notin J, \sharp(J)\le K_n-1} |n^{-1}\sum_{t=1}^n x_{tj}\hat{x}_{ti;J}^{\perp} - E(x_j x_{i;J}^{\perp})| > \lambda(\log\frac{p_n}{n})^{1/2}\Big) = o(1). \tag{A.17}$$

To prove (A.16), let $\mathbf{x}_t(J)$ be a subvector of $\mathbf{x}_t$, with J denoting the corresponding subset of indices, and denote $\hat{\mathbf{\Gamma}}_n(J)$ by $\hat{\mathbf{\Gamma}}(J)$ for simplicity. Note

that

$$\max_{\sharp(J)\le K_n-1,i\notin J}|n^{-1}\sum_{t=1}^n\varepsilon_t\hat{x}^{\perp}_{ti;J}|\le\max_{1\le i\le p_n}|n^{-1}\sum_{t=1}^n\varepsilon_t x_{ti}|$$
$$+\max_{1\le\sharp(J)\le K_n-1,i\notin J}|(n^{-1}\sum_{t=1}^n x^{\perp}_{ti:J}\mathbf{x}_t(J))^{\top}\hat{\mathbf{\Gamma}}^{-1}(J)(n^{-1}\sum_{t=1}^n\varepsilon_t\mathbf{x}_t(J))|$$
$$+\max_{1\le\sharp(J)\le K_n-1,i\notin J}|\mathbf{g}_i^{\top}(J)\mathbf{\Gamma}^{-1}(J)(n^{-1}\sum_{t=1}^n\varepsilon_t\mathbf{x}_t(J))|:=S_{1,n}+S_{2,n}+S_{3,n},\quad(A.18)$$

where $x^{\perp}_{ti:J}=x_{ti}-\mathbf{g}_i^{\top}(J)\mathbf{\Gamma}^{-1}(J)\mathbf{x}_t(J)$. Since

$$S_{3,n}\le\max_{1\le i\le p_n}|n^{-1}\sum_{t=1}^n\varepsilon_t x_{ti}|\max_{1\le\sharp(J)\le K_n-1,i\notin J}\|\mathbf{\Gamma}^{-1}(J)\mathbf{g}_i(J)\|_1,$$

and since we can bound $\max_{1\le\sharp(J)\le K_n-1,i\notin J}\|\mathbf{\Gamma}^{-1}(J)\mathbf{g}_i(J)\|_1$ above by M for all large n in view of (3.2), it follows from Lemma A.4 that there exists $\beta>0$ such that for any $b>\beta^{-1/2}(M+1)$ and all large n,

$$P(S_{1,n}+S_{3,n}>b(\log\frac{p_n}{n})^{1/2},\Omega_n)\le p_n\exp\Big(-\beta b^2\log\frac{p_n}{(M+1)^2}\Big)=o(1).\quad(A.19)$$

Since $K_n=O((n/\log n)^{1/2})$, there exists $D>0$ such that for all large n, $K_n\le D(n/\log p_n)^{1/2}$. As shown in the proof of Lemma A.2, $\max_{1\le\sharp(J)\le K_n}\|\mathbf{\Gamma}^{-1}(J)\|\le\delta^{-1}$ for all large n. In view of this and (A.6), there exists $c>0$ such that for any $d>(2D^2/c)^{1/2}$ and all large n,

$$P(\max_{1\le\sharp(J)\le K_n}\|\hat{\mathbf{\Gamma}}^{-1}(J)\|>\delta^{-1}+d)\le P(\max_{1\le\sharp(J)\le K_n}\|\hat{\mathbf{\Gamma}}^{-1}(J)-\mathbf{\Gamma}^{-1}(J)\|>d)$$
$$\le p_n^2\exp\Big(\frac{-cnd^2}{K_n^2}\Big)=o(1).\quad(A.20)$$

Since $\max_{1\le\sharp(J)\le K_n-1,i\notin J}\|n^{-1}\sum_{t=1}^n\varepsilon_t\mathbf{x}_t(J)\|\le K_n^{1/2}\max_{1\le i\le p_n}|n^{-1}\sum_{t=1}^n\varepsilon_t x_{ti}|$ and

$$\max_{1\le\sharp(J)\le K_n-1,i\notin J}\|n^{-1}\sum_{t=1}^n x^{\perp}_{ti;J}\mathbf{x}_t(J)\|$$
$$\le K_n^{1/2}\max_{1\le i,j\le p_n}|n^{-1}\sum_{t=1}^n x_{ti}x_{tj}-\sigma_{ij}|(1+\max_{1\le\sharp(J)\le K_n,i\notin J}\|\mathbf{\Gamma}^{-1}(J)\mathbf{g}_i(J)\|_1),\quad(A.21)$$

it follows from (3.2) that for all large n,

$$S_{2,n}\le(\max_{1\le\sharp(J)\le K_n}\|\hat{\mathbf{\Gamma}}^{-1}(J)\|)K_n(1+M)$$
$$\times(\max_{1\le i,j\le p_n}|n^{-1}\sum_{t=1}^n x_{ti}x_{tj}-\sigma_{ij}|)(\max_{1\le i\le p_n}|n^{-1}\sum_{t=1}^n\varepsilon_t x_{ti}|).\quad(A.22)$$

Take $d > (2D^2/c)^{1/2}$ and let $c_1 = \delta^{-1} + d$. By (A.4), (A.20), (A.22), and Lemma A.4, there exists sufficiently large c_2 such that

$$\begin{aligned} P(S_{2,n} > c_2(\log \frac{p_n}{n})^{1/2}, \Omega_n) &\le P(\max_{1\le \sharp(J)\le K_n} \|\hat{\mathbf{\Gamma}}^{-1}(J)\| > c_1) \\ &+P\left(\max_{1\le i\le p_n} |n^{-1}\sum_{t=1}^n \varepsilon_t x_{ti}| > \frac{c_2^{1/2}(\log \frac{p_n}{n})^{1/4}}{(K_n c_1(1+M))^{1/2}}, \Omega_n\right) \\ &+P\left(\max_{1\le i,j\le p_n} |n^{-1}\sum_{t=1}^n x_{ti}x_{tj} - \sigma_{ij}| > \frac{c_2^{1/2}(\log \frac{p_n}{n})^{1/4}}{(K_n c_1(1+M))^{1/2}}\right) = o(1). \quad \text{(A.23)} \end{aligned}$$

From (A.18), (A.19), and (A.23), (A.16) follows if λ is sufficiently large.

To prove (A.17), we use the bound

$$\begin{aligned} &\max_{\sharp(J)\le K_n-1, i,j\notin J} |n^{-1}\sum_{t=1}^n x_{tj}\hat{x}_{ti;J}^{\perp} - E(x_j x_{i;J}^{\perp})| \\ &\le \max_{1\le i,j\le p_n} |n^{-1}\sum_{t=1}^n x_{tj}x_{ti} - \sigma_{ij}| \\ &\quad + \max_{1\le \sharp(J)\le K_n-1, i,j\notin J} |\mathbf{g}_j^{\top}(J)\mathbf{\Gamma}^{-1}(J) n^{-1}\sum_{t=1}^n x_{ti;J}^{\perp}\mathbf{x}_t(J)| \\ &\quad + \max_{1\le \sharp(J)\le K_n-1, i,j\notin J} |\mathbf{g}_i^{\top}(J)\mathbf{\Gamma}^{-1}(J)(n^{-1}\sum_{t=1}^n x_{tj}\mathbf{x}_t(J) - \mathbf{g}_j(J))| \\ &\quad + \max_{1\le \sharp(J)\le K_n-1, i,j\notin J} \|\hat{\mathbf{\Gamma}}^{-1}(J)\| \|n^{-1}\sum_{t=1}^n x_{ti;J}^{\perp}\mathbf{x}_t(J)\| \|n^{-1}\sum_{t=1}^n x_{tj;J}^{\perp}\mathbf{x}_t(J)\| \\ &:= S_{4,n} + S_{5,n} + S_{6,n} + S_{7,n}. \quad \text{(A.24)} \end{aligned}$$

Analogous to (A.21) and (A.22), it follows from (3.2) that for all large n,

$$\begin{aligned} S_{5,n} &\le \max_{1\le i,j\le p_n} |n^{-1}\sum_{t=1}^n x_{tj}x_{ti} - \sigma_{ij}|(M + M^2),\ S_{6,n} \\ &\le \max_{1\le i,j\le p_n} |n^{-1}\sum_{t=1}^n x_{tj}x_{ti} - \sigma_{ij}|M. \end{aligned}$$

Combining this with (A.4) yields that for some $c_3 > 0$,

$$P\{S_{4,n} + S_{5,n} + S_{6,n} > c_3\Big(\log \frac{p_n}{n}\Big)^{1/2}\} = o(1). \quad \text{(A.25)}$$

In view of (A.21) and (3.2), it follows that for all large n,

$$S_{7,n} \le (\max_{1\le \sharp(J)\le K_n} \|\hat{\mathbf{\Gamma}}^{-1}(J)\|) K_n(1+M)^2 \max_{1\le i,j\le p_n} (n^{-1}\sum_{t=1}^n x_{tj}x_{ti} - \sigma_{ij})^2.$$

Therefore by (A.4) and (A.20), there exists $c_4 > 0$ such that for all large n,

$$P\{S_{7,n} > c_4(\log \frac{p_n}{n})^{1/2}\} \le P(\max_{1\le \sharp(J)\le K_n} \|\hat{\mathbf{\Gamma}}^{-1}(J)\| > c_1) + P\left(\max_{1\le i,j\le p_n} (n^{-1}\sum_{t=1}^n x_{tj}x_{ti} - \sigma_{ij})^2 > \frac{c_4(\log \frac{p_n}{n})^{1/2}}{c_1 K_n(1+M)^2}\right) = o(1). \quad (A.26)$$

From (A.24)−(A.26), (A.17) follows if λ is sufficiently large.

Proof of (3.13)**.** Let $\mathbf{q}(J) = E(y\mathbf{x}_J)$ and $\mathbf{Q}(J) = n^{-1}\sum_{t=1}^n (y_t - \mathbf{x}_t^\top(J)\mathbf{\Gamma}^{-1}(J)\mathbf{q}(J))\mathbf{x}_t(J)$. Then

$$\|\mathbf{Q}(\hat{J}_m)\|^2 \le 2m \max_{1\le i\le p_n}(n^{-1}\sum_{t=1}^n \varepsilon_t x_{ti})^2 + 2m \max_{1\le i,j\le p_n}(n^{-1}\sum_{t=1}^n x_{ti}x_{tj} - \sigma_{ij})^2 \times (\sum_{j=1}^{p_n}|\beta_j|)^2(1 + \max_{1\le \sharp(J)\le K_n, 1\le l\le p_n}\|\mathbf{\Gamma}^{-1}(J)\mathbf{g}_l(J)\|_1)^2.$$

Combining this bound with (3.2), (C4), (A.4), and (A.11) yields

$$\max_{1\le m\le K_n}\left\{\frac{n\|\mathbf{Q}(\hat{J}_m)\|^2}{m\log p_n}\right\} = O_p(1). \quad (A.27)$$

Moreover, by (A.5) and (A.6), $\max_{1\le m\le K_n}\|\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_m)\| = O_p(1)$ and $\max_{1\le m\le K_n}\|\hat{\mathbf{\Gamma}}(\hat{J}_m) - \mathbf{\Gamma}(\hat{J}_m)\| = O_p(1)$. The desired conclusion (3.13) follows from this, (A.27), and

$$\begin{aligned} &E_n(\hat{y}_m(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x}))^2 \\ &\quad = \mathbf{Q}^\top(\hat{J}_m)\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_m)\mathbf{\Gamma}(\hat{J}_m)\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_m)\mathbf{Q}(\hat{J}_m) \\ &\quad \le \|\mathbf{Q}(\hat{J}_m)\|^2\|\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_m)\|^2\|\hat{\mathbf{\Gamma}}(\hat{J}_m) - \mathbf{\Gamma}(\hat{J}_m)\| + \|\mathbf{Q}(\hat{J}_m)\|^2\|\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_m)\|. \end{aligned}$$

Proof of (4.15)**.** Denote $\lfloor an^\gamma\rfloor$ in (4.10) by m_0. By (3.2) and an argument similar to that to derive (A.20), there exists $d_1 > 0$ such that for all large n,

$$P(\max_{1\le\sharp(J)\le m_0}\|\hat{\mathbf{\Gamma}}^{-1}(J)\| > 2\delta^{-1}) \le p_n^2\exp(-d_1 n^{1-2\gamma}) = o(1). \quad (A.28)$$

Defining Ω_n as in Lemma A.3, it follows from (A.16) and (C5) that

$$P(|\hat{B}_n| \ge \theta n^{-\gamma/2}, \mathcal{D}_n, \Omega_n) \le P(\max_{\sharp(J)\le m_0-1, i\notin J}|n^{-1}\sum_{t=1}^n \varepsilon_t \hat{x}_{ti;J}^\perp| \ge \theta n^{-\gamma/2}, \Omega_n) = o(1). \quad (A.29)$$

Since (4.8) implies that $n^{1-2\gamma}/(w_n \log p_n) \to \infty$, it follows from Lemma A.4, (3.2), (4.8), and (A.28) that

$$P(|\hat{C}_n| \geq \frac{\theta n^{1-2\gamma}}{w_n \log p_n}, \mathcal{D}_n, \Omega_n) \leq P(|n^{-1}\sum_{t=1}^{n} \varepsilon_t^2 - \sigma^2| \geq \frac{\theta}{2})$$
$$+P(\max_{1\leq \sharp(J)\leq m_0} \|\hat{\mathbf{\Gamma}}^{-1}(J)\| m_0 \max_{1\leq j\leq p_n} (n^{-1}\sum_{t=1}^{n} \varepsilon_t x_{tj})^2 \geq \frac{\theta}{2}, \Omega_n) = o(1). \quad \text{(A.30)}$$

As noted in the proof of (3.8), $P(\Omega_n) = 1 + o(1)$. Moreover, by (4.8) and (A.5) with K_n replaced by m_0, there exists $d_2 > 0$ such that for all large n,

$$P(\hat{A}_n < \frac{v_n}{2}, \mathcal{D}_n) \leq P(\lambda_{\min}(\hat{\mathbf{\Gamma}}(\hat{J}_{\tilde{k}_n})) < \frac{v_n}{2}, \mathcal{D}_n) \leq P(\lambda_{\min}(\hat{\mathbf{\Gamma}}(\hat{J}_{m_0})) < \frac{v_n}{2})$$
$$\leq P(\max_{1\leq \sharp(J)\leq m_0} \|\hat{\mathbf{\Gamma}}(J) - \mathbf{\Gamma}(J)\| > \frac{\delta}{2}) \leq p_n^2 \exp(-d_2 n^{1-2\gamma}) = o(1), \quad \text{(A.31)}$$

recalling that $v_n > \delta$ for all large n by (3.2). From (A.29)−(A.31), (4.15) follows.

Proof of (4.20)**.** Since $\hat{k} \leq K_n \leq D(n/\log p_n)^{1/2}$ for some $D > 0$, $\log p_n = o(n)$ and $w_n \to \infty$, there exist $\eta > 0$ and $\zeta_n \to \infty$ such that on $\{\hat{k} > \tilde{k}\}$,

$$\frac{n\{1 - \exp(n^{-1}w_n(\hat{k} - \tilde{k})\log p_n)\}}{\hat{k} - \tilde{k}} \geq \eta \min\{(n\log p_n)^{1/2}, w_n \log p_n\}$$
$$\geq \zeta_n \log p_n. \quad \text{(A.32)}$$

From (4.18), (A.18), and (A.32), we obtain the bound

$$P\{(\hat{a}_n + \hat{b}_n)(\hat{k} - \tilde{k}) \geq \theta n[1 - \exp(-n^{-1}w_n(\hat{k} - \tilde{k})\log p_n)], \hat{k} > \tilde{k}\}$$
$$\leq P(\Omega_n^c) + P(\|\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_{K_n})\| \geq \delta^{-1} + d)$$
$$+P(\max_{1\leq j\leq p_n} (n^{-1/2}\sum_{t=1}^{n} x_{tj}\varepsilon_t)^2 \geq \frac{\theta\zeta_n(\log p_n)}{2(\delta^{-1}+d)}, \Omega_n)$$
$$+P(n(S_{2,n} + S_{3,n})^2 \geq \frac{\theta\zeta_n(\log p_n)}{2(\delta^{-1}+d)}, \Omega_n), \quad \text{(A.33)}$$

where $S_{2,n}$ and $S_{3,n}$ are defined in (A.18) and d is the same as that in (A.20). By Lemma A.4, there exists $\beta > 0$ such that for all large n,

$$P(\max_{1\leq j\leq p_n} (n^{-1/2}\sum_{t=1}^{n} x_{tj}\varepsilon_t)^2 \geq \frac{\theta\zeta_n(\log p_n)}{2(\delta^{-1}+d)}, \Omega_n)$$
$$\leq p_n \max_{1\leq j\leq p_n} P\Big(|n^{-1/2}\sum_{t=1}^{n} x_{tj}\varepsilon_t| \geq \Big\{\frac{\theta\zeta_n(\log p_n)}{2(\delta^{-1}+d)}\Big\}^{1/2}, \Omega_n\Big)$$
$$\leq p_n \exp(-\beta\zeta_n \log p_n) = o(1).$$

An argument similar to that in (A.19) and (A.23) yields

$$P(n(S_{2,n}+S_{3,n})^2 \geq \frac{\theta\zeta_n(\log p_n)}{2(\delta^{-1}+d)}, \Omega_n) = o(1).$$

Moreover, as already shown in the proof of (3.8), $P(\Omega^c) = o(1)$ and $P(\|\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_{K_n})\| \geq \delta^{-1}+d) = o(1)$; see (A.20). Adding these bounds for the summands in (A.33) yields the first conclusion in (4.20). The second conclusion $P(|\hat{C}_n| \geq \theta) = o(1)$ can be derived similarly from (A.30) and (4.10).

Appendix B: Proof of Theorem 2

Note that when the regressors are nonrandom, the population version of OGA is the "noiseless" OGA that replaces y_t in OGA by its mean $y(\mathbf{x}_t)$. Let $\boldsymbol{\mu} = (y(\mathbf{x}_1), \ldots, y(\mathbf{x}_n))^\top$. Let $\mathbf{H}_J$ denote the projection matrix associated with the projection into the linear space spanned by $\mathbf{X}_j, j \in J \subseteq \{1, \ldots, p\}$. Let $\mathbf{U}^{(0)} = \boldsymbol{\mu}$, $\tilde{j}_1 = \arg\max_{1\leq j\leq p} |(\mathbf{U}^{(0)})^\top \mathbf{X}_j| / \|\mathbf{X}_j\|$ and $\mathbf{U}^{(1)} = (\mathbf{I} - \mathbf{H}_{\{\tilde{j}_1\}})\boldsymbol{\mu}$. Proceeding inductively yields

$$\tilde{j}_m = \arg\max_{1\leq j\leq p} \frac{|(\mathbf{U}^{(m-1)})^\top \mathbf{X}_j|}{\|\mathbf{X}_j\|}, \ \mathbf{U}^{(m)} = (\mathbf{I} - \mathbf{H}_{\{\tilde{j}_1,\ldots,\tilde{j}_m\}})\boldsymbol{\mu}.$$

When the procedure stops after m iterations, the noiseless OGA determines an index set $\tilde{J}_m = \{\tilde{j}_1, \ldots, \tilde{j}_m\}$ and approximates $\boldsymbol{\mu}$ by $\mathbf{H}_{\tilde{J}_m}\boldsymbol{\mu}$. A generalization of noiseless OGA takes $0 < \xi \leq 1$ and replaces $\tilde{j}_i$ by $\tilde{j}_{i,\xi}$, where $\tilde{j}_{i,\xi}$ is any $1 \leq l \leq p$ satisfying

$$\frac{|(\mathbf{U}^{(i-1)})^\top \mathbf{X}_l|}{\|\mathbf{X}_l\|} \geq \xi \max_{1\leq j\leq p} \frac{|(\mathbf{U}^{(i-1)})^\top \mathbf{X}_j|}{\|\mathbf{X}_j\|}. \tag{B.1}$$

We first prove an inequality for the generalization (B.1) of noiseless OGA.

Lemma B.1. *Let $0 < \xi \leq 1$, $m \geq 1$, $\tilde{J}_{m,\xi} = \{\tilde{j}_{1,\xi}, \ldots, \tilde{j}_{m,\xi}\}$ and $\hat{\sigma}_j^2 = n^{-1}\sum_{t=1}^n x_{tj}^2$. Then*

$$\|(\mathbf{I} - \mathbf{H}_{\tilde{J}_{m,\xi}})\boldsymbol{\mu}\|^2 \leq n(\inf_{\mathbf{b}\in\mathbf{B}} \sum_{j=1}^p |b_j\hat{\sigma}_j|)^2(1+m\xi^2)^{-1}.$$

Proof. For $J \subseteq \{1, \ldots, p\}$, $i \in \{1, \ldots, p\}$ and $m \geq 1$, define $\nu_{J,i} = (\mathbf{X}_i)^\top(\mathbf{I} - \mathbf{H}_J)\boldsymbol{\mu}/(n^{1/2}\|\mathbf{X}_i\|)$. Note that

$$\begin{aligned}
&\|(\mathbf{I} - \mathbf{H}_{\tilde{J}_{m,\xi}})\boldsymbol{\mu}\|^2 \\
&\leq \|(\mathbf{I} - \mathbf{H}_{\tilde{J}_{m-1,\xi}})\boldsymbol{\mu} - \frac{\boldsymbol{\mu}^\top(\mathbf{I} - \mathbf{H}_{\tilde{J}_{m-1,\xi}})\mathbf{X}_{\tilde{j}_{m,\xi}}}{\|\mathbf{X}_{\tilde{j}_{m,\xi}}\|^2}\mathbf{X}_{\tilde{j}_{m,\xi}}\|^2 \\
&\leq \|(\mathbf{I} - \mathbf{H}_{\tilde{J}_{m-1,\xi}})\boldsymbol{\mu}\|^2 - n\nu^2_{\tilde{J}_{m-1,\xi},\tilde{j}_{m,\xi}} \leq \|(\mathbf{I} - \mathbf{H}_{\tilde{J}_{m-1,\xi}})\boldsymbol{\mu}\|^2 - n\xi^2 \max_{1\leq j\leq p} \nu^2_{\tilde{J}_{m-1,\xi},j},
\end{aligned} \tag{B.2}$$

in which $\mathbf{H}_{\tilde{J}_{0,\xi}} = \mathbf{0}$. Moreover, for any $\mathbf{b} = (b_1, \ldots, b_p)^\top \in \mathbf{B}$,

$$\|(\mathbf{I} - \mathbf{H}_{\tilde{J}_{m-1,\xi}})\boldsymbol{\mu}\|^2 = n^{1/2} \sum_{j=1}^{p} b_j \|\mathbf{X}_j\| \nu_{\tilde{J}_{m-1,\xi},j} \le (\max_{1\le j\le p} |\nu_{\tilde{J}_{m-1,\xi},j}|) n \sum_{j=1}^{p} |b_j \hat{\sigma}_j|. \tag{B.3}$$

Let $S = n(\sum_{j=1}^{p} |b_j \hat{\sigma}_j|)^2$. It follows from (B.2) and (B.3) that

$$\|(\mathbf{I} - \mathbf{H}_{\tilde{J}_{m,\xi}})\boldsymbol{\mu}\|^2 \le \|(\mathbf{I} - \mathbf{H}_{\tilde{J}_{m-1,\xi}})\boldsymbol{\mu}\|^2 \{1 - \frac{\xi^2}{S} \|(\mathbf{I} - \mathbf{H}_{\tilde{J}_{m-1,\xi}})\boldsymbol{\mu}\|^2\}. \tag{B.4}$$

By Minkowski's inequality, $\|(\mathbf{I} - \mathbf{H}_{\tilde{J}_{0,\xi}})\boldsymbol{\mu}\|^2 = \|\boldsymbol{\mu}\|^2 \le S$. Combining this with (B.4) and Temlyakov's (2000) Lemma 3.1 yields the desired conclusion.

Proof of Theorem 2. For the given $0 < \xi < 1$, let $\tilde{\xi} = 2/(1-\xi)$,

$$A = \{\max_{(J,i):\sharp(J)\le m-1, i\notin J} |\hat{\mu}_{J,i} - \nu_{J,i}| \le C\sigma(n^{-1}\log p)^{1/2}\},$$
$$B = \{\min_{0\le i\le m-1} \max_{1\le j\le p} |\nu_{\hat{J}_i,j}| > \tilde{\xi} C\sigma(n^{-1}\log p)^{1/2}\},$$

recalling that $\hat{\mu}_{J,i}$ is defined in (3.4) and that $\nu_{J,i}$ is introduced in the proof of Lemma B.1. Note that $\nu_{J,i}, A$ and B play the same roles as those of $\mu_{J,i}, A_n(m)$, and $B_n(m)$ in the proof of Theorem 1. By an argument similar to that used to prove (3.10), we have for all $1 \le q \le m$,

$$|\nu_{\hat{J}_{q-1},\hat{j}_q}| \ge \xi \max_{1\le j\le p} |\nu_{\hat{J}_{q-1},j}| \text{ on } A \bigcap B, \tag{B.5}$$

which implies that on the set $A \bigcap B$, $\hat{J}_m$ is the index set chosen by the generalization (B.1) of the noiseless OGA. Therefore, it follows from Lemma B.1 that

$$\|(\mathbf{I} - \mathbf{H}_{\hat{J}_m})\boldsymbol{\mu}\|^2 I_{A\bigcap B} \le n(\inf_{\mathbf{b}\in\mathbf{B}} \|\mathbf{b}\|_1)^2 (1 + m\xi^2)^{-1}. \tag{B.6}$$

Moreover, for $0 \le i \le m-1$, $\|(\mathbf{I} - \mathbf{H}_{\hat{J}_m})\boldsymbol{\mu}\|^2 \le \|(\mathbf{I} - \mathbf{H}_{\hat{J}_i})\boldsymbol{\mu}\|^2$, and therefore

$$\begin{aligned}
\|(\mathbf{I} - \mathbf{H}_{\hat{J}_m})\boldsymbol{\mu}\|^2 &\le \min_{0\le i\le m-1} \sum_{j=1}^{p} b_j \mathbf{X}_j^\top (\mathbf{I} - \mathbf{H}_{\hat{J}_i})\boldsymbol{\mu} \\
&\le (\min_{0\le i\le m-1} \max_{1\le j\le p} |\nu_{\hat{J}_i,j}|) n \|\mathbf{b}\|_1 \\
&\le \tilde{\xi} C\sigma(n\log p)^{1/2} \|\mathbf{b}\|_1 \text{ on } B^c.
\end{aligned} \tag{B.7}$$

Since A decreases as m increases, it follows from (3.18), (B.6), and (B.7) that

$$n^{-1}\|(\mathbf{I} - \mathbf{H}_{\hat{J}_m})\boldsymbol{\mu}\|^2 I_{\mathcal{A}} \le \omega_{m,n} \text{ for all } 1 \le m \le \lfloor \frac{n}{\log p} \rfloor, \tag{B.8}$$

where $\mathcal{A}$ denotes the set A with $m = \lfloor n/\log p \rfloor$. Moreover, as is shown below,

$$P(\mathcal{A}^c) \le a^* := p\,\exp\Big\{\frac{-2^{-1}C^2(\log p)}{(1+M)^2}\Big\}, \tag{B.9}$$

$$P(\mathcal{E}^c) \le b^* := \frac{\tilde{r}_p^{1/2} p^{-(sr_p-1)}}{1-\tilde{r}_p^{1/2} p^{-(sr_p-1)}}, \tag{B.10}$$

where r_p and $\tilde{r}_p$ are defined in (3.16) and

$$\mathcal{E} = \{\boldsymbol{\varepsilon}^\top \mathbf{H}_{\hat{J}_m}\boldsymbol{\varepsilon} \le s\sigma^2 m \log p \text{ for all } 1 \le m \le \lfloor \frac{n}{\log p} \rfloor\}.$$

By (B.8)–(B.10) and that $\|\hat{y}_m(\cdot) - y(\cdot)\|_n^2 = n^{-1}(\|(\mathbf{I} - \mathbf{H}_{\hat{J}_m})\boldsymbol{\mu}\|^2 + \boldsymbol{\varepsilon}^\top \mathbf{H}_{\hat{J}_m}\boldsymbol{\varepsilon})$, (3.19) holds on the set $\mathcal{A}\bigcap\mathcal{E}$, whose probability is at least $1 - a^* - b^*$, proving the desired conclusion.

Proof of (B.9)**.** Since $\hat{\mu}_{J,i} = (\mathbf{X}_i)^\top(\mathbf{I} - \mathbf{H}_J)\mathbf{Y}/(n^{1/2}\|\mathbf{X}_i\|)$ and $n^{-1}\sum_{t=1}^n x_{tj}^2 = 1$ for all $1 \le j \le p$, we have for any $J \subseteq \{1,\dots,p\}$, $1 \le i \le p$ and $i \notin J$,

$$|\hat{\mu}_{J,i} - \nu_{J,i}| \le \max_{1\le i\le p} |n^{-1}\sum_{t=1}^n x_{ti}\varepsilon_t|(1 + \inf_{\boldsymbol{\theta}_{J,i}\in\mathbf{B}_{J,i}} \|\boldsymbol{\theta}_{i,J}\|_1),$$

setting $\|\boldsymbol{\theta}_{J,i}\|_1 = 0$ if $J = \emptyset$. This and (3.17) yield

$$\max_{\sharp(J)\le\lfloor n/\log p\rfloor - 1, i\notin J} |\hat{\mu}_{J,i} - \nu_{J,i}| \le \max_{1\le i\le p} |n^{-1}\sum_{t=1}^n x_{ti}\varepsilon_t|(1+M). \tag{B.11}$$

By (B.11) and the Gaussian assumption on ε_t,

$$\begin{aligned} P(\mathcal{A}^c) &\le P\{\max_{1\le i\le p} |n^{-1/2}\sum_{t=1}^n x_{ti}\varepsilon_t/\sigma| > C(\log p)^{1/2}(1+M)^{-1}\} \\ &\le p\,\exp(-\{C^2\log p/[2(1+M)^2]\}). \end{aligned}$$

Proof of (B.10)**.** Clearly $P(\mathcal{E}^c) \le \sum_{m=1}^{\lfloor n/\log p\rfloor} p^m \max_{\sharp(J)=m} P(\boldsymbol{\varepsilon}^\top \mathbf{H}_J\boldsymbol{\varepsilon} > s\sigma^2 m \log p)$. Moreover, we can make use of the χ^2-distribution to obtain the bound

$$\max_{\sharp(J)=m} P(\boldsymbol{\varepsilon}^\top \mathbf{H}_J\boldsymbol{\varepsilon} > s\sigma^2 m\log p) \le (1-2r)^{-m/2}\exp(-rsm\log p) \tag{B.12}$$

for any $0 < r < 1/2$. With $r = r_p$ and $s > \{1 + (2\log p)^{-1}\log\tilde{r}_p\}\ /r_p$ in (B.12), we can use (B.12) to bound $P(\mathcal{E}^c)$ by $\sum_{m=1}^{\lfloor n/\log p\rfloor} g^m \le g/(1-g)$, where $g = \tilde{r}_p^{1/2} p^{-(sr_p-1)} < 1$.

References

Barron, A. R., Cohen, A., Dahmen, W., and DeVore, R. (2008). Approximation and learning by greedy algorithms. *Ann. Statist.* **36**, 64-94.

Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37**, 1705-1732.

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist.* **34**, 559-583.

Bühlmann, P. and Yu, B. (2003). Boosting with the L_2 loss: regression and classification. *J. Amer. Statist. Assoc.* **98**, 324-339.

Bunea, F., Tsybakov, A. B. and Wegkamp, M. H. (2007). Sparsity oracle inequalities for the Lasso. *Electr. J. Statist.* **1**, 169-194.

Candes, E. J. and Plan, Y. (2009). Near-ideal model selection by l_1 minimization. *Ann. Statist.* **37**, 2145-2177.

Candes, E. J. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313-2351.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759-771.

Chen, R.-B., Ing, C.-K. and Lai, T. L. (2011). An efficient pathwise variable selection criterion in weakly sparse regression models. Tech. Report, Dept. Statistics, Stanford Univ.

Donoho, D. L., Elad, M., and Temlyakov, V. N. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory* **52**, 6-18.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.

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

Fan, J., Feng, Y., Samworth, R., and Wu, Y. (2010). SIS: Sure Independence Screening. `R` package version 0.4. `http://cran.r-project.org/web/packages/SIS/index.html`.

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

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

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20**, 101-148.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. Preprint. `http://arxiv.org/abs/0903.5255v4`.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947-1975.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). glmnet: Lasso and elastic-net regularized generalized linear models. `R` package version 1.1-5. `http://cran.r-project.org/web/pack-ages/glmnet/index.html`.

Hannan, E. J. and Quinn, B. C. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B* **41**, 190-195.

Hastie, T. and Efron, B. (2007). lars: Least Angle Regression, Lasso and Forward Stagewise. `R` package version 0.9-7. `http://cran.r-project.org/web/packages/lars/index.html`.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statist. Sinica* **18**, 1603-1618.

Kraemer, N. and Schaefer, J. (2010). parcor: Regularized estimation of partial correlation matrices. `R` package version 0.2-2. `http://cran.r-project.org/web/packages/parcor/index.html`.

Leng, C., Lin, Y. and Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statist. Sinica* **16**, 1273-1284.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34**, 1436-1462.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Shao, J. and Chow, S.-C. (2007). Variable screening in predicting clinical outcome with high-dimensional microarrays. *J. Multivariate Anal.* **98**, 1529-1538.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. B* **58**, 267-288.

Tropp, J. A. (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory* **50**, 2231-2242.

Tropp, J. A. and Gilbert, A. C. (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inform. Theory* **53**, 4655-4666.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55**, 2183-2202.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening *J. Amer. Statist. Assoc.* **104**, 1512-1524.

Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37**, 2178-2201.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567-1594.

Zhang, Y, Li, R., and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *J. Amer. Statist. Assoc.* **105**, 312-323.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Machine Learning Res.* **7**, 2541-2563.

Zhou, S., van de Geer, S., and Bühlmann, P. (2009). Adaptive Lasso for high dimensional regression and Gaussian graphical modeling. To appear in *Ann. Statist.*

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

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.* **36**, 1509-1566.

Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, R.O.C.

E-mail: cking@stat.sinica.edu.tw

Department of Statistics, Sequoia Hall, 390 Serra Mall, Stanford University, Stanford, CA 94305-4065, U.S.A.

E-mail: lait@stat.stanford.edu

(Received April 2010; accepted August 2010)