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

By Ching-Kang Ing [*] and Tze Leung Lai [†]

Academia Sinica and Stanford University

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 $p >> n$, the sample size) sequentially so that the selected variable at each step minimizes the residual sum squares. We derive the convergence rate of OGA as $m = m_n$ becomes infinite, and also develop a consistent model selection procedure along the OGA path so that the resultant regression estimate has the oracle property of being equivalent to least squares regression on an asymptotically minimal set of relevant regressors under a strong sparsity condition.

1. Introduction. Consider the linear regression model

$$(1.1) \qquad y_t = \alpha + \sum_{j=1}^{p} \beta_j x_{tj} + \varepsilon_t, \ t = 1, 2, \cdots, n,$$

with p predictor variables $x_{t1}, x_{t2}, \cdots, x_{tp}$ 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 [2], [3], [14], LARS [11], and Lasso [22] 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

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

is still possible under a sparseness condition on the regression coefficients. In particular, by assuming the "weak sparsity" condition that the regression coefficients are absolutely summable, Bühlmann [2] showed that for $p = \exp(O(n^\xi))$ with $0 < \xi < 1$, the conditional mean squared prediction error

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

1

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. Fan and Lv [12] recently introduced another method, called *sure independence screening* (SIS), to screen out variables from high-dimensional regression models, and showed that the probability of including all relevant regressors by using SIS approaches 1 as the sample size becomes infinite. The most comprehensive theory to date for these high-dimensional regression methods has been developed for Lasso and a closely related method, the Dantzig selector, introduced by Candes and Tao [5]. In particular, oracle inequalities for the Dantzig selector and Lasso have been established in [1], [4], [5], [6], [26] under certain sparsity conditions.

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, (b) a stopping criterion 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 ($p >> n$) under certain sparsity assumptions. We also demonstrate its advantages in simulation studies of its finite-sample performance and prove an inequality, which is similar to the oracle inequality for Lasso in [1], for the stepwise regression procedure.

The forward stepwise component of this procedure is called the *orthogonal greedy algorithm* (OGA) or *orthogonal matching pursuit* in information theory, signal processing and approximation theory, which focuses on approximations in noiseless models (i.e., $\varepsilon_t = 0$ in model (1.1)); see [21], [23], [24]. In Section 2 we review this literature and describe OGA as a greedy forward stepwise variable selection method to enter the input variables in regression models. In this connection we also consider the L_2-boosting procedure of Bühlmann and Yu [3], which corresponds to the *pure greedy algorithm* (PGA) or *matching pursuit* in approximation theory [17], [21]. An apparent computational advantage of PGA over OGA is that PGA does not require matrix inversion involved in the least squares square estimates used by OGA. In Section 2, however, we develop a fast iterative procedure for updating OGA that uses componentwise linear regression similar to PGA and does not require matrix inversion. Section 3 gives an oracle-type inequality and the rate of convergence of the squared prediction error (1.3) in which $\hat{y}_m(\cdot)$ is the OGA predictor, under a *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 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 estab-

lished in Theorem 3.1, 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. This result shows that OGA is a reliable variable screening method, and hence we only need to focus on the variables chosen along the OGA path. The sharp convergence rate in Theorem 3.1 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 [16]. We call such modification a *high-dimensional information criterion* (HDIC), which we use to choose the smallest correct model along the OGA path. This combined estimation and variable selection scheme, which we denote by OGA+HDIC, is shown in Theorem 4.2 to select the smallest set of all relevant variables with probability approaching 1 (and is therefore variable selection consistent). Since OGA is essentially an implementation of ordinary least squares after stepwise variable selection, Theorem 4.2 implies that OGA+HDIC is asymptotically equivalent to an optimal backward elimination procedure following forward stepwise addition to come up with a minimal set of regressors under the strong sparsity condition. In the fixed design case, Chen and Chen [7] proposed an "extended Bayesian information criterion" whose penalty is considerably larger than that of BIC, and showed that the criterion is consistent when $p = O(n^{\kappa})$ for some $\kappa > 0$ and the number of nonzero regression coefficients does not exceed a prescribed constant K. By applying a larger penalty, the HDIC introduced in Section 4 yields consistent variable selection in more general sparse high-dimensional regression models with $\log p = o(n)$.

Zhao and Yu [27] have shown that Lasso is variable selection consistent for non-random high-dimensional regressors under an "irrepresentable condition" (IC) on the sample covariance matrix and regression coefficients. For random regressors, Meinshausen and Bühlmann [18] have proved this consistency result for Lasso under an extension of IC, called the "neighborhood stability condition" (NSC); moreover, IC or NSC is almost necessary for Lasso to be consistent (see [27]) and therefore cannot be weakened. Although SIS has been shown by Fan and Lv [12] to include all relevant regressors (with probability approaching 1) when NSC does not hold, it requires a lower bound on the absolute values of the covariances between the outcome and relevant regressors, besides an assumption on the maximum eigenvalue of the covariance matrix of the candidate regressors that can fail to hold in situations where all regressors are equally correlated; see Section 5 for details. Note that this equi-correlation structure has been adopted as a benchmark example in [7], [20], [27] and other papers. The simulation studies in Section 5 on the finite-sample performance of OGA+HDIC demonstrate its advantages in this and other settings.

2. L_2-boosting, forward stepwise regression and Tymlyakov's greedy algorithms. We begin this section by reviewing Bühlmann and Yu's [3] 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 [21] 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 $\mathbf{x}_t = (x_{t1}, \cdots, x_{tp})^\top$.

2.1. *PGA iterations.* Bühlmann and Yu's [3] 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 $U_t^{(k)} := y_t - \hat{y}_k(\mathbf{x}_t), 1 \leq t \leq n$, 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

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

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

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

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 $\hat{y}_{k+1}(\mathbf{x}_t) = \hat{y}_k(\mathbf{x}_t) + \delta \tilde{\beta}_{\hat{j}_{k+1}}^{(k)} x_{t,\hat{j}_{k+1}}$, with $0 < \delta \leq 1$, during the iterations; see [2, 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 $\mathbf{X}_{\hat{j}_{k+1}}^\perp$, where $\mathbf{X}_j = (x_{1j}, \cdots, x_{nj})^\top$. 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*. With the orthogonal vectors $\mathbf{X}_{\hat{j}_1}, \mathbf{X}_{\hat{j}_2}^\perp, \cdots, \mathbf{X}_{\hat{j}_k}^\perp$ 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 $\mathbf{X}_{\hat{j}_1}, \mathbf{X}_{\hat{j}_2}^\perp, \cdots, \mathbf{X}_{\hat{j}_k}^\perp$

by adding the k projections into the respective one-dimensional linear spaces (i.e., componentwise linear regression of $x_{t,\hat{j}_{k+1}}$ on $x_{t,\hat{j}_i}^\perp$). This also yields the residual vector $\mathbf{X}_{\hat{j}_{k+1}}^\perp = \mathbf{X}_{\hat{j}_{k+1}} - \hat{\mathbf{X}}_{\hat{j}_{k+1}}$. With $\mathbf{X}_{\hat{j}_{k+1}}^\perp = (x_{1,\hat{j}_{k+1}}^\perp, \cdots, x_{n,\hat{j}_{k+1}}^\perp)^\top$ thus computed, OGA uses the following update in lieu of (2.2):

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

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

Note that OGA is equivalent to the least squares regression of y_t on $(x_{t,\hat{j}_1}, \cdots, x_{t,\hat{j}_{k+1}})^\top$ 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 $\sum_{t=1}^n U_t^{(k)} x_{tj}^\perp = \sum_{t=1}^n U_t^{(k)} x_{tj}$, $\tilde{\beta}_j^{(k)}$ and $\hat{\beta}_j^{(k)}$ only differ in their denominators $\sum_{t=1}^n x_{tj}^2$ and $\sum_{t=1}^n (x_{tj}^\perp)^2$. 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 $(x_{t,\hat{j}_1}, \cdots, x_{t,\hat{j}_k})^\top$ for OGA, the corresponding variable selector for $\hat{j}_{k+1}$ in (2.1) can be restricted to $j \notin \{\hat{j}_1, \cdots, \hat{j}_k\}$. Therefore, unlike PGA for which the same predictor variable can be entered repeatedly, OGA excludes variables that are already included from further consideration in (2.1).

While (2.3) evaluates $\hat{y}_{k+1}(\mathbf{x})$ at $\mathbf{x} \in \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, estimating $y(\mathbf{x})$ at general $\mathbf{x}$ requires the OLS estimate $\hat{\boldsymbol{\beta}}_{k+1}$ of $\boldsymbol{\beta}(\hat{J}_{k+1})$, where $\hat{J}_{k+1} = \{\hat{j}_1, \cdots, \hat{j}_{k+1}\}$ and $\boldsymbol{\beta}(\hat{J}_{k+1}) = (\beta_{\hat{j}_1}, \cdots, \beta_{\hat{j}_{k+1}})^\top$. Thus, the OGA analog of (2.2) is

$$\text{(2.4)} \qquad \hat{y}_{k+1}(\mathbf{x}) = (x_{\hat{j}_1}, \cdots, x_{\hat{j}_{k+1}}) \hat{\boldsymbol{\beta}}_{k+1}.$$

We next describe a recursive algorithm to compute $\hat{\boldsymbol{\beta}}_{k+1}$. This involves the following recursive formula for the coefficients $b_{\nu i}, 1 \leq i \leq \nu - 1$, in the representation

$$\text{(2.5)} \qquad \mathbf{X}_{\hat{j}_\nu}^\perp = \mathbf{X}_{\hat{j}_\nu} + b_{\nu 1} \mathbf{X}_{\hat{j}_1} + \cdots + b_{\nu,\nu-1} \mathbf{X}_{\hat{j}_{\nu-1}}.$$

First recall that $\hat{\mathbf{X}}_{\hat{j}_k} = \mathbf{X}_{\hat{j}_k} - \mathbf{X}_{\hat{j}_k}^\perp$ is the projection $a_{k1} \mathbf{X}_{\hat{j}_1} + a_{k2} \mathbf{X}_{\hat{j}_2}^\perp + \cdots + a_{k,k-1} \mathbf{X}_{\hat{j}_{k-1}}^\perp$ of $\mathbf{X}_{\hat{j}_k}$ into the linear space spanned by the orthogonal vectors $\mathbf{X}_{\hat{j}_1}, \mathbf{X}_{\hat{j}_2}^\perp, \cdots, \mathbf{X}_{\hat{j}_{k-1}}^\perp$, and therefore $a_{ki} = \mathbf{X}_{\hat{j}_k}^\top \mathbf{X}_{\hat{j}_i}^\perp / \|\mathbf{X}_{\hat{j}_i}^\perp\|^2$ for $1 \leq i \leq k$, replacing $\mathbf{X}_{\hat{j}_i}^\perp$ by $\mathbf{X}_{\hat{j}_i}$ when $i = 1$. The recursion

$$\text{(2.6)} \quad b_{k,k-1} = -a_{k,k-1}, b_{ki} = -a_{ki} - \sum_{j=i+1}^{k-1} a_{kj} b_{ji} \text{ for } 1 \leq i < k - 1$$

follows by combining (2.5) with $\mathbf{X}_{\hat{j}_k}^{\perp} = \mathbf{X}_{\hat{j}_k} - a_{k1}\mathbf{X}_{\hat{j}_1} - a_{k2}\mathbf{X}_{\hat{j}_2}^{\perp} - \cdots - a_{k,k-1}\mathbf{X}_{\hat{j}_{k-1}}^{\perp}$.
We use the $\hat{\beta}_{\hat{j}_{k+1}}^{(k)}$ in (2.3) and $\mathbf{b}_{k+1} := (b_{k+1,1}, \cdots, b_{k+1,k})$ to compute $\hat{\boldsymbol{\beta}}_{k+1}$ recursively as follows, initializing with $\hat{\boldsymbol{\beta}}_1 = (\sum_{t=1}^{n} x_{t,\hat{j}_1} y_t)/(\sum_{t=1}^{n} x_{t,\hat{j}_1}^2)$:

$$(2.7) \qquad \hat{\boldsymbol{\beta}}_{k+1}^{\top} = (\hat{\boldsymbol{\beta}}_k^{\top} + \hat{\beta}_{\hat{j}_{k+1}}^{(k)} \mathbf{b}_{k+1}, \hat{\beta}_{\hat{j}_{k+1}}^{(k)}).$$

Note that (2.7) follows from (2.5) with $\nu = k+1$ and the fact that the components of $\hat{\boldsymbol{\beta}}_k$ are the coefficients in the projection of $(y_1, \cdots, y_n)^{\top}$ into the linear space spanned by $\mathbf{X}_{\hat{j}_1}, \cdots, \mathbf{X}_{\hat{j}_k}$.

2.3. *Population version of OGA.* Let $y, z_1, \cdots, z_p$ be square integrable random variables having zero means and such that $E(z_i^2) = 1$. Let $\mathbf{z} = (z_1, \cdots, z_p)^{\top}$. The population version of OGA, which is a special case of Temlyakov's [21] greedy algorithms, is an iterative scheme which chooses $j_1, j_2, \cdots$ sequentially by

$$(2.8) \qquad j_{k+1} = \arg \max_{1 \leq j \leq p} |E(u_k z_j)|, \text{ where } u_k = y - \tilde{y}_k(\mathbf{z}),$$

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

3. An oracle-type inequality and convergence rates under weak sparsity.
In this section, we first prove convergence rates for OGA in linear regression models in which the number of regressors is allowed to grow exponentially with 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 assumption (A1) in [2] for PGA. Moreover, as in [2], we assume that the $(\varepsilon_t, \mathbf{x}_t)$ in (1.1) are i.i.d. and such that ε_t is independent of $\mathbf{x}_t$ and

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

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 $s > 0$ such that

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

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) $\sup_{n\geq 1}\sum_{j=1}^{p_n}|\beta_j\sigma_j| < \infty$.

Conditions (C1)-(C4) are closely related to Bühlmann's assumptions (A1)-(A4) in [2]. In particular, condition (C4) is somewhat weaker than (A2). While Bühlmann replaces (C2) by a weaker moment condition (A4), his assumption (A3) requires $x_j(1 \leq j \leq p_n)$ to be uniformly bounded random variables while (C3) is considerably weaker. The second part of this section gives an inequality for OGA similar to Bickel, Ritov and Tysbakov's [1] oracle inequality for Lasso. In this connection we also review related inequalities in the approximation theory literature dealing with the noiseless case $\varepsilon_t = 0$ for all t. When there are no random disturbances, the approximation error comes from the bias in using m (instead of p) input variables. When random disturbances ε_t are present, the inequality we prove for the performance of OGA reflects a bias-variance tradeoff similar to that in the oracle inequality for Lasso in [1].

3.1. *Uniform convergence rates.* Let K_n denote a prescribed upper bound on the number m of OGA iterations. Let

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

where $\mathbf{z}(J)$ is a subvector of $(z_1, \cdots, z_p)^\top$ and J denotes the associated subset of indices $1, \cdots, p$. We assume that for some positive constant M independent of n,

$$(3.2) \quad \liminf_{n\to\infty} \min_{1\leq\sharp(J)\leq K_n} \lambda_{\min}(\mathbf{\Gamma}(J)) > 0, \quad \max_{1\leq\sharp(J)\leq K_n, i\notin J} \|\mathbf{\Gamma}^{-1}(J)\mathbf{g}_i(J)\|_1 < M,$$

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

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

The following theorem gives the rate of convergence, which holds uniformly over $1 \leq m \leq K_n$, for the CPE (defined in (1.3)) of OGA provided that the correlation matrix of the regressors selected sequentially by OGA satisfies (3.2).

THEOREM 3.1. *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,

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

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 $\{1, \cdots, p_n\}$. Let J_k be the set of input variables selected by the population version of OGA at the end of stage k.

8

Then by Theorem 3 of [21], 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 larger variance in the least squares estimates $\hat{\beta}_{\hat{j}_i}, i = 1, \cdots, m$. 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 $\{1, \cdots, p_n\}$. 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 $E[\cdot|y_1, \mathbf{x}_1, \cdots, y_n, \mathbf{x}_n]$. Theorem 3.1 says that uniformly in $m \le (n/\log p_n)^{1/2}$, 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.

PROOF OF THEOREM 3.1. Let $\hat{J}_k = \{\hat{j}_1, \cdots, \hat{j}_k\}$ 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 $i \notin J$, define

$$
\mu_{J,i} = E[\{y(\mathbf{x}) - y_J(\mathbf{x})\}x_i],
$$

(3.4)
$$
\hat{\mu}_{J,i} = \{n^{-1}\sum_{t=1}^{n}(y_t - \hat{y}_{t;J})x_{ti}\}/(n^{-1}\sum_{t=1}^{n}x_{ti}^2)^{1/2},
$$

where $\hat{y}_{t;J}$ denotes the fitted value of y_t when $\mathbf{Y} = (y_1, \cdots, y_n)^\top$ 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 $i \notin J$,

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

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.,

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

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 $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^\top$ 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

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

In Appendix A, we make use of (C2), (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

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

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

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

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

(3.10)
$$
|\mu_{\hat{J}_{q-1},\hat{j}_q}| \geq \xi \max_{1 \leq i \leq p_n} |\mu_{\hat{J}_{q-1},i}| \text{ on } A_n(m) \bigcap B_n(m),
$$

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

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

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

Consider the "semi-population version" of OGA that uses the variable selector $(\hat{j}_1, \hat{j}_2, \cdots)$ but still approximates $y(\mathbf{x})$ by $\sum_{j \in \hat{J}_{k+1}} c_j x_j$, where the c_j are the same as those for the population version of OGA. In view of (3.10), this semi-population

version is a "weak orthogonal greedy algorithm" introduced by Temlyakov [21, pp. 216-217], to which we can apply Theorem 3 of [21] to conclude that

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

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

$$E_n[\{y(\mathbf{x}) - y_{\hat{j}_m}(\mathbf{x})\}^2] \leq \min_{0 \leq i \leq m-1} E_n\{(y(\mathbf{x}) - y_{\hat{j}_i}(\mathbf{x}))(\sum_{j=1}^{p_n} \beta_j x_j)\}$$

$$\leq \min_{0 \leq i \leq m-1} \max_{1 \leq j \leq p_n} |\mu_{\hat{j}_i, j}| \sum_{j=1}^{p_n} |\beta_j| \leq \tilde{\xi} s (n^{-1} \log p_n)^{1/2} \sum_{j=1}^{p_n} |\beta_j| \text{ on } B_n^c(m).$$

Combining this with (C4), (3.11) and the assumption that $m \leq K_n$ yields

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

for some constant $C^* > 0$, since $K_n = O((n/\log p_n)^{1/2})$. Moreover, since $A_n(K_n) \subseteq A_n(m)$, it follows from (3.8) and (3.12) that $\max_{1 \leq m \leq K_n} m E_n[\{y(\mathbf{x}) - y_{\hat{j}_m}(\mathbf{x})\}^2] = O_p(1)$. Theorem 3.1 follows from this and

$$(3.13) \quad \max_{1 \leq m \leq K_n} \frac{n E_n[\{\hat{y}_m(\mathbf{x}) - y_{\hat{j}_m}(\mathbf{x})\}^2]}{m \log p_n} = O_p(1),$$

whose proof is given in Appendix A, noting that

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

3.2. *A bias-variance bound.* In this section, we assume that x_{tj} in (1.1) are nonrandom constants and develop an upper bound for the empirical norm

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

of OGA, providing an analog of the oracle inequalities of [1], [5], [6] and [26] for Lasso and the Dantzig selector. 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 [23]. When ε_t are nonzero but nonrandom, a bound for the bias of the OGA estimate has also been given by Donoho, Elad and Temlyakov [9]. 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

$$(3.15) \qquad \mathbf{B} = \{\mathbf{b} : \mathbf{Xb} = (y(\mathbf{x}_1), \cdots, y(\mathbf{x}_n))^\top\},$$

where $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_p)$ is $n \times p$. In addition, for $J \subseteq \{1, \cdots, p\}$ 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, define

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

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

THEOREM 3.2. *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*

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

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

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

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

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

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 $s \log p$ in order to take into account the worst possible case; see Foster and George [13] for a detailed discussion of the idea of risk inflation. As will be explained in Appendix B that gives the proof of Theorem 3.2, 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 decreases as m increases, 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.

The $\|\boldsymbol{\theta}_{J,i}\|_1$ function in (3.17) is closely related to the cumulative coherence function introduced in Tropp [23]. Since Theorem 3.2 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). 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

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

where d does not depend on n and p. We can regard (3.20) as an analog of the oracle inequality of Bickel et al. [1, Theorem 6.2] for the Lasso predictor $\hat{y}_{\text{Lasso}(r)}(\mathbf{x}_t) = \mathbf{x}_t^\top\hat{\boldsymbol{\beta}}_{\text{Lasso}(r)}$, where

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

with $r > 0$. Let $M(\mathbf{b})$ denote the number of nonzero components of $\mathbf{b} \in \mathbf{B}$, and define $\bar{Q} = \inf_{\mathbf{b}\in\mathbf{B}} M(\mathbf{b})$. Instead of (3.17), Bickel et al. [1] assume that σ is known and $\mathbf{X}^\top\mathbf{X}$ satisfies a restricted eigenvalue assumption $\text{RE}(\bar{Q}, 3)$, and show under the same assumptions of Theorem 3.2 (except for (3.17)) that for $r = A\sigma(n^{-1}\log p)^{1/2}$ with $A > 2\sqrt{2}$,

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

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 engenvalue of $\text{RE}(\bar{Q}, 3)$.

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 $\bar{Q} << (n/\log p)^{1/2}$ (or $\bar{Q} >> (n/\log)^{1/2}$). To see this, note that when the underlying model has a sparse representation in the sense that $\bar{Q} << (n/\log p)^{1/2}$, Lasso provides a sparse solution that can substantially reduce the approximation error. However, when the model does not have a sparse representation, Lasso tends to give a relatively *non-sparse* solution in order to control the approximation error, which also results in large variance. In addition, while the RE assumption is quite flexible, the value of σ, which is required for obtaining $\hat{\boldsymbol{\beta}}_{\text{Lasso}(r)}$ with $r = A\sigma(n^{-1}\log p)^{1/2}$, is usually

unknown in practice. On the other hand, although the ability of OGA to reduce the approximation error is not as good as Lasso due to its stepwise *local* optimization feature, it has advantages in non-sparse situations by striking a good balance between the variance and squared bias with the choice $m = D(n/\log p)^{1/2}$ in (3.20).

Note that Theorems 3.1 and 3.2 consider weakly sparse regression models that allow all p regression coefficients to be nonzero. 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 $p >> n$, as shown in [8] that proposes to modify these criteria by including a factor $\log p$ in order to alleviate overfitting problems.

4. Consistent model selection under strong sparsity. In the case of a fixed number p (not growing with n) of input variables, only q $(< p)$ of which have nonzero regression coefficients, it is well known that AIC or CV does not penalize the residual sum of squares enough for consistent variable selection, for which a larger penalty (e.g., that used in BIC) is needed. When $p = p_n \to \infty$ but the number of nonzero regression coefficients remains bounded, Chen and Chen [7] have shown that BIC does not penalize enough and have introduced a considerably larger penalty to add to the residual sum of squares so that their "extended BIC" is variable selection consistent when $p_n = O(n^\kappa)$ for some $\kappa > 0$ and an asymptotic identifiability condition is satisfied. As in Section 3.1, we assume here that $\log p_n = o(n)$ and also allow the number of nonzero regression coefficients to approach ∞ as $n \to \infty$. 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. This consideration leads to the following strong sparsity condition:

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

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

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 [7], we show under the strong sparsity condition (C5) that all relevant regressors (i.e., those with nonzero regression coefficients) are included by OGA, with probability near 1 if the number K_n of iterations is large enough. Letting

(4.1) $\qquad N_n = \{1 \le j \le p_n : \beta_j \ne 0\}, \ D_n = \{N_n \subseteq \hat{J}_{K_n}\},$

note that the set $\hat{J}_{K_n}$ of variables selected along an OGA path includes all relevant regressors on D_n.

THEOREM 4.1. *Assume* (C1)-(C5) *and* (3.2). *Suppose* $K_n/n^\gamma \to \infty$ *and* $K_n = O((n/\log p_n)^{1/2})$. *Then* $\lim_{n\to\infty} P(D_n) = 1$.

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),

$$
\begin{gathered}
\lim_{n\to\infty} P(A_n^c(m)) \le \lim_{n\to\infty} P(A_n^c(K_n)) = 0, \\
E_n\{[y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})]^2\}I_{A_n(m)} \le C^*m^{-1}.
\end{gathered}
$$
(4.2)

From (4.2), it follows that

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

For $J \subseteq \{1, \cdots, p_n\}$ 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 $E(y - \sum_{i\in J} c_i x_i)^2$. Define $\tilde{\beta}_j(J) = 0$ if $j \notin J$. Note that

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

From (C4) and (C5), it follows that $\sharp(N_n) = o(n^{\gamma/2})$, yielding $\sharp(\hat{J}_m \cup N_n) = o(K_n)$. Let $v_n = \min_{1\le\sharp(J)\le K_n} \lambda_{\min}(\mathbf{\Gamma}(J))$. Then we obtain from (4.4) that for all large n,

(4.5) $\quad E_n[\{y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})\}^2] \ge v_n \min_{j\in N_n} \beta_j^2 \text{ on } \{N_n \cap \hat{J}_m^c \ne \emptyset\}.$

Combining (4.5) with (C5) and (3.2) then yields for some $b > 0$ and all large n, $E_n[\{y(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x})\}^2] \ge bn^{-\gamma}$ on $\{N_n \cap \hat{J}_m^c \ne \emptyset\}$. By choosing the a in $m = \lfloor an^\gamma \rfloor$ large enough, we have $bn^{-\gamma} > C^*m^{-1}$, implying that $\{N_n \cap \hat{J}_m^c \ne \emptyset\} \subseteq F_n$, where F_n is defined in (4.3). Hence by (4.3), $\lim_{n\to\infty} P(N_n \subseteq \hat{J}_m) = 1$. Therefore, the OGA that terminates after $m = \lfloor an^\gamma \rfloor$ iterations contains all relevant regressors with probability approaching 1. This is also true for the OGA that terminates after K_n iterations if $K_n/m \to \infty$. $\quad\square$

Under strong sparsity, we next show that the best fitting model can be chosen along an OGA path by minimizing a high-dimensional information criterion

(HDIC) that uses a heavier penalty than the extended BIC of [7]. For $J \subseteq \{1, \cdots, p_n\}$, define

$$(4.6) \qquad \text{HDIC}(J) = \sum_{t=1}^{n} (y_t - \hat{y}_{t;J})^2 (1 + \frac{\sharp(J) w_n \log p_n}{n}),$$

where $\hat{y}_{t;J}$ is defined below (3.4) and w_n are positive numbers satisfying $w_n \to \infty$ and $w_n = o(n)$. In particular, the choice $w_n = \log n$ corresponds to HDBIC (high-dimensional BIC) while $w_n = c \log \log n$ with $c > 2$ corresponds to HDHQ (high-dimensional Hannan-Quinn criterion).

The next theorem establishes, under strong sparsity, consistency of variable selection along OGA paths by HDIC. Define the minimal ("oracle") number of relevant regressors along an OGA path by

$$(4.7) \qquad \widetilde{k}_n = \min\{k : 1 \leq k \leq K_n, N_n \subseteq \hat{J}_k\} \ (\min \emptyset = K_n).$$

Although N_n is unobserved, Theorem 4.1 says that if K_n is sufficiently large, then $\hat{J}_{K_n}$ contains N_n with probability approaching 1 as $n \to \infty$. Define

$$(4.8) \qquad \hat{k}_n = \arg \min_{1 \leq k \leq K_n} \text{HDIC}(\hat{J}_k).$$

The factor $\log p_n$ in the definition (4.6) of HDIC is used to exclude irrelevant variables that may exhibit spurious sample correlations with y_t because of an overly large number ($p_n >> n$) of such variables. Suppose $x_1, \cdots, x_{p_n}$ are uncorrelated, i.e., $\mathbf{\Gamma}(J) = \mathbf{I}$, for which the following "hard thresholding" [10] method 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

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

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 $x_1, \cdots, x_{p_n}$ are correlated and therefore so are the $\hat{\beta}_j$, we make use of the assumption on $\lambda_{\min}(\mathbf{\Gamma}(J))$ in (3.2). Regarding 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 leads to the criterion (4.6)

and suggests selecting the regressor set J that minimizes HDIC(J). Under (C5) and (3.2), it can be shown that if w_n is so chosen that

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

then the minimizer of HDIC(J) over J with $J \subseteq \hat{J}_{K_n}$ is consistent for variable selection. Note that the penalty used in HDIC is heavier than that used in the extended BIC of [7], and instead of all subset models used in [7], HDIC is applied to models *after* greedy forward screening, which may result in stronger spurious correlations among competing models. The proof is similar to that of the following theorem, which focuses on consistent variable selection along OGA paths, $\hat{J}_1, \cdots, \hat{J}_{K_n}$. In practice, it is not feasible to minimize HDIC(J) over all subsets J of $\hat{J}_{K_n}$ when K_n is large, and a practical alternative is to minimize HDIC($\hat{J}_k$) instead.

THEOREM 4.2. *With the same notation and assumptions as in Theorem* 4.1, *suppose* (4.9) *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. As in the proof of Theorem 4.1, assume $\sigma_j^2 = 1$. For notational simplicity, we drop the subscript n in $\tilde{k}_n$ and $\hat{k}_n$. We first show that $P(\hat{k} < \tilde{k}) = o(1)$. As shown in the proof of Theorem 4.1, for sufficiently large a,

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

On $\{\hat{k} < \tilde{k}\}$, HDIC($\hat{J}_{\hat{k}}$) $\le$ HDIC($\hat{J}_{\tilde{k}}$) and $\sum_{t=1}^n (y_t - \hat{y}_{t;\hat{j}_{\hat{k}}})^2 \ge \sum_{t=1}^n (y_t - \hat{y}_{t;\hat{j}_{\tilde{k}-1}})^2$, so

(4.11)
$$\sum_{t=1}^n (y_t - \hat{y}_{t;\hat{j}_{\tilde{k}-1}})^2 - \sum_{t=1}^n (y_t - \hat{y}_{t;\hat{j}_{\tilde{k}}})^2$$
$$\le n^{-1} w_n (\log p_n)(\tilde{k} - \hat{k}) \sum_{t=1}^n (y_t - \hat{y}_{t;\hat{j}_{\tilde{k}}})^2.$$

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

(4.12)
$$n^{-1} \{ \sum_{t=1}^n (y_t - \hat{y}_{t;\hat{j}_{\tilde{k}-1}})^2 - \sum_{t=1}^n (y_t - \hat{y}_{t;\hat{j}_{\tilde{k}}})^2 \}$$
$$= n^{-1} (\beta_{\hat{j}_{\tilde{k}}} \mathbf{X}_{\hat{j}_{\tilde{k}}} + \boldsymbol{\varepsilon})^\top (\mathbf{H}_{\hat{J}_{\tilde{k}}} - \mathbf{H}_{\hat{J}_{\tilde{k}-1}})(\beta_{\hat{j}_{\tilde{k}}} \mathbf{X}_{\hat{j}_{\tilde{k}}} + \boldsymbol{\varepsilon})$$
$$= \{ \beta_{\hat{j}_{\tilde{k}}} \mathbf{X}_{\hat{j}_{\tilde{k}}}^\top (\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}-1}}) \mathbf{X}_{\hat{j}_{\tilde{k}}} $$
$$+ \mathbf{X}_{\hat{j}_{\tilde{k}}}^\top (\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}-1}}) \boldsymbol{\varepsilon} \}^2 / \{ n \mathbf{X}_{\hat{j}_{\tilde{k}}}^\top (\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}-1}}) \mathbf{X}_{\hat{j}_{\tilde{k}}} \}.$$

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

$$\hat{A}_n = n^{-1} \mathbf{X}_{\hat{j}_{\tilde{k}}}^\top (\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}-1}}) \mathbf{X}_{\hat{j}_{\tilde{k}}}, \quad \hat{B}_n = n^{-1} \mathbf{X}_{\hat{j}_{\tilde{k}}}^\top (\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}-1}}) \varepsilon,$$

(4.13)
$$\hat{C}_n = n^{-1} \sum_{t=1}^{n} (y_t - \hat{y}_{t;\hat{j}_{\tilde{k}}})^2 - \sigma^2.$$

Hence, (4.12) and (4.11) yield $\beta_{\hat{j}_{\tilde{k}}}^2 \hat{A}_n + 2\beta_{\hat{j}_{\tilde{k}}} \hat{B}_n + \hat{A}_n^{-1} \hat{B}_n^2 \le n^{-1} w_n (\log p_n)(\tilde{k} - \hat{k})(\hat{C}_n + \sigma^2)$ on $\{\hat{k} < \tilde{k}\}$, which implies that

(4.14)
$$2\beta_{\hat{j}_{\tilde{k}}} \hat{B}_n - n^{-1} w_n (\log p_n) \lfloor an^\gamma \rfloor |\hat{C}_n|$$
$$\le -\beta_{\hat{j}_{\tilde{k}}}^2 \hat{A}_n + n^{-1} w_n (\log p_n) \lfloor an^\gamma \rfloor \sigma^2 \text{ on } \{\hat{k} < \tilde{k}\} \bigcap \mathcal{D}_n.$$

Define $v_n = \min_{1 \le \sharp(J) \le \lfloor an^\gamma \rfloor} \lambda_{\min}(\mathbf{\Gamma}(J))$. Note that by (3.2), there exists $c_0 > 0$ such that for all large n, $v_n > c_0$. In Appendix A, it will be shown that for any $\theta > 0$,

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

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

It remains to prove $P(\hat{k} > \tilde{k}) = o(1)$. Noting that $\tilde{k} < K_n$ and $\beta_j = 0$ for $j \notin \hat{J}_{\tilde{k}}$, we have the following counterpart of (4.11) and (4.12) on $\{K_n \ge \hat{k} > \tilde{k}\}$:

$$\varepsilon^\top (\mathbf{I} - \mathbf{H}_{\hat{J}_{\hat{k}}}) \varepsilon (1 + n^{-1} \hat{k} w_n \log p_n) \le \varepsilon^\top (\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}}}) \varepsilon (1 + n^{-1} \tilde{k} w_n \log p_n),$$

which implies

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

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

$$\varepsilon^\top (\mathbf{H}_{\hat{J}_{\hat{k}}} - \mathbf{H}_{\hat{J}_{\tilde{k}}}) \varepsilon$$
$$= \varepsilon^\top (\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}}}) \mathbf{F}_{\hat{k},\tilde{k}} \{\mathbf{F}_{\hat{k},\tilde{k}}^\top (\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}}}) \mathbf{F}_{\hat{k},\tilde{k}}\}^{-1} \mathbf{F}_{\hat{k},\tilde{k}}^\top (\mathbf{I} - \mathbf{H}_{\hat{J}_{\tilde{k}}}) \varepsilon$$

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

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

$$\hat{a}_n = \|\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_{K_n})\| \max_{1 \le j \le p_n} (n^{-1/2} \sum_{t=1}^{n} x_{tj} \varepsilon_t)^2,$$

$$\hat{b}_n = \|\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_{K_n})\| \max_{1 \le \sharp(J) \le \tilde{k}, i \notin J} (n^{-1/2} \sum_{t=1}^{n} \varepsilon_t \hat{x}_{ti;J})^2.$$

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

(4.18)
$$2(\hat{k} - \tilde{k})(\hat{a}_n + \hat{b}_n)(1 + n^{-1}\hat{k}w_n \log p_n) + |\hat{C}_n|(\hat{k} - \tilde{k})w_n \log p_n$$
$$\ge \sigma^2(\hat{k} - \tilde{k})w_n \log p_n.$$

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

(4.19)
$$P\{(\hat{a}_n + \hat{b}_n)(1 + n^{-1}K_n w_n \log p_n) \ge \theta w_n \log p_n\}$$
$$+ P\{|\hat{C}_n| \ge \theta\} = o(1).$$

From (4.18) and (4.19), the desired conclusion $P(\hat{k} > \tilde{k}) = o(1)$ follows. $\qquad \square$

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

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

where $\beta_{q+1} = \cdots = \beta_p = 0$, $p >> n$, ε_t are i.i.d. $\mathrm{N}(0, \sigma^2)$ and are independent of the x_{tj}. Examples 1 and 2 consider the case

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

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

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

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

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

As noted in Section 1, Fan and Lv [12] require $\lambda_{\max}(\mathbf{\Gamma}(\{1, \cdots, p\})) \leq cn^r$ for some $c > 0$ and $0 \leq r < 1$ 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 $p >> n$, in view of (5.4). 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 [25] have shown that under the sparse Riesz condition $c_* \leq \min_{1 \leq \sharp(J) \leq q^*} \lambda_{\min}(\hat{\mathbf{\Gamma}}(J)) \leq \max_{1 \leq \sharp(J) \leq q^*} \lambda_{\max}(\hat{\mathbf{\Gamma}}(J)) \leq c^*$ for some $c^* \geq c_* > 0$ and $q^* \geq \{2 + (4c^*/c_*)\}q + 1$, 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), the requirement that $q^* \geq \{2 + (4c^*/c_*)\}q + 1$ in the sparse Riesz condition is difficult to meet when $q \geq (2\eta)^{-2}$.

Example 1. Consider (5.1) with $q = 5$, $(\beta_1, \cdots, \beta_5) = (3, -3.5, 4, -2.8, 3.2)$ 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 [20] to illustrate the performance of their variable screening method. The cases $\eta = 0, 2$ and $(n, p) = (50, 1000), (100, 2000), (200, 4000)$ are considered here to accommodate a much larger number of candidate variables and allow substantial correlations among them. In light of Theorem 4.2 which requires the number K_n of iterations to satisfy $K_n = O((n/\log p_n)^{1/2})$, we choose $K_n = 5(n/\log p_n)^{1/2}$. Table 1 shows that OGA+HDBIC and OGA+HDHQ perform well, in agreement with the asymptotic theory of Theorem 4.2. Each result is based on 1000 simulations. Here and in the sequel, we choose $c = 2.01$ for HDHQ. For comparison, the performance of OGA+BIC is also included in the table.

In the simulations for $n \geq 100$, 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 ($\eta = 2$). In contrast, OGA+BIC always chooses the model obtained at the last iteration. The mean squared prediction errors (MSPE) of OGA+BIC are at least 20 times higher than those of OGA+HDBIC or OGA+HDHQ when $n = 100$, and are at least 35 times higher when $n = 200$, where

$$\text{MSPE} = \frac{1}{1000} \sum_{l=1}^{1000} \left(\sum_{j=1}^{p} \beta_j x_{n+1,j}^{(l)} - \hat{y}_{n+1}^{(l)}\right)^2,$$

in which $x_{n+1,1}^{(l)}, \cdots, x_{n+1,p}^{(l)}$ 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 $y_{n+1}^{(l)}$. Note that the MSPEs of OGA+HDBIC and OGA+HDHQ are close to the ideal value $5\sigma^2/n$ of (1.3) for the least squares predictor based on only the $q = 5$ relevant variables in (5.1).

In the case of $n = 50$ and $p = 1000$, OGA can include all relevant regressors (within K_n iterations) about 94% of the time when $\eta = 0$; this ratio decreases to 83% when $\eta = 2$. As shown in Table 2 of [20], in the case of $n = 100$ and $p = 400$, the variable screening method of Shao and Chow [20], used in conjunction with AIC or BIC, can only identify the smallest correct model about 50% of the time even when $\eta = 0$. To examine the sensitivity of K_n in our simulation study, we have varied the value of D in $K_n = D(n/\log p_n)^{1/2}$ from 3 to 10. The performance of OGA+HDBIC and OGA+HDHQ remains similar to the case of $D = 5$, but that of OGA+BIC becomes worse as D becomes larger.

Example 2. Consider (5.1) with $q = 9$, $n = 400$, $p = 4000$, $(\beta_1, \cdots, \beta_q)$=(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 $\eta = 1$. Table 2 summarizes the results of 1000 simulations on the performance of Lasso, OGA+HDBIC and OGA+HDHQ, using $K_n = 5(n/\log p)^{1/2}$ iterations for OGA. The Lasso estimate $\hat{\boldsymbol{\beta}}^L(\lambda) = (\hat{\beta}_1^L(\lambda), \cdots, \hat{\beta}_p^L(\lambda))^\top$ is defined as the minimizer of $n^{-1} \sum_{t=1}^n (y_t - \mathbf{x}_t^\top \mathbf{c})^2 + \lambda\|\mathbf{c}\|_1$ over $\mathbf{c}$. To implement Lasso, we use the Glmnet package of [15] and conduct 5-fold cross-validation to select the optimal λ, yielding the estimate "Lasso" in Table 2. In addition, we have also traced the Lasso path using a modification of the LARS algorithm [11, Section 3] based on the MATLAB code of [19], and the tracing is stopped when the value of the regularization parameter defined in [19] along the path is smaller than the default tolerance of 10^{-5} or 10^{-4}, yielding the estimate "LARS(10^{-5})" or "LARS(10^{-4})" in Table 2. We have also tried other default tolerance values $10^{-1}, 10^{-2}, \cdots, 10^{-8}$, but the associated performance is worse than that of LARS(10^{-4}) or LARS(10^{-5}). Table 2 shows that OGA+HDBIC and OGA+HDHQ can identify the smallest correct model in 98% of the time and choose slightly overfitting models in 2% of the time. Their MSPEs are near the oracle value $q\sigma^2/n = 0.051$. Although Lasso or LARS-based implementation of Lasso can include the 9 relevant variables in all simulation runs, they encounter overfitting problems. The model chosen by Lasso is usually more parsimonious than that chosen by LARS (10^{-4}) or LARS(10^{-5}). However, the MSPE of Lasso is larger than those of LARS(10^{-5}) and Lasso(10^{-4}), which are about 10-12 times larger than $q\sigma^2/n$ due to overfitting.

This example satisfies the neighborhood stability condition, introduced in [18], which requires that for some $0 < \delta < 1$ and all $i = q + 1, \cdots, p$,

$$(5.5) \qquad |\mathbf{c}_{qi}'\mathbf{R}^{-1}(q)\mathrm{sign}(\boldsymbol{\beta}(q))| < \delta,$$

where $\mathbf{x}_t(q) = (x_{t1}, \cdots, x_{tq})^\top$, $\mathbf{c}_{qi} = E(\mathbf{x}_t(q)x_{ti})$, $\mathbf{R}(q) = E(\mathbf{x}_t(q)\mathbf{x}_t^\top(q))$ and $\mathrm{sign}(\boldsymbol{\beta}(q)) = (\mathrm{sign}(\beta_1), \cdots, \mathrm{sign}(\beta_q))^\top$. To show that (5.5) holds in this example, 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 $\mathrm{sign}(\boldsymbol{\beta}(q)) = \mathbf{1}_q$, where $\mathbf{1}_q$ is the q-dimensional vector of 1's. Therefore, for all $i = q + 1, \cdots, p$, $|\mathbf{c}_{qi}'\mathbf{R}^{-1}(q)\mathrm{sign}(\boldsymbol{\beta}(q))| = \eta^2 q/(1 + \eta^2 q) < 1$.

TABLE 1

Frequency, in 1000 simulations, of selecting all 5 relevant variables in Example 1

η	n	p	Method	E	E+1	E+2	E+3	E+4	E+5	E*	Correct	MSPE
0	50	1000	OGA+HDBIC	902	33	7	0	0	0	0	942	2.99
			OGA+HDHQ	864	64	13	3	1	0	0	945	2.90
			OGA+BIC	0	0	0	0	0	0	945	945	4.87
	100	2000	OGA+HDBIC	1000	0	0	0	0	0	0	1000	0.053
			OGA+HDHQ	994	6	0	0	0	0	0	1000	0.055
			OGA+BIC	0	0	0	0	0	0	1000	1000	1.245
	200	4000	OGA+HDBIC	1000	0	0	0	0	0	0	1000	0.023
			OGA+HDHQ	997	2	1	0	0	0	0	1000	0.025
			OGA+BIC	0	0	0	0	0	0	1000	1000	0.895
2	50	1000	OGA+HDBIC	663	121	41	5	0	0	0	830	9.43
			OGA+HDHQ	643	134	44	17	4	1	0	843	8.36
			OGA+BIC	0	0	0	0	0	0	843	843	12.59
	100	2000	OGA+HDBIC	994	6	0	0	0	0	0	1000	0.054
			OGA+HDHQ	993	7	0	0	0	0	0	1000	0.055
			OGA+BIC	0	0	0	0	0	0	1000	1000	1.121
	200	4000	OGA+HDBIC	1000	0	0	0	0	0	0	1000	0.024
			OGA+HDHQ	996	4	0	0	0	0	0	1000	0.025
			OGA+BIC	0	0	0	0	0	0	1000	1000	0.881

1. E denotes the frequency of selecting *exactly* all relevant variables.

2. E+i denotes the frequency of selecting all relevant variables and i additional irrelevant variables.

3. E* denotes the frequency of selecting the largest model along the OGA path which includes all relevant variables.

4. Correct denotes the frequency of including all relevant variables.

TABLE 2

Frequency, in 1000 simulations, of selecting all 9 relevant variables in Example 2

Method						
	E	E+1	E+2	E+3	Correct	MSPE
OGA+HDBIC	980	19	1	0	1000	0.052
OGA+HDHQ	980	19	1	0	1000	0.052
	Range		Average		Correct	MSPE
LARS(10^{-5})	[44, 389]		124.3		1000	0.602
LARS(10^{-4})	[30, 227]		82.6		1000	0.513
Lasso	[25, 84]		50.17		1000	0.864

1. The lower and upper limits of range are the smallest and largest numbers of selected variables in 1000 runs.

2. Average is the mean number of selected variables over 1000 runs.

3. See the footnotes of Table 1 for the notations E, E+i, and Correct.

Frequency, in 1000 simulations, of selecting all 10 relevant variables in Example 3 using

the same notation as that in Table 2						
Method						
	E	E+1	E+2	E+3	Correct	MSPE
OGA+HDBIC	0	43	939	18	1000	0.033
OGA+HDHQ	0	43	939	18	1000	0.033
	Range		Average		Correct	MSPE
LARS(10^{-7})	[325, 439]		408.3		116	10.52
LARS(10^{-5})	[81, 427]		331.9		34	1.39
Lasso	[163, 236]		199.1		0	0.92

Under (5.4) and some other conditions, Meinshausen and Bühlmann [19, Theorems 1 and 2] have shown that if $\lambda = \lambda_n$ in $\hat{\boldsymbol{\beta}}^L(\lambda)$ converges to 0 at a rate slower than $n^{-1/2}$, then $\lim_{n\to\infty} P(\hat{L}_n(\lambda) = N_n) = 1$, where $\hat{L}_n(\lambda) = \{i : \hat{\beta}_i^L(\lambda) \neq 0\}$.

Example 3. Consider (5.1) with $q = 10$, $n = 400$, $p = 4000$ and $(\beta_1, \cdots, \beta_q)=$ $(3, 3.75, 4.5, 5.25, 6, 6.75, 7.5, 8.25, 9, 9.75)$. Assume that $\sigma = 1$, that $x_{t1}, \cdots, x_{tq}$ are i.i.d. standard normal, and that

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

where $b = (3/4q)^{1/2}$ and $(d_{t(q+1)}, \cdots, d_{tp})^\top$ are i.i.d. multivariate normal with mean $\mathbf{0}$ and covariance matrix $(1/4)\mathbf{I}$ and are independent of x_{tj} for $1 \leq j \leq q$. Using the same notation as in the last paragraph of Example 2, straightforward calculations show that for $q + 1 \leq j \leq p$, $\mathbf{c}_{qj} = (b, \cdots, b)^\top$, $\mathbf{R}(q) = \mathbf{I}$ and $\text{sign}(\boldsymbol{\beta}(q)) = (1, \cdots, 1)^\top$. Therefore, for $q+1 \leq j \leq p$, $|\mathbf{c}_{qj}^\top \mathbf{R}^{-1}(q)\text{sign}(\boldsymbol{\beta}(q))| = (3q/4)^{1/2} = (7.5)^{1/2} > 1$, and hence (5.5) is violated. On the other hand, it is not difficult to show that (3.2) is satisfied in this example.

Table 3 gives the results of 1000 simulations on the performance of Lasso, LARS(10^{-7}), LARS(10^{-5}), OGA+HDBIC and OGA+HDHQ, using $K_n = 5(n/\log p)^{1/2}$ iterations for OGA. The table shows that LARS (10^{-7}) only includes all 10 relevant regressors in 116 of the 1000 simulations, LARS (10^{-5}) includes them in only 34 simulations and Lasso fails to include them in all simulations. This result suggests that violation of (5.5) seriously degrades Lasso's ability in variable selection. Despite its inferiority in including all relevant variables, Lasso has MSPE equal to 0.92, which is two thirds of the MSPE of LARS(10^{-5}) and is about 11 times lower than the MSPE of LARS(10^{-7}). In contrast, OGA+HDBIC and OGA+HDHQ can pick up all 10 relevant regressors and include at most three irrelevant ones. Moreover, the MSPEs of OGA+HDBIC and OGA+HDHQ are only slightly larger than $q\sigma^2/n = 0.025$.

In Example 1, p is 20 times larger than n and in Examples 2 and 3, $(n, p) =$

$(400, 4000)$. In these examples, the number of relevant regressors ranges from 5 to 10 and OGA includes nearly all of them after only $5(n/\log p)^{1/2}$ iterations. This computationally economical feature is attributable to the greedy variable selector (2.1) and the fact that, unlike PGA, a selected variable enters the model only once and is not re-selected later. Although OGA requires least squares regression of y_t on all selected regressors, instead of regressing $U_t^{(k)}$ on the most recently selected regressor as in PGA, the iterative nature of OGA enables us to avoid the computational task of solving a large system of linear normal equations for the least squares estimates, by sequently orthogonalizing the input variables so that the OGA update (2.3) is basically componentwise linear regression, similar to PGA.

The HDIC used in conjunction with OGA can be viewed as backward elimination after forward stepwise inclusion of variables by OGA. Example 1 shows that if BIC is used in lieu of HDBIC, then the backward elimination still leaves all included variables in the model. This is attributable to p being 20 times larger than 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 (4.6) of HDIC can be regarded as such adjustment, as explained in the paragraph preceding Theorem 4.2 that establishes the oracle property of OGA+HDIC under the strong sparsity assumption (C5).

It should be mentioned that since Lasso can include all relevant variables in Example 2, the overfitting problem of Lasso can be substantially rectified if one performs Zou's [28] adaptive Lasso procedure that uses Lasso to choose an initial set of variables for further refinement. On the other hand, in situations where it is difficult for Lasso to include all relevant variables as in Example 3, such two-stage modification does not help to improve Lasso.

APPENDIX A: PROOF OF (3.8), (3.13), (4.15) AND (4.19)

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 3.2. The proof of (3.13) will also make use of a similar representation. The following four lemmas give exponential bounds for moderate deviation probabilities of (i) and (ii).

LEMMA A.1. *Let $\varepsilon, \varepsilon_1, \cdots, \varepsilon_n$ 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*

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

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

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

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

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

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). $\qquad\square$

LEMMA A.2. *With the same notation and assumptions as in Theorem* 3.1 *and assuming that $\sigma_j = 1$ for all j so that $z_j = x_j$, there exists $C > 0$ such that*

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

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,

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

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

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

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

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

By (3.2), $\min_{1\leq\sharp(J)\leq K_n}\lambda_{\min}(\boldsymbol{\Gamma}) \geq c_0$ for some positive constant c_0 and all large n. Therefore, $\max_{1\leq\sharp(J)\leq K_n}\|\boldsymbol{\Gamma}^{-1}\| \leq c_0^{-1}$. Let $D = \sup_{n\geq 1} K_n\delta_n$ and note that $1-(D+1)^{-1}K_n\delta_n \geq (D+1)^{-1}$. We use (A.7) to bound $P\{\max_{1\leq\sharp(J)\leq K_n}\|\hat{\boldsymbol{\Gamma}}^{-1} - \boldsymbol{\Gamma}^{-1}\| > K_n\delta_n\}$ by

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

Since $\max_{1\leq\sharp(J)\leq K_n}\|\boldsymbol{\Gamma}^{-1}\|^2 \leq c_0^{-2}$, combining (A.8) with (A.5) (in which δ_n is replaced by $c_0^2\delta_n/\{(D+1)\}$ for the first summand in (A.8), and by $c_0\delta_n/\{(D+1)\}$ for the second) yields (A.6) with $c < Cc_0^4/(D+1)^2$. $\qquad\square$

LEMMA A.3. *With the same notation and assumptions as in Theorem* 3.1 *and assuming $\sigma_j = 1$ for all j, let $n_1 = \sqrt{n}/(\log n)^2$ and $n_{k+1} = \sqrt{n_k}$ for $k \geq 1$. 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 $\Omega_n = \{\max_{1\leq t\leq n}|\varepsilon_t| < (\log n)^2\}$. Then there exists $\alpha > 0$ such that for all large n,*

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

(A.10)
$$
\begin{aligned}
&\max_{1\leq k\leq K, 1\leq i\leq p_n} P\left(\sum_{t=1}^n |\varepsilon_t x_{ti}|I_{\{n_{k+1}\leq|x_{ti}|<n_k\}} \geq u_n(\log n)^2, \Omega_n\right)\\
&\leq \exp(-\alpha u_n).
\end{aligned}
$$

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

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

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

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

where $\pi_{n,k,i} = P(|x_i| \geq n_{k+1}) \leq \exp(-cn_{k+1}^2) = \exp(-cn_k)$ 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 $u_n/n_1 \to \infty$. $\qquad\square$

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,

(A.11) $$\max_{1 \leq i \leq p_n} P\left(|\sum_{t=1}^{n} \varepsilon_t x_{ti}| \geq n\delta_n, \Omega_n\right) \leq \exp(-\beta n\delta_n^2).$$

PROOF. Let $n_i, i \geq 1$ 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 $2 \leq k \leq K$, $A^{(K+1)} = [0, n_K)$, note that

(A.12)
$$\begin{aligned}
&P\left(|\sum_{t=1}^{n} \varepsilon_t x_{ti}| \geq n\delta_n, \Omega_n\right) \\
&\leq \sum_{k=1}^{K+1} P\left(|\sum_{t=1}^{n} \varepsilon_t x_{ti} I_{\{|x_{ti}| \in A^{(k)}\}}| \geq n\delta_n/(K+1), \Omega_n\right) \\
&\leq P(\max_{1 \leq t \leq n} |x_{ti}| \geq n_1) \\
&\quad + \sum_{k=2}^{K+1} P\left(|\sum_{t=1}^{n} \varepsilon_t x_{ti} I_{\{|x_{ti}| \in A^{(k)}\}}| \geq n\delta_n/(K+1), \Omega_n\right).
\end{aligned}$$

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

(A.13)
$$\begin{aligned}
&\max_{1 \leq i \leq p_n} P\left(|\sum_{t=1}^{n} \varepsilon_t x_{ti} I_{\{|x_{ti}| \in A^{(k)}\}}| \geq n\delta_n/(K+1), \Omega_n\right) \\
&\leq \exp\left\{-\frac{\alpha n\delta_n}{(K+1)(\log n)^2}\right\},
\end{aligned}$$

where α is some positive constant. Moreover, by (A.9), $\max_{1\leq i\leq p_n} P(\max_{1\leq t\leq n} |x_{ti}| \geq n_1) \leq \exp\{-\alpha n/(\log n)^4\}$, 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

$$
\text{(A.14)} \qquad \max_{1\leq i\leq p_n} P\left(|\sum_{t=1}^n \varepsilon_t x_{ti}I_{\{|x_{ti}|\in A^{(K+1)}\}}| \geq n\delta_n/(K+1), \Omega_n\right)
$$
$$
\leq \exp(-c_1 n\delta_n^2).
$$

Let $0 < d_1 < 1$ and $L_i = \{d_1 \leq n^{-1}\sum_{t=1}^n x_{ti}^2 I_{\{|x_{ti}|\in A^{(K+1)}\}} < d_1^{-1}\}$. By an argument similar to that used in the proof of (A.4), it can be shown that there exists $c_2 > 0$ such that for all large n, $\max_{1\leq i\leq p_n} P(L_i^c) \leq \exp(-c_2 n)$. Application of Lemma A.1 after conditioning on $\mathbf{X}_i$, which is independent of $(\varepsilon_1, \cdots, \varepsilon_n)^\top$, shows that there exists $c_3 > 0$ for which

$$
\max_{1\leq i\leq p_n} P(|\sum_{t=1}^n \varepsilon_t x_{ti}I_{\{|x_{ti}|\in A^{(K+1)}\}}| \geq n\delta_n/(K+1), L_i) \leq \exp(-c_3 n\delta_n^2),
$$

for all large n. This completes the proof of (A.14). $\qquad\square$

PROOF OF (3.8). Let $\Omega_n = \{\max_{1\leq t\leq n} |\varepsilon_t| < (\log n)^2\}$. 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 $d^2 C > 1$ and all large n,

$$
\text{(A.15)} \qquad P(\max_{1\leq i\leq p_n} |n^{-1}\sum_{t=1}^n x_{ti}^2 - 1| > d(\log p_n/n)^{1/2}))
$$
$$
\leq p_n\exp(-Cd^2\log p_n) = \exp\{\log p_n - Cd^2\log p_n\} = o(1).
$$

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 $d > 0$,

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

and

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

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 $\hat{\Gamma}_n(J)$ by $\hat{\Gamma}(J)$ for simplicity. Note that

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

(A.18)

where $x_{ti:J}^\perp = x_{ti} - \mathbf{g}_i^\top(J)\Gamma^{-1}(J)\mathbf{x}_t(J)$. Since $S_{3,n} \leq \max_{1 \leq i \leq p_n}|n^{-1}\sum_{t=1}^n \varepsilon_t x_{ti}|$ $\max_{\sharp(J) \leq K_n - 1, i \notin J} \|\Gamma^{-1}(J)\mathbf{g}_i(J)\|_1$, by (3.2) and Lemma A.4, there exists $\beta > 0$ such that for any $d > \beta^{-1/2}(M+1)$ and all large n,

(A.19)
$$
\begin{aligned}
&P(S_{1,n} + S_{3,n} > d(\log p_n / n)^{1/2}, \Omega_n) \\
&\leq p_n \exp(-\beta d^2 \log p_n / (M+1)^2) = o(1).
\end{aligned}
$$

Since $K_n = O((n/\log n)^{1/2})$, there exists $c_1 > 0$ such that for all large n, $K_n \leq c_1(n/\log p_n)^{1/2}$. As shown in the proof of Lemma A.2, $\max_{1 \leq \sharp(J) \leq K_n} \|\Gamma^{-1}(J)\| \leq c_0^{-1}$ for all large n. In view of this and (A.6), there exists $c > 0$ such that for any $\bar{M} > (2c_1^2/c)^{1/2}$ and all large n,

(A.20)
$$
\begin{aligned}
&P(\max_{1 \leq \sharp(J) \leq K_n} \|\hat{\Gamma}^{-1}(J)\| > c_0^{-1} + \bar{M}) \\
&\leq P(\max_{1 \leq \sharp(J) \leq K_n} \|\hat{\Gamma}^{-1}(J) - \Gamma^{-1}(J)\| > \bar{M}) \\
&\leq p_n^2 \exp(-cn\bar{M}^2 / K_n^2) = o(1).
\end{aligned}
$$

By observing

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

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

it follows from (3.2) that

$$S_{2,n} \leq (\max_{1 \leq \sharp(J) \leq K_n} \|\hat{\mathbf{\Gamma}}^{-1}(J)\|) K_n (1 + M)$$

(A.22)

$$\times (\max_{1 \leq i,j \leq p_n} |n^{-1} \sum_{t=1}^{n} x_{ti} x_{tj} - \sigma_{ij}|)(\max_{1 \leq i \leq p_n} |n^{-1} \sum_{t=1}^{n} \varepsilon_t x_{ti}|).$$

Define $\Omega_{1,n} = \{\max_{1 \leq \sharp(J) \leq K_n} \|\hat{\mathbf{\Gamma}}^{-1}(J)\| \leq c_2 = c_0^{-1} + \bar{M}\}$. Then by Lemma A.4, (A.20), (A.22), (A.4) and $K_n = O((n/\log p_n)^{1/2})$, there exists sufficiently large $d > 0$ such that

(A.23)

$$P(S_{2,n} > d(\log p_n/n)^{1/2}, \Omega_n) \leq P(\Omega_{1,n}^c)$$

$$+ P\left(\max_{1 \leq i \leq p_n} |n^{-1} \sum_{t=1}^{n} \varepsilon_t x_{ti}| > \frac{d^{1/2}(\log p_n/n)^{1/4}}{(K_n c_2(1+M))^{1/2}}, \Omega_n\right)$$

$$+ P\left(\max_{1 \leq i,j \leq p_n} |n^{-1} \sum_{t=1}^{n} x_{ti} x_{tj} - \sigma_{ij}| > \frac{d^{1/2}(\log p_n/n)^{1/4}}{(K_n c_2(1+M))^{1/2}}\right)$$

$$= o(1).$$

From (A.18), (A.19) and (A.23), (A.16) follows.

We prove (A.17) by using the bound

(A.24)

$$\max_{\sharp(J) \leq K_n-1, i,j \notin J} |n^{-1} \sum_{t=1}^{n} x_{tj} \hat{x}_{ti;J}^{\perp} - E(x_j x_{i;J}^{\perp})|$$

$$\leq \max_{1 \leq i,j \leq p_n} |n^{-1} \sum_{t=1}^{n} x_{tj} x_{ti} - \sigma_{ij}|$$

$$+ \max_{\sharp(J) \leq K_n-1, i,j \notin J} |\mathbf{g}_j^\top(J) \mathbf{\Gamma}^{-1}(J) n^{-1} \sum_{t=1}^{n} x_{ti;J}^{\perp} \mathbf{x}_t(J)|$$

$$+ \max_{\sharp(J) \leq K_n-1, i,j \notin J} |\mathbf{g}_i^\top(J) \mathbf{\Gamma}^{-1}(J)(n^{-1} \sum_{t=1}^{n} x_{tj} \mathbf{x}_t(J) - \mathbf{g}_j(J))|$$

$$+ \max_{\sharp(J) \leq K_n-1, i,j \notin J} \|\hat{\mathbf{\Gamma}}^{-1}(J)\| \|n^{-1} \sum_{t=1}^{n} x_{ti;J}^{\perp} \mathbf{x}_t(J)\| \|n^{-1} \sum_{t=1}^{n} x_{tj;J}^{\perp} \mathbf{x}_t(J)\|$$

$$:= S_{4,n} + S_{5,n} + S_{6,n} + S_{7,n}.$$

It follows from (3.2) that $S_{5,n} \leq \max_{1 \leq i,j \leq p_n} |n^{-1} \sum_{t=1}^{n} x_{tj} x_{ti} - \sigma_{ij}|(1+M)M$ and $S_{6,n} \leq \max_{1 \leq i,j \leq p_n} |n^{-1} \sum_{t=1}^{n} x_{tj} x_{ti} - \sigma_{ij}|M$. Combining this with (A.4) yields that for some $d > 0$,

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

In view of (A.21) and (3.2),

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

Therefore, by (A.4) and (A.20), there exists $d > 0$ such that

$$\begin{aligned}
P(S_{7,n} &> d(\log p_n/n)^{1/2}) \le P(\Omega_{1,n}^c) \\
\text{(A.26)} \quad &+ P\left(\max_{1 \le i,j \le p_n} (n^{-1} \sum_{t=1}^{n} x_{tj} x_{ti} - \sigma_{ij})^2 > \frac{d(\log p_n/n)^{1/2}}{c_2 K_n (1+M)^2} \right) \\
&= o(1).
\end{aligned}$$

Consequently, (A.17) follows from (A.24)-(A.26). □

PROOF OF (3.13). Let $\mathbf{q}(J) = E(y\mathbf{x}_J)$ and

$$\mathbf{Q}(J) = n^{-1} \sum_{t=1}^{n} (y_t - \mathbf{x}_t^\top(J)\boldsymbol{\Gamma}^{-1}(J)\mathbf{q}(J))\mathbf{x}_t(J).$$

Then, $E_n(\hat{y}_m(\mathbf{x}) - y_{\hat{J}_m}(\mathbf{x}))^2 = \mathbf{Q}^\top(\hat{J}_m)\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_m)\boldsymbol{\Gamma}(\hat{J}_m)\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_m)\mathbf{Q}(\hat{J}_m)$. We first show that

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

By observing

$$\|\mathbf{Q}(\hat{J}_m)\|^2 \le 2m \max_{1 \le i \le p_n} (n^{-1} \sum_{t=1}^{n} \varepsilon_t x_{ti})^2 + 2m \max_{1 \le i,j \le p_n} (n^{-1} \sum_{t=1}^{n} x_{ti} x_{tj} - \sigma_{ij})^2$$

$$\times (\sum_{j=1}^{p_n} |\beta_j|)^2 (1 + \max_{1 \le \sharp(J) \le K_n, 1 \le l \le p_n} \|\boldsymbol{\Gamma}^{-1}(J)\mathbf{g}_l(J)\|_1)^2,$$

(A.27) follows from (C4), (3.2), (A.4) and Lemma A.4. By (A.5) and (A.6),

$$\max_{1 \le m \le K_n} \|\hat{\boldsymbol{\Gamma}}^{-1}(\hat{J}_m)\| = O_p(1), \quad \max_{1 \le m \le K_n} \|\hat{\boldsymbol{\Gamma}}(\hat{J}_m) - \boldsymbol{\Gamma}(\hat{J}_m)\| = O_p(1).$$

This, (A.27) and the fact that

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

together yield the desired conclusion (3.13). □

PROOF OF (4.15). Denote $\lfloor an^\gamma \rfloor$ in (4.10) by m_0. By (3.2) and an argument similar to that used to prove (A.6) and in (A.20), there exists $c_1 > 0$ such that

$$(A.28) \qquad P(\max_{1 \le \sharp(J) \le m_0} \|\hat{\mathbf{\Gamma}}^{-1}(J)\| > 2v_n^{-1}) \le p_n^2 \exp(-c_1 n^{1-2\gamma}) = o(1).$$

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

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

Since (4.9) implies that $n^{1-2\gamma}/(w_n \log p_n) \to \infty$, it follows from Lemma A.4, (4.9), (3.2) and (A.28) that for all large n,

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

As noted above in the proof of (3.8), $P(\Omega_n) = 1 + o(1)$. Moreover, by (4.9) and (A.5) with K_n replaced by m_0, there exists $c_2 > 0$ such that

$$(A.31)$$
$$P(\hat{A}_n < v_n/2, \mathcal{D}_n) \le P(\lambda_{\min}(\hat{\mathbf{\Gamma}}(\hat{J}_{\tilde{k}_n})) < v_n/2, \mathcal{D}_n)$$
$$\le P(\lambda_{\min}(\hat{\mathbf{\Gamma}}(\hat{J}_{m_0})) < v_n/2)$$
$$\le P(\max_{1 \le \sharp(J) \le m_0} \|\hat{\mathbf{\Gamma}}(J) - \mathbf{\Gamma}(J)\| > c_0/2)$$
$$\le p_n^2 \exp(-c_2 n^{1-2\gamma}) = o(1),$$

where c_0 is defined in the line following (4.14). In view of (A.29)-(A.31), the desired conclusion follows. □

PROOF OF (4.19). Letting $\bar{M}$ be defined as in (A.20) and $\pi_n = w_n(\log p_n)/(1 + w_n(\log p_n)K_n n^{-1})$, note that

$$(A.32)$$
$$P\{(\hat{a}_n + \hat{b}_n)(1 + n^{-1}K_n w_n \log p_n) \ge \theta w_n \log p_n\}$$
$$\le P(\Omega_n^c) + P(\|\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_{K_n})\| \ge c_0^{-1} + \bar{M})$$
$$+ P(\max_{1 \le j \le p_n} (n^{-1/2} \sum_{t=1}^n x_{tj}\varepsilon_t)^2 \ge \pi_n \theta/\{2(c_0^{-1} + \bar{M})\}, \Omega_n)$$
$$+ P(2n(S_{2,n} + S_{3,n})^2 \ge \pi_n \theta/\{2(c_0^{-1} + \bar{M})\}, \Omega_n),$$

where $S_{2,n}$ and $S_{3,n}$ are defined in (A.18). Since $K_n = O((n/\log p_n)^{1/2})$, $(n/\log p_n)^{1/2} \to \infty$ and $w_n \to \infty$, there exists $l_n \to \infty$ such that

$$(A.33) \qquad \pi_n \geq 2^{-1}\min\{w_n \log p_n, nK_n^{-1}\} \geq l_n \log p_n.$$

By (A.33) and Lemma A.4, we bound $P(\max_{1 \leq j \leq p_n}(n^{-1/2}\sum_{t=1}^n x_{tj}\varepsilon_t)^2 \geq \pi_n\theta/\{2(c_0^{-1} + \bar{M})\}, \Omega_n)$ by

$$p_n \max_{1 \leq j \leq p_n} P(|n^{-1/2}\sum_{t=1}^n x_{tj}\varepsilon_t| \geq (\pi_n\theta/\{2(c_0^{-1} + \bar{M})\})^{1/2}, \Omega_n)$$

$$\leq p_n\exp(-c_1 l_n \log p_n) = o(1),$$

for some $c_1 > 0$. By (A.33) and an argument similar to that used in (A.19) and (A.23), it follows that

$$P(2n(S_{2,n} + S_{3,n})^2 \geq \pi_n\theta/\{2(c_0^{-1} + \bar{M})\}, \Omega_n) = o(1).$$

In view of (A.20), one also has

$$P(\|\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_{K_n})\| \geq c_0^{-1} + \bar{M}) = o(1).$$

Adding these bounds for the summands in (A.32), we obtain

$$(A.34) \quad P\{(\hat{a}_n + \hat{b}_n)(1 + n^{-1}K_n w_n \log p_n) \geq \theta w_n \log p_n\} = o(1).$$

Form (A.30), (4.10) and $P(\Omega_n) = 1 + o(1)$, it is straightforward to see that $P(|\hat{C}_n| \geq \theta) = o(1)$. Combining this with (A.34) yields (4.19). $\qquad\square$

APPENDIX B: PROOF OF THEOREM 3.2

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 $\boldsymbol{\mu} = (y(\mathbf{x}_1), \cdots, y(\mathbf{x}_n))^\top$. Let $\mathbf{H}_J$ denotes the projection matrix associated with the projection into the linear space spanned by $\mathbf{X}_j, j \in J \subseteq \{1, \cdots, p\}$. 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

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

When the procedure stops after m iterations, the noiseless OGA determines an index set $\tilde{J}_m = \{\tilde{j}_1, \cdots, \tilde{j}_m\}$ 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 $|(\mathbf{U}^{(i-1)})^\top\mathbf{X}_l|/\|\mathbf{X}_l\| \geq \xi\max_{1 \leq j \leq p} |(\mathbf{U}^{(i-1)})^\top\mathbf{X}_j|/\|\mathbf{X}_j\|$. We first prove the following inequality for such generalization of the noiseless OGA.

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

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

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

(B.1)
$$\|(\mathbf{I} - \mathbf{H}_{\tilde{J}_{m,\xi}})\boldsymbol{\mu}\|^2$$
$$\le \|(\mathbf{I} - \mathbf{H}_{\tilde{J}_{m-1,\xi}})\boldsymbol{\mu} - \frac{(\boldsymbol{\mu})^\top (\mathbf{I} - \mathbf{H}_{\tilde{J}_{m-1,\xi}})\mathbf{X}_{\tilde{j}_{m,\xi}}}{\|\mathbf{X}_{\tilde{j}_{m,\xi}}\|^2}\mathbf{X}_{\tilde{j}_{m,\xi}}\|^2$$
$$\le \|(\mathbf{I} - \mathbf{H}_{\tilde{J}_{m-1,\xi}})\boldsymbol{\mu}\|^2 - n\nu_{\tilde{J}_{m-1,\xi},\tilde{j}_{m,\xi}}^2$$
$$\le \|(\mathbf{I} - \mathbf{H}_{\tilde{J}_{m-1,\xi}})\boldsymbol{\mu}\|^2 - n\xi^2 \max_{1 \le j \le p} \nu_{\tilde{J}_{m-1,\xi},j}^2,$$

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

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

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

(B.3)
$$\|(\mathbf{I} - \mathbf{H}_{\tilde{J}_{m,\xi}})\boldsymbol{\mu}\|^2 \le \|(\mathbf{I} - \mathbf{H}_{\tilde{J}_{m-1,\xi}})\boldsymbol{\mu}\|^2\{1 - (\xi^2\|(\mathbf{I} - \mathbf{H}_{\tilde{J}_{m-1,\xi}})\boldsymbol{\mu}\|^2/Q^2)\}.$$

By Minkowski's inequality, $\|(\mathbf{I} - \mathbf{H}_{\tilde{J}_{0,\xi}})\boldsymbol{\mu}\|^2 = \|\boldsymbol{\mu}\|^2 \le Q^2$. From this, (B.3) and Lemma 3.1 of [21], the stated inequality follows. □

PROOF OF THEOREM 3.2. 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 $\nu_{J,i}$ 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 3.1. By an argument similar to that used to prove (3.10), we have for all $1 \leq q \leq m$,

$$(\text{B.4}) \qquad |\nu_{\hat{J}_{q-1},\hat{j}_q}| \geq \xi \max_{1 \leq j \leq p} |\nu_{\hat{J}_{q-1},j}| \text{ on } A \bigcap B,$$

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

$$(\text{B.5}) \qquad \|(\mathbf{I} - \mathbf{H}_{\hat{J}_m})\boldsymbol{\mu}\|^2 I_{A \bigcap B} \leq n (\inf_{\mathbf{b} \in \mathbf{B}} \|\mathbf{b}\|_1)^2 (1 + m\xi^2)^{-1}.$$

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

$$
\begin{aligned}
(\text{B.6}) \qquad \|(\mathbf{I} - \mathbf{H}_{\hat{J}_m})\boldsymbol{\mu}\|^2 &\leq \min_{0 \leq i \leq m-1} \sum_{j=1}^{p} b_j \mathbf{X}_j^{\top} (\mathbf{I} - \mathbf{H}_{\hat{J}_i})\boldsymbol{\mu} \\
&\leq (\min_{0 \leq i \leq m-1} \max_{1 \leq j \leq p} |\nu_{\hat{J}_i,j}|) n \|\mathbf{b}\|_1 \leq \tilde{\xi} C \sigma (n \log p)^{1/2} \|\mathbf{b}\|_1 \text{ on } B^c.
\end{aligned}
$$

Since A decreases as m increases, it follows from (3.18), (B.5) and (B.6) that

$$(\text{B.7}) \qquad n^{-1} \|(\mathbf{I} - \mathbf{H}_{\hat{J}_m})\boldsymbol{\mu}\|^2 I_{\mathcal{A}} \leq \omega_{m,n} \text{ for all } 1 \leq m \leq \lfloor n/\log p \rfloor,$$

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

$$(\text{B.8}) \qquad P(\mathcal{A}^c) \leq a^* := p \exp\{-2^{-1}C^2(\log p)/(1 + M)^2\},$$

$$(\text{B.9}) \qquad P(\mathcal{E}^c) \leq b^* := \frac{\tilde{r}_p^{1/2} p^{-(sr_p - 1)}}{1 - \tilde{r}_p^{1/2} p^{-(sr_p - 1)}},$$

where $\mathcal{E} = \{\boldsymbol{\varepsilon}^{\top} \mathbf{H}_{\hat{J}_m} \boldsymbol{\varepsilon} \leq s\sigma^2 m \log p \text{ for all } 1 \leq m \leq \lfloor n/\log p \rfloor\}$ and r_p and $\tilde{r}_p$ are defined in (3.16). By (B.7)-(B.9) and observing 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 $1 - a^* - b^*$. Hence the desired conclusion follows. $\qquad \square$

PROOF OF (B.8). 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 $1 \leq j \leq p$, it follows that for any $J \subseteq \{1, \cdots, p\}$, $1 \leq i \leq p$ and $i \notin J$,

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

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

$$(\text{B.10}) \qquad \max_{\sharp(J) \leq \lfloor n/\log p \rfloor - 1, i \notin J} |\hat{\mu}_{J,i} - \nu_{J,i}| \leq \max_{1 \leq i \leq p} |n^{-1} \sum_{t=1}^{n} x_{ti}\varepsilon_t| (1 + M).$$

By (B.10) and the Gaussian assumption on ε_t,

$$P(\mathcal{A}^c) \leq P\{\max_{1 \leq i \leq p} |n^{-1/2} \sum_{t=1}^{n} x_{ti}\varepsilon_t/\sigma| > C(\log p)^{1/2}(1+M)^{-1}\}$$

$$\leq p \exp(-\{C^2 \log p/[2(1+M)^2]\}).$$

$\square$

PROOF OF (B.9). Clearly $P(\mathcal{E}^c) \leq \sum_{m=1}^{\lfloor n/\log p \rfloor} p^m \max_{\sharp(J)=m} P(\varepsilon^\top \mathbf{H}_J \varepsilon > s\sigma^2 m \log p)$. Moreover, we can make use of the χ^2-distribution to obtain the bound

(B.11) $$\max_{\sharp(J)=m} P(\varepsilon^\top \mathbf{H}_J \varepsilon > s\sigma^2 m \log p) \leq \exp(-rsm \log p)(1-2r)^{-m/2}$$

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.11), we can use (B.11) to bound $P(\mathcal{E}^c)$ by $\sum_{m=1}^{\lfloor n/\log p \rfloor} g^m \leq g/(1-g)$, where $g = \tilde{r}_p^{1/2} p^{-(sr_p-1)} < 1$, yielding (B.9). $\square$

REFERENCES

[1] BICKEL, P., RITOV, Y. and TSYBAKOV, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, to appear.

[2] BÜHLMANN, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist.* **34** 559-583.

[3] BÜHLMANN, P. and YU, B. (2003). Boosting with the L_2 loss: regression and classification. *J. Amer. Statist. Assoc.* **98** 324-339.

[4] BUNEA F., TSYBAKOV, A.B. and WEGKAMP, M.H. (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics* **1** 169-194.

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

[6] CANDES, E.J. and PLAN, Y. (2009). Near-ideal model selection by l_1 minimization. *Ann. Statist.* **37** 2145-2177.

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

[8] CHEN, R.-B., ING, C.-K., LAI, T.L. and ZHANG, F. (2009). Cross-validation in high-dimensional linear regression models. Tech. Report, Dept. Statistics, Stanford Univ.

[9] DONOHO, D.L., ELAD, M. and TEMLYAKOV, V.N. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Info. Theory* **52** 6-18.

[10] DONOHO, D.L. and JOHNSTONE, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425-455.

[11] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32** 407-499.

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

[13] FOSTER, D.P. and GEORGE, E.I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947-1975.

[14] FRIEDMAN, J. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Statist.* **29** 1189-1232.

[15] FRIEDMAN, J. , HASTIE, T. and TIBSHIRANI, R (2009). Regularized Paths for Generalized Linear Models via Coordinate Descent. *Technical Report*.

[16] 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.

[17] MALLAT, S. and ZHANG, Z. (1993). Matching pursuit with time-frequency dictionaries. *IEEE Trans. Signal Process* **41** 3397V3415.

[18] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436-1462.

[19] ROCHA, G.V. and ZHAO, P. (2006). An implementation of the LARS algorithm for getting the Lasso path in MATLAB. http://www.stat.berkeley.edu/~gvrocha.

[20] SHAO, J. and CHOW, S.-C. (2007). Variable screening in predicting clinical outcome with high-dimensional microarrays. *J. Multivariate Anal.* **98** 1529-1538.

[21] TEMLYAKOV, V.N. (2000). Weak greedy algorithms. *Adv. Comput. Math.* **12** 213-227.

[22] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267-288.

[23] TROPP, J.A. (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Info. Theory* **50** 2231-2242.

[24] TROPP, J.A. and GILBERT, A.C. (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Info. Theory* **53** 4655-4666.

[25] 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.

[26] ZHANG, T. (2009). Some sharp performance bounds for least squares regression with L_1 regularization. *Ann. Statist.*, to appear.

[27] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Machine Learning Research* **7** 2541-2563.

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

CHING-KANG ING
INSTITUTE OF STATISTICAL SCIENCE
ACADEMIA SINICA
TAIPEI 115, TAIWAN, ROC
E-MAIL: cking@stat.sinica.edu.tw

TZE LEUNG LAI
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD CA94305-4065, USA
E-MAIL: lait@stanford.edu