Orthogonal Matching Pursuit for Sparse Signal Recovery With Noise

T. Tony Cai and Lie Wang

***Abstract*—We consider the orthogonal matching pursuit (OMP) algorithm for the recovery of a high-dimensional sparse signal based on a small number of noisy linear measurements. OMP is an iterative greedy algorithm that selects at each step the column, which is most correlated with the current residuals. In this paper, we present a fully data driven OMP algorithm with explicit stopping rules. It is shown that under conditions on the mutual incoherence and the minimum magnitude of the nonzero components of the signal, the support of the signal can be recovered exactly by the OMP algorithm with high probability. In addition, we also consider the problem of identifying significant components in the case where some of the nonzero components are possibly small. It is shown that in this case the OMP algorithm will still select all the significant components before possibly selecting incorrect ones. Moreover, with modified stopping rules, the OMP algorithm can ensure that no zero components are selected.**

***Index Terms*—ℓ_1 minimization, compressed sensing, mutual incoherence, orthogonal matching pursuit (OMP), signal reconstruction, support recovery.**

I. Introduction

RECOVERY of a high-dimensional sparse signal based on a small number of linear measurements, possibly corrupted by noise, is a fundamental problem in signal processing. Specifically, one considers the following model:

$$y = X\beta + \epsilon \tag{1}$$

where the observation $y \in \mathbb{R}^n$, the matrix $X \in \mathbb{R}^{n\times p}$ and the measurement errors $\epsilon \in \mathbb{R}^n$. Suppose $X = (X_1, X_2, \ldots, X_p)$ where X_i denotes the ith column of X. Throughout the paper we shall assume that the columns of X are normalized, i.e., $\|X_i\|_2 = 1$ for $i = 1, 2, \ldots, p$. The goal is to reconstruct the unknown vector $\beta \in \mathbb{R}^p$ based on y and X. A setting that is of significant interest and challenge is when the dimension p of the signal is much larger than the number of measurements n. This and other related problems have received much recent attention in a number of fields including applied mathematics, electrical engineering and statistics.

Manuscript received March 17, 2010; revised February 04, 2011; accepted February 09, 2011. Date of current version June 22, 2011. T. T. Cai was supported in part by NSF FRG Grant DMS-0854973. L. Wang was supported by NSF Grant DMS-1005539.

T. T. Cai is with the Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: tcai@wharton.upenn.edu).

L. Wang is with the Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: liewang@math.mit.edu).

Communicated by J. Romberg, Associate Editor for Signal Processing.

Digital Object Identifier 10.1109/TIT.2011.2146090

For a vector $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$, the support of β is defined to be the set $\mathrm{supp}(\beta) = \{i : \beta_i \neq 0\}$ and β is said to be k-*sparse* if $|\mathrm{supp}(\beta)| \leq k$. A widely used framework for sparse signal recovery is the *Mutual Incoherence Property* (MIP) introduced in Donoho and Huo (2001). The *mutual incoherence* is defined by

$$\mu = \max_{i\neq j} |\langle X_i, X_j\rangle|. \tag{2}$$

The MIP requires the mutual incoherence μ to be small. Other conditions used in the compressed sensing literature include the Restricted Isometry Property (RIP) and Exact Recovery Condition (ERC). See, for example, Candes and Tao (2005) and Tropp (2004). In contrast to the MIP, these conditions are not computationally feasible to verify for a given matrix X. On the other hand, the MIP condition is stronger than both RIP and ERC: The MIP implies RIP and ERC but the converse is not true. However, it should be emphasized here that although we focus our attention under the MIP because the condition is more intuitive, all the results given in this paper hold under the ERC. See Section IV for further discussions.

In the present paper we consider the orthogonal matching pursuit (OMP) algorithm for the recovery of the support of the k-sparse signal β under the model (1). OMP is an iterative greedy algorithm that selects at each step the column of X which is most correlated with the current residuals. This column is then added into the set of selected columns. The algorithm updates the residuals by projecting the observation y onto the linear subspace spanned by the columns that have already been selected and the algorithm then iterates. Compared with other alternative methods, a major advantage of the OMP is its simplicity and fast implementation. This method has been used for signal recovery and approximation, for example, in Davis, Mallat, and Avellaneda (1997), Tropp (2004, 2006) and Barron *et al.* (2008). In particular, support recovery has been considered in the noiseless case by Tropp (2004), where it was shown that $\mu < \frac{1}{2k-1}$ is a sufficient condition for recovering a k-sparse β exactly in the noiseless case. Results in Cai, Wang and Xu (2010a) imply that this condition is in fact sharp.

In this paper we consider the OMP algorithm in the general setting where noise is present. Note that the residuals after each step in the OMP algorithm are orthogonal to all the selected columns of X, so no column is selected twice and the set of selected columns grows at each step. One of the key components of an iterative procedure like OMP is the stopping rule. Specific stopping rules are given for the OMP algorithm in both bounded noise and Gaussian noise cases. The algorithm is then fully data-driven. Our results show that under the MIP condition $\mu < \frac{1}{2k-1}$ and a condition on the minimum magnitude of

the nonzero coordinates of β, the support of β can be recovered exactly by the OMP algorithm in the bounded noise cases and with high probability in the Gaussian case. In fact, it can be seen from our discussion in Section III that a more general condition than $\mu < \frac{1}{2k-1}$ can guarantee the recovery of the support with high probability. In particular, all the main results hold under the Exact Recovery Condition (ERC).

In many applications, the focus is often on identifying significant components, i.e., coordinates of β with large magnitude, instead of the often too ambitious goal of recovering the whole support of β exactly. In this paper, we also consider the problem of identifying large coordinates of β in the case where some of the nonzero coordinates are possibly small. It is shown that in this case the OMP algorithm will still select all the most important components before possibly selecting incorrect ones. In addition, with modified stopping rules, the OMP algorithm can ensure that no zero components are selected.

Besides OMP, several other methods for sparse signal recovery have been proposed and extensively studied in the literature. In particular, it is now well understood that ℓ_1 minimization methods provide effective ways for reconstructing a sparse signal. For example, the ℓ_1 penalized least squares (Lasso) estimator has been studied in Tibshirani (1996), Chen, Donoho, and Saunders (1998) and Efron *et al.* (2004). Zhao and Yu (2006) considered support recovery using the Lasso. In addition, two specific constrained ℓ_1 minimization methods have been well studied. Donoho, Elad and Temlyakov (2006) considered constrained ℓ_1 minimization under an ℓ_2 constraint. Candes and Tao (2007) introduced the Dantzig Selector, which is a constrained ℓ_1 minimization method under an ℓ_∞ constraint. A particularly simple and elementary analysis of constrained ℓ_1 minimization methods is given in Cai, Wang, and Xu (2010b). Bickel, Ritov, and Tsybakov (2009) gives a unified treatment of the Lasso and Dantzig Selector.

Compared with the known results on the model selection consistency of the Lasso in the Gaussian noise case given in Zhao and Yu (2006), the condition on the minimum magnitude of the nonzero coordinates of β is much weaker for OMP than that for the Lasso. More detailed discussion can be found in Section III. This together with the computational simplicity make OMP a very appealing method for support recovery.

The rest of the paper is organized as follows. We will begin in Section II with a detailed description of the OMP algorithm. The stopping rules and the properties of the algorithm are considered in Section III for both bounded noise cases and Gaussian noise case. The theoretical results are first formulated under the MIP. Section IV discusses the corresponding results under the ERC and compares our results with some of the existing ones in the literature. Section V provides some technical analysis of the OMP algorithm which sheds light on how and when the OMP algorithm works properly. The proofs of the main results are contained in Section VI.

II. The OMP Algorithm

In this section we give a detailed description of the orthogonal matching pursuit (OMP) algorithm. We assume that the columns of X are normalized so that $\|X_i\|_2 = 1$ for $i = 1, 2, \ldots, p$. For any subset $S \subseteq \{1, 2, \ldots, p\}$, denote by $X(S)$ a submatrix of X consisting of the columns X_i with $i \in S$. In this paper we shall also call columns of X *variables* by following the convention in statistics. Thus we use X_i to denote the both ith column of X and the ith variable of the model. Following the same convention, we shall call X_i a correct variable if the corresponding $\beta_i \neq 0$ and call X_i an incorrect variable otherwise. With slight abuse of notation, we shall use $X(S)$ to denote both the subset of columns of X with indices in S and the corresponding submatrix of X.

The OMP algorithm can be stated as follows.

- Step 1: Initialize the residual $r_0 = y$ and initialize the set of selected variables $X(c_0) = \emptyset$. Let the iteration counter $i = 1$.
- Step 2: Find the variable X_{t_i} that solves the maximization problem

$$\max_t |X_t' r_{i-1}|$$

  and add the variable X_{t_i} to the set of selected variables. Update $c_i = c_{i-1} \cup \{t_i\}$.
- Step 3: Let $P_i = X(c_i)(X(c_i)'X(c_i))^{-1}X(c_i)'$ denote the projection onto the linear space spanned by the elements of $X(c_i)$. Update $r_i = (I - P_i)y$.
- Step 4: If the stopping condition is achieved, stop the algorithm. Otherwise, set $i = i + 1$ and return to Step 2.

The OMP is a stepwise forward selection algorithm and is easy to implement. A key component of OMP is the stopping rule which depends on the noise structure. In the noiseless case the natural stopping rule is $r_i = 0$. That is, the algorithm stops whenever $r_i = 0$ is achieved. In this paper, we shall consider several different noise structures. To be more specific, two types of bounded noise are considered. One is ℓ_2 bounded noise, i.e., $\|\epsilon\|_2 \leq b_2$ for some constant $b_2 > 0$. Another is ℓ_∞ bounded noise where $\|X'\epsilon\|_\infty \leq b_\infty$ for some constant $b_\infty > 0$. In addition, we shall also consider the important case of Gaussian noise where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. The stopping rule for each case and the properties of the resulting procedure will be discussed in Section III.

III. The OMP Algorithm: Stopping Rules and Properties

In this section we discuss the stopping rules and investigate the properties of the OMP algorithm for the bounded noise cases as well as the Gaussian noise case. Results for the noiseless case can be found in Tropp (2004).

We begin with the basic notation and definitions. The mutual incoherence of X, defined in (2), is the maximum magnitude of the pairwise correlation between the columns of X. Let $T = \{i : \beta_i \neq 0\}$ be the support of β and let $X(T)$ be the set of columns of X corresponding to the support T. Define

$$M = \max_{x \in X \setminus X(T)} \{\|(X(T)'X(T))^{-1}X(T)'x\|_1\}. \quad (3)$$

The condition

$$M < 1$$

is called the Exact Recovery Condition (ERC) in Tropp (2004). It was shown in Tropp (2004) that the ERC is a sufficient condition for the exact recovery of the support of the signal β in

the noiseless case. It is easy to see that the value of M is not computable as it depends on the unknown support T. However, it can be easily bounded in terms of the mutual incoherence μ.

Lemma 1: If $\mu < \frac{1}{2k-1}$, then $M \le \frac{k\mu}{1-(k-1)\mu} < 1$.

This lemma is a special case of Theorem 3.5 in Tropp (2004). The extreme eigenvalues of $X(T)'X(T)$ are also useful. Denote the minimum and maximum eigenvalues of $X(T)'X(T)$ by $\lambda_{\min}$ and $\lambda_{\max}$ respectively. The minimum eigenvalue $\lambda_{\min}$ is a key quantity to the sparse signal recovery problem. It has been used in, for example, Zhao and Yu (2006) and Cai, Wang, and Xu (2010b). Note that $\lambda_{\min}$ is usually assumed to be bounded away from zero. In particular, the ERC $M < 1$ requires $\lambda_{\min} > 0$. The following lemma shows that $\lambda_{\min}$ and $\lambda_{\max}$ can also be bounded in terms of μ. A similar, but slightly weaker, result was given in Needell and Tropp (2008).

Lemma 2: Suppose $\mu < \frac{1}{k-1}$, then $1-(k-1)\mu \le \lambda_{\min} \le \lambda_{\max} \le 1+(k-1)\mu$, where k denotes the cardinality of T.

It is easy to see that, in order for any variable selection procedure to work properly, both the degree of collinearity among the columns of X and the signal-to-noise ratio need to be properly controlled. Generally speaking, to recover accurately the support of the unknown signal, the degree of linear dependency among the X_i's needs to be small, otherwise the effects of the variables cannot be well separated. On the other hand, the signal-to-noise ratio needs to be sufficiently high in order for the significant variables to be selected. In the case of OMP, the performance of the algorithm depends on the probability of selecting a correct variable at each step. This probability is affected by the degree of collinearity among the variables and the noise structure.

We shall begin with the bounded noise cases and then consider the Gaussian case. As mentioned in Section II, two types of bound noise are considered: $\|\epsilon\|_2 \le b_2$ and $\|X'\epsilon\|_\infty \le b_\infty$. Once the bounded noise cases are understood, the Gaussian case follows easily. In what follows, our analysis of the OMP algorithm will be carried out in terms of the mutual incoherence μ. However, all the main results also hold under the ERC with essentially the same proofs. We shall focus on the MIP in the next section and discuss the results under the ERC $M < 1$ in Section IV.

A. ℓ_2 Bounded Noise

We first consider the case where the error vector ϵ is bounded in ℓ_2 norm with $\|\epsilon\|_2 \le b_2$. In this case we set the stopping rule as $\|r_i\|_2 \le b_2$. It is intuitively easy to see that this rule is reasonable because in the special case of $\beta \equiv 0$ the stopping rule will guarantee that OMP does not select any incorrect variables. We have the following result for OMP with this stopping rule.

Theorem 1: Suppose $\|\epsilon\|_2 \le b_2$ and $\mu < \frac{1}{2k-1}$. Then the OMP algorithm with the stopping rule $\|r_i\|_2 \le b_2$ recovers exactly the true subset of correct variables $X(T)$ if all the nonzero coefficients β_i satisfy $|\beta_i| \ge \frac{2b_2}{1-(2k-1)\mu}$.

Theorem 1 and other main results given in this paper can also be stated under the ERC $M < 1$. We formally restate Theorem 1 under the ERC below and only make brief remarks for the other results later. See Section IV for more discussions.

Proposition 1: Suppose $\|\epsilon\|_2 \le b_2$ and $M < 1$. Then the OMP algorithm with the stopping rule $\|r_i\|_2 \le b_2$ recovers exactly the true subset of correct variables $X(T)$ if all the nonzero coefficients β_i satisfy $|\beta_i| \ge \frac{2b_2}{(1-M)\lambda_{\min}}$.

This follows from essentially the same argument as the proof of Theorem 1 given in Section VI.

It is worth noting that after the OMP algorithm returns the true subset $X(T)$, the signal β can be easily estimated, for example, by using the ordinary least squares regression on the subset of variables $X(T)$.

Theorem 1 has two conditions, $\mu < \frac{1}{2k-1}$ and $|\beta_i| \ge \frac{2b_2}{1-(2k-1)\mu}$, which together ensure the OMP algorithm to recover exactly the true support of the signal. The condition $\mu < \frac{1}{2k-1}$ was shown to be sharp in the noisy case in Cai, Wang and Xu (2010a). The other condition $|\beta_i| \ge \frac{2b_2}{1-(2k-1)\mu}$ for all nonzero coefficient β_i is to ensure that all significant variables are selected.

In many applications, the focus is often on identifying coordinates of β with large magnitude or equivalently variables with significant effects, instead of the often too ambitious goal of recovering the whole support of β exactly. So a practically interesting question is: Can OMP identify coordinates with large magnitude when some of the nonzero coordinates β_i are small? The following result shows that the OMP algorithm with the same stopping rule will still select all the most important variables before it possibly also selects incorrect ones.

Theorem 2: Suppose $\|\epsilon\|_2 \le b_2$ and $\mu < \frac{1}{2k-1}$. Let

$$S = \left\{ X_i : 1 \le i \le p, |\beta_i| \ge \frac{2\sqrt{k}b_2}{1-(2k-1)\mu} \right\}.$$

Then the OMP algorithm with the stopping rule $\|r_i\|_2 \le b_2$ selects a correct variable at each step until all the variables in S are selected.

Remark 1: Similar to Theorem 1, Theorem 2 can also be stated under the ERC with the condition $\mu < \frac{1}{2k-1}$ replaced by $M < 1$ and the condition on the minimum magnitude of β_i in the set S replaced by $|\beta_i| > \frac{2\sqrt{k}b_2}{(1-M)\lambda_{\min}}$. See Section IV for more discussions.

In many applications, it is often desirable to select a subset of the support of the signal β without incorrectly selecting any coordinates outside of the support. The OMP algorithm with the stopping rule $\|r_i\|_2 \le b_2$ does not rule out the possibility of incorrectly selecting a zero coordinate after all the significant ones are selected. The OMP with a modified stopping rule can ensure that no zero coordinates are selected. We have the following result.

Theorem 3: Suppose $\|\epsilon\|_2 \le b_2$ and $\mu < \frac{1}{2k-1}$. Let us see the first equation at the bottom of the next page. Then OMP with the stopping rule $\|r_i\|_2 \le (1 + \frac{(1+(k-1)\mu)2\sqrt{k}}{1-(2k-1)\mu})b_2$ selects a subset $\hat{T}$ such that $S \subset \hat{T} \subset T$.

Hence, all the significant variables in S are selected by the algorithm and all the selected coordinates are in the support of β.

B. ℓ_∞ Bounded Noise

We now turn to the case where the noise ϵ is assumed to satisfy $\|X'\epsilon\|_\infty \le b_\infty$. The stopping rule in this case is $\|X'r_i\|_\infty \le b_\infty$. Similar to the previous case this is a natural stopping rule which ensures that no incorrect variables are selected in the special case of $\beta \equiv 0$. We have the following result for OMP with this stopping rule.

Theorem 4: Suppose $\|X'\epsilon\|_\infty \le b_\infty$ and $\mu < \frac{1}{2k-1}$. Moreover, assume all the nonzero coefficients β_i satisfy

$$|\beta_i| \ge \frac{2b_\infty}{1-(2k-1)\mu}\left(1+\frac{\sqrt{k}}{\sqrt{1-(k-1)\mu}}\right).$$

Then OMP with the stopping rule $\|X'r_i\|_\infty \le b_\infty$ will return the true subset $X(T)$.

Remark 2: Note that $\mu < \frac{1}{2k-1}$ implies $(k-1)\mu < \frac{1}{2}$. So a special case of the previous theorem is that when

$$|\beta_i| \ge \frac{2(1+\sqrt{2k})b_\infty}{1-(2k-1)\mu},$$

the OMP algorithm selects the true subset of significant variables $X(T)$.

As in the ℓ_2 bounded noise case, when some of the nonzero coordinates are small, OMP can also identify all the large components in this case. To be more precise, we have the following result.

Theorem 5: Suppose $\|X'\epsilon\|_\infty \le b_\infty$ and $\mu < \frac{1}{2k-1}$. Let us see the second equation at the bottom of the page. Then the OMP algorithm selects a correct variable at each step until all the variables in S are selected.

In addition, with a modified stopping rule, OMP can also ensure that no incorrect variables are selected in this case.

Theorem 6: Suppose $\|X'\epsilon\|_\infty \le b_\infty$ and $\mu < \frac{1}{2k-1}$. Let us see the third equation at the bottom of the page. Then OMP with the stopping rule

$$\|X'r_i\|_\infty \le \left(1+\frac{2\sqrt{k}(1+(k-1)\mu)}{1-(2k-1)\mu}\right)C_{k,\mu}b_\infty,$$

selects a subset $\hat{T}$ such that $S \subset \hat{T} \subset T$, where $C_{k,\mu} = 1 + \frac{\sqrt{k}}{\sqrt{1-(k-1)\mu}}$.

Remark 3: It will be shown that in fact a stronger result holds. Theorem 6 is true with the set S enlarged to (see the fourth equation at the bottom of the page), where $C_{k,\mu} = 1 + \frac{\sqrt{k}}{\sqrt{1-(k-1)\mu}}$.

C. Gaussian Noise

The Gaussian noise case is of particular interest in statistics. The results on the bounded noise cases given earlier are directly applicable to the case where noise is Gaussian. This is due to the fact that Gaussian noise is "essentially bounded."

Suppose now the noise vector ϵ follows a Gaussian distribution, $\epsilon \sim N(0, \sigma^2 I_n)$. Define two bounded sets

$$B_2 = \left\{\epsilon : \|\epsilon\|_2 \le \sigma\sqrt{n+2\sqrt{n\log n}}\right\} \quad \text{and}$$
$$B_\infty(\eta) = \left\{\epsilon :\ \|X^T\epsilon\|_\infty \le \sigma\sqrt{2(1+\eta)\log p}\right\}$$

where $\eta \ge 0$. The following result, which follows from standard probability calculations, shows that the Gaussian noise z

---

$$S = \left\{X_i : 1 \le i \le p, |\beta_i| \ge \left(\frac{(1+(k-1)\mu)2\sqrt{k}}{(1-(k-1)\mu)(1-(2k-1)\mu)} + \frac{2}{1-(k-1)\mu}\right)b_2\right\}.$$

$$S = \left\{X_i : 1 \le i \le p, |\beta_i| \ge \frac{2\sqrt{k}b_\infty}{1-(2k-1)\mu}\left(1+\frac{\sqrt{k}}{\sqrt{1-(k-1)\mu}}\right)\right\}.$$

$$S = \left\{X_i : 1 \le i \le p, |\beta_i| \ge \left(\frac{6k}{1-(2k-1)\mu} + 4\sqrt{k}\right)(1+\sqrt{2k})b_\infty\right\}.$$

$$S = \left\{X_i : 1 \le i \le p, |\beta_i| \ge \left(\frac{2k(1+(k-1)\mu)}{(1-(k-1)\mu)(1-(2k-1)\mu)} + \frac{2\sqrt{k}}{1-(k-1)\mu}\right)C_{k,\mu}b_\infty\right\}$$

is essentially bounded. The readers are referred to Cai, Xu, and Zhang (2009) for a proof.

Lemma 3: The Gaussian error $\epsilon \sim N(0, \sigma^2 I_n)$ satisfies

$$P(\epsilon \in B_2) \geq 1 - \frac{1}{n} \quad \text{and}$$
$$P(\epsilon \in B_\infty(\eta)) \geq 1 - \frac{1}{2p^\eta \sqrt{\pi \log p}}. \tag{4}$$

The following result is a direct consequence of the results for the ℓ_2 bounded noise case.

Theorem 7: Suppose $\epsilon \sim N(0, \sigma^2 I_n)$, $\mu < \frac{1}{2k-1}$ and all the nonzero coefficients β_i satisfy

$$|\beta_i| \geq \frac{2\sigma\sqrt{n + 2\sqrt{n \log n}}}{1 - (2k-1)\mu}. \tag{5}$$

Then OMP with the stopping rule $\|r_i\|_2 \leq \sigma\sqrt{n + 2\sqrt{n \log n}}$ selects the true subset $X(T)$ with probability at least $1 - 1/n$.

One can also directly apply the results for the ℓ_∞ bounded noise case to the Gaussian case. In fact, a stronger result holds.

Theorem 8: Suppose $\epsilon \sim N(0, \sigma^2 I_n)$, $\mu < \frac{1}{2k-1}$ and all the nonzero coefficients β_i satisfy

$$|\beta_i| \geq \frac{2\sigma\sqrt{2(1+\eta)\log p}}{1 - (2k-1)\mu} \tag{6}$$

for some $\eta \geq 0$. Then OMP with the stopping rule $\|X' r_i\|_\infty \leq \sigma\sqrt{2(1+\eta)\log p}$ selects exactly the correct subset $X(T)$ with probability at least $1 - k/p^\eta\sqrt{2\log p}$.

Remark 4: The conditions in the previous Theorem can also be reformulated under the ERC and $\lambda_{\min} > 0$. Suppose $M < 1$ and $\lambda_{\min} > 0$, then the OMP algorithm can recover the true support of β with high probability when each nonzero coefficient β_i satisfies

$$|\beta_i| \geq \frac{2\sigma\sqrt{2(1+\eta)\log p}}{(1-M)\lambda_{\min}}. \tag{7}$$

Remark 5: After the OMP algorithm returns the estimated subset, one can use the ordinary least squares to further estimate the values of the nonzero coordinates of β. Then with high probability, the mean squared error of the resulting estimator will be the same as the case when the true support of β were known.

It is interesting to compare the results given above with some of the known results in the literature based on other methods. As mentioned in the introduction, ℓ_1 minimization methods are widely used for reconstructing a sparse signal as well as for support recovery. In particular, Zhao and Yu (2006) considered the model selection consistency of the Lasso and introduced the Irrepresentable Condition. First, it is worth noting that if the Irrepresentable Condition holds for every k-sparse signal $\beta \in \mathbb{R}^p$, then it is equivalent to the ERC. This can be explained as follows. The Irrepresentable Condition requires $\|X(U)'X(T)(X(T)'X(T))^{-1}\text{sign}(\beta(T))\|_\infty < 1$, where $U = \{i : \beta_i = 0\}$ and $\beta(T)$ denotes the k dimensional subvector that only keeps the nonzero coordinates of β. If the Irrepresentable Condition holds for every $\beta(T) \in \mathbb{R}^k$, then the sum of the absolute values of the entries in each column of the matrix $(X(T)'X(T))^{-1}X(T)'X(U)$ must be less than 1, which is equivalent to the ERC. Also, for the Lasso estimator to be sign consistent, the minimum eigenvalue $\lambda_{\min}$ must be positive as we remarked earlier. In Zhao and Yu (2006), the order of magnitude of all the nonzero coefficients β_i are required to be at least $n^{(1+c)/2}$ for some $c > 0$. This condition is much stronger than Condition (6) that is required in Theorem 8 or Condition (7) under the ERC. It is also stronger than Condition (5) used in Theorem 7.

If not all nonzero coordinates of β are large, then the OMP algorithm can still select all the significant coordinates of β with high probability. More specifically, we have the following result.

Theorem 9: Suppose $\epsilon \sim N(0, \sigma^2 I_n)$, $\mu < \frac{1}{2k-1}$ and let

$$S = \left\{ X_i : 1 \leq i \leq p, |\beta_i| \geq \frac{2\sqrt{k}\sigma\sqrt{2(1+\eta)\log p}}{1 - (2k-1)\mu} \right\}.$$

Then the OMP algorithm selects a correct variable at each step with probability at least $1 - 1/p^\eta\sqrt{2\log p}$ until all the variables in S are selected.

As mentioned earlier, it is sometimes desirable to select a subset of the significant variables without selecting any incorrect variables. By modifying the stopping rule in the Gaussian noise case, it is possible to ensure that with high probability OMP only selects the significant variables and does not select incorrect variables. More specifically, we have the following theorem.

Theorem 10: Suppose $\epsilon \sim N(0, \sigma^2 I_n)$, $\mu < \frac{1}{2k-1}$ and let us see equation at the bottom of the page. Then the OMP algorithm returns a subset $\hat{T}$ such that $S \subset \hat{T} \subset T$ with probability at least $1 - 1/p^\eta\sqrt{2\log p}$.

IV. Discussions

The analysis of the OMP algorithm given in Section III is given under the MIP condition $\mu < \frac{1}{2k-1}$. As mentioned earlier, the main results can all be reformulated under the ERC $M < 1$. The reason we use the MIP condition is that the mutual incoherence μ is a computable quantity, while M is not as it depends on the unknown support T. A precise restatement of Theorem 1 under the ERC was given in Proposition 1 and a brief comment

$$S = \left\{ X_i : 1 \leq i \leq p, |\beta_i| \geq \left(\frac{6k}{1 - (2k-1)\mu} + 4\sqrt{k} \right) (1 + \sqrt{2k})\sqrt{2(1+\eta)\log p} \right\}.$$

was given for Theorem 2. We now discuss other results under the ERC.

Theorem 3 holds under the ERC and the result can be restated as follows. Suppose $\|\epsilon\|_2 \le b_2$ and $M < 1$. Let

$$S = \left\{ X_i : 1 \le i \le p, |\beta_i| \ge \left(\frac{2\sqrt{k}\lambda_{\max}}{(1-M)\lambda_{\min}^2} + \frac{2}{\lambda_{\min}} \right) b_2 \right\}.$$

Then the OMP algorithm with the stopping rule $\|r_i\|_2 \le (1 + \frac{2\sqrt{k}\lambda_{\max}}{(1-M)\lambda_{\min}})b_2$ selects a subset $\hat{T}$ such that $S \subset \hat{T} \subset T$. Similar to Theorem 1, Theorem 4 is also true if the MIP condition $\mu < \frac{1}{2k-1}$ is replaced by $M < 1$ and the lower bound on the magnitude of the nonzero β_i is changed to $|\beta_i| \ge \frac{2(1+\sqrt{2k})b_\infty}{(1-M)\lambda_{\min}}$. Other main results can also be restated in terms of the ERC in a similar way.

It is useful to compare our results with some of the known results in the literature. Donoho, Elad, and Temlyakov (2006) considered the OMP algorithm for the noiseless and ℓ_2 bounded noise cases. It was shown that OMP can recover the support of the signal when $\mu \le \frac{1}{2k-1}(1 - \frac{2b_2}{\beta_{\min}})$, where $\beta_{\min} = \min_i\{|\beta_i| : \beta_i \ne 0\}$, whereas only $\mu < \frac{1}{2k-1}$ is required in all of our results. As shown in Cai, Wang and Xu (2010a) the condition $\mu < \frac{1}{2k-1}$ is sharp in the sense that there exists a design matrix X with the mutual incoherence $\mu = \frac{1}{2k-1}$ such that certain k-sparse signals are not identifiable based on y and X in the noiseless case. Moreover, we also considered the case where no lower bound is assumed on the magnitude of the nonzero coordinates of β. It is shown in this setting that OMP is still able to identify the significant components before possibly selecting the incorrect ones.

Zhang (2009) considered model selection consistency of the OMP algorithm and showed that under suitable stopping conditions, OMP will return a subset of the true support and the number of unselected nonzero components can be bounded. In the present paper we show that under a different stopping rule, OMP not only returns a subset of the true support, but also guarantees that all the significant components are selected. The advantage is that with our stopping rule, the algorithm will not ignore any components with large values. This is an important property for many applications. Moreover, with the same probability of identifying the true support, the lower bound on the magnitude of the nonzero coordinates for our method is smaller than what is required in Zhang (2009). For example, when the probability of identifying the true support is set to be $1 - k/p^\eta\sqrt{2\log p}$, then the lower bound of nonzero $|\beta_i|$ is $\frac{2\sigma\sqrt{2(1+\eta)\log p}}{(1-M)\lambda_{\min}}$ (see Theorem 8), while the lower bound given in Zhang (2009) is $\frac{3\sigma\sqrt{2(1+\eta)\log p + \log(4\sqrt{2\log p}/k)}}{(1-M)\lambda_{\min}}$.

Finally, we note that Lounici (2008) considered the properties of the LASSO and Dantzig selector under the MIP conditions. It was showed that when the mutual incoherence is sufficiently small both the LASSO and Dantzig selector have desirable variable selection properties. The MIP condition used in Lounici (2008) is $\mu < \frac{1}{3k}$ for the Dantzig selector and $\mu < \frac{1}{5k}$ for the LASSO. In comparison, our condition, $\mu < \frac{1}{2k-1}$, is clearly weaker than both of them and as we mentioned earlier this condition is sharp. In addition, the analysis given in the present paper on variable selection is much more detailed.

V. Understanding the OMP Algorithm

We will prove all the main results in Section VI. To gain insight on the OMP algorithm and to illustrate the main ideas behind the proofs, it is instructive to provide some technical analysis of the algorithm. The analysis sheds light on how and when the OMP algorithm works properly.

Note that the support $T = \{i : \beta_i \ne 0\}$ and the set of significant or "correct" variables is $X(T) = \{X_i : i \in T\}$. At each step of the OMP algorithm, the residual vector is projected onto the space spanned by the selected variables (columns of X). Suppose the algorithm selects the correct variables at the first t steps and the set of all selected variables at the current step is $X(c_t)$. Then $X(c_t)$ contains t variables and $X(c_t) \subset X(T)$. Recall that $P_t = X(c_t)(X(c_t)'X(c_t))^{-1}X(c_t)'$ is the projection operator onto the linear space spanned by the elements of $X(c_t)$. Then the residual after t steps can be written as

$$r_t = (I - P_t)y = (I - P_t)X\beta + (I - P_t)\epsilon \equiv s_t + n_t$$

where $s_t = (I - P_t)X\beta$ is the signal part of the residual and $n_t = (I - P_t)\epsilon$ is the noise part of the residual. Let

$$M_{t,1} = \max_{x \in X(T)} \{|x's_t|\}, \quad M_{t,2} = \max_{x \in X \setminus X(T)} \{|x's_t|\} \tag{8}$$

and

$$N_t = \max_{x \in X} \{|x'n_t|\}.$$

It is clear that in order for OMP to select a correct variable at this step, it is necessary to have $\max_{x \in X(T)}\{|x'r_t|\} > \max_{x \in X \setminus X(T)}\{|x'r_t|\}$. A sufficient condition is $M_{t,1} - M_{t,2} > 2N_t$. This is because $M_{t,1} - M_{t,2} > 2N_t$ implies

$$\max_{x \in X(T)} \{|x'r_t|\} \ge M_{t,1} - N_t > M_{t,2} + N_t \ge \max_{x \in X \setminus X(T)} \{|x'r_t|\}.$$

We first focus on the value of $M_{t,1} - M_{t,2}$. The following result is due to Tropp (2004).

Lemma 4: Let M be defined as in (3) and let $M_{t,1}$ and $M_{t,2}$ be defined as in (8). Then $MM_{t,1} > M_{t,2}$ for all t.

Note that $M < 1$ is the Exact Recovery Condition. From this lemma, we know that $M_{t,1} - M_{t,2} > (1 - M)M_{t,1}$. The previous discussion shows that $M_{t,1} > \frac{2}{1-M}N_t$ is a sufficient condition under which OMP will make a correct decision. Then from Lemma 1 the condition

$$M_{t,1} > 2\frac{1 - (k-1)\mu}{1 - (2k-1)\mu}N_t \tag{9}$$

guarantees that the OMP algorithm selects a correct variable at the current step. Let $X(u_t) = X(T) \setminus X(c_t)$ denote the set of significant variables that are yet to be selected and let $\beta(u_t)$ denote the corresponding linear coefficients, then $M_{t,1} = \|X(u_t)'s_t\|_\infty$. Note that

$$\begin{aligned} M_{t,1} &= \|X(u_t)'s_t\|_\infty = \|X(u_t)'(I - P_t)X\beta\|_\infty \\ &= \|X(u_t)'(I - P_t)X(u_t)\beta(u_t)\|_\infty. \end{aligned}$$

The following lemma, which is proved in Section VI, can be used to further bound $M_{t,1}$.

Lemma 5: The minimum eigenvalue of $X(T)'X(T)$ is less than or equal to the minimum eigenvalue of $X(u_t)'(I - P_t)X(u_t)$. The maximum eigenvalue of $X(T)'X(T)$ is greater than or equal to the maximum eigenvalue of $X(u_t)'(I - P_t)X(u_t)$.

It then follows immediately that $\|X(u_t)'(I - P_t)X(u_t)\beta(u_t)\|_2 \geq \lambda_{\min}\|\beta(u_t)\|_2$. Lemma 2 now yields

$$\begin{aligned} M_{t,1} &\geq (k-t)^{-1/2}\|X(u_t)'s_t\|_2 \\ &\geq (k-t)^{-1/2}\lambda_{\min}\|\beta(u_t)\|_2 \\ &\geq (k-t)^{-1/2}(1-(k-1)\mu)\|\beta(u_t)\|_2. \end{aligned}$$

This and equation (9) show that a sufficient condition for selecting a correct variable at the current step is

$$\|\beta(u_t)\|_2 > \frac{2\sqrt{k-t}N_t}{1-(2k-1)\mu}. \tag{10}$$

Or more generally,

$$\|\beta(u_t)\|_2 > \frac{2\sqrt{k-t}N_t}{(1-M)\lambda_{\min}}. \tag{11}$$

This means that if any of the remaining coefficients is large enough, then OMP will select a correct variable at this step. For example, if there exists an unselected variable β_i with $|\beta_i| > \frac{2\sqrt{k}N_t}{1-(2k-1)\mu}$, then a correct variable would be selected. Also, if all the remaining coefficients are relatively large, i.e., $|\beta_i| > \frac{2N_t}{1-(2k-1)\mu}$ for all $i \in T$, then (10) is satisfied and OMP will select a correct variable at this step. The value of N_t depends on the noise structure and different bounds will be used for different cases in Section VI.

VI. Proofs

In this section we shall prove the main results in the order of Theorems 1, 3, 4, 6, and 8. The proofs of the other theorems are similar and are thus omitted. Some of the technical lemmas are proved at the end.

A. Proof of Theorem 1

It follows from the assumption $\|\epsilon\|_2 \leq b_2$ that

$$\|n_t\|_2 = \|(I - P_t)\epsilon\|_2 \leq \|\epsilon\|_2 \leq b_2.$$

Let X_i be any column of X. Then

$$|X_i'n_t| \leq \|X_i\|_2\|n_t\|_2 \leq b_2.$$

This means $N_t \leq b_2$. It follows from (10) that for any $t < k$, $\|\beta(u_t)\|_2 > \frac{2\sqrt{k-t}N_t}{1-(2k-1)\mu}$ implies that a correct variable will be selected at this step. So $|\beta_i| \geq \frac{2b_2}{1-(2k-1)\mu}$ for all nonzero coefficients β_i ensures that all the k correct variables will be selected in the first k steps.

Let us now turn to the stopping rule. Let P_k denote the projection onto the linear space spanned by $X(T)$. Then $\|(I - P_k)\epsilon\|_2 \leq \|\epsilon\|_2 \leq b_2$. So when all the k correct variables are selected, the ℓ_2 norm of the residual will be less than b_2 and hence the algorithm stops. It remains to be shown that the OMP algorithm does not stop early.

Suppose the algorithm has run t steps for some $t < k$. We will verify that $\|r_t\|_2 > b_2$ and so OMP does not stop at the current step. Again, let $X(u_t)$ denote the set of unselected but correct variable and $\beta(u_t)$ be the corresponding coefficients. Note that

$$\begin{aligned} \|r_t\|_2 &= \|(I - P_t)X\beta + (I - P_t)\epsilon\|_2 \\ &\geq \|(I - P_t)X\beta\|_2 - \|(I - P_t)\epsilon\|_2 \\ &\geq \|(I - P_t)X(u_t)\beta(u_t)\|_2 - b_2. \end{aligned}$$

It follows from Lemma 5 that

$$\begin{aligned} \|(I - P_t)X(u_t)\beta(u_t)\|_2 &\geq \lambda_{\min}\|\beta(u_t)\|_2 \\ &\geq (1-(k-1)\mu)\frac{2b_2}{1-(2k-1)\mu} > 2b_2. \end{aligned}$$

So

$$\|r_t\|_2 \geq \|(I - P_t)X(u_t)\beta(u_t)\|_2 - b_2 > b_2$$

and the theorem is proved. ∎

B. Proof of Theorem 3

From the proof of Theorem 1, we know that

$$\|r_t\|_2 \geq \lambda_{\min}\|\beta(u_t)\|_2 - b_2 \geq (1-(k-1)\mu)\|\beta(u_t)\|_2 - b_2.$$

On the other hand,

$$\begin{aligned} \|r_t\|_2 &= \|(I - P_t)X\beta + (I - P_t)\epsilon\|_2 \\ &\leq \|(I - P_t)X(u_t)\beta(u_t)\|_2 + b_2 \\ &\leq \lambda_{\max}\|\beta(u_t)\|_2 + b_2 \\ &\leq (1+(k-1)\mu)\|\beta(u_t)\|_2 + b_2. \end{aligned}$$

So

$$\frac{\|r_t\|_2 - b_2}{1+(k-1)\mu} \leq \|\beta(u_t)\|_2 \leq \frac{\|r_t\|_2 + b_2}{1-(k-1)\mu}.$$

Since the stopping rule is to check whether $\|r_t\|_2 \leq \left(1 + \frac{(1+(k-1)\mu)2\sqrt{k}}{1-(2k-1)\mu}\right)b_2$, we know that when the stopping rule is not satisfied

$$\|\beta(u_t)\|_2 \geq \frac{2\sqrt{k}b_2}{1-(2k-1)\mu}.$$

From the previous discussion, this means OMP will select a correct variable at this step. When the stopping rule is satisfied

$$\|\beta(u_t)\|_2 \leq \left(\frac{(1+(k-1)\mu)2\sqrt{k}}{(1-(k-1)\mu)(1-(2k-1)\mu)} + \frac{2}{1-(k-1)\mu}\right)b_2$$

and so all the variables in the set

$$S = \left\{X_i : 1 \leq i \leq p, |\beta_i| \geq \left(\frac{(1+(k-1)\mu)2\sqrt{k}}{(1-(k-1)\mu)(1-(2k-1)\mu)} + \frac{2}{1-(k-1)\mu}\right)b_2\right\}$$

have been selected. ∎

C. Proof of Theorem 4

Since $\|X'\epsilon\|_\infty \le b_\infty$ and $\lambda_{\min} \ge 1-(k-1)\mu$, for any $t < k$

$$\begin{aligned}\|P_t\epsilon\|_2^2 &= \epsilon' X(c_t)(X(c_t)'X(c_t))^{-1}X(c_t)'\epsilon \\ &\le \frac{1}{\lambda_{\min}}\|X(c_t)'\epsilon\|_2^2 \le \frac{tb_\infty^2}{1-(k-1)\mu}.\end{aligned}$$

Let X_i be any column of X. Then

$$\begin{aligned}|X_i'n_t| &= |X_i'(I-P_t)\epsilon| \\ &\le |X_i'\epsilon| + |X_i'P_t\epsilon| \le b_\infty + \frac{\sqrt{t}b_\infty}{\sqrt{1-(k-1)\mu}}\end{aligned}$$

which implies

$$N_t \le \left(1+\frac{\sqrt{t}}{\sqrt{1-(k-1)\mu}}\right)b_\infty.$$

Now since

$$|\beta_i| \ge \frac{2b_\infty}{1-(2k-1)\mu}\left(1+\frac{\sqrt{k}}{\sqrt{1-(k-1)\mu}}\right)$$

we have $\|\beta(u_t)\|_2 > \frac{2\sqrt{k-t}N_t}{1-(2k-1)\mu}$, which ensures that OMP selects a correct variable at this step.

We now turn to the stopping rule. It suffices to prove that for any $t < k$, $\|X'r_t\|_\infty > b_\infty$ and so the algorithm does not stop early. It can be seen that

$$\begin{aligned}\|X'r_t\|_\infty &= \|X'(I-P_t)X\beta + X'(I-P_t)\epsilon\|_\infty \\ &\ge \|X(u_t)'(I-P_t)X(u_t)\beta(u_t)\|_\infty \\ &\quad - \|X(u_t)'(I-P_t)\epsilon\|_\infty \\ &\ge \frac{1}{\sqrt{k-t}}\|X(u_t)'(I-P_t)X(u_t)\beta(u_t)\|_2 \\ &\quad - \left(1+\frac{\sqrt{t}}{\sqrt{1-(k-1)\mu}}\right)b_\infty \\ &\ge \frac{1}{\sqrt{k-t}}\lambda_{\min}\|\beta(u_t)\|_2 \\ &\quad - \left(1+\frac{\sqrt{t}}{\sqrt{1-(k-1)\mu}}\right)b_\infty \\ &\ge \left(1+\frac{\sqrt{t}}{\sqrt{1-(k-1)\mu}}\right)b_\infty > b_\infty\end{aligned}$$

and the theorem then follows. ■

D. Proof of Theorem 6

The proof of this theorem is similar to that of Theorem 3. Note that

$$\frac{\|X'r_t\|_\infty - C_{k,\mu}b_\infty}{1+(k-1)\mu} \le \|\beta(u_t)\|_2 \le \frac{\|X'r_t\|_\infty + C_{k,\mu}b_\infty}{1-(k-1)\mu}\sqrt{k},$$

where $C_{k,\mu} = 1+\frac{\sqrt{k}}{\sqrt{1-(k-1)\mu}}$. This ensures that if the stopping rule is not satisfied, i.e.,

$$\|X'r_i\|_\infty > \left(1+\frac{2\sqrt{k}(1+(k-1)\mu)}{1-(2k-1)\mu}\right)C_{k,\mu}b_\infty,$$

then $\|\beta(u_t)\|_2 > \frac{2\sqrt{k}C_{k,\mu}}{1-(2k-1)\mu}b_\infty$ and so OMP will select a correct variable at this step. On the other hand, when the stopping rule is satisfied, all the variables in the equation at the bottom of the page are selected. ■

E. Proof of Theorem 8

First, we will prove that with high probability $(1-M)M_{t1} > 2N_t$ at any step $t < k$. It can be seen that

$$M_{t1} \ge (k-t)^{-1/2}\|X(u_t)'s_t\|_2 \ge (k-t)^{-1/2}\lambda_{\min}\|\beta(u_t)\|_2.$$

Since for any $t < k$, $N_t = X'(I-P_t)\epsilon$, it follows that

$$P(N_t \le \sigma\sqrt{2(1+\eta)\log p}) \ge 1-\frac{1}{p^\eta\sqrt{2\log p}}.$$

This means if

$$\|\beta(u_t)\|_2 \ge \frac{2\sqrt{k-t}\sigma\sqrt{2(1+\eta)\log p}}{(1-M)\lambda_{\min}}$$

with probability at least $1-\frac{1}{p^\eta\sqrt{2\log p}}$, $(1-M)M_{t1} > 2N_t$ and hence a correct variable is selected at the current step. This is true when

$$|\beta_i| \ge \frac{2\sigma\sqrt{2(1+\eta)\log p}}{(1-M)\lambda_{\min}}$$

for some $i \in T$. Therefore under the conditions of the theorem, we can select all the k correct variables at the first k steps with probability at least $1-\frac{k}{p^\eta\sqrt{2\log p}}$.

We now consider the stopping rule. Suppose $N_t \le \sigma\sqrt{2(1+\eta)\log p}$ for $t = 1,2,\ldots,k$, which means the algorithm makes correct decisions at the first k steps. Now using the same argument as in the proof of Theorem 4, it can

$$S = \left\{X_i : 1 \le i \le p,\ |\beta_i| \ge \left(\frac{2k(1+(k-1)\mu)}{(1-(k-1)\mu)(1-(2k-1)\mu)} + \frac{2\sqrt{k}}{1-(k-1)\mu}\right)C_{k,\mu}b_\infty\right\}$$

be shown that for any $t < k$, when $N_t \le \sigma\sqrt{2(1+\eta)\log p}$, the algorithm will not stop early. And when $t = k$ since $N_k \le \sigma\sqrt{2(1+\eta)\log p}$ and all the correct variables have been selected, the stopping rule is satisfied and hence the algorithm stops. So the probability of selecting exactly the correct subset is at least

$$P(N_t \le \sigma\sqrt{2(1+\eta)\log p} \text{ for } t = 1, 2, \ldots, k)$$
$$\ge 1 - \sum_{t=1}^{k} P(N_t > \sigma\sqrt{2(1+\eta)\log p}) \ge 1 - \frac{k}{p^{\eta}\sqrt{2\log p}}.$$

■

F. Proofs of the Technical Lemmas

Proof of Lemma 2: To prove the lower bound on $\lambda_{\min}$, it suffices to show that when $\mu < 1/(k-1)$, the matrix $X(T)'X(T) - \lambda I$ is nonsingular for any $\lambda < 1 - (k-1)\mu$. This is equivalent to showing that for any nonzero vector $c = (c_1, c_2, \ldots, c_k)' \in R^k$, $(X(T)'X(T) - \lambda I)c \ne 0$. Without loss of generality, suppose $|c_1| \ge |c_2| \ge \cdots \ge |c_k|$ and $X(T) = (X_{T1}, X_{T2}, \ldots, X_{Tk})$. Then the first coordinate of the vector $(X(T)'X(T) - \lambda I)c$ satisfies

$$\begin{aligned}
&|\{(X(T)'X(T) - \lambda I)c\}_1| \\
&\quad = |(1-\lambda)c_1 + X'_{T1}X_{T2}c_2 + \cdots + X'_{T1}X_{Tk}c_k| \\
&\quad \ge (1-\lambda)|c_1| - \mu(|c_2| + \cdots + |c_k|) \\
&\quad > (k-1)\mu|c_1| - \mu(|c_2| + \cdots + |c_k|) \ge 0.
\end{aligned}$$

This means $(X(T)'X(T) - \lambda I)c \ne 0$ and hence $\lambda_{\min} \ge 1 - (k-1)\mu$ is proved. By the same argument, it can be shown that $\lambda_{\max} \le 1 + (k-1)\mu$. ■

Proof of Lemma 5: Without loss of generality, we can write $X(T) = (X(u_t), X(c_t))$, then we can partition matrix $X(T)'X(T)$ into blocks

$$X(T)'X(T) = \begin{pmatrix} X(u_t)'X(u_t) & X(u_t)'X(c_t) \\ X(c_t)'X(u_t) & X(c_t)'X(c_t) \end{pmatrix}.$$

Suppose we partition matrix $(X(T)'X(T))^{-1}$ into blocks the same way as $X(T)'X(T)$. Then by the standard result on the inverse of a block matrix, the up left block of $(X(T)'X(T))^{-1}$ is $(X(u_t)'(I - P_t)X(u_t))^{-1}$. Therefore the maximum eigenvalue of $(X(u_t)'(I - P_t)X(u_t))^{-1}$ is less than or equal to the maximum eigenvalue of $(X(T)'X(T))^{-1}$. Also the minimum eigenvalue of $(X(u_t)'(I-P_t)X(u_t))^{-1}$ is greater than or equal to the minimum eigenvalue of $(X(T)'X(T))^{-1}$. The lemma then follows. ■

ACKNOWLEDGMENT

The authors would like to thank the Associate Editor and two referees for their helpful comments which have led to an improvement in the presentation of the paper.

REFERENCES

[1] P. Bickel, Y. Ritov, and A. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Statist.*, vol. 37, pp. 1705–1732, 2009.
[2] A. Barron, A. Cohen, W. Dahmen, and R. DeVore, "Approximation and learning by greedy algorithms," *Ann. Statist.*, vol. 36, pp. 64–94, 2008.
[3] T. Cai, L. Wang, and G. Xu, "Stable recovery of sparse signals and an oracle inequality," *IEEE Trans. Inf. Theory*, vol. 56, pp. 3516–3522, 2010a.
[4] T. Cai, L. Wang, and G. Xu, "New bounds for restricted isometry constants," *IEEE Trans. Inf. Theory*, vol. 56, pp. 4388–4394, 2010b.
[5] T. Cai, G. Xu, and J. Zhang, "On recovery of sparse signals via l_1 minimization," *IEEE Trans. Inf. Theory.*, vol. 55, pp. 3388–3397, 2009.
[6] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, pp. 4203–4215, 2005.
[7] E. Candes and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, pp. 2313–2351, 2007.
[8] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, pp. 33–61, 1998.
[9] G. Davis, S. Mallat, and M. Avellaneda, "Greedy adaptive approximation," *J. Constr. Approx.*, vol. 13, pp. 57–98, 1997.
[10] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory.*, vol. 52, pp. 6–18, 2006.
[11] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inf. Theory.*, vol. 47, pp. 2845–2862, 2001.
[12] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, pp. 407–451, 2004.
[13] K. Lounici, "Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators," *Electronic Journal of Statistics*, vol. 2, pp. 90–102, 2008.
[14] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comp. Harmonic Anal.*, vol. 26, pp. 301–321, 2008.
[15] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B*, vol. 58, pp. 267–288, 1996.
[16] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, pp. 2231–2242, 2004.
[17] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals," *IEEE Trans. Inf. Theory*, vol. 51, pp. 1030–1051, 2006.
[18] T. Zhang, "On the consistency of feature selection using greedy least squares regression," *J. Machine Learning Res.*, vol. 10, pp. 555–568, 2009.
[19] P. Zhao and B. Yu, "On model selection consistency of Lasso," *J. Machine Learning Res.*, vol. 7, pp. 2541–2567, 2006.

T. Tony Cai received the Ph.D. degree from Cornell University, Ithaca, NY, in 1996.

His research interests include high-dimensional inference, large-scale multiple testing, nonparametric function estimation, functional data analysis and statistical decision theory. He is the Dorothy Silberberg Professor of Statistics at the Wharton School of the University of Pennsylvania, Philadelphia.

Dr. Cai is the recipient of the 2008 COPSS Presidents' Award and a fellow of the Institute of Mathematical Statistics. He is also the current editor of the *Annals of Statistics*.

Lie Wang received the Ph.D. degree in statistics from the University of Pennsylvania, Philadelphia, in 2008.

He is now with the Department of Mathematics, Massachusetts Institute of Technology, Cambridge. His research interests include nonparametric function estimation, high-dimensional sparse regression, semiparametric models and functional data regression.