

Model Selection: Two Fundamental Measures of Coherence and Their Algorithmic Significance

Waheed U. Bajwa, Robert Calderbank, and Sina Jafarpour
Princeton University, Princeton, NJ 08544

Abstract—The problem of model selection arises in a number of contexts, such as compressed sensing, subset selection in linear regression, estimation of structures in graphical models, and signal denoising. This paper generalizes the notion of *incoherence* in the existing literature on model selection and introduces two fundamental measures of coherence—termed as the worst-case coherence and the average coherence—among the columns of a design matrix. In particular, it utilizes these two measures of coherence to provide an in-depth analysis of a simple one-step thresholding (OST) algorithm for model selection. One of the key insights offered by the ensuing analysis is that OST is feasible for model selection as long as the design matrix obeys an easily verifiable property. In addition, the paper also characterizes the model-selection performance of OST in terms of the worst-case coherence, μ , and establishes that OST performs near-optimally in the low signal-to-noise ratio regime for $N \times C$ design matrices with $\mu \approx O(N^{-1/2})$. Finally, in contrast to some of the existing literature on model selection, the analysis in the paper is nonasymptotic in nature, it does not require knowledge of the true model order, it is applicable to generic (random or deterministic) design matrices, and it neither requires submatrices of the design matrix to have full rank, nor does it assume a statistical prior on the values of the nonzero entries of the data vector.

I. INTRODUCTION

In information processing problems involving high-dimensional data, the “curse of dimensionality” can often be broken by exploiting the fact that real-world data tend to live in low-dimensional manifolds. This phenomenon is exemplified by the important special case in which a data vector $\alpha \in \mathbb{C}^C$ satisfies $\|\alpha\|_0 \doteq \sum_{i=1}^C 1_{\{|\alpha_i|>0\}} \leq k \ll C$ and is observed according to the linear measurement model $f = \Phi\alpha + \eta$. Here, Φ is an $N \times C$ (real- or complex-valued) matrix called the *measurement* or *design matrix*, while $\eta \in \mathbb{C}^N$ represents noise in the measurement system. In this problem, the fact that α is “ k -sparse” allows one to operate in the so-called “compressed” setting, $k < N \ll C$, thereby enabling tasks that might be deemed prohibitive otherwise.

Fundamentally, given a measurement vector $f = \Phi\alpha + \eta$ in the compressed setting, there are two complementary—but nonetheless distinct—questions that one needs to answer:

[Estimation] Under what conditions can a k -sparse α be reliably and efficiently reconstructed from f ?

[Model Selection] Under what conditions can the locations of the nonzero entries of a k -sparse α be reliably and efficiently recovered from f ?

A number of researchers have successfully addressed the estimation question over the past few years under the rubric of *compressed sensing*. In many application areas, however, the

model-selection question is equally—if not more—important than the estimation question. In particular, the problem of model selection (sometimes also known as *variable selection* or *sparsity pattern recovery*) arises indirectly in a number of contexts, such as subset selection in linear regression [1], estimation of structures in graphical models [2], and signal denoising [3]. In addition, solving the model-selection problem sometimes also enables one to solve the estimation problem.

In this paper, we study the problem of *polynomial-time model selection in a compressed setting for the case when the true model order k is unknown*. Despite being well-motivated by applications, this problem has received less attention compared to its estimation counterpart in the compressed sensing literature; the most notable exceptions here being [2], [4]–[8]. In particular, the results reported in [2], [4] establish that the lasso [9] asymptotically identifies the correct model under certain conditions on the design matrix Φ and the sparse vector α . Later, Wainwright in [5] strengthens the results of [2], [4] and makes explicit the dependence of model selection using the lasso on the smallest (in magnitude) nonzero entry of α . However, apart from the fact that the results reported in [2], [4], [5] are asymptotic in nature, the main limitation of these works is that explicit verification of the conditions that Φ needs to satisfy is computationally intractable for $k \gtrsim \sqrt{N}$.

The most general (and nonasymptotic) results for model selection using the lasso have been reported in [6]. Specifically, Candès and Plan establish in [6] that the lasso correctly identifies *most* models with probability $1 - O(C^{-1})$ under certain conditions on the smallest nonzero entry of α provided: (i) the spectral norm (the largest singular value) and the worst-case coherence (the maximum absolute innerproduct between the columns) of Φ are not too large, and (ii) the values of the nonzero entries of α are statistically independent (and statistically symmetric around zero). The main limitation of this work is that statistical independence among the nonzero entries of α can be difficult to ensure in many applications.

Finally, as opposed to the approach taken in [2], [4]–[6], the focus in [7], [8] is on model selection using a simple thresholding algorithm. In particular, it is shown in both [7], [8] that model selection using thresholding is asymptotically optimal in the low signal-to-noise ratio (SNR) regime. However, one of the main limitations of these works is that the reported results are mainly asymptotic in nature and rely on having some knowledge of the true model order. In addition, the analysis carried out in [7] is for the specific case of an independent and identically distributed (i.i.d.) Gaussian design

matrix, while the analysis carried out in [8] is for the specific case of α with i.i.d. Gaussian nonzero entries.

A. Our Contributions

We begin by assuming that the design matrix Φ has unit ℓ_2 -norm columns and introducing two fundamental measures of coherence among the columns $\{\varphi_i \in \mathbb{C}^N\}$ of Φ :

- **Worst-Case Coherence:** $\mu \doteq \max_{i,j:i \neq j} |\langle \varphi_i, \varphi_j \rangle|$, and
- **Average Coherence:** $\nu \doteq \frac{1}{\mathcal{C}-1} \max_i \left| \sum_{j:j \neq i} \langle \varphi_i, \varphi_j \rangle \right|$.

In words, worst-case coherence is a similarity measure between the columns of a design matrix and average coherence is a measure of the spread of the columns of a design matrix within the N -dimensional unit ball. Our main objective in this paper is to make use of these two measures of coherence in order to analyze the *one-step thresholding* (OST) algorithm (see Algorithm 1) for model selection. Algorithmically, this makes our approach to model selection somewhat similar to the one studied by Fletcher, Rangan, and Goyal [7] and Reeves and Gastpar [8]. Analytically, however, the results reported in this paper are more general in nature than the ones in [7], [8]; in particular, the asymptotic results of [7], [8] for thresholding can be obtained as a special case of Theorem 1 in Section II.

More specifically, Theorem 1 holds for any (random or deterministic) design matrix with sufficiently small values of the worst-case and average coherence, and the stated result in that case is completely nonasymptotic in nature. Equally importantly, unlike the case of [7], [8], the threshold value in Theorem 1 is completely independent of the model order k and relies only on the knowledge of μ, \mathcal{C} , and SNR. In addition, Theorem 1 can also be combined with the necessary conditions for asymptotically consistent model selection reported in [7], [10] to conclude that model selection using the OST is asymptotically optimal in the low SNR regime for any design matrix that has $\mu \approx O(N^{-1/2})$ and $\nu \approx O(N^{-1})$.

Finally, in order to compare the results obtained in this paper for model selection using the OST with the nonasymptotic results reported in [6, Theorem 1.3] for the lasso, Theorem 2 rederives Theorem 1 in terms of conditions on the model order k and the smallest nonzero entry of α . In particular, it can be easily concluded from Theorem 2 and [6, Theorem 1.3] that the OST—despite being computationally primitive—performs as well as the lasso for model selection in the low SNR regime provided the design matrix has $\mu \approx O(N^{-1/2})$ and $\nu \approx O(N^{-1})$. In addition, unlike the assumptions made in [6], the OST achieves this without requiring that most $N \times k$ submatrices of Φ be well-conditioned and the nonzero entries of α be statistically independent.

II. MAIN RESULT

A. Problem Setup

Before proceeding with presenting the main result of this paper, we need to be precise about our problem formulation. To this end, we begin by reconsidering the measurement model $f = \Phi\alpha + \eta$ in the compressed setting ($k < N \ll \mathcal{C}$) and take

Algorithm 1 The One-Step Thresholding (OST) Algorithm for Model Selection

Input: An $N \times \mathcal{C}$ (real- or complex-valued) matrix Φ , a vector $f \in \mathbb{C}^N$, and a thresholding parameter $\lambda > 0$.

Output: Compute $y \doteq \Phi^H f$ and return an estimate of the model $\hat{\mathcal{S}} \doteq \{i \in \{1, \dots, \mathcal{C}\} : |y_i| > \lambda\}$.

the noise vector η to be distributed as $\text{CN}(0, \sigma^2 I)$, although the results can be readily generalized for other noise distributions. We also assume without loss of generality that Φ has unit ℓ_2 -norm columns and $\|\alpha\|_2 = 1$, since any scaling of Φ and α can be accounted for in the scaling of σ . In addition, we do not impose any prior distribution on the design matrix Φ and the nonzero entries of α . Finally, we use the notation $\text{supp}(\alpha)$ for the set containing the locations of the nonzero entries of α and assume, similar to the case of [6], [8], that $\text{supp}(\alpha)$ is a uniformly random k -subset of $\{1, \dots, \mathcal{C}\}$. In other words, we have a uniform prior on the model $\text{supp}(\alpha)$.

B. The Coherence Property and Its Implications

It is often realized in the literature that successful model selection requires the columns of the design matrix to be *incoherent*; see, e.g., [2], [4], [6]. Below, we mathematically formalize this notion in terms of the coherence property.

Definition 1 (The Coherence Property). A matrix Φ is said to obey the coherence property if the following conditions hold:

$$\text{(CP-1)} \quad \mu \leq \frac{1}{\sqrt{10 \log \mathcal{C}}}, \quad \text{and} \quad \text{(CP-2)} \quad \nu \leq \frac{12\mu}{\sqrt{N}}.$$

Notice that the coherence property can be easily verified in polynomial time since it only requires checking that $\|\Phi^H \Phi - I\|_{\max} \leq (10 \log \mathcal{C})^{-1/2}$ and $\|(\Phi^H \Phi - I)\mathbf{1}\|_{\infty} \leq 12(\mathcal{C} - 1)N^{-1/2} \|\Phi^H \Phi - I\|_{\max}$.

Before proceeding with describing the implications of the coherence property, it is instructive to first define three fundamental quantities as follows:

$$\alpha_{\min} \doteq \min_{i:|\alpha_i|>0} |\alpha_i|, \quad \text{SNR}_{\min} \doteq \frac{\alpha_{\min}^2}{\mathbb{E}[\|\eta\|_2^2]/k}, \quad \text{MAR} \doteq \frac{\alpha_{\min}^2}{1/k}.$$

In words, α_{\min} is the magnitude of the smallest nonzero entry of α , SNR_{\min} is the ratio of the energy in the smallest nonzero entry of α and the average noise energy per nonzero entry, and MAR—which is termed as the *minimum-to-average ratio* [7]—is the ratio of the energy in the smallest nonzero entry of α and the average signal energy per nonzero entry. We are now ready to state the main result of this paper.

Theorem 1. *Suppose that Φ obeys the coherence property and write its worst-case coherence as $\mu = c_1 N^{-1/\beta}$ for some $c_1 > 0$ (which may depend on $\log \mathcal{C}$) and $\beta \in (1, \infty]$. Next, choose the threshold $\lambda = 4 \max\{12\mu\sqrt{2 \log \mathcal{C}}, \sqrt{\sigma^2 \log \mathcal{C}}\}$. Then the OST satisfies $\Pr(\hat{\mathcal{S}} \neq \text{supp}(\alpha)) \leq 9\mathcal{C}^{-1}$ as long as the number of measurements*

$$N > \max \left\{ 2k \log \mathcal{C}, \frac{64}{\text{SNR}_{\min}} k \log \mathcal{C}, \left(\frac{2c_2}{\text{MAR}} k \log \mathcal{C} \right)^{\beta/2} \right\}.$$

Here, the quantity $c_2 > 0$ is defined as $c_2 = (96 c_1)^2$.

Remark 1. The constants in the second and third terms in the \max expression can be significantly reduced if one is only interested in showing the model-selection consistency of OST; that is, $\lim_{C \rightarrow \infty} \Pr(\hat{\mathcal{S}} \neq \text{supp}(\alpha)) = 0$. One should be particularly vigilant of this fact while comparing these results to the asymptotic ones reported in [7] for thresholding.

Note that there are two fundamental but complementary approaches that can be taken while analyzing an algorithm for model selection, namely, the *minimum measurement resources* approach and the *permissible signal class* approach. The statement of Theorem 1 helps us analyze the OST for model selection using the former approach and is best suited for comparing our results with those in [5], [7], [10]. On the other hand, Candès and Plan in [6] take the latter approach while analyzing the lasso for model selection and the following result is best suited for comparison purposes in this regard.

Theorem 2. *Suppose that Φ obeys the coherence property and choose the threshold $\lambda = 4 \max \{12\mu\sqrt{2\log C}, \sqrt{\sigma^2 \log C}\}$. Then, as long as $k \leq N/(2\log C)$ and*

$$\alpha_{\min} > \max \left\{ 8\sqrt{\sigma^2 \log C}, 96\mu\sqrt{2\log C} \right\}$$

the OST satisfies $\Pr(\hat{\mathcal{S}} \neq \text{supp}(\alpha)) \leq 9C^{-1}$.

C. Discussion

The statements of Theorem 1 and Theorem 2 can be best put into perspective by considering specific examples of design matrices. Because of space constraints, we only consider here the case when Φ is an (appropriately normalized) i.i.d. Gaussian matrix. It is a well-known fact in the literature that the worst-case coherence of Φ in this case is roughly $\mu \approx \sqrt{2\log C/N}$ with high probability; see, e.g., [6]. In addition, it can also be shown using the Bernstein inequality that $\nu \leq 12N^{-1}\sqrt{2\log C}$ with high probability in this case. It therefore follows that a Gaussian design matrix obeys the coherence property with high probability.

Theorem 1 therefore implies that the OST identifies the correct model in this case with high probability as long as $N \gtrsim \max \left\{ \frac{k \log C}{\text{SNR}_{\min}}, \frac{k \log^2 C}{\text{MAR}} \right\}$. In particular, this expression reduces to $N \gtrsim \frac{k \log C}{\text{SNR}_{\min}}$ for the case of $\text{SNR} \doteq 1/\mathbb{E}[\|\eta\|_2^2] \ll 1$. On the other hand, we have from [7], [10] that no scheme can asymptotically identify the correct model if $N \not\gtrsim \frac{k \log C}{\text{SNR}_{\min}}$. This proves the near-optimality of the OST for model selection in the low SNR regime for any design matrix that has $\mu \approx O(N^{-1/2})$ and $\nu \approx O(N^{-1})$. Finally, note that we could have made a similar conclusion by focusing on Theorem 2 and comparing the conditions in the low SNR regime in that case with those in [6, Theorem 1.3].

III. PROOFS

The general roadmap for the proofs of Theorem 1 and Theorem 2 is as follows. Below, we first introduce the notion of (k, ϵ, δ) -statistical orthogonality condition (StOC). Next, we establish in Lemma 1 that if Φ satisfies the StOC then

OST recovers the support of α with high probability provided α_{\min} is large enough. Subsequently, we establish in Lemma 2 and Lemma 3 the relationship between the StOC parameters and the worst-case and average coherence of Φ . The proofs of Theorem 1 and Theorem 2 then follow by judiciously combining the results of these three lemmas using the coherence property.

Definition 2 ((k, ϵ, δ) -StOC). Let $\Pi \doteq (\pi_1, \dots, \pi_k)$ be a uniformly random (ordered) k -subset of $\{1, \dots, C\}$ and let $\Pi^c \doteq \{1, \dots, C\} - \Pi$. Then, given $\epsilon, \delta \in [0, 1]$, Φ is said to satisfy the (k, ϵ, δ) -statistical orthogonality condition if the following inequalities

$$\text{(StOC-1)} \quad \|(\Phi_{\Pi}^H \Phi_{\Pi} - I)z\|_{\infty} \leq \epsilon \|z\|_2$$

$$\text{(StOC-2)} \quad \|\Phi_{\Pi^c}^H \Phi_{\Pi} z\|_{\infty} \leq \epsilon \|z\|_2$$

hold for every fixed $z \in \mathbb{C}^k$ with probability exceeding $1 - \delta$ (with respect to the choice of Π).

Remark 2. Note that the StOC derives its name from the fact that if Φ is an orthogonal matrix then it trivially satisfies the StOC for every k with $\epsilon = \delta = 0$.

Lemma 1. *Let $\Pi \doteq \text{supp}(\alpha)$ be a uniformly random k -subset of $\{1, \dots, C\}$. Further, suppose that the matrix Φ satisfies the (k, ϵ, δ) -StOC and choose the threshold as $\lambda = 2 \max \{\epsilon, 2\sqrt{\sigma^2 \log C}\}$. Then, under the assumption that $\alpha_{\min} > 2\lambda$, the OST satisfies*

$$\Pr(\hat{\mathcal{S}} \neq \Pi) \leq \delta + 2 \left(\sqrt{2\pi \log C} \cdot C \right)^{-1}.$$

Proof: We begin by defining $z^T \doteq [\alpha_{\pi_1} \ \dots \ \alpha_{\pi_k}]$ and writing the vector $y = \Phi^H f$ as $y = \Phi^H \Phi_{\Pi} z + \Phi^H \eta$. Now, let $\Pi^c \doteq \{1, \dots, C\} - \Pi$ and note that in order to establish that $\hat{\mathcal{S}} = \Pi$ we need to show that $\|y_{\Pi^c}\|_{\infty} \leq \lambda$ and $\min_i |y_{\pi_i}| > \lambda$.

In this regard, first note that $\tilde{\eta} \doteq \Phi^H \eta$ is a complex Gaussian random vector whose entries are identically (although not independently) distributed as $\text{CN}(0, \sigma^2)$. It therefore follows from the tail bound on the maximum of C arbitrary complex Gaussian random variables that $\|\tilde{\eta}\|_{\infty} \leq 2\sqrt{\sigma^2 \log C}$ with probability exceeding $1 - 2 \left(\sqrt{2\pi \log C} \cdot C \right)^{-1}$. Further, define

$$\mathcal{G} \doteq \left\{ \|\tilde{\eta}\|_{\infty} \leq 2\sqrt{\sigma^2 \log C} \right\} \cap \left\{ \text{(StOC-1)} \cap \text{(StOC-2)} \right\}$$

and notice that, since the noise is independent of Π , we have $\Pr(\mathcal{G}) > 1 - \delta - 2 \left(\sqrt{2\pi \log C} \cdot C \right)^{-1}$. In addition, conditioned on the event \mathcal{G} , we have

$$\begin{aligned} \|y_{\Pi^c}\|_{\infty} &\stackrel{(a)}{\leq} \|\Phi_{\Pi^c}^H \Phi_{\Pi} z\|_{\infty} + \|\tilde{\eta}\|_{\infty} \\ &\stackrel{(b)}{\leq} \epsilon + 2\sqrt{\sigma^2 \log C} \stackrel{(c)}{\leq} \lambda \end{aligned} \quad (1)$$

where (a) follows from the triangle inequality, (b) is a consequence of the conditioning on \mathcal{G} , and (c) follows from the fact that $\lambda = 2 \max \{\epsilon, 2\sqrt{\sigma^2 \log C}\}$.

Finally, in order to show that $\min_i |y_{\pi_i}| > \lambda$, we define $r = (\Phi_{\Pi}^H \Phi_{\Pi} - I)z$ and note that, conditioned on the event \mathcal{G} ,

we have for any $i \in \{1, \dots, k\}$

$$|y_{\pi_i}| = |\alpha_{\pi_i} + r_i + \tilde{\eta}_{\pi_i}| \geq |\alpha_{\pi_i}| - \|r\|_\infty - \|\tilde{\eta}\|_\infty \stackrel{(d)}{>} 2\lambda - \epsilon - 2\sqrt{\sigma^2 \log \mathcal{C}} \stackrel{(e)}{\geq} \lambda \quad (2)$$

where (d) follows from the conditioning on \mathcal{G} and the assumption that $\alpha_{\min} > 2\lambda$, while (e) is a simple consequence of the choice of λ . This completes the proof of the lemma since we have now shown that $\Pr(\hat{\mathcal{S}} \neq \Pi) \leq \Pr(\mathcal{G}^c)$. ■

Having established Lemma 1, our next goal is to relate the StOC parameters with the worst-case and average coherence.

Lemma 2. *Let $\Pi \doteq (\pi_1, \dots, \pi_k)$ be a uniformly random (ordered) k -subset of $\{1, \dots, \mathcal{C}\}$. Then, for any fixed $z \in \mathbb{C}^k$, $\epsilon \geq 0$, and $k \leq \min\{\epsilon^2 \nu^{-2}/4, \mathcal{C}/2\}$, we have*

$$\Pr(\{\Phi \text{ does not satisfy (StOC-1)}\}) \leq 4k \exp\left(-\frac{\epsilon^2 \mu^{-2}}{576}\right).$$

Proof: The proof of this lemma relies heavily on the so-called *method of bounded differences* (MOBD) [11]. Specifically, note that $\|(\Phi_{\Pi}^H \Phi_{\Pi} - I)z\|_\infty = \max_i \left| \sum_{j \neq i} z_j \langle \varphi_{\pi_i}, \varphi_{\pi_j} \rangle \right|$ and define $\Pi^{-i} \doteq (\pi_1, \dots, \pi_{i-1}, \pi_{i+1}, \dots, \pi_k)$. Then for a fixed index i , and conditioned on the event $\mathcal{A}_{i'} \doteq \{\pi_i = i'\}$, we have the following equality from basic probability theory

$$\Pr\left(\left|\sum_{\substack{j=1 \\ j \neq i}}^k z_j \langle \varphi_{\pi_i}, \varphi_{\pi_j} \rangle\right| > \epsilon \|z\|_2 \middle| \mathcal{A}_{i'}\right) = \Pr\left(\left|\sum_{\substack{j=1 \\ j \neq i}}^k z_j \langle \varphi_{i'}, \varphi_{\pi_j} \rangle\right| > \epsilon \|z\|_2 \middle| \mathcal{A}_{i'}\right). \quad (3)$$

Next, in order to apply the MOBD, we construct a Doob's martingale sequence $(M_0, M_1, \dots, M_{k-1})$ as follows:

$$M_0 = \mathbb{E}\left[\sum_{\substack{j=1 \\ j \neq i}}^k z_j \langle \varphi_{i'}, \varphi_{\pi_j} \rangle \middle| \mathcal{A}_{i'}\right], \text{ and} \\ M_\ell = \mathbb{E}\left[\sum_{\substack{j=1 \\ j \neq i}}^k z_j \langle \varphi_{i'}, \varphi_{\pi_j} \rangle \middle| \pi_{1 \rightarrow \ell}^{-i}, \mathcal{A}_{i'}\right], \ell = 1, \dots, k-1 \quad (4)$$

where $\pi_{1 \rightarrow \ell}^{-i}$ is the first ℓ coordinates of Π^{-i} . Here, note that

$$|M_0| \leq \sum_{j \neq i} |z_j| \left| \mathbb{E}[\langle \varphi_{i'}, \varphi_{\pi_j} \rangle | \mathcal{A}_{i'}] \right| \stackrel{(a)}{\leq} \sum_{j \neq i} |z_j| \left| \sum_{\substack{q=1 \\ q \neq i'}}^{\mathcal{C}} \frac{1}{\mathcal{C}-1} \langle \varphi_{i'}, \varphi_q \rangle \right| \stackrel{(b)}{\leq} \sqrt{k} \nu \|z\|_2 \quad (5)$$

where (a) follows since, conditioned on $\mathcal{A}_{i'}$, π_j has a uniform distribution over $\{1, \dots, \mathcal{C}\} - \{i'\}$, while (b) mainly follows from the definition of average coherence. Further, if we define

$$M_\ell(x) \doteq \mathbb{E}\left[\sum_{\substack{j=1 \\ j \neq i}}^k z_j \langle \varphi_{i'}, \varphi_{\pi_j} \rangle \middle| \pi_{1 \rightarrow \ell-1}^{-i}, \pi_\ell^{-i} = x, \mathcal{A}_{i'}\right] \quad (6)$$

then, since $(M_0, M_1, \dots, M_{k-1})$ is a Doob's martingale sequence, it can be easily verified that $|M_\ell - M_{\ell-1}|$ is upper-bounded by $\sup_{x,y} [M_\ell(x) - M_\ell(y)]$ (see, e.g., [12]).

Next, in order to upperbound $\sup_{x,y} [M_\ell(x) - M_\ell(y)]$, we first define $d_{\ell,j} \doteq \mathbb{E}[\langle \varphi_{i'}, \varphi_{\pi_j} \rangle | \pi_{1 \rightarrow \ell-1}^{-i}, \pi_\ell^{-i} = x, \mathcal{A}_{i'}] - \mathbb{E}[\langle \varphi_{i'}, \varphi_{\pi_j} \rangle | \pi_{1 \rightarrow \ell-1}^{-i}, \pi_\ell^{-i} = y, \mathcal{A}_{i'}]$ and then notice that

$$|M_\ell(x) - M_\ell(y)| \leq \sum_{\substack{j \leq \ell+1 \\ j \neq i}} |z_j| |d_{\ell,j}| + \sum_{\substack{j > \ell+1 \\ j \neq i}} |z_j| |d_{\ell,j}|. \quad (7)$$

In addition, we have that for every $j > \ell+1$, $j \neq i$, the random variable π_j has a uniform distribution over $\{1, \dots, \mathcal{C}\} - \{\pi_{1 \rightarrow \ell-1}^{-i}, x, i'\}$ when conditioned on $\{\pi_{1 \rightarrow \ell-1}^{-i}, \pi_\ell^{-i} = x, i'\}$, while π_j has a uniform distribution over $\{1, \dots, \mathcal{C}\} - \{\pi_{1 \rightarrow \ell-1}^{-i}, y, i'\}$ when conditioned on $\{\pi_{1 \rightarrow \ell-1}^{-i}, \pi_\ell^{-i} = y, i'\}$. Therefore, we have for every $j > \ell+1$, $j \neq i$ that

$$|d_{\ell,j}| = \frac{1}{\mathcal{C} - \ell - 1} |\langle \varphi_{i'}, \varphi_y \rangle - \langle \varphi_{i'}, \varphi_x \rangle| \leq \frac{2\mu}{\mathcal{C} - k}. \quad (8)$$

Similarly, it can be argued that $\sum_{\substack{j \leq \ell+1 \\ j \neq i}} |z_j| |d_{\ell,j}| \leq |z_{\ell+1}| 2\mu$ when $i \leq \ell$, $\sum_{\substack{j \leq \ell+1 \\ j \neq i}} |z_j| |d_{\ell,j}| \leq |z_\ell| 2\mu$ when $i = \ell+1$, and $\sum_{\substack{j \leq \ell+1 \\ j \neq i}} |z_j| |d_{\ell,j}| \leq (|z_\ell| + \frac{|z_{\ell+1}|}{\mathcal{C}-k}) 2\mu$ when $i > \ell+1$. Consequently, it can be easily verified that

$$\sup_{x,y} [M_\ell(x) - M_\ell(y)] \leq 2\mu \underbrace{\left(|z_\ell| + |z_{\ell+1}| + \frac{\|z\|_1}{\mathcal{C} - k}\right)}_{\doteq c_\ell}. \quad (9)$$

We have now established that $(M_0, M_1, \dots, M_{k-1})$ is a bounded-difference martingale with $|M_\ell - M_{\ell-1}| \leq c_\ell$ for $\ell = 1, \dots, k-1$. Further, it can also be verified from (9) that $\sum_{\ell=1}^{k-1} c_\ell^2 \leq 36\mu^2 \|z\|_2^2$ since $k \leq \mathcal{C}/2$. In addition, since $|M_0| \leq \sqrt{k} \nu \|z\|_2$ and $k \leq \epsilon^2 \nu^{-2}/4$, we have from the Azuma inequality for bounded-difference martingale sequences [13] adapted to the complex-valued setup that

$$\Pr\left(\left|\sum_{\substack{j=1 \\ j \neq i}}^k z_j \langle \varphi_{i'}, \varphi_{\pi_j} \rangle\right| > \epsilon \|z\|_2 \middle| \mathcal{A}_{i'}\right) \leq \Pr\left(|M_{k-1} - M_0| > \frac{\epsilon \|z\|_2}{2} \middle| \mathcal{A}_{i'}\right) \leq 4 \exp\left(-\frac{\epsilon^2 \mu^{-2}}{576}\right). \quad (10)$$

Combining all these facts together, we therefore finally obtain

$$\Pr\left(\|(\Phi_{\Pi}^H \Phi_{\Pi} - I)z\|_\infty > \epsilon \|z\|_2\right) \stackrel{(c)}{\leq} k \sum_{i'=1}^{\mathcal{C}} \Pr\left(\left|\sum_{\substack{j=1 \\ j \neq i'}}^k z_j \langle \varphi_{i'}, \varphi_{\pi_j} \rangle\right| > \epsilon \|z\|_2 \middle| \mathcal{A}_{i'}\right) \Pr(\mathcal{A}_{i'}) \stackrel{(d)}{\leq} 4k \exp\left(-\frac{\epsilon^2 \mu^{-2}}{576}\right) \quad (11)$$

where (c) follows from the union bound and the fact that the π_i 's are identically (although not independently) distributed,

while (d) follows from (10) and the fact that π_i has a uniform distribution over $\{1, \dots, \mathcal{C}\}$. ■

Lemma 3. *Let $\Pi \doteq (\pi_1, \dots, \pi_k)$ be a uniformly random (ordered) k -subset of $\{1, \dots, \mathcal{C}\}$. Further, define the random subset $\Pi^c \doteq \{1, \dots, \mathcal{C}\} - \Pi$. Then, for any fixed $z \in \mathbb{C}^k$, $\epsilon \geq 0$, and $k \leq \min\{\epsilon^2 \nu^{-2}/4, \mathcal{C}/2\}$, we have*

$$\Pr(\{\Phi \text{ does not satisfy (StOC-2)}\}) \leq 4\mathcal{C} \exp\left(-\frac{\epsilon^2 \mu^{-2}}{256}\right).$$

Proof Sketch: The proof of this lemma also relies on the MOBD and is very similar to that of Lemma 2. As such, we only provide a sketch of the proof here. To begin with, we note that $\|\Phi_{\Pi^c}^H \Phi_{\Pi} z\|_{\infty} = \max_{i \in [\mathcal{C}-k]} \left| \sum_j z_j \langle \varphi_{\pi_i^c}, \varphi_{\pi_j} \rangle \right|$, where $[\mathcal{C}-k] \doteq \{1, \dots, \mathcal{C}-k\}$ and π_i^c denotes the i^{th} coordinate of Π^c . Then for a fixed index $i \in [\mathcal{C}-k]$, and conditioned on the event $\mathcal{A}_{i'}$ with $\pi_i^c = i'$, we have the following equality

$$\Pr\left(\left|\sum_{j=1}^k z_j \langle \varphi_{\pi_i^c}, \varphi_{\pi_j} \rangle\right| > \epsilon \|z\|_2 \middle| \mathcal{A}_{i'}\right) = \Pr\left(\left|\sum_{j=1}^k z_j \langle \varphi_{i'}, \varphi_{\pi_j} \rangle\right| > \epsilon \|z\|_2 \middle| \mathcal{A}_{i'}\right). \quad (12)$$

Next, as in the case of Lemma 2, we construct a Doob's martingale sequence (M_0, M_1, \dots, M_k) as follows:

$$M_0 = \mathbb{E}\left[\sum_{j=1}^k z_j \langle \varphi_{i'}, \varphi_{\pi_j} \rangle \middle| \mathcal{A}_{i'}\right], \text{ and}$$

$$M_{\ell} = \mathbb{E}\left[\sum_{j=1}^k z_j \langle \varphi_{i'}, \varphi_{\pi_j} \rangle \middle| \pi_{1 \rightarrow \ell}, \mathcal{A}_{i'}\right], \ell = 1, \dots, k \quad (13)$$

where $\pi_{1 \rightarrow \ell}$ now denotes the first ℓ coordinates of Π . It can now be argued (as in Lemma 2) that: (i) $|M_0| \leq \sqrt{k} \nu \|z\|_2$, (ii) $|M_{\ell} - M_{\ell-1}| \leq 2\mu \left(|z_{\ell}| + \frac{\|z\|_1}{\mathcal{C}-k}\right) \doteq c_{\ell}$, and (iii) $\sum_{\ell=1}^k c_{\ell}^2 \leq 16\mu^2 \|z\|_2^2$. Therefore, since $k \leq \epsilon^2 \nu^{-2}/4$, we once again have from the (complex) Azuma inequality that (12) is upperbounded by $4 \exp\left(-\frac{\epsilon^2 \mu^{-2}}{256}\right)$. Combining all these facts together, we finally obtain the claimed result as follows

$$\Pr\left(\|\Phi_{\Pi^c}^H \Phi_{\Pi} z\|_{\infty} > \epsilon \|z\|_2\right) \stackrel{(a)}{\leq} 4\mathcal{C} \exp\left(-\frac{\epsilon^2 \mu^{-2}}{256}\right) \quad (14)$$

where (a) mainly follows from the union bound and the fact that π_i^c has a uniform distribution over $\{1, \dots, \mathcal{C}\}$. ■

We are finally ready to prove the main results of this paper using Lemmata 1–3.

Proof of Theorem 1: Note that Lemma 2 and Lemma 3 imply that if Φ has worst-case coherence μ and average coherence ν then, as long as $k \leq \mathcal{C}/2$, Φ satisfies the (k, ϵ, δ) -StOC for any $\epsilon \in [2\sqrt{k}\nu, 1)$ with $\delta = 8\mathcal{C} \exp\left(-\frac{\epsilon^2 \mu^{-2}}{576}\right)$.

Now let $k \leq N/(2 \log \mathcal{C})$ and define $\epsilon' \doteq 24\mu\sqrt{2 \log \mathcal{C}}$. Then, since Φ satisfies the coherence property, we have $2\sqrt{k}\nu \leq \epsilon' < 1$ and therefore Φ satisfies the (k, ϵ', δ') -StOC with $\delta' \doteq 8\mathcal{C}^{-1}$. Consequently, Lemma 1 states that

$\Pr(\widehat{\mathcal{S}} \neq \text{supp}(\alpha)) \leq 9\mathcal{C}^{-1}$ as long as $N \geq 2k \log \mathcal{C}$, $\alpha_{\min} > 4\epsilon'$, and $\alpha_{\min} > 8\sqrt{\sigma^2 \log \mathcal{C}}$. Further, note that

$$\alpha_{\min} > 8\sqrt{\sigma^2 \log \mathcal{C}} \iff N > \frac{64}{\text{SNR}_{\min}} k \log \mathcal{C}, \text{ and}$$

$$\alpha_{\min} > 4\epsilon' \iff N > \left(\frac{2c_2}{\text{MAR}} k \log \mathcal{C}\right)^{\beta/2}.$$

This completes the proof of the theorem. ■

Proof of Theorem 2: The proof of this theorem follows along similar lines as that of Theorem 1 and is therefore omitted here for the sake of brevity. ■

IV. CONCLUSION

In this paper, we have analyzed the one-step thresholding (OST) algorithm for model selection in terms of the worst-case and average coherence of the design matrix. In stark contrast to the existing work on model selection using thresholding, our analysis is completely nonasymptotic in nature, it does not require knowledge of the true model order, and it is applicable to arbitrary (random or deterministic) design matrices. In particular, we have established in the paper that the OST can be used for model selection as long as the design matrix obeys an easily verifiable property. Further, we have specified the dependence of the OST performance on the worst-case coherence of the design matrix and shown that it performs near-optimally in the low SNR regime for design matrices with $O(N^{-1/2})$ worst-case coherence. Finally, unlike the assumptions made in [6], our analysis also does not require that most $N \times k$ submatrices of Φ be well-conditioned and the nonzero entries of the data vector be statistically independent.

REFERENCES

- [1] A. Miller, *Subset Selection in Regression*. Chapman and Hall, 1990.
- [2] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *Ann. Statist.*, vol. 34, no. 3, pp. 1436–1462, Jun. 2006.
- [3] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Scientific Comput.*, vol. 20, no. 1, pp. 33–61, Jan. 1998.
- [4] P. Zhao and B. Yu, "On model selection consistency of lasso," *J. Machine Learning Res.*, vol. 7, pp. 2541–2563, 2006.
- [5] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (Lasso)," *IEEE Trans. Inform. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.
- [6] E. J. Candès and Y. Plan, "Near-ideal model selection by ℓ_1 minimization," *Ann. Statist.*, vol. 37, no. 5A, pp. 2145–2177, Oct. 2009.
- [7] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Necessary and sufficient conditions for sparsity pattern recovery," *IEEE Trans. Inform. Theory*, vol. 55, no. 12, pp. 5758–5772, Dec. 2009.
- [8] G. Reeves and M. Gastpar, "A note on optimal support recovery in compressed sensing," in *Proc. 43rd Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, Nov. 2009.
- [9] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.
- [10] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices," submitted. Available: arXiv:0806.0604v1
- [11] C. McDiarmid, "On the method of bounded differences," in *Surveys in Combinatorics*, J. Siemons, Ed. Cambridge University Press, 1989, pp. 148–188.
- [12] R. Motwani and P. Raghavan, *Randomized Algorithms*. New York, NY: Cambridge University Press, 1995.
- [13] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Math. J.*, vol. 19, no. 3, pp. 357–367, 1967.