Performance Analysis for Sparse Support Recovery

Gongguo Tang, *Student Member, IEEE*, and Arye Nehorai, *Fellow, IEEE*

***Abstract*—The performance of estimating the common support for jointly sparse signals based on their projections onto lower-dimensional space is analyzed. Support recovery is formulated as a multiple-hypothesis testing problem. Both upper and lower bounds on the probability of error are derived for general measurement matrices, by using the Chernoff bound and Fano's inequality, respectively. The upper bound shows that the performance is determined by a quantity measuring the measurement matrix incoherence, while the lower bound reveals the importance of the total measurement gain. The lower bound is applied to derive the minimal number of samples needed for accurate direction-of-arrival (DOA) estimation for a sparse representation based algorithm. When applied to Gaussian measurement ensembles, these bounds give necessary and sufficient conditions for a vanishing probability of error for majority realizations of the measurement matrix. Our results offer surprising insights into sparse signal recovery. For example, as far as support recovery is concerned, the well-known bound in Compressive Sensing with the Gaussian measurement matrix is generally not sufficient unless the noise level is low. Our study provides an alternative performance measure, one that is natural and important in practice, for signal recovery in Compressive Sensing and other application areas exploiting signal sparsity.**

***Index Terms*—Chernoff bound, compressive sensing, Fano's inequality, jointly sparse signals, multiple hypothesis testing, probability of error, support recovery.**

I. Introduction

SUPPORT recovery for jointly sparse signals concerns accurately estimating the nonzero component locations shared by a set of sparse signals based on a limited number of noisy linear observations. More specifically, suppose that $\{x(t) \in \mathbb{F}^N, t = 1, 2, \ldots, T\}$, $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, is a sequence of jointly sparse signals (possibly under a sparsity-inducing basis Φ instead of the canonical domain) with a common support S, which is the index set indicating the nonvanishing signal coordinates. This model is the same as the joint sparsity model 2 (JSM-2) in [1]. The observation model is linear

$$y(t) = Ax(t) + w(t) \quad t = 1, 2, \ldots, T. \tag{1}$$

In (1), $A \in \mathbb{F}^{M \times N}$ is the measurement matrix, $y(t) \in \mathbb{F}^M$ the noisy data vector, and $w(t) \in \mathbb{F}^M$ an additive noise. In most cases, the sparsity level $K \triangleq |S|$ and the number of observations M are far less than N, the dimension of the ambient space. This problem arises naturally in several signal processing areas such as Compressive Sensing [2]–[6], source localization [7]–[10], sparse approximation and signal denoising [11].

Compressive Sensing [2]–[4], a recently developed field exploiting the sparsity property of most natural signals, shows great promise to reduce signal sampling rate. In the classical setting of Compressive Sensing, only one snapshot is considered; i.e., $T = 1$ in (1). The goal is to recover a long vector $x := x(1)$ with a small fraction of nonzero coordinates from the much shorter observation vector $y := y(1)$. Since most natural signals are compressible under some basis and are well approximated by their K–sparse representations [12], this scheme, if properly justified, will reduce the necessary sampling rate beyond the limit set by Nyquist and Shannon [5], [6]. Surprisingly, for exact K–sparse signals, if $M = O(K \log(\frac{N}{K})) \ll N$ and the measurement matrix is generated randomly from, for example, a Gaussian distribution, we can recover x exactly in the noise-free setting by solving a linear programming task. Besides, various methods have been designed for the noisy case [13]–[17]. Along with these algorithms, rigorous theoretic analysis is provided to guarantee their effectiveness in terms of, for example, various l_p-norms of the estimation error for x [13]–[17]. However, these results offer no guarantee that we can recover the support of a sparse signal correctly.

The accurate recovery of signal support is crucial to Compressive Sensing both in theory and in practice. Since for signal recovery it is necessary to have $K \leq M$, signal component values can be computed by solving a least squares problem once its support is obtained. Therefore, support recovery is a stronger theoretic criterion than various l_p-norms. In practice, the success of Compressive Sensing in a variety of applications relies on its ability for correct support recovery because the nonzero component indices usually have significant physical meanings. The support of temporally or spatially sparse signals reveals the timing or location for important events such as anomalies. The indices for nonzero coordinates in the Fourier domain indicate the harmonics existing in a signal [18], which is critical for tasks such as spectrum sensing for cognitive radios [19]. In compressed DNA microarrays for bio-sensing, the existence of certain target agents in the tested solution is reflected by the locations of nonvanishing coordinates, while the magnitudes are determined by their concentrations [20]–[23]. For compressive radar imaging, the sparsity constraints are usually imposed on the discretized time-frequency domain. The distance and velocity of an object have a direct correspondence to its coordinate in the time-frequency domain. The magnitude determined by coefficients of reflection is of less physical significance[24]–[26]. In sparse linear regression [27], the recovered parameter support corresponds to the few factors that explain

Manuscript received June 05, 2009; revised November 04, 2009. Current version published March 10, 2010. This work was supported by the Department of Defense under the Air Force Office of Scientific Research MURI Grant FA9550-05-1-0443, and ONR Grant N000140810849. The material in this paper was presented in part at the 47th Annual Allerton Conference on Communication, Control, Monticello, IL, September 2009.

The authors are with the Department of Electrical and Systems Engineering, Washington University in St. Louis, St. Louis, MO 63130 USA (e-mail: gt2@ese.wustl.edu; nehorai@ese.wustl.edu).

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

Digital Object Identifier 10.1109/TIT.2009.2039039

the data. In all these applications, the supports are physically more significant than the component values.

Our study of sparse support recovery is also motivated by the recent reformulation of the source localization problem as one of sparse spectrum estimation. In [7], the authors transform the process of source localization using sensory arrays into the task of estimating the spectrum of a sparse signal by discretizing the parameter manifold. This method exhibits super-resolution in the estimation of direction of arrival (DOA) compared with traditional techniques such as beamforming [28], Capon [29], and MUSIC [30], [31]. Since the basic model employed in [7] applies to several other important problems in signal processing (see [32] and references therein), the principle is readily applicable to those cases. This idea is later generalized and extended to other source localization settings in [8]–[10]. For source localization, the support of the sparse signal reveals the DOA of sources. Therefore, the recovery algorithm's ability of exact support recovery is key to the effectiveness of the method. We also note that usually multiple temporal snapshots are collected, which results in a jointly sparse signal sets as in (1). In addition, since M is the number of sensors while T is the number of temporal samples, it is far more expensive to increase M than T. The same comments apply to several other examples in the Compressive Sensing applications discussed in the previous paragraph, especially the compressed DNA microarrays, spectrum sensing for cognitive radios, and Compressive Sensing radar imaging.

The signal recovery problem with joint sparsity constraint [33]–[36], also termed the multiple measurement vector (MMV) problem [37]–[41], has been considered in a line of previous works. Several algorithms, among them Simultaneous Orthogonal Matching Pursuit (SOMP) [34], [37], [40]; convex relaxation [41]; ℓ_1–minimization [38], [39]; and M-FOCUSS [37], are proposed and analyzed, either numerically or theoretically. These algorithms are multiple-dimension extensions of their one-dimension counterparts. Most performance measures of the algorithms are concerned with bounds on various norms of the difference between the true signals and their estimates or their closely related variants. The performance bounds usually involve the mutual coherence between the measurement matrix A and the basis matrix Φ under which the measured signals $x(t)$ have a jointly sparse representation. However, with joint sparsity constraints, a natural measure of performance would be the model (1) 's potential for correctly identifying the true common support, and, hence, the algorithm's ability to achieve this potential. As part of their research, Chen and Huo derived, in a noiseless setting, sufficient conditions on the uniqueness of solutions to (1) under ℓ_0 and ℓ_1 minimization. In [37], Cotter *et al.* numerically compared the probabilities of correctly identifying the common support by basic matching pursuit, orthogonal matching pursuit, FOCUSS, and regularized FOCUSS in the multiple-measurement setting with a range of SNRs and different numbers of snapshots.

The availability of multiple temporal samples offers serval advantages to the single-sample case. As suggested by the upper bound (26) on the probability of error, increasing the number of temporal samples drives the probability of error to zero exponentially fast as long as certain condition on the inconsistency property of the measurement matrix is satisfied. The probability of error is driven to zero by scaling the SNR according to the signal dimension in [42], which is not very natural compared with increasing the samples, however. Our results also show that under some conditions increasing temporal samples is usually equivalent to increasing the number of observations for a single snapshot. The later is generally much more expensive in practice. In addition, when there is considerable noise and the columns of the measurement matrix are normalized to one, it is necessary to have multiple temporal samples for accurate support recovery as discussed in Sections IV and V-B.

Our work has several major differences compared to related work [42] and [43], which also analyze the performance bounds on the probability of error for support recovery using information theoretic tools. The first difference is in the way the problem is modeled: In [42] and [43], the sparse signal is deterministic with known smallest absolute value of the nonzero components while we consider a random signal model. This leads to the second difference: We define the probability of error over the signal and noise distributions with the measurement matrix fixed; In [42] and [43], the probability of error is taken over the noise, the Gaussian measurement matrix and the signal support. Most of the conclusions in this paper apply to general measurement matrices and we only restrict ourselves to the Gaussian measurement matrix in Section V. Therefore, although we use a similar set of theoretical tools, the exact details of applying them are quiet different. In addition, we consider a multiple measurement model while only one temporal sample is available in [42], [43]. In particular, to get a vanishing probability of error, Aeron *et al.* [42] require to scale the SNR according to the signal dimension, which has a similar effect to having multiple temporal measurements in our paper. Although the first two differences make it difficult to compare corresponding results in these two papers, we will make some heuristic comments in Section V.

The contribution of our work is threefold. First, we introduce a hypothesis-testing framework to study the performance for multiple support recovery. We employ well-known tools in statistics and information theory such as the Chernoff bound and Fano's inequality to derive both upper and lower bounds on the probability of error. The upper bound we derive is for the *optimal* decision rule, in contrast to performance analysis for specific suboptimal reconstruction algorithms [13]–[17]. Hence, the bound can be viewed as a measure of the measurement system's ability to correctly identify the true support. Our bounds isolate important quantities that are crucial for system performance. Since our analysis is based on measurement matrices with as few assumptions as possible, the results can be used as a guidance in system design. Second, we apply these performance bounds to other more specific situations and derive necessary and sufficient conditions in terms of the system parameters to guarantee a vanishing probability of error. In particular, we study necessary conditions for accurate source localization by the mechanism proposed in [7]. By restricting our attention to Gaussian measurement matrices, we derive a result parallel to those for classical Compressive Sensing [2], [3], namely, the number of measurements that are sufficient for signal reconstruction. Even if we adopt the probability of error as the performance criterion, we get the same bound on

M as in [2], [3]. However, our result suggests that generally it is impossible to obtain the true support accurately with only one snapshot when there is considerable noise. We also obtain a necessary condition showing that the $\log \frac{N}{K}$ term cannot be dropped in Compressive Sensing. Last but not least, in the course of studying the performance bounds we explore the eigenvalue structure of a fundamental matrix in support recovery hypothesis testing for both general measurement matrices and the Gaussian measurement ensemble. These results are of independent interest.

The paper is organized as follows. In Section II, we introduce the mathematical model and briefly review the fundamental ideas in hypothesis testing. Section III is devoted to the derivation of upper bounds on the probability of error for general measurement matrices. We first derive an upper bound on the probability of error for the binary support recovery problem by employing the well-known Chernoff bound in detection theory [44] and extend it to multiple support recovery. We also study the effect of noise on system performance. In Section IV, an information theoretic lower bound is given by using the Fano's inequality, and a necessary condition is shown for the DOA problem considered in [7]. We focus on the Gaussian ensemble in Section V. Necessary and sufficient conditions on system parameters for accurate support recovery are given and their implications discussed. The paper is concluded in Section VI.

II. Notations, Models, and Preliminaries

A. Notations

We first introduce some notations used throughout this paper. Suppose $x \in \mathbb{F}^N$ is a column vector. We denote by $S = \mathrm{supp}(x) \subseteq \{1,\ldots,N\}$ the support of x, which is defined as the set of indices corresponding to the nonzero components of x. For a matrix X, $S = \mathrm{supp}\,(X)$ denotes the index set of nonzero rows of X. Here the underlying field $\mathbb{F}$ can be assumed as $\mathbb{R}$ or $\mathbb{C}$. We consider both real and complex cases simultaneously. For this purpose, we denote a constant $\kappa = 1/2$ or 1 for the real or complex case, respectively.

Suppose S is an index set. We denote by $|S|$ the number of elements in S. For any column vector $x \in \mathbb{F}^N$, x^S is the vector in $\mathbb{F}^{|S|}$ formed by the components of x indicated by the index set S; for any matrix B, B^S denotes the submatrix formed by picking the rows of B corresponding to indices in S, while B_S is the submatrix with columns from B indicated by S. If I and J are two index sets, then $B_J^I = (B^I)_J$, the submatrix of B with rows indicated by I and columns indicated by J.

Transpose of a vector or matrix is denoted by $'$ while conjugate transpose by †. $A \otimes B$ represents the Kronecker product of two matrices. For a vector v, $\mathrm{diag}(v)$ is the diagonal matrix with the elements of v in the diagonal. The identity matrix of dimension M is I_M. The trace of matrix A is given by $\mathrm{tr}(A)$, the determinant by $|A|$, and the rank by $\mathrm{rank}(A)$. Though the notation for determinant is inconsistent with that for cardinality of an index set, the exact meaning can always be understood from the context.

Bold symbols are reserved for *random* vectors and matrices. We use $\mathbb{P}$ to denote the probability of an event and $\mathbb{E}$ the expectation. The underlying probability space can be inferred from the context. Gaussian distribution for a random vector in field $\mathbb{F}$ with mean μ and covariance matrix Σ is represented by $\mathbb{F}\mathcal{N}(\mu,\Sigma)$. Matrix variate Gaussian distribution [45] for $\boldsymbol{Y} \in \mathbb{F}^{M\times T}$ with mean $\Theta \in \mathbb{F}^{M\times T}$ and covariance matrix $\Sigma \otimes \Psi$, where $\Sigma \in \mathbb{F}^{M\times M}$ and $\Psi \in \mathbb{F}^{T\times T}$, is denoted by $\mathbb{F}\mathcal{N}_{M,T}(\Theta, \Sigma \otimes \Psi)$

Suppose $\{f_n\}_{n=1}^\infty, \{g_n\}_{n=1}^\infty$ are two positive sequences, $f_n = o(g_n)$ means that $\lim_{n\to\infty} \frac{f_n}{g_n} = 0$. An alternative notation in this case is $g_n \gg f_n$. We use $f_n = O(g_n)$ to denote that there exists an $N \in \mathbb{N}$ and $C > 0$ independent of N such that $f_n \le Cg_n$ for $n \ge N$. Similarly, $f_n = \Omega(g_n)$ means $f_n \ge Cg_n$ for $n \ge N$. These simple but expedient notations introduced by G. H. Hardy greatly simplify derivations [46].

B. Models

Next, we introduce our mathematical model. Suppose $\boldsymbol{x}(t) \in \mathbb{F}^N, t = 1,\ldots,T$ are jointly sparse signals with common support; that is, only a few components of $\boldsymbol{x}(t)$ are nonzero and the indices corresponding to these nonzero components are the same for all $t = 1,\ldots,T$. The common support $S = \mathrm{supp}\,(\boldsymbol{x}(t))$ has known size $K = |S|$. We assume that the vectors $\boldsymbol{x}^S(t), t = 1,\ldots,T$ formed by the nonzero components of $\boldsymbol{x}(t)$ follow i.i.d. $\mathbb{F}\mathcal{N}(0,\mathrm{I}_K)$. The measurement model is as follows:

$$\boldsymbol{y}(t) = A\boldsymbol{x}(t) + \boldsymbol{w}(t), t = 1,2,\ldots,T \tag{2}$$

where A is the measurement matrix and $\boldsymbol{y}(t) \in \mathbb{F}^M$ the measurements. The additive noise $\boldsymbol{w}(t) \in \mathbb{F}^N$ is assumed to follow i.i.d. $\mathbb{F}\mathcal{N}\left(0,\sigma^2\mathrm{I}_M\right)$. Note that assuming unit variance for signals loses no generality since only the ratio of signal variance to noise variance appears in all subsequent analyses. In this sense, we view $1/\sigma^2$ as the signal-to-noise ratio (SNR).

Let $\boldsymbol{X} = [\boldsymbol{x}(1) \quad \boldsymbol{x}(2) \quad \cdots \quad \boldsymbol{x}(T)]$ and $\boldsymbol{Y}$, $\boldsymbol{W}$ be defined in a similar manner. Then we write the model in the more compact matrix form

$$\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{W}. \tag{3}$$

We start our analysis for general measurement matrix A. For an arbitrary measurement matrix $A \in \mathbb{F}^{M\times N}$, if every $M\times M$ submatrix of A is nonsingular, we then call A a *nondegenerate* measurement matrix. In this case, the corresponding linear system $Ax = b$ is said to have the Unique Representation Property (URP), the implication of which is discussed in [13]. While most of our results apply to general nondegenerate measurement matrices, we need to impose more structure on the measurement matrices in order to obtain more profound results. In particular, we will consider Gaussian measurement matrix $\boldsymbol{A}$ whose elements $\boldsymbol{A}_{mn}$ are generated from i.i.d. $\mathbb{F}\mathcal{N}(0,1)$. However, since our performance analysis is carried out by conditioning on a particular realization of $\boldsymbol{A}$, we still use nonbold A except in Section V. The role played by the variance of $\boldsymbol{A}_{mn}$ is indistinguishable from that of a signal variance and hence can be combined to $1/\sigma^2$, the SNR, by the note in the previous paragraph.

We now consider two hypothesis-testing problems. The first one is a binary support recovery problem

$$\begin{cases} \mathrm{H}_0 : \mathrm{supp}\,(\boldsymbol{X}) = S_0 \\ \mathrm{H}_1 : \mathrm{supp}\,(\boldsymbol{X}) = S_1 \end{cases}. \tag{4}$$

The results we obtain for binary support recovery (4) offer insight into our second problem: the multiple support recovery. In the multiple support recovery problem we choose one among $\binom{N}{K}$ distinct candidate supports of $\boldsymbol{X}$, which is a multiple-hypothesis testing problem

$$\begin{cases} \mathrm{H}_0: & \mathrm{supp}\,(\boldsymbol{X}) = S_0 \\ \mathrm{H}_1: & \mathrm{supp}\,(\boldsymbol{X}) = S_1 \\ & \vdots \\ \mathrm{H}_{L-1}: & \mathrm{supp}\,(\boldsymbol{X}) = S_{L-1}. \end{cases} \tag{5}$$

C. Preliminaries for Hypothesis Testing

We now briefly introduce the fundamentals of hypothesis testing. The following discussion is based mainly on [44]. In a simple binary hypothesis test, the goal is to determine which of two candidate distributions is the true one that generates the data matrix (or vector) $\boldsymbol{Y}$

$$\begin{cases} \mathrm{H}_0: \boldsymbol{Y} \sim p\,(\boldsymbol{Y}|\mathrm{H}_0) \\ \mathrm{H}_1: \boldsymbol{Y} \sim p\,(\boldsymbol{Y}|\mathrm{H}_1) \end{cases}. \tag{6}$$

There are two types of errors when one makes a choice based on the observed data $\boldsymbol{Y}$. A *false alarm* corresponds to choosing H_1 when H_0 is true, while a *miss* happens by choosing H_0 when H_1 is true. The probabilities of these two types of errors are called the probability of a false alarm and the probability of a miss, which are denoted by

$$P_{\mathrm{F}} = \mathbb{P}\,(\mathrm{Choose}\,\mathrm{H}_1|\mathrm{H}_0) \tag{7}$$

$$P_{\mathrm{M}} = \mathbb{P}\,(\mathrm{Choose}\,\mathrm{H}_0|\mathrm{H}_1) \tag{8}$$

respectively. Depending on whether one knows the prior probabilities $\mathbb{P}(\mathrm{H}_0)$ and $\mathbb{P}(\mathrm{H}_1)$ and assigns losses to errors, different criteria can be employed to derive the optimal decision rule. In this paper we adopt the probability of error with equal prior probabilities of H_0 and H_1 as the decision criterion; that is, we try to find the optimal decision rule by minimizing

$$P_{\mathrm{err}} = P_{\mathrm{F}}\mathbb{P}(\mathrm{H}_0) + P_{\mathrm{M}}\mathbb{P}(\mathrm{H}_1) = \frac{1}{2}P_{\mathrm{F}} + \frac{1}{2}P_{\mathrm{M}}. \tag{9}$$

The optimal decision rule is then given by the likelihood ratio test

$$\ell(\boldsymbol{Y}) = \log\frac{p\,(\boldsymbol{Y}|\mathrm{H}_1)}{p\,(\boldsymbol{Y}|\mathrm{H}_0)} \underset{\mathrm{H}_0}{\overset{\mathrm{H}_1}{\gtrless}} 0 \tag{10}$$

where $\log(\cdot)$ is the natural logarithm function. The probability of error associated with the optimal decision rule, namely, the likelihood ratio test (10), is a measure of the best performance a system can achieve. In many cases of interest, the simple binary hypothesis testing problem (6) is derived from a signal-generation system. For example, in a digital communication system, hypotheses H_0 and H_1 correspond to the transmitter sending digit 0 and 1, respectively, and the distributions of the observed data under the hypotheses are determined by the modulation method of the system. Therefore, the minimal probability of error achieved by the likelihood ratio test is a measure of the performance of the modulation method. For the problem addressed in this paper, the minimal probability of error reflects the measurement matrix's ability to distinguish different signal supports.

The Chernoff bound [44] is a well-known tight upper bound on the probability of error. In many cases, the optimum test can be derived and implemented efficiently but an exact performance calculation is impossible. Even if such an expression can be derived, it is too complicated to be of practical use. For this reason, sometimes a simple bound turns out to be more useful in many problems of practical importance. The Chernoff bound, based on the moment generating function of the test statistic $\ell(\boldsymbol{Y})$ (10), provides an easy way to compute such a bound.

Define $\mu(s)$ as the logarithm of the moment generating function of $\ell(\boldsymbol{Y})$

$$\begin{aligned} \mu(s) &\triangleq \log\int_{-\infty}^{\infty} e^{s\ell(\boldsymbol{Y})} p(\boldsymbol{Y}|\mathrm{H}_0)d\boldsymbol{Y} \\ &= \log\int_{-\infty}^{\infty} [p(\boldsymbol{Y}|\mathrm{H}_1)]^s [p(\boldsymbol{Y}|\mathrm{H}_0)]^{1-s} d\boldsymbol{Y}. \end{aligned} \tag{11}$$

Then the Chernoff bound states that

$$P_{\mathrm{F}} \le \exp[\mu(s_m)] \le \exp[\mu(s)] \tag{12}$$

$$P_{\mathrm{M}} \le \exp[\mu(s_m)] \le \exp[\mu(s)] \tag{13}$$

and

$$P_{\mathrm{err}} \le \frac{1}{2}\exp[\mu(s_m)] \le \frac{1}{2}\exp[\mu(s)] \tag{14}$$

where $0 \le s \le 1$ and $s_m = \mathrm{argmin}_{0\le s\le 1}\mu(s)$. Note that a refined argument gives the constant 1/2 in (14) instead of 1 as obtained by direct application of (12) and (13) [44]. We use these bounds to study the performance of the support recovery problem.

We next extend to multiple-hypothesis testing the key elements of the binary hypothesis testing. The goal in a simple multiple-hypothesis testing problem is to make a choice among L distributions based on the observations

$$\begin{cases} \mathrm{H}_0: & \boldsymbol{Y} \sim p\,(\boldsymbol{Y}|\mathrm{H}_0) \\ \mathrm{H}_1: & \boldsymbol{Y} \sim p\,(\boldsymbol{Y}|\mathrm{H}_1) \\ & \vdots \\ \mathrm{H}_{L-1}: & \boldsymbol{Y} \sim p\,(\boldsymbol{Y}|\mathrm{H}_{L-1}) \end{cases}. \tag{15}$$

Using the total probability of error as a decision criterion and assuming equal prior probabilities for all hypotheses, we obtain the optimal decision rule given by

$$\mathrm{H}^* = \mathrm{argmax}_{0\le i\le L-1} p\,(\boldsymbol{Y}|\mathrm{H}_i). \tag{16}$$

Application of the union bound and the Chernoff bound (14) shows that the total probability of error is bounded as follows:

$$\begin{aligned} P_{\mathrm{err}} &= \sum_{i=0}^{L-1} \mathbb{P}(\mathrm{H}^* \ne \mathrm{H}_i|\mathrm{H}_i)\mathbb{P}(\mathrm{H}_i) \\ &\le \frac{1}{2L}\sum_{i=0}^{L-1}\sum_{\substack{j=0 \\ j\ne i}}^{L-1} \exp[\mu(s;\mathrm{H}_i,\mathrm{H}_j)],\ 0 \le s \le 1 \end{aligned} \tag{17}$$

where $\exp[\mu(s;\mathrm{H}_i,\mathrm{H}_j)]$ is the moment-generating function in the binary hypothesis testing problem for H_i and H_j. Hence, we

obtain an upper bound for multiple-hypothesis testing from that for binary hypothesis testing.

III. Upper Bound on Probability of Error for Nondegenerate Measurement Matrices

In this section, we apply the general theory for hypothesis testing, the Chernoff bound on the probability of error in particular, to the support recovery problems (4) and (5). We first study binary support recovery, which lays the foundation for the general support recovery problem.

A. Binary Support Recovery

Under model (3) and the assumptions pertaining to it, observations $\boldsymbol{Y}$ follow a matrix variate Gaussian distribution [45] when the true support is S

$$\boldsymbol{Y}|S \sim \mathbb{FN}_{M,T}(0, \Sigma_S \otimes \mathrm{I}_T) \tag{18}$$

with the probability density function (pdf) given by

$$p\left(\boldsymbol{Y}|S\right) = \frac{1}{(\pi/\kappa)^{\kappa MT}|\Sigma_S|^{\kappa T}} \exp\left[-\kappa \mathrm{tr}\left(\boldsymbol{Y}^\dagger \Sigma_S^{-1} \boldsymbol{Y}\right)\right] \tag{19}$$

where $\Sigma_S = A_S A_S^\dagger + \sigma^2 \mathrm{I}_M$ is the common covariance matrix for each column of $\boldsymbol{Y}$. The binary support recovery problem (4) is equivalent to a linear Gaussian binary hypothesis testing problem

$$\begin{cases} \mathrm{H}_0 : \boldsymbol{Y} \sim \mathbb{FN}_{M,T}(0, \Sigma_{S_0} \otimes \mathrm{I}_T) \\ \mathrm{H}_1 : \boldsymbol{Y} \sim \mathbb{FN}_{M,T}(0, \Sigma_{S_1} \otimes \mathrm{I}_T) \end{cases}. \tag{20}$$

From now on, for notation simplicity we will denote Σ_{S_i} by Σ_i. The optimal decision rule with minimal probability of error given by the likelihood ratio test $\ell(\boldsymbol{Y})$ (10) reduces to

$$-\kappa \mathrm{tr}\left[\boldsymbol{Y}^\dagger \left(\Sigma_1^{-1} - \Sigma_0^{-1}\right) \boldsymbol{Y}\right] - \kappa T \log \frac{|\Sigma_1|}{|\Sigma_0|} \underset{\mathrm{H}_0}{\overset{\mathrm{H}_1}{\gtrless}} 0. \tag{21}$$

To analyze the performance of the likelihood ratio test (21), we first compute the log-moment-generating function of $\ell(\boldsymbol{Y})$ according to (11)

$$\begin{aligned} &\mu(s) \qquad (22)\\ &= \log \int \left[p\left(\boldsymbol{Y}|\mathrm{H}_1\right)\right]^s \left[p\left(\boldsymbol{Y}|\mathrm{H}_0\right)\right]^{1-s} d\boldsymbol{Y} \\ &= \log \left[\frac{1}{(\pi/\kappa)^{\kappa MT}|\Sigma_1|^{\kappa sT}|\Sigma_0|^{\kappa(1-s)T}} \right. \\ &\quad \left. \times \int \exp\left\{-\kappa \mathrm{tr}\left[\boldsymbol{Y}^\dagger \left(s\Sigma_1^{-1} + (1-s)\Sigma_0^{-1}\right)\boldsymbol{Y}\right]\right\} d\boldsymbol{Y}\right] \\ &= \log \left[\frac{\left|s\Sigma_1^{-1} + (1-s)\Sigma_0^{-1}\right|^{-\kappa T}}{|\Sigma_1|^{\kappa sT}|\Sigma_0|^{\kappa(1-s)T}}\right] \\ &= -\kappa T \log\left|sH^{1-s} + (1-s)H^{-s}\right|, \quad 0 \le s \le 1 \qquad (23) \end{aligned}$$

where $H = \Sigma_0^{1/2}\Sigma_1^{-1}\Sigma_0^{1/2}$. The computation of the exact minimizer $s_m = \mathrm{argmin}_{0\le s\le 1}\mu(s)$ is nontrivial and will lead to an expression of $\mu(s_m)$ too complicated to handle. When $|S_0| = |S_1|$ and the columns of A are not highly correlated, for example in the case of A with i.i.d. elements, $s_m \approx \frac{1}{2}$. We then take $s = \frac{1}{2}$ in the Chernoff bounds (12), (13), and (14). Whereas the bounds obtained in this way may not be the absolute best ones, they are still valid.

As positive definite Hermitian matrices, H and H^{-1} can be simultaneously diagonalized by a unitary transformation. Suppose that the eigenvalues of H are $\lambda_1 \ge \cdots \ge \lambda_{k_0} > 1 = \cdots = 1 > \sigma_1 \ge \cdots \ge \sigma_{k_1}$ and $D = \mathrm{diag}[\lambda_1, \ldots, \lambda_{k_0}, 1, \ldots, 1, \sigma_1, \ldots, \sigma_{k_1}]$. Then it is easy to show that

$$\begin{aligned} \mu(1/2) &= -\kappa T \log\left|\frac{D^{1/2} + D^{-1/2}}{2}\right| \\ &= -\kappa T\left[\sum_{j=1}^{k_0} \log\left(\frac{\sqrt{\lambda_j} + 1/\sqrt{\lambda_j}}{2}\right)\right. \\ &\quad \left. + \sum_{j=1}^{k_1} \log\left(\frac{\sqrt{\sigma_j} + 1/\sqrt{\sigma_j}}{2}\right)\right]. \end{aligned} \tag{24}$$

Therefore, it is necessary to count the numbers of eigenvalues of H that are greater than 1, equal to 1 and less than 1, i.e., the values of k_0 and k_1 for general nondegenerate measurement matrix A. We have the following theorem on the eigenvalue structure of H.

Proposition 1: For any nondegenerate measurement matrix A, let $H = \Sigma_0^{1/2}\Sigma_1^{-1}\Sigma_0^{1/2}$, $k_\mathrm{i} = |S_0 \cap S_1|$, $k_0 = |S_0 \backslash S_1| = |S_0| - k_\mathrm{i}$, $k_1 = |S_1 \backslash S_0| = |S_1| - k_\mathrm{i}$ and assume $M \geqslant k_0 + k_1$; then k_0 eigenvalues of matrix H are greater than 1, k_1 less than 1, and $M - (k_0 + k_1)$ equal to 1.

Proof: See Appendix A.

For binary support recovery (4) with $|S_0| = |S_1| = K$, we have $k_0 = k_1 \triangleq k_\mathrm{d}$. The subscripts i and d in k_i and k_d are short for "intersection" and "difference," respectively. Employing the Chernoff bounds (14) and Proposition 1, we have the following proposition.

Proposition 2: If $M \ge 2k_\mathrm{d}$, the probability of error for the binary support recovery problem (4) is bounded by

$$P_\mathrm{err} \le \frac{1}{2}\left[\frac{\bar{\lambda}_{S_0,S_1}\bar{\lambda}_{S_1,S_0}}{16}\right]^{-\kappa k_\mathrm{d} T/2} \tag{25}$$

where $\bar{\lambda}_{S_i,S_j}$ is the geometric mean of the eigenvalues of $H = \Sigma_i^{1/2}\Sigma_j^{-1}\Sigma_i^{1/2}$ that are greater than one.

Proof: According to (14) and (24), we have

$$\begin{aligned} P_\mathrm{err} &\le \frac{1}{2}\exp\left[\mu\left(\frac{1}{2}\right)\right] \\ &\le \frac{1}{2}\left[\prod_{j=1}^{k_\mathrm{d}}\left(\frac{\sqrt{\lambda_j}}{2}\right)\prod_{j=1}^{k_\mathrm{d}}\left(\frac{1}{\sqrt{\sigma_j}/2}\right)\right]^{-\kappa T} \\ &= \frac{1}{2}\left[\frac{\left(\prod_{j=1}^{k_\mathrm{d}}\lambda_j\right)^{1/k_\mathrm{d}}\left(\prod_{j=1}^{k_\mathrm{d}}\frac{1}{\sigma_j}\right)^{1/k_\mathrm{d}}}{16}\right]^{-\kappa k_\mathrm{d} T/2}. \end{aligned}$$

Define $\bar{\lambda}_{S_i,S_j}$ as the geometric mean of the eigenvalues of $H = \Sigma_i^{1/2}\Sigma_j^{-1}\Sigma_i^{1/2}$ that are greater than one. Then obviously we have $\bar{\lambda}_{S_0,S_1} = (\prod_{j=1}^{k_\mathrm{d}} \lambda_j)^{1/k_\mathrm{d}}$. Since H^{-1} and $\Sigma_1^{1/2}\Sigma_0^{-1}\Sigma_1^{1/2}$ have the same set of eigenvalues, $1/\sigma_j, j = 1, \ldots, k_\mathrm{d}$ are the eigenvalues of $\Sigma_1^{1/2}\Sigma_0^{-1}\Sigma_1^{1/2}$ that are greater than 1. We conclude that $\bar{\lambda}_{S_1,S_0} = (\prod_{j=1}^{k_\mathrm{d}} 1/\sigma_j)^{1/k_\mathrm{d}}$. □

Note that $\bar{\lambda}_{S_0,S_1}$ and $\bar{\lambda}_{S_1,S_0}$ completely determine the measurement system (3) 's performance in differentiating two different signal supports. It must be larger than the constant 16 for a vanishing bound when more temporal samples are taken. Once the threshold 16 is exceeded, taking more samples will drive the probability of error to 0 exponentially fast. From numerical simulations and our results on the Gaussian measurement matrix, $\bar{\lambda}_{S_i,S_j}$ does not vary much when S_i, S_j and k_d change, as long as the elements in the measurement matrix A are highly uncorrelated.[1] Therefore, quite appealing to intuition, the larger the size k_d of the difference set between the two candidate supports, the smaller the probability of error.

B. Multiple Support Recovery

Now we are ready to use the union bound (17) to study the probability of error for the multiple support recovery problem (5). We assume each candidate support S_i has known cardinality K, and we have $L = \binom{N}{K}$ such supports. Our general approach is also applicable to cases for which we have some prior information on the structure of the signal's sparsity pattern, for example the setup in model-based Compressive Sensing[47]. In these cases, we usually have $L \ll \binom{N}{K}$ supports, and a careful examination on the intersection pattern of these supports will give a better bound. However, in this paper we will not address this problem and will instead focus on the full support recovery problem with $L = \binom{N}{K}$. Defining $\bar{\lambda} = \min_{i\neq j}\{\bar{\lambda}_{S_i,S_j}\}$, we have the following theorem.

Theorem 1: If $M \geq 2K$ and $\bar{\lambda} > 4[K(N-K)]^{\frac{1}{\kappa T}}$, then the probability of error for the full support recovery problem (5) with $|S_i| = K$ and $L = \binom{N}{K}$ is bounded by

$$P_\mathrm{err} \leq \frac{1}{2}\frac{\frac{K(N-K)}{\left(\bar{\lambda}/4\right)^{\kappa T}}}{1-\frac{K(N-K)}{\left(\bar{\lambda}/4\right)^{\kappa T}}}. \tag{26}$$

Proof: Combining the bound in Proposition 2 and (17), we have

$$\begin{aligned} P_\mathrm{err} &\leq \frac{1}{2L}\sum_{i=0}^{L-1}\sum_{\substack{j=1\\ j\neq i}}^{L-1}\left[\frac{\bar{\lambda}_{S_i,S_j}\bar{\lambda}_{S_j,S_i}}{16}\right]^{-\kappa k_\mathrm{d} T/2} \\ &\leq \frac{1}{2L}\sum_{i=0}^{L-1}\sum_{\substack{j=1\\ j\neq i}}^{L-1}\left(\frac{\bar{\lambda}}{4}\right)^{-\kappa k_\mathrm{d} T}. \end{aligned}$$

Here k_d depends on the supports S_i and S_j. For fixed S_i, the number of supports that have a difference set with S_i with cardinality k_d is $\binom{K}{k_\mathrm{d}}\binom{N-K}{k_\mathrm{d}}$. Therefore, using $\binom{K}{k_\mathrm{d}} \leq K^{k_\mathrm{d}}$ and $\binom{N-K}{k_\mathrm{d}} \leq (N-K)^{k_\mathrm{d}}$ and the summation formula for geometric series, we obtain

$$\begin{aligned} P_\mathrm{err} &\leq \frac{1}{2L}\sum_{i=0}^{L-1}\sum_{k_\mathrm{d}=1}^{K}\binom{K}{k_\mathrm{d}}\binom{N-K}{k_\mathrm{d}}\left(\frac{\bar{\lambda}}{4}\right)^{-\kappa k_\mathrm{d} T} \\ &\leq \frac{1}{2}\sum_{k_\mathrm{d}=1}^{K}\left[\frac{K(N-K)}{\left(\bar{\lambda}/4\right)^{\kappa T}}\right]^{k_\mathrm{d}} \\ &\leq \frac{1}{2}\frac{\frac{K(N-K)}{\left(\bar{\lambda}/4\right)^{\kappa T}}}{1-\frac{K(N-K)}{\left(\bar{\lambda}/4\right)^{\kappa T}}}. \end{aligned}$$

□

We make several comments here. First, $\bar{\lambda}$ depends only on the measurement matrix A and the variance of noise σ^2. Compared with the results in [43], where the bounds involve the signal, we get more insight into what quantity of the measurement matrix is important in support recovery. This information is obtained by modelling the signals $\boldsymbol{x}(t)$ as Gaussian random vectors. The quantity $\bar{\lambda}$ effectively characterizes system (3)'s ability to distinguish different supports. Clearly, $\bar{\lambda}$ is related to the restricted isometry property (RIP), which guarantees stable sparse signal recovery in Compressive Sensing [4]–[6]. We discuss the relationship between RIP and $\bar{\lambda}$ for the special case with $K = 1$ at the end of Section III-C. However, a precise relationship for the general case is yet to be discovered.

Second, we observe that increasing the number of temporal samples plays two roles simultaneously in the measurement system. For one thing, it decreases the the threshold $4[K(N-K)]^{\frac{1}{\kappa T}}$ that $\bar{\lambda}$ must exceed for the bound (26) to hold. However, since $\lim_{T\to\infty} 4[K(N-K)]^{\frac{1}{\kappa T}} = 4$ for fixed K and N, increasing temporal samples can reduce the threshold only to a certain limit. For another, since the bound (26) is proportional to $e^{-T\log(\bar{\lambda}/4)}$, the probability of error turns to 0 exponentially fast as T increases, as long as $\bar{\lambda} > 4[K(N-K)]^{\frac{1}{\kappa T}}$ is satisfied.

In addition, the final bound (26) is of the same order as the probability of error when $k_\mathrm{d} = 1$. The probability of error P_err is dominated by the probability of error in cases for which the estimated support differs by only one index from the true support, which are the most difficult cases for the decision rule to make a choice. However, in practice we can imagine that these cases induce the least loss. Therefore, if we assign weights/costs to the errors based on k_d, then the weighted probability of error or average cost would be much lower. For example, we can choose the costs to exponentially decrease when k_d increases. Another possible choice of cost function is to assume zero cost when k_d is below a certain critical number. Our results can be easily extended to these scenarios.

Finally, note that our bound (26) applies to any nondegenerate matrix. In Section V, we apply the bound to Gaussian measurement matrices. The additional structure allows us to derive more profound results on the behavior of the bound.

C. The Effect of Noise

In this subsection, we explore how the noise variance affects the probability of error, which is equivalent to analyzing the behavior of $\bar{\lambda}_{S_i,S_j}$ and $\bar{\lambda}$ as indicated in (25) and (26).

[1]Unfortunately, this is not the case when the columns of A are samples from uniform linear sensor array manifold.

We now derive bounds on the eigenvalues of H. The lower bound is expressed in terms of the QR decomposition of a submatrix of the measurement matrix with the noise variance σ^2 isolated.

Proposition 3: For any nondegenerate measurement matrix A, let $H = \Sigma_0^{1/2}\Sigma_1^{-1}\Sigma_0^{1/2}$ with $\Sigma_i = A_{S_i}A_{S_i}^\dagger + \sigma^2 \mathrm{I}_M$, $k_\mathrm{i} = |S_0 \cap S_1|$, $k_0 = |S_0 \backslash S_1| = |S_0| - k_\mathrm{i}$, $k_1 = |S_1 \backslash S_0| = |S_1| - k_\mathrm{i}$. We have the following:

1) if $M \geqslant k_0 + k_1$, then the sorted eigenvalues of H that are greater than 1 are lower bounded by the corresponding eigenvalues of $\mathrm{I}_{k_0} + \frac{1}{\sigma^2} R_{33} R_{33}^\dagger$, where R_{33} is the $k_0 \times k_0$ submatrix at the lower-right corner of the upper triangle matrix in the QR decomposition of $[\, A_{S_1 \backslash S_0} \quad A_{S_1 S_0} \quad A_{S_0 \backslash S_1} \,]$;
2) the eigenvalues of H are upper bounded by the corresponding eigenvalues of $\mathrm{I}_M + \frac{1}{\sigma^2} A_{S_0 \backslash S_1} A_{S_0 \backslash S_1}^\dagger$; in particular, the sorted eigenvalues of H that are greater than 1 are upper bounded by the corresponding ones of $\mathrm{I}_{k_0} + \frac{1}{\sigma^2} A_{S_0 \backslash S_1}^\dagger A_{S_0 \backslash S_1}$.

Proof: See Appendix B.

The importance of this proposition is twofold. First, by isolating the noise variance from the expression of matrix H, this theorem clearly shows that when noise variance decreases to zero, the relatively large eigenvalues of H will blow up, which results in increased performance in support recovery. Second, the bounds provide ways to analyze special measurement matrices, especially the Gaussian measurement ensemble discussed in Section V.

We have the following corollary.

Corollary 1: For support recovery problems (4) and (5) with support size K, suppose $M \geq 2K$; then there exist constants $c_1, c_2 > 0$ that depend only on the measurement matrix A such that

$$1 + \frac{c_2}{\sigma^2} \geq \bar{\lambda} \geq 1 + \frac{c_1}{\sigma^2}. \tag{27}$$

From (25) and (26), we then conclude that for any temporal sample size T

$$\lim_{\sigma^2 \to 0} P_\mathrm{err} = 0 \tag{28}$$

and the speed of convergence is approximately $(\sigma^2)^{\kappa k_\mathrm{d} T}$ and $(\sigma^2)^{\kappa T}$ for the binary and multiple cases, respectively.

Proof: According to Proposition 3, for any fixed S_i, S_j, the eigenvalues of $H = \Sigma_i^{1/2}\Sigma_j^{-1}\Sigma_i^{1/2}$ that are greater than 1 are lower bounded by those of $\mathrm{I}_{k_\mathrm{d}} + \frac{1}{\sigma^2} R_{33} R_{33}^\dagger$; hence, we have

$$\begin{aligned}
\bar{\lambda}_{S_i,S_j} &\geq \left| \mathrm{I}_{k_\mathrm{d}} + \frac{1}{\sigma^2} R_{33} R_{33}^\dagger \right|^{1/k_\mathrm{d}} \\
&\geq \left| \mathrm{I}_{k_\mathrm{d}} + \frac{1}{\sigma^2} R_{33}^2 \right|^{1/k_\mathrm{d}} \\
&= \left[\prod_{l=1}^{k_\mathrm{d}} \left(1 + \frac{1}{\sigma^2} r_{ll}^2 \right) \right]^{1/k_\mathrm{d}} \\
&\geq 1 + \frac{1}{\sigma^2} \left(\prod_{l=1}^{k_\mathrm{d}} r_{ll}^2 \right)^{1/k_\mathrm{d}}
\end{aligned} \tag{29}$$

where r_{ll} is the lth diagonal element of R_{33}. For the second inequality we have used [48, Fact 8. 11. 20]. Since A is nondegenerate and $M \geq 2K$, $[\, A_{S_j \backslash S_i} \quad A_{S_j S_i} \quad A_{S_i \backslash S_j} \,]$ is of full rank and $r_{ll}^2 > 0, 0 \leq l \leq k_\mathrm{d}$ for all S_i, S_j. Defining c_1 as the minimal value of $\left(\prod_{l=1}^{k_\mathrm{d}} r_{ll}^2 \right)^{1/k_\mathrm{d}}$'s over all possible support pairs S_i, S_j, we then have $c_1 > 0$ and

$$\bar{\lambda} \geq 1 + \frac{c_1}{\sigma^2}.$$

On the other hand, the upper bound on the eigenvalues of H yields

$$\begin{aligned}
\bar{\lambda}_{S_i,S_j} &\leq \left| \mathrm{I}_{k_\mathrm{d}} + \frac{1}{\sigma^2} A_{S_i \backslash S_j}^\dagger A_{S_i \backslash S_j} \right|^{1/k_\mathrm{d}} \\
&\leq 1 + \frac{1}{\sigma^2 k_\mathrm{d}} \mathrm{tr} \left(A_{S_i \backslash S_j}^\dagger A_{S_i \backslash Sj} \right) \\
&= 1 + \frac{1}{\sigma^2 k_\mathrm{d}} \sum_{\substack{1 \leq m \leq M \\ n \in S_i \backslash S_j}} |A_{mn}|^2 .
\end{aligned} \tag{30}$$

Therefore, we have

$$\bar{\lambda} \leq 1 + \frac{c_2}{\sigma^2}$$

with $c_2 = \max_{S:|S| \leq K} \frac{1}{|S|} \sum_{\substack{1 \leq m \leq M \\ n \in S}} |A_{mn}|^2$. All other statements in the theorem follows immediately from (25) and (26). □

Corollary 1 suggests that in the limiting case where there is no noise, $M \geq 2K$ is sufficient to recover a K–sparse signal. This fact has been observed in [4]. Our result also shows that the optimal decision rule, which is unfortunately inefficient, is robust to noise. Another extreme case is when the noise variance σ^2 is very large. Then from $\log(1+x) \approx x, 0 < x \ll 1$, the bounds in (25) and (26) are approximated by $e^{-\kappa k_\mathrm{d} T/\sigma^2}$ and $e^{-\kappa T/\sigma^2}$. Therefore, the convergence exponents for the bounds are proportional to the SNR in this limiting case.

The diagonal elements of R_{33}, r_{ll}'s, have clear meanings. Since QR factorization is equivalent to the Gram-Schmidt orthogonalization procedure, r_{11} is the distance of the first column of A_{S_i/S_j} to the subspace spanned by the columns of A_{S_j}; r_{22} is the distance of the second column of A_{S_i/S_j} to the subspace spanned by the columns of A_{S_j} plus the first column of A_{S_i/S_j}, and so on. Therefore, $\bar{\lambda}_{S_i,S_j}$ is a measure of how well the columns of A_{S_i/S_j} can be expressed by the columns of A_{S_j}, or, put another way, a measure of the incoherence between the columns of A_{S_i} and A_{S_j}. Similarly, $\bar{\lambda}$ is an indicator of the incoherence of the entire matrix A of order K. To relate $\bar{\lambda}$ with the incoherence, we consider the case with $K = 1$ and $\mathbb{F} = \mathbb{R}$. By restricting our attention to matrices with *unit* columns, the above discussion implies that a better bound is achieved if the minimal distance of all pairs of column vectors of matrix A is maximized. Finding such a matrix A is equivalent to finding a matrix with the inner product between columns as large as possible, since the distance between two unit vectors u and v is $2 - 2|< u, v >|$ where $< u, v >= u'v$ is the inner product

between u and v. For each integer s, the RIP constant δ_s is defined as the smallest number such that [4], [5]

$$1-\delta_s \le \frac{\|Ax\|_2^2}{\|x\|_2^2} \le 1+\delta_s, \qquad |\mathrm{supp}(x)| = s. \tag{31}$$

A direct computation shows that δ_2 is equal to the minimum of the absolute values of the inner products between all pairs of columns of A. Hence, the requirements of finding the smallest δ_2 that satisfies (31) and maximizing $\bar{\lambda}$ coincide when $K=1$. For general K, Milenkovic *et al.* established a relationship between δ_2 and δ_K via Geršgorin's disc theorem [49] and discussed them as well as some coding theoretic issues in Compressive Sensing context [50].

IV. An Information Theoretic Lower Bound on Probability of Error

In this section, we derive an information theoretic lower bound on the probability of error for *any* decision rule in the multiple support recovery problem. The main tool is a variant of the well-known Fano's inequality[51]. In the variant, the average probability of error in a multiple-hypothesis testing problem is bounded in terms of the Kullback-Leibler divergence[52]. Suppose that we have a random vector or matrix $\boldsymbol{Y}$ with L possible densities $f_0,\ldots,f_{L-1}$. Denote the average of the Kullback-Leibler divergence between any pair of densities by

$$\beta = \frac{1}{L^2}\sum_{i,j} D_{KL}(f_i\|f_j). \tag{32}$$

Then by Fano's inequality [43], [53], the probability of error (17) for *any* decision rule to identify the true density is lower bounded by

$$P_{\mathrm{err}} \ge 1 - \frac{\beta+\log 2}{\log L}. \tag{33}$$

Since in the multiple support recovery problem (5), all the distributions involved are matrix variate Gaussian distributions with zero mean and different variances, we now compute the Kullback-Leibler divergence between two matrix variate Gaussian distributions. Suppose $f_i = \mathbb{F}\mathcal{N}_{M,T}(0,\Sigma_i\otimes I_T)$, $f_j = \mathbb{F}\mathcal{N}_{M,T}(0,\Sigma_j\otimes I_T)$, the Kullback-Leibler divergence has closed form expression

$$\begin{aligned} &D_{KL}(f_i\|f_j)\\ &\quad=\mathbb{E}_{f_i}\log\frac{f_i}{f_j}\\ &\quad=\frac{1}{2}\,\mathbb{E}_{f_i}\left[-\kappa\mathrm{tr}\left[\boldsymbol{Y}^\dagger\left(\Sigma_i^{-1}-\Sigma_j^{-1}\right)\boldsymbol{Y}\right]-\kappa T\log\frac{|\Sigma_i|}{|\Sigma_j|}\right]\\ &\quad=\frac{1}{2}\kappa T\left[\mathrm{tr}\left(H_{i,j}-\mathrm{I}_M\right)+\log\frac{|\Sigma_j|}{|\Sigma_i|}\right] \end{aligned}$$

where $H_{i,j}=\Sigma_i^{1/2}\Sigma_j^{-1}\Sigma_i^{1/2}$. Therefore, we obtain the average Kullback-Leibler divergence (32) for the multiple support recovery problem as

$$\begin{aligned} \beta &= \frac{1}{L^2}\sum_{S_i,S_j}\frac{1}{2}\kappa T\left[\mathrm{tr}(H_{i,j})-M+\log\frac{|\Sigma_j|}{|\Sigma_i|}\right]\\ &= \frac{\kappa T}{2L^2}\sum_{S_i,S_j}\left[\mathrm{tr}(H_{i,j})-M\right] \end{aligned}$$

where the $\log\frac{|\Sigma_j|}{|\Sigma_i|}$ terms all cancel out and $L=\binom{N}{K}$. Invoking the second part of Proposition 3, we get

$$\begin{aligned} \mathrm{tr}\,(H_{i,j}) &\le \mathrm{tr}\left(\mathrm{I}_M+\frac{1}{\sigma^2}A_{S_i\setminus S_j}A_{S_i\setminus S_j}^\dagger\right)\\ &= M+\frac{1}{\sigma^2}\sum_{\substack{1\le m\le M\\ n\in S_i\setminus S_j}}|A_{mn}|^2. \end{aligned}$$

Therefore, the average Kullback-Leibler divergence is bounded by

$$\beta \le \frac{\kappa T}{2\sigma^2L^2}\sum_{S_i,S_j}\sum_{\substack{1\le m\le M\\ n\in S_i\setminus S_j}}|A_{mn}|^2.$$

Due to the symmetry of the right-hand side, it must be of the form $a\sum_{\substack{1\le m\le M\\ 1\le n\le N}}|A_{mn}|^2 = a\,\|A\|_{\mathrm{F}}^2$, where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. Setting all $A_{mn}=1$ gives

$$\begin{aligned} &\frac{\kappa T}{2\sigma^2L^2}\sum_{S_i,S_j}\sum_{\substack{1\le m\le M\\ n\in S_i\setminus S_j}}1\\ &\quad=\frac{\kappa T}{2\sigma^2L^2}\sum_{i=0}^{L-1}\sum_{k_{\mathrm{d}}=1}^{K}\binom{K}{k_{\mathrm{d}}}\binom{N-K}{k_{\mathrm{d}}}k_{\mathrm{d}}M\\ &\quad=aMN. \end{aligned}$$

Therefore, we get $a=\frac{\kappa TK(N-K)}{2\sigma^2N^2}$ using the mean expression for hypergeometric distribution

$$\sum_{k_{\mathrm{d}}=1}^{K}\frac{\binom{K}{k_{\mathrm{d}}}\binom{N-K}{k_{\mathrm{d}}}}{\binom{N}{K}}k_{\mathrm{d}} = \frac{K\,(N-K)}{N}.$$

Hence, we have

$$\beta \le \frac{\kappa TK\,(N-K)}{2\sigma^2N^2}\,\|A\|_{\mathrm{F}}^2.$$

Therefore, the probability of error is lower bounded by

$$P_{\mathrm{err}} \ge 1-\frac{\frac{\kappa TK(N-K)}{2\sigma^2N^2}\,\|A\|_{\mathrm{F}}^2+\log 2}{\log L}. \tag{34}$$

We conclude with the following theorem.

Theorem 2: For multiple support recovery problem (5), the probability of error for any decision rule is lower bounded by

$$P_{\mathrm{err}} \geq 1 - \frac{\kappa T \frac{K}{N}\left(1-\frac{K}{N}\right)\|A\|_{\mathrm{F}}^2}{2\sigma^2 \log\binom{N}{K}} + o(1). \tag{35}$$

Each term in bound (35) has clear meanings. The Frobenius norm of measurement matrix $\|A\|_{\mathrm{F}}^2$ is total gain of system (2). Since the measured signal is K–sparse, only a fraction of the gain plays a role in the measurement, and its average over all possible K–sparse signals is $\frac{K}{N}\|A\|_{\mathrm{F}}^2$. While an increase in signal energy enlarges the distances between signals, a penalty term $\left(1-\frac{K}{N}\right)$ is introduced because we now have more signals. The term $\log L = \log\binom{N}{K}$ is the total uncertainty or entropy of the support variable $\boldsymbol{S}$, since we impose a uniform prior on it. As long as $K \leq \frac{N}{2}$, increasing K increases both the average gain exploited by the measurement system, and the entropy of the support variable $\boldsymbol{S}$. The overall effect, quite counterintuitively, is a decrease of the lower bound in (35). Actually, the term involving K, $\frac{\frac{K}{N}(1-\frac{K}{N})}{\log\binom{N}{K}}$, is approximated by an increasing function $\frac{\alpha(1-\alpha)}{NH(\alpha)}$ with $\alpha = \frac{K}{N}$ and the binary entropy function $H(\alpha) = -\alpha\log\alpha - (1-\alpha)\log(1-\alpha)$. The reason for the decrease of the bound is that the bound only involves the *effective* SNR without regard to any inner structure of A (e.g., the incoherence) and the effective SNR increases with K. To see this, we compute the effective SNR as $\mathbb{E}_{\boldsymbol{x}}\left[\frac{\|A\boldsymbol{x}\|_{\mathrm{F}}^2}{M\sigma^2}\right] = \frac{\frac{K}{N}\|A\|_{\mathrm{F}}^2}{M\sigma^2}$. If we scale down the effective SNR through increasing the noise energy σ^2 by a factor of K, then the bound is strictly increasing with K.

The above analysis suggests that the lower bound (35) is weak as it disregards any incoherence property of the measurement matrix A. For some cases, the bound reduces to $2\kappa\sigma^2\log\frac{N}{K}$ (refer to Corollary 2, Theorem 3 and 4) and is less than K when the noise level or K is relatively large. Certainly recovering the support is not possible with fewer than K measurements. The bound is loose also in the sense that when T, $\|A\|_{\mathrm{F}}^2$, or the SNR $1/\sigma^2$ is large enough the bound becomes negative, but when there is noise, perfect support recovery is generally impossible. While the original Fano's inequality

$$H(P_{\mathrm{err}}) + P_{\mathrm{err}}\log(L-1) \geq H(\boldsymbol{S}|\boldsymbol{Y}) \tag{36}$$

is tight in some cases[51], the adoption of the average divergence (32) as an upper bound on the mutual information $I(\boldsymbol{S};\boldsymbol{Y})$ between the random support $\boldsymbol{S}$ and the observation $\boldsymbol{Y}$ reduces the tightness (see the proof of [54, (33)]). Due to the difficulty of computing $H(\boldsymbol{S}|\boldsymbol{Y})$ and $I(\boldsymbol{S};\boldsymbol{Y})$ analytically, it is not clear whether a direct application of (36) results in a significantly better bound.

Despite its drawbacks, the bound (35) identifies the importance of the gain $\|A\|_{\mathrm{F}}^2$ of the measurement matrix, a quantity usually ignored in, for example, Compressive Sensing. We can also draw some interesting conclusions from (35) for measurement matrices with special properties. In particular, in the following corollary, we consider measurement matrices with rows or columns normalized to one. The rows of a measurement matrix are normalized to one in sensor network scenario (SNET) where each sensor is power limited while the columns are sometimes normalized to one in Compressive Sensing (refer to [42] and references therein).

Corollary 2: In order to have a probability of error $P_{\mathrm{err}} < \varepsilon$ with $0 < \varepsilon < 1$, the number of measurements must satisfy

$$\begin{aligned} MT &\geq (1-\varepsilon)\frac{2\sigma^2 K\log\frac{N}{K}}{\kappa\frac{K}{N}(1-\frac{K}{N})} + o(1) \\ &\geq (1-\varepsilon)\frac{8\sigma^2}{\kappa}K\log\frac{N}{K} + o(1) \end{aligned} \tag{37}$$

if the rows of A have unit norm; and

$$\begin{aligned} T &\geq (1-\varepsilon)\frac{2\sigma^2\log\frac{N}{K}}{\kappa\left(1-\frac{K}{N}\right)} + o(1) \\ &\geq (1-\varepsilon)\frac{2\sigma^2}{\kappa}\log\frac{N}{K} + o(1) \end{aligned} \tag{38}$$

if the columns of A have unit norm.

Note that the necessary condition (37) has the same critical quantity as the sufficient condition in Compressive Sensing. The inequality in (38) is independent of M. Therefore, if the columns are normalized to have unit norm, it is necessary to have multiple temporal measurements for a vanishing probability of error. Refer to Theorem 3 and 4 and discussions following them.

In the work of [7], each column of A is the array manifold vector function evaluated at a sample of the DOA parameter. The implication of the bound (35) for optimal design is that we should construct an array whose geometry leads to maximal $\|A\|_{\mathrm{F}}^2$. However, under the narrowband signal assumption and narrowband array assumption [55], the array manifold vector for isotropic sensor arrays always has norm $\sqrt{M}$ [56], which means that $\|A\|_{\mathrm{F}}^2 = MN$. Hence, in this case, the probability of error is always bounded by

$$P_{\mathrm{err}} \geq 1 - \frac{T\frac{K}{N}\left(1-\frac{K}{N}\right)MN}{2\sigma^2\log\binom{N}{K}} + o(1). \tag{39}$$

Therefore, we have the following theorem.

Theorem 3: Under the narrowband signal assumption and narrowband array assumption, for an isotropic sensor array in the DOA estimation scheme proposed in [7], in order to let the probability of error $P_{\mathrm{err}} < \varepsilon$ with $0 < \varepsilon < 1$ for any decision rule, the number of measurements must satisfy the following:

$$\begin{aligned} MT &\geq (1-\varepsilon)\frac{2\sigma^2\log\binom{N}{K}}{K\left(1-\frac{K}{N}\right)} + o(1) \\ &\geq (1-\varepsilon)\,2\sigma^2\log\frac{N}{K} + o(1). \end{aligned} \tag{40}$$

We comment that the same lower bound applies to Fourier measurement matrix (not normalized by $1/\sqrt{M}$) due to the same line of argument. We will not explicitly present this result in the current paper.

Since in radar and sonar applications the number of targets K is usually small, our result shows that the number of samples is lower bounded by $\log N$. Note that N is the number of intervals we use to divide the whole range of DOA; hence, it is a measure

of resolution. Therefore, the number of samples only needs to increase in the logarithm of N, which is very desirable. The symmetric roles played by M and T are also desirable since M is the number of sensors and is expensive to increase. As a consequence, we simply increase the number of samples to achieve a desired probability of error. In addition, unlike the upper bound of Theorem 1, we do not need to assume that $M \geq 2K$ in Theorem 2 and 3. Actually, Malioutov *et al.* made the empirical observation that ℓ_1–SVD technique can resolve $M-1$ sources if they are well separated [7]. Theorem 3 still applies to this extreme case.

Analysis of support recovery problem with measurement matrix obtained from sampling a manifold has considerable complexity compared with the Gaussian case. For example, it presents significant challenge to estimate $\bar{\lambda}_{S_i,S_j}$ in the DOA problem except for a few special cases that we discuss in [57]. As aforementioned, unlike the Gaussian case, $\bar{\lambda}_{S_i,S_j}$ for uniform linear arrays varies greatly with S_i and S_j. Therefore, even if we can compute $\bar{\lambda}_{S_i,S_j}$, replacing it with $\bar{\lambda}$ in the upper bound of Theorem 1 would lead to a very loose bound. On the other hand, the lower bound of Theorem 2 only involves the Frobenius norm of the measurement matrix, so we apply it to the DOA problem effortlessly. However, the lower bound is weak as it does not exploit any inner structure of the measurement matrix.

Donoho *et al.* considered the recovery of a "sparse" wide-band signal from narrow-band measurements [58], [59], a problem with essentially the same mathematical structure when we sample the array manifold uniformly in the wave number domain instead of the DOA domain. It was found that the spectral norm of the product of the band-limiting and time-limiting operators is crucial to stable signal recovery measured by the l_2 norm. In [58], Donoho and Stark bounded the spectral norm using the Frobenius norm, which leads to the well-known uncertainty principle. The authors commented that the uncertainty principle condition demands an extreme degree of sparsity for the signal. However, this condition can be relaxed if the signal support are widely scattered. In [7], Malioutov *et al.* also observed from numerical simulations that the ℓ_1–SVD algorithm performs much better when the sources are well separated than when they are located close together. In particular, they observed that presence of bias is mitigated greatly when sources are far apart. Donoho and Logan [59] explored the effect of the scattering of the signal support by using the "analytic principle of the large sieve." They bounded the spectral norm for the limiting operator by the maximum Nyquist density, a quantity that measures the degree of scattering of the signal support. We expect that our results can be improved in a similar manner. The challenges include using support recovery as a performance measure, incorporating multiple measurements, as well as developing the whole theory within a probabilistic framework.

V. Support Recovery for the Gaussian Measurement Ensemble

In this section, we refine our results in previous sections from general nondegenerate measurement matrices to the Gaussian ensemble. Unless otherwise specified, we always assume that the elements in a measurement matrix $\boldsymbol{A}$ are i.i.d. samples from unit variance real or complex Gaussian distributions. The Gaussian measurement ensemble is widely used and studied in Compressive Sensing [2]–[6]. The additional structure and the theoretical tools available enable us to derive deeper results in this case. In this section, we assume general scaling of (N, M, K, T). We do not find in our results a clear distinction between the regime of sublinear sparsity and the regime of linear sparsity as the one discussed in [43].

We first show two corollaries on the eigenvalue structure for the Gaussian measurement ensemble. Then we derive sufficient and necessary conditions in terms of M, N, K and T for the system to have a vanishing probability of error.

A. Eigenvalue Structure for a Gaussian Measurement Matrix

First, we observe that a Gaussian measurement matrix is nondegenerate with probability one, since any $p \leq M$ random vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_p$ from $\mathbb{F}\mathcal{N}(0, \Sigma)$ with $\Sigma \in \mathbb{R}^{M\times M}$ positive definite are linearly independent with probability one (refer to [45, Theorem 3.2.1]). As a consequence, we have the following.

Corollary 3: For Gaussian measurement matrix $\boldsymbol{A}$, let $\boldsymbol{H} = \boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0^{1/2}$, $k_{\mathrm{i}} = |S_0 \cap S_1|$, $k_0 = |S_0 \backslash S_1| = |S_0| - k_{\mathrm{i}}$, $k_1 = |S_1 \backslash S_0| = |S_1| - k_{\mathrm{i}}$. If $M \geqslant k_0 + k_1$, then with probability one, k_0 eigenvalues of matrix $\boldsymbol{H}$ are greater than 1, k_1 less than 1, and $M - (k_0 + k_1)$ equal to 1.

We refine Proposition 3 based on the well-known QR factorization for Gaussian matrices [45], [60].

Corollary 4: With the same notations as in Corollary 3, then with probability one, we have:

1) if $M \geqslant k_0 + k_1$, then the sorted eigenvalues of $\boldsymbol{H}$ that are greater than 1 are lower bounded by the corresponding ones of $\mathrm{I}_{k_0} + \frac{1}{\sigma^2}\boldsymbol{R}_{33}\boldsymbol{R}_{33}^{\dagger}$, where the elements of $\boldsymbol{R}_{33} = (r_{mn})_{k_0\times k_0}$ satisfy
$$\begin{aligned} 2\kappa r_{mm}^2 &\sim \chi^2_{2\kappa(M-k_1-k_{\mathrm{i}}-m+1)}, \qquad 1 \leq m \leq k_0 \\ r_{mn} &\sim \mathbb{F}\mathcal{N}(0,1), \qquad 1 \leq m < n \leq k_0. \end{aligned}$$
2) the eigenvalues of $\boldsymbol{H}$ are upper bounded by the corresponding eigenvalues of $\mathrm{I}_M + \frac{1}{\sigma^2}\boldsymbol{A}_{S_0\backslash S_1}\boldsymbol{A}_{S_0\backslash S_1}^{\dagger}$; in particular, the sorted eigenvalues of $\boldsymbol{H}$ that are greater than 1 are upper bounded by the corresponding ones of $\mathrm{I}_{k_0} + \frac{1}{\sigma^2}\boldsymbol{A}_{S_0\backslash S_1}^{\dagger}\boldsymbol{A}_{S_0\backslash S_1}$.

Now with the distributions on the elements of the bounding matrices, we can give sharp estimate on $\bar{\lambda}_{S_i,S_j}$. In particular, we have the following proposition.

Proposition 4: For Gaussian measurement matrix $\boldsymbol{A}$, suppose S_i and S_j are a pair of distinct supports with the same size K. Then we have
$$1 + \frac{M}{\sigma^2} \geq \mathbb{E}\bar{\lambda}_{S_i,S_j} \geq 1 + \frac{M - K - k_{\mathrm{d}}}{\sigma^2}.$$

Proof: We copy (29), (30) on $\bar{\lambda}_{S_i,S_j}$ here
$$1 + \frac{1}{\sigma^2 k_{\mathrm{d}}} \sum_{\substack{1\leq m\leq M \\ n\in S_i\backslash S_j}} |A_{mn}|^2 \geq \bar{\lambda}_{S_i,S_j} \geq 1 + \frac{1}{\sigma^2}\left(\prod_{m=1}^{k_{\mathrm{d}}} |r_{mm}|^2\right)^{1/k_{\mathrm{d}}}.$$

The proof then reduces to the computation of two expectations, one of which is trivial

$$\mathbb{E}\frac{1}{\sigma^2 k_{\mathrm{d}}}\sum_{\substack{1\le m\le M\\ n\in S_0\setminus S_1}}|\boldsymbol{A}_{mn}|^2=\frac{M}{\sigma^2}.$$

Next, the independence of the r_{nn}'s and the convexity of exponential functions together with Jensen's inequality yield

$$\begin{aligned}&\mathbb{E}\frac{1}{\sigma^2}\left(\prod_{n=1}^{k_{\mathrm{d}}}r_{nn}^2\right)^{1/k_{\mathrm{d}}}\\&\quad=\frac{1}{2\kappa\sigma^2}\mathbb{E}\exp\left[\frac{1}{k_{\mathrm{d}}}\sum_{n=1}^{k_{\mathrm{d}}}\log\left(2\kappa r_{nn}^2\right)\right]\\&\quad\ge\frac{1}{2\kappa\sigma^2}\exp\left[\frac{1}{k_{\mathrm{d}}}\sum_{n=1}^{k_{\mathrm{d}}}\mathbb{E}\log\left(2\kappa r_{nn}^2\right)\right].\end{aligned}$$

Since $\left(2\kappa r_{nn}^2\right)\sim\chi^2_{2\kappa(M-K-n+1)}$, the expectation of logarithm is $\mathbb{E}\log\left(2\kappa r_{nn}^2\right)=\log 2+\psi\left(\kappa(M-K-n+1)\right)$, where $\psi(z)=\frac{\Gamma'(z)}{\Gamma(z)}$ is the digamma function. Note that $\psi(z)$ is increasing and satisfies $\psi(z+1)\ge\log z$. Therefore, we have

$$\begin{aligned}&\mathbb{E}\frac{1}{\sigma^2}\left(\prod_{n=1}^{k_{\mathrm{d}}}r_{nn}^2\right)^{1/k_{\mathrm{d}}}\\&\quad\ge\frac{1}{2\kappa\sigma^2}\exp\left[\log 2+\frac{1}{k_{\mathrm{d}}}\sum_{n=1}^{k_{\mathrm{d}}}\psi\left(\kappa(M-K-n+1)\right)\right]\\&\quad\ge\frac{1}{\kappa\sigma^2}\exp\left[\psi\left(\kappa(M-K-k_{\mathrm{d}}+1)\right)\right]\\&\quad\ge\frac{1}{\kappa\sigma^2}\exp\left[\log\left(\kappa(M-K-k_{\mathrm{d}})\right)\right]\\&\quad\ge\frac{M-K-k_{\mathrm{d}}}{\sigma^2}.\end{aligned}$$

□

The expected value of the critical quantity $\bar{\lambda}_{S_i,S_j}$ lies between $1+\frac{M-2K}{\sigma^2}$ and $1+\frac{M}{\sigma^2}$, linearly proportional to M. Note that in conventional Compressive Sensing, the variance of the elements of $\boldsymbol{A}$ is usually taken to be $\frac{1}{M}$, which is equivalent to scaling the noise variance σ^2 to $M\sigma^2$ in our model. The resultant $\bar{\lambda}_{S_i,S_j}$ is then centered between $1+\frac{1-2\frac{K}{M}}{\sigma^2}$ and $1+\frac{1}{\sigma^2}$.

B. Necessary Condition

One fundamental problem in Compressive Sensing is how many samples should the system take to guarantee a stable reconstruction. Although many sufficient conditions are available, nontrivial necessary conditions are rare. Besides, in previous works, stable reconstruction has been measured in the sense of l_p norms between the reconstructed signal and the true signal. In this section, we derive two necessary conditions on M and T in terms of N and K in order to guarantee respectively that, first, $\mathbb{E}P_{\mathrm{err}}$ turns to zero and, second, for majority realizations of $\boldsymbol{A}$, the probability of error vanishes. More precisely, we have the following theorem.

Theorem 4: In the support recovery problem (5), for any $\varepsilon,\delta>0$, a necessary condition of $\mathbb{E}P_{\mathrm{err}}<\varepsilon$ is

$$MT\ge(1-\varepsilon)\frac{2\sigma^2\log\binom{N}{K}}{\kappa K\left(1-\frac{K}{N}\right)}+o(1)\tag{41}$$

$$\ge(1-\varepsilon)\frac{2\sigma^2}{\kappa}\log\frac{N}{K}+o(1)\tag{42}$$

and a necessary condition of $\mathbb{P}\left\{P_{\mathrm{err}}\left(\boldsymbol{A}\right)\le\varepsilon\right\}\ge1-\delta$ is

$$MT\ge(1-\varepsilon-\delta)\frac{2\sigma^2\log\binom{N}{K}}{\kappa K\left(1-\frac{K}{N}\right)}+o(1)\tag{43}$$

$$\ge(1-\varepsilon-\delta)\frac{2\sigma^2}{\kappa}\log\frac{N}{K}+o(1).\tag{44}$$

Proof: Equation (35) and $\mathbb{E}\left\|A\right\|_{\mathrm{F}}^2=\sum_{m,l}\mathbb{E}\left|\boldsymbol{A}_{ml}\right|^2=MN$ give

$$\mathbb{E}P_{\mathrm{err}}\ge1-\frac{\kappa T\frac{K}{N}\left(1-\frac{K}{N}\right)MN}{2\sigma^2\log\binom{N}{K}}+o\left(1\right).\tag{45}$$

Hence, $\mathbb{E}P_{\mathrm{err}}<\varepsilon$ entails (41) and (42).

Denote by E the event $\{\boldsymbol{A}:P_{\mathrm{err}}\left(\boldsymbol{A}\right)\le\varepsilon\}$; then $\mathbb{P}\left\{E^c\right\}\le\delta$ and we have

$$\begin{aligned}\mathbb{E}P_{\mathrm{err}}&=\int_E P_{\mathrm{err}}\left(\boldsymbol{A}\right)+\int_{E^c}P_{\mathrm{err}}\left(A\right)\\&\le\varepsilon\mathbb{P}\left(E\right)+\mathbb{P}\left(E^c\right)\\&\le\varepsilon+\delta.\end{aligned}$$

Therefore, from the first part of the theorem, we obtain (43) and (44). □

We compare our results with those of [43] and [42]. As we mentioned in the introduction, the differences in problem modeling and the definition of the probability of error make a direct comparison difficult. We first note that [43, Theorem 2] is established for the restricted problem where it is known *a priori* that all nonzero components in the sparse signal are equal. Because the set of signal realizations with equal nonzero components is a rare event in our signal model, it is not fitting to compare our result with the corresponding one in [43] by computing the distribution of the smallest on-support element, e.g., the expectation. Actually, the square of the smallest on-support element for the restricted problem, $\mathcal{M}^2(\beta^*)$ (or β in [42]), is equivalent to the signal variance in our model: both are measures of the signal energy. If we take into account the noise variance and replace $\mathcal{M}^2(\beta^*)$ (or $\beta^2\mathrm{SNR}$ in [42]) with $1/\sigma^2$, the necessary conditions in these papers coincide with ours when only one temporal sample is available. Our result shows that as far as support recovery is concerned, one cannot avoid the $\log\frac{N}{K}$ term when only given one temporal sample. Worse, for conventional Compressive Sensing with a measurement matrix generated from a Gaussian random variable with variance $1/M$, the necessary condition becomes

$$\begin{aligned}T&\ge\frac{2\sigma^2\log\binom{N}{K}}{\kappa K\left(1-\frac{K}{N}\right)}+o(1)\\&\ge\frac{2\sigma^2}{\kappa}\log\frac{N}{K}+o(1)\end{aligned}$$

which is independent of M. Therefore, when there is *considerable* noise ($\sigma^2 > \kappa/(2\log\frac{N}{K})$) , it is impossible to have a vanishing $\mathbb{E}P_{\text{err}}$ no matter how large an M one takes. Basically this situation arises because while taking more samples, one scales down the measurement gains A_{ml}, which effectively reduces the SNR and thus is not helpful in support recovery. As discussed, after Theorem 3, $\log\binom{N}{K}$ is the uncertainty of the support variable S, and $\log\frac{N}{K}$ actually comes from it. Therefore, it is no surprise that the number of samples is determined by this quantity and cannot be made independent of it.

C. Sufficient Condition

We derive a sufficient condition in parallel with sufficient conditions in Compressive Sensing. In Compressive Sensing, when only one temporal sample is available, $M = \Omega\left(K\log\frac{N}{K}\right)$ is enough for stable signal reconstruction for the majority of the realizations of measurement matrix $\boldsymbol{A}$ from a Gaussian ensemble with variance $\frac{1}{M}$. As shown in the previous subsection, if we take the probability of error for support recovery as a performance measure, it is impossible in this case to recover the support with a vanishing probability of error unless the noise is small. Therefore, we consider a Gaussian ensemble with unit variance. We first establish a lemma to estimate the lower tail of the distribution for $\bar{\lambda}_{S_i,S_j}$. We have shown that the $\mathbb{E}\left(\bar{\lambda}_{S_i,S_j}\right)$ lies between $1+\frac{M-2K}{\sigma^2}$ and $1+\frac{M}{\sigma^2}$. When γ is much less than $1+\frac{M-2K}{\sigma^2}$, we expect that $\mathbb{P}\left\{\bar{\lambda}_{S_i,S_j}\le\gamma\right\}$ decays quickly. More specifically, we have the following large deviation lemma:

Lemma 1: Suppose that $\gamma = \frac{1}{3}\frac{M-2K}{\sigma^2}$. Then there exists constant $c>0$ such that for $M-2K$ sufficiently large, we have

$$\mathbb{P}\left\{\bar{\lambda}_{S_i,S_j}\le\gamma\right\}\le\exp\left[-c\left(M-2K\right)\right].$$

This large deviation lemma together with the union bound yield the following sufficient condition for support recovery.

Theorem 5: Suppose that

$$M=\Omega\left(K\log\frac{N}{K}\right) \tag{46}$$

and

$$\kappa T\log\frac{M}{\sigma^2}\gg\log\left[K\left(N-K\right)\right]. \tag{47}$$

Then given any realization of measurement matrix $\boldsymbol{A}$ from a Gaussian ensemble, the optimal decision rule (16) for multiple support recovery problem (5) has a vanishing P_{err} with probability turning to one. In particular, if $M=\Omega\left(K\log\frac{N}{K}\right)$ and

$$T\gg\frac{\log N}{\log\log N} \tag{48}$$

then the probability of error turns to zero as N turns to infinity.

Proof: Denote $\gamma=\frac{1}{3}\frac{M-2K}{\sigma^2}$. Then according to the union bound, we have

$$\begin{aligned}&\mathbb{P}\left\{\bar{\lambda}\le\gamma\right\}\\&=\mathbb{P}\left\{\bigcup_{S_i\neq S_j}\left[\bar{\lambda}_{S_i,S_j}\le\gamma\right]\right\}\\&\le\sum_{S_i\neq S_j}\mathbb{P}\left\{\bar{\lambda}_{S_i,S_j}\le\gamma\right\}.\end{aligned}$$

Therefore, application of Lemma 1 gives

$$\begin{aligned}&\mathbb{P}\left\{\bar{\lambda}\le\gamma\right\}\\&\le\binom{N}{K}^2K\exp\left\{-c\left(M-2K\right)\right\}\\&\le\exp\left[-c\left(M-2K\right)+2K\log\frac{N}{K}+\log K\right].\end{aligned}$$

Hence, as long as $M=\Omega\left(K\log\frac{N}{K}\right)$, we know that the exponent turns to $-\infty$ as $N\longrightarrow\infty$. We now define $E=\left\{\boldsymbol{A}:\bar{\lambda}\left(\boldsymbol{A}\right)>\gamma\right\}$, where $\mathbb{P}\left\{E\right\}$ approaches one as N turns to infinity. Now the upper bound (26) becomes

$$\begin{aligned}P_{\text{err}}&=O\left(\frac{K\left(N-K\right)}{\left(\frac{\bar{\lambda}}{12\sigma^2}\right)^{\kappa T}}\right)\\&=O\left(\frac{K\left(N-K\right)}{\left(\frac{M}{\sigma^2}\right)^{\kappa T}}\right).\end{aligned}$$

Hence, if $\kappa T\log\frac{M}{\sigma^2}\gg\log\left[K\left(N-K\right)\right]$, we get a vanishing probability of error. In particular, under the assumption that $M\ge\Omega\left(K\log\frac{N}{K}\right)$, if $T\gg\frac{\log N}{\log\log N}$, then $\frac{\log\left[K(N-K)\right]}{\log\left[K\log\frac{N}{K}\right]}\le\frac{\log N}{\log\log N}$ implies that $K\left(N-K\right)\ll O\left(\left(\frac{K\log\frac{N}{K}}{\sigma^2}\right)^{\kappa T}\right)=O\left(\frac{K(N-K)}{\left(\frac{M}{\sigma^2}\right)^{\kappa T}}\right)$ for suitably selected constants. □

We now consider several special cases and explore the implications of the sufficient conditions. The discussions are heuristic in nature and their validity requires further checking.

If we set $T=1$, then we need M to be much greater than N to guarantee a vanishing probability P_{err}. This restriction suggests that even if we have more observations than the original signal length N, in which case we can obtain the original sparse signal by solving a least squares problem, we still might not be able to get the correct support because of the noise, as long as M is not sufficiently large compared to N. We discussed in the introduction that for many applications, the support of a signal has significant physical implications and its correct recovery is of crucial importance. Therefore, without multiple temporal samples and with moderate noise, the scheme proposed by Compressive Sensing is questionable as far as support recovery is concerned. Worse, if we set the variance for the elements in $\boldsymbol{A}$ to be $1/M$ as in Compressive Sensing, which is equivalent to replacing σ^2 with $M\sigma^2$, even increasing the number of temporal samples will not improve the probability of error significantly unless the noise variance is very small. Hence, using support recovery as a criterion, one cannot expect the Compressive Sensing scheme to work very well in the low SNR case. This conclusion is not a surprise, since we reduce the number of samples to achieve compression.

Another special case is when $K=1$. In this case, the sufficient condition becomes $M\ge\log N$ and $\kappa T\log\frac{M}{\sigma^2}\gg\log N$. Now the number of total samples should satisfy $MT\gg\frac{(\log N)^2}{\log\log N}$ while the necessary condition states that $MT=\Omega(\log N)$. The smallest gap between the necessary condition and sufficient condition is achieved when $K=1$.

From a denoising perspective, Fletcher *et al.* [61] upper bounded and approximated the probability of error for support

recovery averaged over the Gaussian ensemble. The bound and its approximation are applicable only to the special case with $K = 1$ and involve complex integrals. The authors obtained a useful SNR threshold as a function of M, N and K through the analytical bound. Note that our bounds are valid for general K and have a simple form. Besides, most of our derivation is conditioned on a realization of the Gaussian measurement ensemble. The conditioning makes more sense than averaging because in practice we usually make observations with fixed sensing matrix and varying signals and noise. The result of Theorem 5 also exhibits several interesting properties in the general case. Compared with the necessary condition (43) and (44), the asymmetry in the sufficient condition is even more desirable in most cases because of the asymmetric cost associated with sensors and temporal samples. Once the threshold $K \log \frac{N}{K}$ of M is exceeded, we can achieve a desired probability of error by taking more temporal samples. If we were concerned only with the total number of samples, we would minimize MT subject to the constraints (46) and (47) to achieve a given level of probability of error. However, in applications for which timing is important, one has to increase the number of sensors to reduce P_{err} to a certain limit.

The sufficient condition (46), (47), and (48) is separable in the following sense. We observe from the proof that the requirement $M = \Omega\left(K \log \frac{N}{K}\right)$ is used only to guarantee that the randomly generated measurement matrix is a good one in the sense that its incoherence $\bar{\lambda}$ is sufficiently large, as in the case of Compressive Sensing. It is in Lemma 1 that we use the Gaussian ensemble assumption. If another deterministic construction procedure (for attempts in this direction, see [62]) or random distribution give measurement matrix with better incoherence $\bar{\lambda}$, it would be possible to reduce the orders for both M and T.

VI. Conclusion

In this paper, we formulated the support recovery problems for jointly sparse signals as binary and multiple-hypothesis testings. Adopting the probability of error as the performance criterion, the optimal decision rules are given by the likelihood ratio test and the maximum *a posteriori* probability estimator. The latter reduces to the maximum likelihood estimator when equal prior probabilities are assigned to the supports. We then employed the Chernoff bound and Fano's inequality to derive bounds on the probability of error. We discussed the implications of these bounds at the end of Sections III-B, III-C, IV, V-B and V-C, in particular when they are applied to the DOA estimation problem considered in [7] and Compressive Sensing with a Gaussian measurement ensemble. We derived sufficient and necessary conditions for Compressive Sensing using Gaussian measurement matrices to achieve a vanishing probability of error in both the mean and large probability senses. These conditions show the necessity of considering multiple temporal samples. The symmetric and asymmetric roles played by the spatial and temporal samples and their implications in system design were discussed. For Compressive Sensing, we demonstrated that it is impossible to obtain accurate signal support with only one temporal sample if the variance for the Gaussian measurement matrix scales with $1/M$ and there is considerable noise.

This research on support recovery for jointly sparse signals is far from complete. Several questions remain to be answered. First, we notice an obvious gap between the necessary and sufficient conditions even in the simplest case with $K = 1$. Better techniques need to be introduced to refine the results. Second, as in the case for RIP, computation of the quantity $\bar{\lambda}$ for an arbitrary measurement matrix is extremely difficult. Although we derive large derivation bounds on $\bar{\lambda}$ and compute the expected value for $\bar{\lambda}_{S_i,S_j}$ for the Gaussian ensemble, its behaviors in both the general and Gaussian cases require further study. Its relationship with RIP also needs to be clarified. Finally, our lower bound derived from Fano's inequality identifies only the effect of the total gain. The effect of the measurement matrix's incoherence is elusive. The answers to these questions will enhance our understanding of the measurement mechanism (2).

Appendix A
Proof of Proposition 1

In this proof, we focus on the case for which both $k_0 \neq 0$ and $k_1 \neq 0$. Other cases have similar and simpler proofs. The eigenvalues of H satisfy $|\lambda \mathrm{I}_M - H| = 0$, which is equivalent to $|\lambda \Sigma_1 - \Sigma_0| = 0$. The substitution $\lambda = \mu + 1$ defines

$$g(\mu) = |(\mu+1)\Sigma_1 - \Sigma_0| = |\mu\Sigma_1 - (\Sigma_0 - \Sigma_1)|.$$

The following algebraic manipulation

$$\begin{aligned} G &\triangleq \Sigma_0 - \Sigma_1 \\ &= A_{S_0} A_{S_0}^{\dagger} - A_{S_1} A_{S_1}^{\dagger} \\ &= \left[A_{S_0 \cap S_1} A_{S_0 \cap S_1}^{\dagger} + A_{S_0 \setminus S_1} A_{S_0 \setminus S_1}^{\dagger}\right] \\ &\quad - \left[A_{S_0 \cap S_1} A_{S_0 \cap S_1}^{\dagger} + A_{S_1 \setminus S_0} A_{S_1 \setminus S_0}^{\dagger}\right] \\ &= A_{S_0 \setminus S_1} A_{S_0 \setminus S_1}^{\dagger} - A_{S_1 \setminus S_0} A_{S_1 \setminus S_0}^{\dagger} \end{aligned}$$

leads to

$$\begin{aligned} g(\mu) &= |\mu\Sigma_1 - G| \\ &= |\Sigma_1|^{\frac{1}{2}} \left|\mu \mathrm{I}_M - \Sigma_1^{-\frac{1}{2}} G \Sigma_1^{-\frac{1}{2}\dagger}\right| |\Sigma_1|^{\frac{1}{2}}. \end{aligned}$$

Therefore, to prove the theorem, it suffices to show that $\Sigma_1^{-\frac{1}{2}} G \Sigma_1^{-\frac{1}{2}\dagger}$ has k_0 positive eigenvalues, k_1 negative eigenvalues and $M-(k_0+k_1)$ zero eigenvalues or, put another way, $\Sigma_1^{-\frac{1}{2}} G \Sigma_1^{-\frac{1}{2}\dagger}$ has inertia $(k_0, k_1, M-(k_0+k_1))$. Sylvester's law of inertia ([49, Theorem 4.5.8, p. 223]) states that the inertia of a symmetric matrix is invariant under congruence transformations. Hence, we need only to show that G has inertia $(k_0, k_1, M-(k_0+k_1))$. Clearly $G = PQ^{\dagger}$ with $P = [\, A_{S_0 \setminus S_1} \quad A_{S_1 \setminus S_0} \,]$ and $Q = [\, A_{S_0 \setminus S_1} \quad -A_{S_1 \setminus S_0} \,]$. To find the number of zero eigenvalues of G, we calculate the rank of G. The nondegenerateness of measurement matrix A implies that $\operatorname{rank}(P) = \operatorname{rank}(Q) = k_0 + k_1$. Therefore, from rank inequality ([49, Theorem 0.4.5, p. 13]),

$$\begin{aligned} &\operatorname{rank}(P) + \operatorname{rank}\left(Q^{\dagger}\right) - (k_0+k_1) \\ &\quad \leq \operatorname{rank}\left(PQ^{\dagger}\right) \\ &\quad \leq \min\left\{\operatorname{rank}(P), \operatorname{rank}\left(Q^{\dagger}\right)\right\} \end{aligned}$$

we conclude that $\operatorname{rank}(G) = k_0 + k_1$.

To count the number of negative eigenvalues of G, we use the Jacobi-Sturm rule ([63, Theorem A.1.4, p. 320]), which states that for an $M \times M$ symmetric matrix whose jth leading principal minor has determinant $d_j, j = 1, \ldots, M$, the number of nonnegative eigenvalues is equal to the number of sign changes of sequence $\{1, d_1, \ldots, d_M\}$. We consider only the first $k_0 + k_1$ leading principal minors, since higher order minors have determinant 0.

Suppose $I = \{1, \ldots, k_0 + k_1\}$ is an index set. Without loss of generality, we assume that P^I is nonsingular. Applying QL factorization (one variation of QR factorization, see [64]) to matrix P^I, we obtain $P^I = OL$, where O is an orthogonal matrix, $OO^\dagger = \mathrm{I}_{k_0+k_1}$, and $L = (l_{ij})_{(k_0+k_1)\times(k_0+k_1)}$ is a lower triangular matrix. The diagonal entries of L are nonzero because P^I is nonsingular. The partition of L into

$$L = [\, L_1 \quad L_2 \,]$$

with $L_1 \in \mathbb{F}^{(k_0+k_1)\times k_0}$, $L_2 \in \mathbb{F}^{(k_0+k_1)\times k_1}$, and $L_2 = \begin{bmatrix} 0 \\ L_3 \end{bmatrix}$ with $L_3 \in \mathbb{F}^{k_1 \times k_1}$ implies

$$G_I^I = P^I (Q_I)^\dagger = O [\, L_1 \quad L_2 \,] \begin{bmatrix} L_1^\dagger \\ -L_2^\dagger \end{bmatrix} O^\dagger.$$

Again using the invariance property of inertia under congruence transformation, we focus on the leading principal minors of

$$U \triangleq [\, L_1 \quad L_2 \,] \begin{bmatrix} L_1^\dagger \\ -L_2^\dagger \end{bmatrix}.$$

Suppose $J = \{1, \ldots, j\}$. For $1 \le j \le k_0$, from the lower triangularity of L, it is clear that

$$\left|(U_J^J)\right| = \left|(L_1)_J^J\right|^2 = \prod_{i=1}^{j} |l_{ii}|^2 > 0.$$

For $k_0 + 1 \le j \le k_0 + k_1$, suppose $J_0 = \{1, \ldots, k_0\}$ and $J_1 = \{1, \ldots, j - k_0\}$. We then have

$$\begin{aligned} \left|U_J^J\right| &= \left|(L_1)_{J_0}^{J_0}\right|^2 \left|(L_3)_{J_1}^{J_1}\right| - \left[(L_3)_{J_1}^{J_1}\right]^\dagger \Big| \\ &= (-1)^{j-k_0} \left|(L_1)_{J_0}^{J_0}\right|^2 \left|(L_3)_{J_1}^{J_1}\right|^2 \\ &= (-1)^{j-k_0} \prod_{i=1}^{j} |l_{ii}|^2. \end{aligned}$$

Therefore, the sequence $1, d_1, d_2, \ldots d_{k_0+k_1}$ has k_1 sign changes, which implies that G_I^I—hence G—has k_1 negative eigenvalues. Finally, we conclude that the theorem holds for H. □

Appendix B
Proof of Proposition 3

We first prove the first claim. From the proof of Proposition 1, it suffices to show that the sorted positive eigenvalues of $\Sigma_1^{-\frac{1}{2}} G \Sigma_1^{-\frac{1}{2}\dagger}$ are greater than those of $\frac{1}{\sigma^2} R_{33} R_{33}^\dagger$, where $G = A_{S_0\setminus S_1} A_{S_0\setminus S_1}^\dagger - A_{S_1\setminus S_0} A_{S_1\setminus S_0}^\dagger$. Since cyclic permutation of a matrix product does not change its eigenvalues, we restrict ourselves to $\Sigma_1^{-1} G$. Consider the QR decomposition

$$\begin{aligned} &[\, A_{S_1\setminus S_0} \quad A_{S_1 S_0} \quad A_{S_0\setminus S_1} \,] \\ &\quad = QR \\ &\quad \triangleq [\, Q_1 \quad Q_2 \quad Q_3 \quad Q_4 \,] \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ 0 & R_{22} & R_{23} \\ 0 & 0 & R_{33} \\ 0 & 0 & 0 \end{bmatrix} \end{aligned}$$

where $Q \in \mathbb{F}^{M\times M}$ is an orthogonal matrix with partitions $Q_1 \in \mathbb{F}^{M\times k_1}$, $Q_2 \in \mathbb{F}^{M\times k_i}$, $Q_3 \in \mathbb{F}^{M\times k_0}$, $R \in \mathbb{F}^{M\times(k_1+k_i+k_0)}$ is an upper triangular matrix with partitions $R_{11} \in \mathbb{F}^{k_1\times k_1}$, $R_{22} \in \mathbb{F}^{k_i\times k_i}$, $R_{33} \in \mathbb{F}^{k_0\times k_0}$, and other submatrices have corresponding dimensions.

First, we note that

$$\begin{aligned} &Q^\dagger G Q \\ &\quad = \begin{bmatrix} R_{13} & R_{11} \\ R_{23} & 0 \\ R_{33} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} R_{13}^\dagger & R_{23}^\dagger & R_{33}^\dagger & 0 \\ -R_{11}^\dagger & 0 & 0 & 0 \end{bmatrix} \\ &\quad = \begin{bmatrix} \begin{bmatrix} R_{13} & R_{11} \\ R_{23} & 0 \end{bmatrix} \begin{bmatrix} R_{13}^\dagger & R_{23}^\dagger \\ -R_{11}^\dagger & 0 \end{bmatrix} & \begin{bmatrix} R_{13}R_{33}^\dagger \\ R_{23}R_{33}^\dagger \end{bmatrix} & 0 \\ \begin{bmatrix} R_{33}R_{13}^\dagger & R_{33}R_{23}^\dagger \end{bmatrix} & R_{33}R_{33}^\dagger & 0 \\ 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

Therefore, the last $M - (k_1 + k_i + k_0)$ rows and columns of $Q^\dagger G Q$—and hence of $(Q^\dagger \Sigma_1 Q)^{-1}(Q^\dagger G Q)$—are zeros, which lead to the $M - (k_1 + k_i + k_0)$ zero eigenvalues of $\Sigma_1^{-\frac{1}{2}} G \Sigma_1^{-\frac{1}{2}\dagger}$. We then drop these rows and columns in all matrices involved in subsequent analysis. In particular, the submatrix of $Q^\dagger \Sigma_1 Q = Q^\dagger(\sigma^2 \mathrm{I}_M + A_{S_1} A_{S_1}^\dagger)Q$ without the last $M - (k_1 + k_i + k_0)$ rows and columns is

$$\begin{aligned} &\sigma^2 \mathrm{I}_M + \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} R_{11}^\dagger & 0 & 0 \\ R_{12}^\dagger & R_{22}^\dagger & 0 \end{bmatrix} \\ &\quad = \begin{bmatrix} \sigma^2 \mathrm{I}_{k_1+k_i} + \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} \begin{bmatrix} R_{11}^\dagger & 0 \\ R_{12}^\dagger & R_{22}^\dagger \end{bmatrix} & 0 \\ 0 & \sigma^2 \mathrm{I}_{k_0} \end{bmatrix} \\ &\quad \triangleq \begin{bmatrix} F & 0 \\ 0 & \sigma^2 \mathrm{I}_{k_0} \end{bmatrix}. \end{aligned}$$

Define

$$\begin{aligned} &\begin{bmatrix} V & K^\dagger \\ K & R_{33}R_{33}^\dagger \end{bmatrix} \\ &\qquad \triangleq \begin{bmatrix} \begin{bmatrix} R_{13} & R_{11} \\ R_{23} & 0 \end{bmatrix} \begin{bmatrix} R_{13}^\dagger & R_{23}^\dagger \\ -R_{11}^\dagger & 0 \end{bmatrix} & \begin{bmatrix} R_{13}R_{33}^\dagger \\ R_{23}R_{33}^\dagger \end{bmatrix} \\ \begin{bmatrix} R_{33}R_{13}^\dagger & R_{33}R_{23}^\dagger \end{bmatrix} & R_{33}R_{33}^\dagger \end{bmatrix}. \end{aligned}$$

Due to the invariance of eigenvalues with respect to orthogonal transformations and switching to the symmetrized version, we focus on

$$\begin{bmatrix} F & 0 \\ 0 & \sigma^2 \mathrm{I}_{k_0} \end{bmatrix}^{-\frac{1}{2}} \begin{bmatrix} V & K^\dagger \\ K & R_{33} R_{33}^\dagger \end{bmatrix} \begin{bmatrix} F & 0 \\ 0 & \sigma^2 \mathrm{I}_{k_0} \end{bmatrix}^{-\frac{1}{2}\dagger} = \begin{bmatrix} F^{-\frac{1}{2}} V F^{-\frac{1}{2}\dagger} & F^{-\frac{1}{2}} K^\dagger \frac{1}{\sigma} \\ \frac{1}{\sigma} K F^{-\frac{1}{2}} & \frac{1}{\sigma^2} R_{33} R_{33}^\dagger \end{bmatrix}.$$

Next we argue that the sorted positive eigenvalues of

$$\begin{bmatrix} F^{-\frac{1}{2}} V F^{-\frac{1}{2}\dagger} & F^{-\frac{1}{2}} K^\dagger \frac{1}{\sigma} \\ \frac{1}{\sigma} K F^{-\frac{1}{2}} & \frac{1}{\sigma^2} R_{33} R_{33}^\dagger \end{bmatrix}$$

are greater than the corresponding sorted eigenvalues of $\frac{1}{\sigma^2} R_{33} R_{33}^\dagger$.

For any $\varepsilon > 0$, we define a matrix

$$M_{\varepsilon,N} = \begin{bmatrix} -N\mathrm{I}_{k_1+k_\mathrm{i}} & 0 \\ 0 & \frac{1}{\sigma^2} R_{33} R_{33}^\dagger - \varepsilon \mathrm{I}_{k_0} \end{bmatrix}.$$

Then we have

$$\begin{bmatrix} F^{-\frac{1}{2}} V F^{-\frac{1}{2}\dagger} & F^{-\frac{1}{2}} K^\dagger \frac{1}{\sigma} \\ \frac{1}{\sigma} K F^{-\frac{1}{2}} & \frac{1}{\sigma^2} R_{33} R_{33}^\dagger \end{bmatrix} - M_{\varepsilon,N} = \begin{bmatrix} F^{-\frac{1}{2}} V F^{-\frac{1}{2}\dagger} + N\mathrm{I}_{k_1+k_\mathrm{i}} & F^{-\frac{1}{2}} K^\dagger \frac{1}{\sigma} \\ \frac{1}{\sigma} K F^{-\frac{1}{2}} & \varepsilon \mathrm{I}_{k_0} \end{bmatrix}.$$

Note that

$$\begin{bmatrix} F^{-\frac{1}{2}} V F^{-\frac{1}{2}\dagger} + N\mathrm{I}_{k_1+k_\mathrm{i}} & F^{-\frac{1}{2}} K^\dagger \frac{1}{\sigma} \\ \frac{1}{\sigma} K F^{-\frac{1}{2}} & \varepsilon \mathrm{I}_{k_0} \end{bmatrix}$$

is congruent to

$$\begin{bmatrix} F^{-\frac{1}{2}} V F^{-\frac{1}{2}\dagger} + N\mathrm{I}_{k_1+k_\mathrm{i}} - \frac{1}{\varepsilon\sigma^2} F^{-\frac{1}{2}} K^\dagger K F^{-\frac{1}{2}} & 0 \\ 0 & \varepsilon \mathrm{I}_{k_0} \end{bmatrix}.$$

Clearly $F^{-\frac{1}{2}} V F^{-\frac{1}{2}\dagger} + N\mathrm{I}_{k_1+k_\mathrm{i}} - \frac{1}{\varepsilon\sigma^2} F^{-\frac{1}{2}} K^\dagger K F^{-\frac{1}{2}}$ is positive definite when N is sufficiently large. Hence, when N is large enough, we obtain

$$\begin{bmatrix} F^{-\frac{1}{2}} V F^{-\frac{1}{2}\dagger} & F^{-\frac{1}{2}} K^\dagger \frac{1}{\sigma} \\ \frac{1}{\sigma} K F^{-\frac{1}{2}} & \frac{1}{\sigma^2} R_{33} R_{33}^\dagger \end{bmatrix} \succ M_{\varepsilon,N}.$$

Using [49, Corollary 4.3.3], we conclude that the eigenvalues of

$$\begin{bmatrix} F^{-\frac{1}{2}} V F^{-\frac{1}{2}\dagger} & F^{-\frac{1}{2}} K^\dagger \frac{1}{\sigma} \\ \frac{1}{\sigma} K F^{-\frac{1}{2}} & \frac{1}{\sigma^2} R_{33} R_{33}^\dagger \end{bmatrix}$$

are greater than those of $M_{\varepsilon,N}$ if sorted. From Proposition 1, we know that

$$\begin{bmatrix} F^{-\frac{1}{2}} V F^{-\frac{1}{2}\dagger} + N\mathrm{I}_{k_1+k_\mathrm{i}} & F^{-\frac{1}{2}} K^\dagger \frac{1}{\sigma} \\ \frac{1}{\sigma} K F^{-\frac{1}{2}} & \varepsilon \mathrm{I}_{k_0} \end{bmatrix}$$

has exactly k_0 positive eigenvalues, which are the only eigenvalues that could be greater than $\lambda\left(\frac{1}{\sigma^2} R_{33} R_{33}^\dagger\right) - \varepsilon$. Since ε is arbitrary, we finally conclude that the positive eigenvalues of $\Sigma_1^{-1} G$ are greater than those of $\frac{1}{\sigma^2} R_{33} R_{33}^\dagger$ if sorted in the same way.

For the second claim, we need some notations and properties of symmetric and Hermitian matrices. For any pair of symmetric (or Hermitian) matrices P and Q, $P \prec Q$ means that $Q - P$ is positive definite and $P \preceq Q$ means $Q - P$ is nonnegative definite. Note that if P and Q are positive definite, then from [49, Corollary 7.7.4] $P \preceq Q$ if and only if $Q^{-1} \preceq P^{-1}$; if $P \preceq Q$ then the eigenvalues of P and Q satisfy $\lambda_k(P) \leq \lambda_k(Q)$, where $\lambda_k(P)$ denotes the kth largest eigenvalue of P; furthermore, $A \preceq B$ implies that $PAP^\dagger \preceq PBP^\dagger$ for any P, square or rectangular. Therefore, $\sigma^2 \mathrm{I}_M + A_{S_0 S_1} A_{S_0 S_1}^\dagger \preceq \sigma^2 \mathrm{I}_M + A_{S_1} A_{S_1}^\dagger = \Sigma_1$ yields

$$\Sigma_0^{1/2} \Sigma_1^{-1} \Sigma_0^{1/2} \preceq \Sigma_0^{1/2} \left(\sigma^2 \mathrm{I}_M + A_{S_0 S_1} A_{S_0 S_1}^\dagger\right)^{-1} \Sigma_0^{1/2}.$$

Recall that from the definition of eigenvalues, the nonzero eigenvalues of AB and BA are the same for any matrices A and B. Since we are interested only in the eigenvalues, a cyclic permutation in the matrix product on the previous inequality's right-hand side gives us

$$\begin{aligned} &\left(\sigma^2 \mathrm{I}_M + A_{S_0 S_1} A_{S_0 S_1}^\dagger\right)^{-\frac{1}{2}} \Sigma_0 \left(\sigma^2 \mathrm{I}_M + A_{S_0 S_1} A_{S_0 S_1}^\dagger\right)^{-\frac{1}{2}} \\ &= \mathrm{I}_M + \left(\sigma^2 \mathrm{I}_M + A_{S_0 S_1} A_{S_0 S_1}^\dagger\right)^{-\frac{1}{2}} A_{S_0 \backslash S_1} \\ &\quad \times A_{S_0 \backslash S_1}^\dagger \left(\sigma^2 \mathrm{I}_M + A_{S_0 S_1} A_{S_0 S_1}^\dagger\right)^{-\frac{1}{2}} \\ &\triangleq \mathrm{I}_M + Q^{-\frac{1}{2}} A_{S_0 \backslash S_1} A_{S_0 \backslash S_1}^\dagger Q^{-\frac{1}{2}} \\ &\triangleq \mathrm{I}_M + P. \end{aligned}$$

Until now we have shown that the sorted eigenvalues of H are less than the corresponding ones of $\mathrm{I}_M + P$. The nonzero eigenvalues of $Q^{-\frac{1}{2}} A_{S_0 \backslash S_1} A_{S_0 \backslash S_1}^\dagger Q^{-\frac{1}{2}}$ are the same as the nonzero eigenvalues of $A_{S_0 \backslash S_1}^\dagger Q^{-1} A_{S_0 \backslash S_1} \preceq \frac{1}{\sigma^2} A_{S_0 \backslash S_1}^\dagger A_{S_0 \backslash S_1}$. Using the same fact again, we conclude that the nonzero eigenvalues of $\frac{1}{\sigma^2} A_{S_0 \backslash S_1}^\dagger A_{S_0 \backslash S_1}$ are the same as the nonzero eigenvalues of $\frac{1}{\sigma^2} A_{S_0 \backslash S_1} A_{S_0 \backslash S_1}^\dagger$. Therefore, we obtain that

$$\begin{aligned} \lambda_k(\Sigma_0^{1/2} \Sigma_1^{-1} \Sigma_0^{1/2}) &\leq \lambda_k(\mathrm{I}_M + P) \\ &\leq \lambda_k\big(\mathrm{I}_M + \frac{1}{\sigma^2} A_{S_0 \backslash S_1} A_{S_0 \backslash S_1}^\dagger\big). \end{aligned}$$

In particular, the eigenvalues of H that are greater than 1 are upper bounded by the corresponding ones of $\mathrm{I}_M + \frac{1}{\sigma^2} A_{S_0 \backslash S_1} A_{S_0 \backslash S_1}^\dagger$ if they are both sorted ascendantly. Hence, we get that the eigenvalues of H that are greater than 1 are less than those of $\mathrm{I}_{k_0} + \frac{1}{\sigma^2} A_{S_0 \backslash S_1}^\dagger A_{S_0 \backslash S_1}$.

Therefore, the conclusion of the second part of the theorem holds. We comment here that usually it is not true that $H \preceq \mathrm{I}_M + \frac{1}{\sigma^2} A_{S_0 \backslash S_1} A_{S_0 \backslash S_1}^\dagger$. Only the inequality on eigenvalues holds. □

APPENDIX C
PROOF OF LEMMA 1

For arbitrary fixed supports S_i, S_j, we have

$$\begin{aligned}\bar{\lambda}_{S_i,S_j} &\geq 1+\frac{1}{2\kappa\sigma^2}\left(\prod_{l=1}^{k_{\mathrm{d}}} 2\kappa r_{ll}^2\right)^{1/k_{\mathrm{d}}} \\ &\geq \frac{1}{2\kappa\sigma^2}\min_{1\leq l\leq k_{\mathrm{d}}} q_l\end{aligned}$$

where $2\kappa r_{ll}^2 \sim \chi^2_{2\kappa(M-K-l+1)}$ can be written as a sum of $2\kappa(M-K-l+1)$ independent squared standard Gaussian random variables and $q_l \sim \chi^2_{2\kappa(M-2K)}$ is obtained by dropping $K-l+1$ of them. Therefore, using the union bound we obtain

$$\begin{aligned}&\mathbb{P}\left\{\bar{\lambda}_{S_i,S_j}\leq\gamma\right\} \\ &\leq \mathbb{P}\left\{\frac{1}{2\kappa\sigma^2}\min_{1\leq l\leq k_{\mathrm{d}}} q_l\leq\gamma\right\} \\ &\leq \mathbb{P}\left\{\bigcup_{1\leq l\leq k_{\mathrm{d}}}\left[q_l\leq 2\kappa\sigma^2\gamma\right]\right\} \\ &\leq k_{\mathrm{d}}\mathbb{P}\left\{q_l\leq 2\kappa\sigma^2\gamma\right\}.\end{aligned}$$

Since $\gamma=\frac{1}{3}\frac{M-2K}{\sigma^2}$ implies that $2\kappa\sigma^2\gamma=\frac{2\kappa}{3}(M-2K)<2\kappa(M-2K)-2$, the mode of $\chi^2_{\kappa(M-2K)}$, when $M-2K$ is sufficiently large, we have

$$\begin{aligned}&\mathbb{P}\left\{q_l\leq 2\kappa\sigma^2\gamma\right\} \\ &=\int_0^{2\kappa\sigma^2\gamma}\frac{(1/2)^{\kappa(M-2K)}}{\Gamma\left(\kappa(M-2K)\right)}x^{\kappa(M-2K)-1}e^{-x/2}dx \\ &\leq\frac{\left[\kappa\sigma^2\gamma\right]^{\kappa(M-2K)}}{\Gamma\left(\kappa(M-2K)\right)}e^{-\kappa\sigma^2\gamma}.\end{aligned}$$

The inequality $\log\Gamma(z)\geq\left(z-\frac{1}{2}\right)\log z-z$ says that when $M-2K$ is large enough,

$$\begin{aligned}&\mathbb{P}\left\{q_l\leq 2\kappa\sigma^2\gamma\right\} \\ &\leq\exp\left\{\kappa(M-2K)\log\left(\kappa\sigma^2\gamma\right)-\kappa\sigma^2\gamma\right. \\ &\quad-\left[\kappa(M-2K)-\frac{1}{2}\right]\log\left[\kappa(M-2K)\right] \\ &\quad\left.+\kappa(M-2K)\right\} \\ &\leq\exp\left\{-c(M-2K)\right\}\end{aligned}$$

where $c<\kappa(\log 3-1)$. Therefore, we have

$$\mathbb{P}\left\{\bar{\lambda}_{S_i,S_j}\leq\gamma\right\}\leq K\exp\left\{-c(M-2K)\right\}. \qquad \square$$

ACKNOWLEDGMENT

The authors thank the anonymous referees for their careful and helpful comments.

REFERENCES

[1] M. B. Wakin, S. Sarvotham, M. F. Duarte, D. Baron, and R. G. Baraniuk, "Recovery of jointly sparse signals from few random projections," in *Proc. Neural Inf. Process. Syst.*, Vancouver, Canada, Dec. 2005, pp. 1435–1442.

[2] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[3] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[4] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[5] E. Candès and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[6] R. Baraniuk, "Compressive sensing [lecture notes]," *IEEE Signal Process. Mag.*, vol. 24, no. 4, pp. 118–121, Jul. 2007.

[7] D. Malioutov, M. Cetin, and A. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, Aug. 2005.

[8] D. Model and M. Zibulevsky, "Signal reconstruction in sensor arrays using sparse representations," *Signal Process.*, vol. 86, no. 3, pp. 624–638, Mar. 2006.

[9] V. Cevher, M. Duarte, and R. G. Baraniuk, "Distributed target localization via spatial sparsity," in *Eur. Signal Process. Conf. (EUSIPCO 2008)*, Lausanne, Switzerland, Aug. 2008.

[10] V. Cevher, P. Indyk, C. Hegde, and R. G. Baraniuk, "Recovery of clustered sparse signals from compressive measurements," in *Proc. Int. Conf. Sampling Theory Appl. (SAMPTA 2009)*, Marseille, France, May 2009, pp. 18–22.

[11] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1998.

[12] S. Mallat, *A Wavelet Tour of Signal Processing*. San Diego, CA: Academic, 1998.

[13] I. Gorodnitsky and B. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.

[14] E. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, pp. 2313–2351, 2007.

[15] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[16] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2230–2249, May 2009.

[17] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, May 2009.

[18] P. Borgnat and P. Flandrin, "Time-frequency localization from sparsity constraints," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP 2008)*, Las Vegas, NV, Apr. 2008, pp. 3785–3788.

[19] Z. Tian and G. Giannakis, "Compressed sensing for wideband cognitive radios," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP 2007)*, Honolulu, HI, Apr. 2007, pp. IV–1357–IV–1360.

[20] M. A. Sheikh, S. Sarvotham, O. Milenkovic, and R. G. Baraniuk, "DNA array decoding from nonlinear measurements by belief propagation," in *Proc. IEEE Workshop Statist. Signal Process. (SSP 2007)*, Madison, WI, Aug. 2007, pp. 215–219.

[21] H. Vikalo, F. Parvaresh, and B. Hassibi, "On recovery of sparse signals in compressed DNA microarrays," in *Proc. Asilomar Conf. Signals, Syst. Comput. (ACSSC 2007)*, Pacific Grove, CA, Nov. 2007, pp. 693–697.

[22] F. Parvaresh, H. Vikalo, S. Misra, and B. Hassibi, "Recovering sparse signals using sparse measurement matrices in compressed DNA microarrays," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 3, pp. 275–285, Jun. 2008.

[23] H. Vikalo, F. Parvaresh, S. Misra, and B. Hassibi, "Sparse measurements, compressed sampling, and DNA microarrays," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP 2008)*, Las Vegas, NV, Apr. 2008, pp. 581–584.

[24] R. Baraniuk and P. Steeghs, "Compressive radar imaging," in *Proc. IEEE Radar Conf.* , Apr. 2007, pp. 128–133.

[25] M. Herman and T. Strohmer, "Compressed sensing radar," in *Proc. IEEE Radar Conf.*, May 2008, pp. 1–6.

[26] M. A. Herman and T. Strohmer, "High-resolution radar via compressed sensing," *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2275–2284, Jun. 2009.

[27] E. G. Larsson and Y. Selen, "Linear regression with a sparse parameter vector," *IEEE Trans. Signal Process.*, vol. 55, no. 2, pp. 451–460, Feb. 2007.

[28] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[29] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, pp. 1408–1418, Aug. 1969.
[30] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.
[31] G. Bienvenu and L. Kopp, "Adaptivity to background noise spatial coherence for high resolution passive methods," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP 1980)*, Denver, CO, Apr. 1980, vol. 5, pp. 307–310.
[32] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramér-Rao bound: Further results and comparisons," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 12, pp. 2140–2150, Dec. 1990.
[33] M. Duarte, S. Sarvotham, D. Baron, M. Wakin, and R. Baraniuk, "Distributed compressed sensing of jointly sparse signals," in *Proc. Asilomar Conf. Signals, Syst. Comput. (ACSSC 2005)*, Pacific Grove, CA, Nov. 2005, pp. 1537–1541.
[34] M. Duarte, M. Wakin, D. Baron, and R. Baraniuk, "Universal distributed sensing via random projections," in *Proc. Int. Conf. Inf. Process. Sens. Netw. (IPSN 2006)*, Nashville, TN, Apr. 2006, pp. 177–185.
[35] M. Fornasier and H. Rauhut, "Recovery algorithms for vector-valued data with joint sparsity constraints," *SIAM J. Numer. Anal.*, vol. 46, no. 2, pp. 577–613, 2008.
[36] M. Mishali and Y. Eldar, "The continuous joint sparsity prior for sparse representations: Theory and applications," in *Proc. IEEE Int. Workshop Computat. Adv. Multi-Sens. Adapt. Process. (CAMPSAP 2007)*, St. Thomas, U.S. Virgin Islands, Dec. 2007, pp. 125–128.
[37] S. Cotter, B. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, Jul. 2005.
[38] J. Chen and X. Huo, "Sparse representations for multiple measurement vectors (MMV) in an over-complete dictionary," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP 2005)*, Philadelphia, PA, Mar. 2005, pp. 257–260.
[39] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Trans. Signal Process.*, vol. 54, no. 12, pp. 4634–4643, Dec. 2006.
[40] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Process.*, vol. 86, no. 3, pp. 572–588, 2006.
[41] J. A. Tropp, "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation," *Signal Process.*, vol. 86, no. 3, pp. 589–602, 2006.
[42] S. Aeron, M. Zhao, and V. Saligrama, Information Theoretic Bounds to Performance of Compressed Sensing and Sensor Networks 2008 [Online]. Available: http://arxiv.org/pdf/0804.3439v4, to be published
[43] M. Wainwright, "Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT 2007)*, Nice, France, Jun. 2007, pp. 961–965.
[44] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*. New York: Wiley-Interscience, 2001.
[45] A. K. Gupta and D. K. Nagar, *Matrix Variate Distributions*. Boca Raton, FL: Chapman and Hall/CRC, 1999.
[46] Y. S. Chow and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales*. New York: Springer-Verlag, 1988.
[47] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, *Model-Based Compressive Sensing*, submitted for publication.
[48] D. S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear Systems Theory*. Princeton, NJ: Princeton Univ. Press, 2005.
[49] R. A. Horn and C. R. Johnson, *Matrix Analysis*. New York: Cambridge Univ. Press, 1990.
[50] O. Milenkovic, H. Pham, and W. Dai, "Sublinear compressive sensing reconstruction via belief propagation decoding," in *Proc. Int. Symp. Inf. Theory (ISIT 2009)*, Seoul, South Korea, Jul. 2009, pp. 674–678.
[51] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley-Interscience, 2006.
[52] S. Kullback, *Information Theory and Statistics*. New York: Dover , 1997.
[53] Y. G. Yatracos, "A lower bound on the error in nonparametric regression type problems," *Ann. Statist.*, vol. 16, no. 3, pp. 1180–1187, 1988.
[54] L. Birge, "Approximation dans les espaces métriques et théorie de l'estimation," *Probabil. Theory and Related Fields*, vol. 65, no. 2, pp. 181–237, Jun. 1983.
[55] A. Dogandzic and A. Nehorai, "Space-time fading channel estimation and symbol detection in unknown spatially correlated noise," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 457–474, Mar. 2002.
[56] H. L. Van Trees, *Optimum Array Processing*. New York: Wiley-Interscience, 2002.
[57] G. Tang and A. Nehorai, "Support recovery for source localization based on overcomplete signal representation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (CASSP 2010)*.
[58] D. L. Donoho and P. B. Stark, "Uncertainty principles and signal recovery," *SIAM J. Appl. Math.*, vol. 49, no. 3, pp. 906–931, 1989.
[59] D. L. Donoho and B. F. Logan, "Signal recovery and the large sieve," *SIAM J. Appl. Math.*, vol. 52, no. 2, pp. 577–591, 1992.
[60] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Hoboken, NJ: Wiley-Interscience, 2003.
[61] A. K. Fletcher, S. Rangan, V. K. Goyal, and K. Ramchandran, "Denoising by sparse approximation: Error bounds based on rate-distortion theory," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–19, 2006.
[62] R. A. DeVore, "Deterministic constructions of compressed sensing matrices," *J. Complexity*, vol. 23, no. 4–6, pp. 918–925, 2007.
[63] I. Gohberg, P. Lancaster, and L. Rodman, *Indefinite Linear Algebra and Applications*. Basel, Switzerland: Birkhäuser, 2005.
[64] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1996.

Gongguo Tang (S'09) received the B.Sc. degree in mathematics from Shandong University, China, in 2003, and the M.Sc. degree in system science from the Chinese Academy of Sciences, in 2006.

Currently, he is a Ph.D. degree candidate with the Department of Electrical and Systems Engineering, Washington University, St. Louis, MO, under the guidance of Dr. Arye Nehorai. His research interests are in the area of compressive sensing, statistical signal processing, detection and estimation, and their applications.

Arye Nehorai (S'80–M'83–SM'90–F'94) received the B.Sc. and M.Sc. degrees in electrical engineering from the Technion-Israel Institute of Technology, Haifa, Israel, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA.

From 1985 to 1995, he was a Faculty Member with the Department of Electrical Engineering, Yale University, New Haven, CT. In 1995, he joined the Department of Electrical Engineering and Computer Science, The University of Illinois at Chicago (UIC), as a Full Professor. From 2000 to 2001, he was Chair of the department's Electrical and Computer Engineering (ECE) Division, which then became a new department. In 2001, he was named University Scholar of the University of Illinois. In 2006, he became Chair of the Department of Electrical and Systems Engineering, Washington University in St. Louis (WUSTL), St. Louis, MO. He is the inaugural holder of the Eugene and Martha Lohman Professorship and the Director of the Center for Sensor Signal and Information Processing (CSSIP) at WUSTL since 2006.

Dr. Nehorai was Editor-in-Chief of the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2000 to 2002. From 2003 to 2005, he was Vice President (Publications) of the IEEE Signal Processing Society (SPS), Chair of the Publications Board, member of the Board of Governors, and member of the Executive Committee of this Society. From 2003 to 2006, he was the founding editor of the special columns on Leadership Reflections in the IEEE SIGNAL PROCESSING MAGAZINE. He was corecipient of the IEEE SPS 1989 Senior Award for Best Paper with P. Stoica, coauthor of the 2003 Young Author Best Paper Award, and corecipient of the 2004 Magazine Paper Award with A. Dogandzic. He was elected Distinguished Lecturer of the IEEE SPS for the term 2004 to 2005. He received the 2006 IEEE SPS Technical Achievement Award and the 2010 IEEE SPS Meritorious Service Award. He is the Principal Investigator of the Multidisciplinary University Research Initiative (MURI) project entitled Adaptive Waveform Diversity for Full Spectral Dominance. He has been a Fellow of the Royal Statistical Society since 1996.