Forty-Seventh Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 30 - October 2, 2009

Performance Analysis of Support Recovery with Joint Sparsity Constraints*

Gongguo Tang
Dept. of Electrical&Systems Engineering
Washington University in St. Louis
St. Louis, MO 63130-1127
gt2@ese.wustl.edu

Arye Nehorai
Dept. of Electrical&Sytems Engineering
Washington University in St. Louis
St. Louis, MO 63130-1127
nehorai@ese.wustl.edu

Abstract

In this paper, we analyze the performance of estimating the common support for jointly sparse signals based on their projections onto lower-dimensional space. We formulate support recovery as a multiple-hypothesis testing problem and derive both upper and lower bounds on the probability of error for general measurement matrices, by using Chernoff bound and Fano's inequality, respectively. When applied to Gaussian measurement ensembles, these bounds give necessary and sufficient conditions to guarantee a vanishing probability of error for majority realizations of the measurement matrix. Our results offer surprising insights into sparse signal reconstruction based on their projections. For example, as far as support recovery is concerned, the well-known bound in compressive sensing is generally not sufficient if the Gaussian ensemble is used. Our study provides an alternative performance measure, one that is natural and important in practice, for signal recovery in compressive sensing as well as other application areas taking advantage of signal sparsity.

1. Introduction

Support recovery for jointly sparse signals concerns accurately estimating the locations of non-zero components shared by a set of sparse signals based on a limited number of noisy linear observations. More specifically, suppose $\{x(t) \in \mathbb{F}^N, t = 1, 2, \ldots, T\}$ with $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$ is a sequence of jointly sparse signals with a common support S and we have a linear observation model:

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

In most cases, the sparsity level $K \triangleq |S|$ and the number of observations M is far less than N, the dimension of the ambient space. This problem arises naturally in several signal processing areas such as compressive sensing [1–3], source localization [4–6], sparse approximation and signal denoising [7].

Compressive Sensing [1–3] is a major field that motivates our study for support recovery. In the classical setting of compressive sensing, only one snapshot is considered; *i.e.*, $T = 1$. The goal is to recover a long vector $x \triangleq x(1)$ with a small fraction of non-zero coordinates from the much shorter observation vector $y \triangleq y(1)$. Since most natural signals are compressible under some basis and can be well approximated by their $K-$sparse representation [8], this scheme, if properly justified, will reduce the necessary sampling rate beyond the limit set by Nyquist and Shannon [9]. Surprisingly, if $M = O(K \ln(\frac{N}{K})) \ll N$ 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 noise cases [10–13]. Along with these algorithms, rigorous theoretical analysis is provided to guarantee their effectiveness in terms of, for example, various l_p-norms between the estimator $\hat{x}$ and the true value of x [10–13] . 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-square problem once its support is obtained. Therefore, support recovery is a stronger theoretical 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 non-zero 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 non-zero indices in the Fourier domain indicate the harmonics existing in a signal [14], which is critical for tasks such as spectrum sensing for cognitive radios [15]. In compressive DNA arrays for bio-sensing, the existence of certain target agents in the tested solution is reflected by the locations of non-vanishing coordinates, while the magnitudes are determined by their

*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.

978-1-4244-5871-4/09/$26.00 ©2009 IEEE

concentrations [16, 17]. 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 locations in the time-frequency domain. The magnitude determined by coefficients of reflection is of less physical significance [18, 19]. In all these applications, the support is physically more significant than the component values.

The contributions of our work are 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 Chernoff bound and Fano's inequality to derive both upper and lower bounds for the probability of error. The upper bound we derive is for the *optimal* decision rule, in contrast to a performance analysis for specific sub-optimal reconstruction algorithms [10–13, 20]. Hence, the bound can be viewed as a measure of the measurement system's ability to correctly identify the true support. Second, we apply these performance bounds to Gaussian measurement ensemble and derive necessary and sufficient conditions in terms of the system parameters to guarantee a vanishing probability of error. By restricting our attention to Gaussian measurement matrices, we derive a result parallel to those for classical compressive sensing [1, 2], 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 [1, 2]. However, our result suggests that generally it is impossible to obtain the true support accurately with only one snapshot. We also obtain a necessary condition showing that the $\ln \frac{N}{K}$ 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 notations, the mathematical model and assumptions. Section III is devoted to the derivation of an upper bound on the probability of error for support recovery assuming general measurement matrices. In Section IV, an information theoretic lower bound is derived by using Fano's inequality. We focus on the Gaussian ensemble in Section V. The paper is concluded in Section VI.

2. Notations and Mathematical Model

2.1. Notations

We first introduce some notations used throughout the 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 non-zero components of x. For a matrix X, $S = \mathrm{supp}(X)$ denotes the index set of non-zero 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 $x \in \mathbb{F}^N$, $x^S \in \mathbb{F}^N$ 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 is the submatrix with columns from B indicated by S.

Transpose conjugate of a vector or matrix is denoted by †. $A \otimes B$ represents the Kronecker product of two matrices. The identity matrix of dimension M is I_M. The trace of matrix A is given by $\mathrm{tr}(A)$.

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 [21] for $Y \in \mathbb{F}^{M\times T}$ 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 $f_n \leq C g_n$ for $n \geq N$. Similarly, $f_n = \Omega(g_n)$ means $f_n \geq C g_n$ for $n \geq N$.

2.2. Mathematical Model

We describe the mathematical model and relevant assumptions in this subsection. Suppose $x(t) \in \mathbb{F}^N, t = 1, \cdots, T$ are jointly sparse signals with common support $S = \mathrm{supp}(x(t))$, whose size $K = |S|$ is known.

In model (1), we assume that the vectors $x^S(t), t = 1, \cdots, T$ formed by the non-zero components of $x(t)$ follow *i.i.d.* $\mathbb{F}\mathcal{N}(0, \mathrm{I}_K)$ and the additive noises $w(t) \in \mathbb{F}^N$ follow *i.i.d.* $\mathbb{F}\mathcal{N}(0, \sigma^2 \mathrm{I}_M)$. Note that assuming unit variance for signals loses no generality since only the ratio of signal variance to noise variance appears in all subsequence analyses. In this sense, we view $1/\sigma^2$ as the signal-to-noise ratio (SNR).

Let $X = \begin{bmatrix} x(1) & x(2) & \cdots & x(T) \end{bmatrix}$ and Y, W be defined in a similar manner. Then we write the model in a more compact matrix form:

$$Y = AX + W. \tag{2}$$

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 non-singular, we call A a *non-degenerate* 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 [10]. While most of our results apply to general non-degenerate measurement matrices, we need to impose more structures on measurement matrices in order to obtain more profound results. In particular, we will consider Gaussian measurement matrix A whose elements 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 A, we still use non-bold A except in Section V. The role played by the variance of 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:

$$\left\{ \begin{array}{l} \mathrm{H}_0 : \mathrm{supp}(X) = S_0 \\ \mathrm{H}_1 : \mathrm{supp}(X) = S_1 \end{array} \right. . \tag{3}$$

The results we obtain for binary binary support recovery (3) offer insight into our second problem: the multiple support recovery. In the multiple support recovery problem we choose one among $L \triangleq \binom{N}{K}$ distinct candidate supports of X, which is a multiple-hypothesis testing problem:

$$\left\{ \begin{array}{ll} \mathrm{H}_0 : & \mathrm{supp}(X) = S_0 \\ \mathrm{H}_1 : & \mathrm{supp}(X) = S_1 \\ & \vdots \\ \mathrm{H}_{L-1} : & \mathrm{supp}(X) = S_{L-1} \end{array} \right. . \tag{4}$$

3. Upper Bound on Probability of Error for Non-degenerate 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 (3) and (4). An excellent introduction to hypothesis testing and Chernoff bound can be found in [22]. We first study binary support recovery, which lays foundation for the general support recovery problem. The result is based on a theorem that counts the eigenvalues of a particular matrix defined by the columns of the measurement matrix corresponding to the two candidate supports. We then obtain a bound for the general support recovery problem via the union bound.

Under model (2) and the assumptions, the observations Y follow a matrix variate Gaussian distribution $Y|S \sim \mathbb{F}\mathscr{N}_{M,T}(0, \Sigma_S \otimes \mathrm{I}_T)$, when the true support is S, and the probability density function is given by

$$p\,(Y|S) = \frac{1}{(\pi/\kappa)^{\kappa MT} |\Sigma_S|^{\kappa T}} \exp\left[-\kappa \mathrm{tr}\left(Y^\dagger \Sigma_S^{-1} Y\right)\right], \tag{5}$$

where $\Sigma_S = A_S A_S^\dagger + \sigma^2 \mathrm{I}_M$ is the common covariance matrix for each column of Y [21]. The binary support recovery problem (3) is equivalent to a linear Gaussian binary hypothesis testing problem with equal prior probability:

$$\left\{ \begin{array}{l} \mathrm{H}_0 : Y \sim \mathbb{F}\mathscr{N}_{M,T}(0, \Sigma_{S_0} \otimes \mathrm{I}_T) \\ \mathrm{H}_1 : Y \sim \mathbb{F}\mathscr{N}_{M,T}(0, \Sigma_{S_1} \otimes \mathrm{I}_T) \end{array} \right. . \tag{6}$$

Note that the notation for Σ_S is inconsistent with our notation convention. We henceforth will denote Σ_{S_i} by Σ_i.

The optimal decision rule with minimal probability of error is given by the likelihood ratio test [22]

$$\ell(Y) = -\kappa \mathrm{tr}\left[Y^\dagger\left(\Sigma_1^{-1} - \Sigma_0^{-1}\right) Y\right] - \kappa T \ln \frac{|\Sigma_1|}{|\Sigma_0|} \underset{\mathrm{H}_0}{\overset{\mathrm{H}_1}{\gtrless}} 0. \tag{7}$$

The Chernoff bound states that the probability of error for the optimal decision rule (7) is upper bounded:

$$P_{\mathrm{err}} \le \frac{1}{2} e^{\mu(s)}, \forall s \in [0,1], \tag{8}$$

where $\mu(s)$ is the log moment generating function for $\ell(Y)$.

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, $s \approx \frac{1}{2}$ gives a good bound. We then take $s = \frac{1}{2}$ in the Chernoff bound (8). Using the simultaneous diagonalization property of H and H^{-1}, it is easy to show that

$$\begin{aligned} \mu(1/2) &= -\kappa T \ln \left| \frac{H^{1/2} + H^{-1/2}}{2} \right| \\ &= -\kappa T \left[\sum_{j=1}^{k_0} \ln\left(\left(\sqrt{\lambda_j} + 1/\sqrt{\lambda_j} \right)/2 \right) \right. \\ &\qquad \left. + \sum_{j=1}^{k_1} \ln\left(\left(\sqrt{\sigma_j} + 1/\sqrt{\sigma_j} \right) 2 \right) \right], \end{aligned} \tag{9}$$

where $\lambda_1 \ge \cdots \ge \lambda_{k_0} > 1 = \cdots = 1 > \sigma_1 \ge \cdots \ge \sigma_{k_1}$ are the eigenvalues of H. 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 non-degenerate measurement matrix A. We present the following two propositions on the eigenvalue structure of H without proofs:

Proposition 1 *For any non-degenerate measurement matrix A, let* $H = \Sigma_0^{1/2} \Sigma_1^{-1} \Sigma_0^{1/2}$ *with* $\Sigma_i = A_{S_i} A_{S_i}^\dagger + \sigma^2 \mathrm{I}_M$, $k_i = |S_0 \cap S_1|, k_0 = |S_0 \backslash S_1| = |S_0| - k_i, k_1 = |S_1 \backslash S_0| = |S_1| - k_i$ *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*.*

Proposition 2 *With the same setup as in Proposition 1,the sorted eigenvalues of H that are greater than one are lower bounded by the corresponding eigenvalues of* $\mathrm{I}_{k_0} + \frac{1}{\sigma^2} R R^\dagger$*, where R is the* $k_0 \times k_0$ *submatrix at the lower-right corner of the upper triangle matrix in the QR decomposition of* $\begin{bmatrix} A_{S_1 \backslash S_0} & A_{S_1 S_0} & A_{S_0 \backslash S_1} \end{bmatrix}$*; and upper bounded by the corresponding eigenvalues of* $\mathrm{I}_{k_0} + \frac{1}{\sigma^2} A_{S_0 \backslash S_1} A_{S_0 \backslash S_1}^\dagger$

For binary support recovery (3) with $|S_0| = |S_1| = K$, we have $k_0 = k_1 \triangleq k_d$. Dropping the term $1/\sqrt{\lambda_j}$ and $\sqrt{\sigma_j}$ in (9), we have

Proposition 3 *The probability of error for the binary support recovery problem* (3) *is bounded by*

$$P_{\mathrm{err}} \quad \le \quad \frac{1}{2} \left[\frac{\bar{\lambda}_{S_0,S_1} \bar{\lambda}_{S_1,S_0}}{16} \right]^{-\kappa k_d T/2}, \tag{10}$$

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.*

Now we turn to analyze the probability of error for the multiple support recovery problem (4). We assume each candidate support S_i has known cardinality K, and we have $L = \binom{N}{K}$ such supports. Under the equal prior probability assumption, the optimal decision rule is given by

$$\mathrm{H}^* = \mathrm{argmax}_{0 \leq i \leq L-1} p\left(Y | \mathrm{H}_i\right). \tag{11}$$

Since the error event decomposes as $\{\mathrm{H}^* \neq \mathrm{H}_i\} = \bigcup_{j \neq i} \{p\left(Y|\mathrm{H}_j\right) > p\left(Y|\mathrm{H}_i\right)\}$, application of the union bound gives:

Theorem 1 *If* $\bar{\lambda} > 4\left[K(N-K)\right]^{\frac{1}{\kappa T}}$ *with* $\bar{\lambda} = \min_{i \neq j}\{\bar{\lambda}_{S_i,S_j}\}$, *then the probability of error for the full support recovery problem* (4) *with* $|S_i| = K$ *and* $L = \binom{N}{K}$ *is bounded by*

$$P_{\mathrm{err}} \leq \frac{1}{2} \frac{\frac{K(N-K)}{\left(\bar{\lambda}/4\right)^{\kappa T}}}{1 - \frac{K(N-K)}{\left(\bar{\lambda}/4\right)^{\kappa T}}}. \tag{12}$$

We note that increasing the number of temporal samples can force the probability of error to 0 exponentially fast as long as $\bar{\lambda}$ exceeds the threshold $4[K(N-K)]^{\frac{1}{\kappa T}}$. The final bound (12) is of the same order as the probability of error when $k_d = 1$. 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.

4. An Information Theoretical Lower Bound on Probability of Error

In this section, we derive an information theoretical 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 [23]. Suppose that we have a random vector Y with L possible densities $f_0, \ldots, f_{L-1}$. Then by Fano's inequality [24], [25], the probability of error for *any* decision rule to identify the true density is lower bounded by

$$P_{\mathrm{err}} \geq 1 - \frac{\beta + \ln 2}{\ln L}, \tag{13}$$

where $\beta = \frac{1}{L^2}\sum_{i,j} D_{KL}(f_i || f_j)$ is the average of the Kullback-Leibler divergence [23] between all pairs of densities.

Since in the multiple support recovery problem (4), all the distributions involved are matrix variate Gaussian distributions with mean 0 and different variances, a direct computation gives the expression of β:

$$\beta = \frac{\kappa T}{2L^2} \sum_{i,j} [\mathrm{tr}(H_{i,j}) - M],$$

where $H_{i,j} = \Sigma_i^{1/2}\Sigma_j^{-1}\Sigma_i^{1/2}$. Invoking Proposition 1 and the second part of Proposition 2, we derive an upper bound on the average Kullback-Leibler divergence

$$\beta \leq \frac{\kappa T}{2L^2\sigma^2} \sum_{i=0}^{L-1} \sum_{k_d=1}^{K} \binom{K}{K-k_d} \binom{N-K}{k_d} \sum_{\substack{1 \leq m \leq M \\ n \in S_i \setminus S_j}} |A_{mn}|^2 .$$

Due to the symmetry of the right-hand side, it must be of the form $\alpha \sum_{\substack{1 \leq m \leq M \\ 1 \leq n \leq N}} |A_{mn}|^2 = \alpha \|A\|_F^2$, where $\|\cdot\|_F$ is the Frobenius norm. Setting all $A_{mn} = 1$ and using the mean expression for hypergeometric distribution gives $\alpha = \frac{\kappa T K(N-K)}{2\sigma^2 N^2}$. Hence, we get

$$\beta \leq \frac{\kappa T K\left(N-K\right)}{2\sigma^2 N^2} \|A\|_F^2 .$$

Therefore, the probability of error is lower bounded by

$$P_{\mathrm{err}} \geq 1 - \frac{\frac{\kappa T K(N-K)}{2\sigma^2 N^2}\|A\|_F^2 + \ln 2}{\ln L}. \tag{14}$$

We conclude with the following theorem:

Theorem 2 *For multiple support recovery problem* (4)*, the probability of error for any decision rule is lower bounded by*

$$P_{\mathrm{err}} \geq 1 - \frac{\kappa T \frac{K}{N}\left(1 - \frac{K}{N}\right)\|A\|_F^2}{2\sigma^2 \ln \binom{N}{K}} + o\left(1\right). \tag{15}$$

Terms in the bound (15) have clear meanings. The Frobenius norm $\|A\|_F^2$ of the measurement matrix is the maximal possible total gain of system (1). Since the measured signal is $K-$sparse, only a fraction of the gain plays a role in the measurements, and its average over all possible $K-$sparse signals is $\frac{K}{N}\left(1 - \frac{K}{N}\right)\|A\|_F^2$. When K is small compared with N, approximately $\frac{K}{N}$ of $\|A\|_F^2$ contributes to the measurements. The term $\ln L = \ln \binom{N}{K}$ is the total uncertainty or entropy of the support variable S, since we impose a uniform prior on it. As long as $K \leq \frac{N}{2}$, increasing K will increase both the average gain exploited by the measurement system, namely, $\frac{K}{N}\left(1 - \frac{K}{N}\right)\|A\|_F^2$, and the entropy of the support variable S. The overall effect is, quite counterintuitively, a decrease of the lower bound in (15). We note that the bound decreases linearly with T, in contrast with (12)'s exponential decrease. This linear decrease will drive the bound to zero with finite samples. We commented previously that $\frac{1}{\sigma^2}$ plays the role of the signal-to-noise ratio. Therefore, increasing the SNR will also force the bound to zero.

5. Support Recovery for the Gaussian Measurement Ensemble

In this section, we refine our results in previous sections from general non-degenerate measurement matrices to the Gaussian ensemble. Unless otherwise specified, we always assume that the elements in a measurement matrix A are *i.i.d.* samples from unit variance real or complex normal distributions. The Gaussian measurement ensemble is widely used and studied in compressive sensing [1–3, 9, 26]. The additional structure and the theoretical tools available enable us to derive deeper results in this case. We first show two results on the eigenvalue structures of the Gaussian measurement matrix. 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.

5.1. Eigenvalue Structure for a Gaussian Measurement Matrix

We present refined results on the eigenvalue behavior for Gaussian measurement matrix in this subsection. First, we observe that a Gaussian measurement matrix is non-degenerate with probability one, since any $p \leq M$ random vectors $x_1, x_2, \ldots, x_p$ 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 Theorem 3.2.1 in [21]). As a consequence, Proposition 1 holds with probability one for Gaussian measurement matrix A. Furthermore, we refine Proposition 2 based on the well-known QR factorization for Gaussian matrices [21].

Corollary 1 *For Gaussian measurement matrix A, let* $H = \Sigma_0^{1/2}\Sigma_1^{-1}\Sigma_0^{1/2}$, $k_i = |S_0 \cap S_1|, k_0 = |S_0 \backslash S_1| = |S_0| - k_i, k_1 = |S_1 \backslash S_0| = |S_1| - k_i$ *and assume* $M \geqslant k_0 + k_1$. *Then with probability one, the sorted eigenvalues of H that are greater than 1 are lower bounded by the corresponding ones of* $\mathrm{I}_{k_0} + \frac{1}{\sigma^2}RR^\dagger$, *where the elements of* $R = (r_{mn})_{k_0\times k_0}$ *satisfy*

$$\begin{aligned} 2\kappa r_{mn}^2 &\sim \chi^2_{2\kappa(M-k_1-k_i-m+1)}, m = 1, \cdots, k_0, \\ r_{mn} &\sim \mathbb{F}\mathcal{N}(0,1), 1 \leq m < n \leq k_0. \end{aligned}$$

Now with the distributions on the elements of the bounding matrices, direct calculation shows that the expected values for 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. When γ is much less than $1 + \frac{M-2K}{\sigma^2}$, we expect that $\mathbb{P}\left\{\bar{\lambda}_{S_i,S_j} \leq \gamma\right\}$ 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*

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

We will use Lemma 1 to establish a sufficient condition for support recovery in Section V-C. We first present a necessary condition in the next section due to its relative simplicity.

5.2. 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, non-trivial 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 in order to guarantee respectively that, first, $\mathbb{E}P_{\mathrm{err}}$ turns to zeros and, second, for majority realizations of A, the probability of error vanishes. More precisely, we have the following theorem:

Theorem 3 *In the support recovery problem* (4), *for any* $\varepsilon, \delta > 0$, *a necessary condition for* $\mathbb{E}P_{\mathrm{err}} < \varepsilon$ *is*

$$\frac{\kappa MT}{\sigma^2} \geq (1-\varepsilon)\frac{2\ln\binom{N}{K}}{K\left(1-\frac{K}{N}\right)} + o(1), \tag{16}$$

and a necessary condition for $\mathbb{P}\left\{P_{\mathrm{err}}(A) \leq \varepsilon\right\} \geq 1-\delta$ *is*

$$\frac{\kappa MT}{\sigma^2} \geq (1-\varepsilon-\delta)\frac{2\ln\binom{N}{K}}{K\left(1-\frac{K}{N}\right)} + o(1). \tag{17}$$

Proof: Equation (16) is obtained by substituting $\mathbb{E}\|A\|_F^2 = \sum_{m,l}\mathbb{E}|A_{ml}|^2 = MN$ into equation (15). Denote by E the event $\{A : P_{\mathrm{err}}(A) \leq \varepsilon\}$; then $\mathbb{P}\{E^c\} \leq \delta$ and the following derivation

$$\mathbb{E}P_{\mathrm{err}} = \int_E P_{\mathrm{err}}(A) + \int_{E^c} P_{\mathrm{err}}(A) \leq \varepsilon\mathbb{P}(E) + \mathbb{P}(E^c)$$

yields equation (17). ■

Our result shows that as far as support recovery is concerned, one cannot avoid the $\ln\frac{N}{K}$ 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

$$T \geq \frac{2\sigma^2\ln\binom{N}{K}}{\kappa K(1-K/N)} + o(1) \geq \frac{2\sigma^2}{\kappa}\ln\frac{N}{K} + o(1),$$

which is independent of M. Therefore, it is impossible to have a vanishing $\mathbb{E}P_{\mathrm{err}}$ 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. We note that $\ln\binom{N}{K}$ is the uncertainty of the support variable S, and $\ln\frac{N}{K}$ 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.

5.3. Sufficient Condition

We derive a sufficient condition in parallel with sufficient conditions in the literature of compressive sensing.

In compressive sensing, when only one temporal sample is available, $M = \Omega\left(K \ln \frac{N}{K}\right)$ is enough for stable signal reconstruction for the majority realizations of measurement matrix 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 vanishing probability of error. Therefore, we consider a Gaussian ensemble with unit variance.

The large deviation lemma 1 together with the union bound gives us the following sufficient condition for support recovery:

Theorem 4 *Suppose that*

$$M = \Omega\left(K \ln \frac{N}{K}\right) \tag{18}$$

and

$$\kappa T \ln \frac{M}{\sigma^2} \gg \ln\left[K(N-K)\right]. \tag{19}$$

Then with probability approaching one, a realization of measurement matrix A from a Gaussian ensemble produces a system (2) *that its optimal decision rule* (11) *for multiple support recovery problem* (4) *has a vanishing* P_{err}. *In particular, if* $M = \Omega\left(K \ln \frac{N}{K}\right)$ *and*

$$T \gg \frac{\ln N}{\ln \ln N}, \tag{20}$$

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, by applying Lemma 1, we have

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

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 $E = \left\{A : \bar{\lambda}(A) > \gamma\right\}$, where $\mathbb{P}\{E\}$ approaches one as N turns to infinity. Now the upper bound (12) becomes

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

Hence, if $\kappa T \ln \frac{M}{\sigma^2} \gg \ln\left[K(N-K)\right]$, we get a vanishing probability of error. In particular, under the assumption that $M \ge \Omega\left(K \log \frac{N}{K}\right)$, if $T \gg \frac{\ln N}{\ln \ln N}$, then $\frac{\ln[K(N-K)]}{\ln\left[K \ln \frac{N}{K}\right]} \le \frac{\ln N}{\ln \ln N}$ implies that $K(N-K) \ll O\left(\left(\frac{K \ln \frac{N}{K}}{\sigma^2}\right)^{\kappa T}\right) = O\left(\frac{K(N-K)}{\left(M/\sigma^2\right)^{\kappa T}}\right)$ 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 square problem, we still might not be able to get the correct support because of the noise, if 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, 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 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 $M \ge \ln N$ and $\kappa T \ln \frac{M}{\sigma^2} \gg \ln N$. Now the number of total samples should satisfy $MT \gg \frac{(\ln N)^2}{\ln \ln N}$ while the necessary condition states that $MT = \Omega(\ln N)$. The smallest gap between the necessary condition and sufficient condition is achieved when $K = 1$.

6. Conclusions

In this paper, we formulated the support recovery problems for jointly sparse signals as multiple-hypothesis testings. We then employed Chernoff bound and Fano's inequality to derive bounds on the probability of error for support recovery. We discussed the implications of these bounds once they are derived. We presented sufficient and necessary conditions to achieve a vanishing probability of error in both the mean and large probability senses for Gaussian measurement ensemble. These conditions show the necessity of considering multiple temporal samples. 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 $1/M$.

References

[1] 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.

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

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

[4] 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.

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

[6] V. Cevher, M. Duarte, and R. G. Baraniuk, "Distributed target localization via spatial sparsity," in *European Signal Processing Conference (EUSIPCO 2008)*, Lausanne, Switzerland, Aug. 2008.

[7] 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.

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

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

[10] 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.

[11] 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.

[12] 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.

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

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

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

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

[17] 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 Processing*, vol. 2, no. 3, pp. 275–285, Jun. 2008.

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

[19] 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.

[20] 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.

[21] A. K. Gupta and D. K. Nagar, *Matrix Variate Distributions*. Boca Raton, FL: Chapman & Hall/CRC, 1999.

[22] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I.* New York: Wiley-Interscience, 2001.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley-Interscience, 2006.

[24] Y. G. Yatracos, "A lower bound on the error in nonparametric regression type problems," *Ann. Statist.*, vol. 16, no. 3, pp. 1180–1187, 1988.

[25] M. Wainwright, "Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting," in *IEEE Int. Symp. Information Theory (ISIT 2007)*, Nice, France, Jun. 2007, pp. 961–965.

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