Necessary and Sufficient Conditions on Sparsity Pattern Recovery

Alyson K. Fletcher, *Member, IEEE,* Sundeep Rangan,
and Vivek K Goyal, *Senior Member, IEEE*

Abstract

The problem of detecting the sparsity pattern of a k-sparse vector in $\mathbb{R}^n$ from m random noisy measurements is of interest in many areas such as system identification, denoising, pattern recognition, and compressed sensing. This paper addresses the scaling of the number of measurements m, with signal dimension n and sparsity-level nonzeros k, for asymptotically-reliable detection. We show a necessary condition for perfect recovery at any given SNR for all algorithms, regardless of complexity, is $m = \Omega(k \log(n - k))$ measurements. Conversely, it is shown that this scaling of $\Omega(k \log(n - k))$ measurements is sufficient for a remarkably simple "maximum correlation" estimator. Hence this scaling is optimal and does not require more sophisticated techniques such as lasso or matching pursuit. The constants for both the necessary and sufficient conditions are precisely defined in terms of the minimum-to-average ratio of the nonzero components and the SNR. The necessary condition improves upon previous results for maximum likelihood estimation. For lasso, it also provides a necessary condition at any SNR and for low SNR improves upon previous work. The sufficient condition provides the first asymptotically-reliable detection guarantee at finite SNR.

Index Terms

compressed sensing, convex optimization, lasso, maximum likelihood estimation, random matrices, random projections, regression, sparse approximation, sparsity, subset selection

I. INTRODUCTION

Suppose one is given an observation $y \in \mathbb{R}^m$ that was generated through $y = Ax + d$, where $A \in \mathbb{R}^{m \times n}$ is known and $d \in \mathbb{R}^m$ is an additive noise vector with a known distribution. It may be desirable for an estimate of x to have a small number of nonzero components. An intuitive example is when one wants to choose a small subset from a large number of possibly-related factors that linearly influence a vector of observed data. Each factor corresponds to a column of A, and one wishes to find a small subset of columns with which to form a linear combination that closely matches the observed data y. This is the subset selection problem in (linear) regression [1], and it gives no reason to penalize large values for the nonzero components.

In this paper, we assume that the true signal x has k nonzero entries and that k is known when estimating x from y. We are concerned with establishing necessary and sufficient conditions for the recovery of the *positions* of the nonzero entries of x, which we call the *sparsity pattern*. Once the sparsity pattern is correct, $n - k$ columns of A can be ignored and the stability of the solution is well understood; however, we do not study any other performance criterion.

A. Previous Work

Sparsity pattern recovery (or more simply, sparsity recovery) has received considerable attention in a variety guises. Most transparent from our formulation is the connection to sparse approximation. In a typical sparse approximation problem, one is given data $y \in \mathbb{R}^m$, dictionary[1] $A \in \mathbb{R}^{m \times n}$, and tolerance $\epsilon > 0$. The aim is to find $\hat{x}$ with the fewest number of nonzero entries among those satisfying $\|A\hat{x} - y\| \leq \epsilon$. This problem is NP-hard [3] but greedy heuristics (matching pursuit [2] and its variants) and convex relaxations (basis pursuit [4], lasso [5] and others) can be effective under certain conditions on A and y [6]–[8]. Scaling laws for sparsity recovery with any A were first given in [9].

More recently, the concept of "sensing" sparse x through multiplication by a suitable random matrix A, with measurement error d, has been termed *compressed sensing* [10]–[12]. This has popularized the study of sparse approximation with respect to random dictionaries, which was considered also in [13]. Results are generally phrased as the asymptotic scaling of the number of measurements m (the length of y) needed for sparsity recovery to succeed with high probability, as a function of the other problem parameters. More specifically, most results are sufficient conditions for specific tractable recovery algorithms to succeed. For example, if A has i.i.d. Gaussian entries and $d = 0$, then $m \asymp 2k \log(n/k)$ dictates the minimum scaling at which basis pursuit succeeds with high probability [14]. With nonzero noise variance, necessary and sufficient conditions for the success of lasso in this setting have the asymptotic scaling [15]

$$m \asymp 2k \log(n - k) + k + 1. \tag{1}$$

To understand the ultimate limits of sparsity recovery, while also casting light on the efficacy of lasso or orthogonal matching pursuit (OMP), it is of interest to determine necessary and sufficient conditions for an optimal recovery algorithm to succeed. Of course, since it is sufficient for lasso, the condition (1) is sufficient for an optimal algorithm. Is it close to a necessary condition? We address precisely this question by proving a necessary condition that differs from (1) by a factor that is *constant with respect to n and k* while depending on the signal-to-noise ratio (SNR) and mean-to-average ratio (MAR), which will be defined precisely in Section II. Furthermore, we present an extremely simple algorithm for which a sufficient condition for sparsity recovery is similarly within a constant factor of (1).

Previous necessary conditions had been based on information-theoretic analyses such as the capacity arguments in [16], [17] and a use of Fano's inequality in [18]. More recent publications with necessary conditions include [19]–[22]. As described in Section III, our new necessary conditions are stronger than the previous results.

Table I previews our main results and places (1) in context. The measurement model and parameters MAR and SNR are defined in the following section. Arbitrarily small constants have been omitted, and the last column—labeled simply $\mathsf{SNR} \to \infty$—is more specifically for $\mathsf{MAR} > \epsilon > 0$ for some fixed ϵ and $\mathsf{SNR} = \Omega(k)$.

B. Paper Organization

The setting is formalized in Section II. In particular, we define our concepts of signal-to-noise ratio and mean-to-average ratio; our results clarify the roles of these quantities in the sparsity recovery problem. Necessary conditions for success of any algorithm are considered in Section III. There we present a new necessary condition and compare it to previous results and numerical experiments. Section IV introduces a very simple recovery algorithm for the purpose of showing that a sufficient condition for its success is rather weak—it has the same dependence on n and k as (1). Conclusions are given in Section V, and proofs appear in the Appendix.

---

[1]The term seems to have originated in [2] and may apply to A or the columns of A as a set.

	finite SNR	SNR $\to \infty$
Any algorithm must fail	$m < \frac{2}{\mathsf{MAR}\cdot\mathsf{SNR}}k\log(n-k)+k-1$ Theorem 1	$m \leq k$ (elementary)
Necessary and sufficient for lasso	unknown (expressions above and right are necessary)	$m \asymp 2k\log(n-k)+k+1$ Wainwright [15]
Sufficient for maximum correlation estimator (8)	$m > \frac{8(1+\mathsf{SNR})}{\mathsf{MAR}\cdot\mathsf{SNR}}k\log(n-k)$ Theorem 2	$m > \frac{8}{\mathsf{MAR}}k\log(n-k)$ from Theorem 2

TABLE I

SUMMARY OF RESULTS ON MEASUREMENT SCALING FOR RELIABLE SPARSITY RECOVERY

(SEE BODY FOR DEFINITIONS AND TECHNICAL LIMITATIONS)

II. PROBLEM STATEMENT

Consider estimating a k-sparse vector $x \in \mathbb{R}^n$ through a vector of observations,

$$y = Ax + d, \tag{2}$$

where $A \in \mathbb{R}^{m \times n}$ is a random matrix with i.i.d. $\mathcal{N}(0, 1/m)$ entries and $d \in \mathbb{R}^m$ is i.i.d. unit-variance Gaussian noise. Denote the sparsity pattern of x (positions of nonzero entries) by the set I_{true}, which is a k-element subset of the set of indices $\{1, 2, \ldots, n\}$. Estimates of the sparsity pattern will be denoted by $\hat{I}$ with subscripts indicating the type of estimator. We seek conditions under which there exists an estimator such that $\hat{I} = I_{\text{true}}$ with high probability.

In addition to the signal dimensions, m, n and k, we will show that there are two variables that dictate the ability to detect the sparsity pattern reliably: the SNR, and what we will call the *minimum-to-average ratio*.

The SNR is defined by

$$\mathsf{SNR} = \frac{\mathbf{E}[\|Ax\|^2]}{\mathbf{E}[\|d\|^2]} = \frac{\mathbf{E}[\|Ax\|^2]}{m}. \tag{3}$$

Since we are considering x as an unknown deterministic vector, the SNR can be further simplified as follows: The entries of A are i.i.d. $\mathcal{N}(0, 1/m)$, so columns $a_i \in \mathbb{R}^m$ and $a_j \in \mathbb{R}^m$ of A satisfy $\mathbf{E}[a_i' a_j] = \delta_{ij}$. Therefore, the signal energy is given by

$$
\begin{aligned}
\mathbf{E}\left[\|Ax\|^2\right] &= \mathbf{E}\left[\|\sum_{j \in I_{\text{true}}} a_j x_j\|^2\right] \\
&= \sum_{i,j \in I_{\text{true}}} \mathbf{E}\left[a_i' a_j x_i x_j\right] \\
&= \sum_{i,j \in I_{\text{true}}} x_i x_j \delta_{ij} = \|x\|^2.
\end{aligned}
$$

Substituting into the definition (3), the SNR is given by

$$\mathsf{SNR} = \frac{1}{m}\|x\|^2. \tag{4}$$

The minimum-to-average ratio of x is defined as

$$\mathsf{MAR} = \frac{\min_{j \in I_{\text{true}}} |x_j|^2}{\|x\|^2/k}. \tag{5}$$

Since $\|x\|^2/k$ is the average of $\{|x_j|^2 \mid j \in I_{\text{true}}\}$, $\mathsf{MAR} \in (0, 1]$ with the upper limit occurring when all the nonzero entries of x have the same magnitude.

Remarks: Other works use a variety of normalizations, e.g.: the entries of A have variance $1/n$ in [12], [20]; the entries of A have unit variance and the variance of d is a variable σ^2 in [15], [18], [21], [22]; and our scaling of A and a noise variance of σ^2 are used in [23]. This necessitates great care in comparing results.

Some results involve

$$\mathsf{MAR} \cdot \mathsf{SNR} = \frac{k}{m} \min_{j \in I_{\text{true}}} |x_j|^2.$$

While a similar quantity affects a regularization weight sequence in [15], there it does not affect the number of measurements required for the success of lasso.[2] The magnitude of the smallest nonzero entry of x is also prominent in the phrasing of results in [21], [22].

III. NECESSARY CONDITION FOR SPARSITY RECOVERY

We first consider sparsity recovery without being concerned with computational complexity of the estimation algorithm. Since the vector $x \in \mathbb{R}^n$ is k-sparse, the vector Ax belongs to one of $L = \binom{n}{k}$ subspaces spanned by k of the n columns of A. Estimation of the sparsity pattern is the selection of one of these subspaces, and since the noise d is Gaussian, the probability of error is minimized by choosing the subspace closest to the observed vector y. This results in the maximum likelihood (ML) estimate.

Mathematically, the ML estimator can be described as follows. Given a subset $J \subseteq \{1, 2, \ldots, n\}$, let $P_J y$ denote the orthogonal projection of the vector y onto the subspace spanned by the vectors $\{a_j \mid j \in J\}$. The ML estimate of the sparsity pattern is

$$\hat{I}_{\text{ML}} = \arg\max_{J \ : \ |J|=k} \|P_J y\|^2,$$

where $|J|$ denotes the cardinality of J. That is, the ML estimate is the set of k indices such that the subspace spanned by the corresponding columns of A contain the maximum signal energy of y.

Since the number of subspaces, L, grows exponentially in n and k, an exhaustive search is computationally infeasible. However, the performance of ML estimation provides a lower bound on the number of measurements needed by any algorithm that cannot exploit a priori information on x other than it being k-sparse.

Theorem 1: Let $k = k(n)$ and $m = m(n)$ vary with n such that $\lim_{n \to \infty} k(n) = \infty$ and

$$m(n) < \frac{2 - \delta}{\mathsf{MAR} \cdot \mathsf{SNR}} k \log(n - k) + k - 1 \tag{6}$$

for some $\delta > 0$. Then even the ML estimator asymptotically cannot detect the sparsity pattern, i.e.,

$$\lim_{n \to \infty} \Pr\left(\hat{I}_{\text{ML}} = I_{\text{true}}\right) = 0.$$

Proof: See Appendix B.   ∎

The theorem shows that for fixed SNR and MAR, the scaling $m = \Omega(k \log(n - k))$ is necessary for reliable sparsity pattern recovery. The next section will show that this scaling can be achieved with an extremely simple method.

Remarks:

1) The theorem applies for any $k(n)$ such that $\lim_{n \to \infty} k(n) = \infty$, including both cases with $k = o(n)$ and $k = \Theta(n)$. In particular, under linear sparsity ($k = \alpha n$ for some constant α), the theorem shows that

$$m \asymp \frac{2\alpha}{\mathsf{MAR} \cdot \mathsf{SNR}} n \log n$$

measurements are necessary for sparsity recovery. Similarly, if m/n is bounded above by a constant, then sparsity recovery will certainly fail unless $k = O(n/\log n)$.

---

[2]The formulation of [15] makes $\mathsf{SNR} = \Theta(n)$, which obscures the effect of the noise level. See also the second remark following Theorem 2.

2) In the case of $\mathsf{MAR} \cdot \mathsf{SNR} < 1$, the bound (6) improves upon the necessary condition of [15] for the asymptotic success of lasso by the factor $(\mathsf{MAR} \cdot \mathsf{SNR})^{-1}$.

3) The bound (6) can be compared against the information-theoretic bounds mentioned earlier. The tightest of these bounds is in [17] and shows that the problem dimensions must satisfy

$$\frac{2}{m} \log_2 \binom{n}{k} \leq \log_2(1 + \mathsf{SNR}) - \alpha \log_2(1 + \frac{\mathsf{SNR}}{\alpha}), \tag{7}$$

where $\alpha = k/n$ is the *sparsity ratio*. For large n and k, the bound can be rearranged as

$$m \geq \frac{2h(\alpha)}{\alpha} \left[\log_2(1 + \mathsf{SNR}) - \alpha \log_2(1 + \frac{\mathsf{SNR}}{\alpha}) \right]^{-1} k,$$

where $h(\cdot)$ is the binary entropy function. In particular, when the sparsity ratio α is fixed, the bound shows only that m needs to grow at least linearly with k. In contrast, Theorem 1 shows that with fixed sparsity ratio $m = \Omega(k \log(n - k))$ is necessary for reliable sparsity recovery. Thus, the bound in Theorem 1 is significantly tighter and reveals that the previous information-theoretic necessary conditions from [16]–[18], [21], [22] are overly optimistic.

4) Results more similar to Theorem 1—based on direct analyses of error events rather than information-theoretic arguments—appeared in [19], [20]. The previous results showed that with fixed SNR as defined here, sparsity recovery with $m = \Theta(k)$ must fail. The more refined analysis in this paper gives the additional $\log(n - k)$ factor and the precise dependence on $\mathsf{MAR} \cdot \mathsf{SNR}$.

5) Theorem 1 is not contradicted by the relevant sufficient condition of [21], [22]. That sufficient condition holds for scaling that gives linear sparsity and $\mathsf{MAR} \cdot \mathsf{SNR} = \Omega(\sqrt{n \log n})$. For $\mathsf{MAR} \cdot \mathsf{SNR} = \sqrt{n \log n}$, Theorem 1 shows that fewer than $m \asymp 2\sqrt{k \log k}$ measurements will cause ML decoding to fail, while [22, Thm. 3.1] shows that a typicality-based decoder will succeed with $m = \Theta(k)$ measurements.

6) Note that the necessary condition of [18] is proven for $\mathsf{MAR} = 1$. Theorem 1 gives a bound that increases for smaller MAR; this suggests (though does not prove, since the condition is merely necessary) that smaller MAR makes the problem harder.

Numerical validation: Computational confirmation of Theorem 1 is technically impossible, and even qualitative support is hard to obtain because of the high complexity of ML detection. Nevertheless, we may obtain some evidence through Monte Carlo simulation.

Fig. 1 shows the probability of success of ML detection for $n = 20$ as k, m, SNR, and MAR are varied, with each point representing at least 500 independent trials. Each subpanel gives simulation results for $k \in \{1, 2, \ldots, 5\}$ and $m \in \{1, 2, \ldots, 40\}$ for one $(\mathsf{SNR}, \mathsf{MAR})$ pair. Signals with $\mathsf{MAR} < 1$ are created by having one small nonzero component and $k - 1$ equal, larger nonzero components. Overlaid on the color-intensity plots is a black curve representing (6).

Taking any one column of one subpanel from bottom to top shows that as m is increased, there is a transition from ML failing to ML succeeding. One can see that (6) follows the failure-success transition qualitatively. In particular, the empirical dependence on SNR and MAR approximately follows (6). Empirically, for the (small) value of $n = 20$, it seems that with $\mathsf{MAR} \cdot \mathsf{SNR}$ held fixed, sparsity recovery becomes easier as SNR increases (and MAR decreases).

Less extensive Monte Carlo simulations for $n = 40$ are reported in Fig. 2. The results are qualitatively similar. As might be expected, the transition from low to high probability of successful recovery as a function of m appears more sharp at $n = 40$ than at $n = 20$.

IV. SUFFICIENT CONDITION WITH MAXIMUM CORRELATION DETECTION

Consider the following simple estimator. As before, let a_j be the jth column of the random matrix A. Define the *maximum correlation (MC) estimate* as

$$\hat{I}_{\mathrm{MC}} = \left\{ j \ : \ |a_j'y| \text{ is one of the } k \text{ largest values of } |a_i'y| \right\}. \tag{8}$$
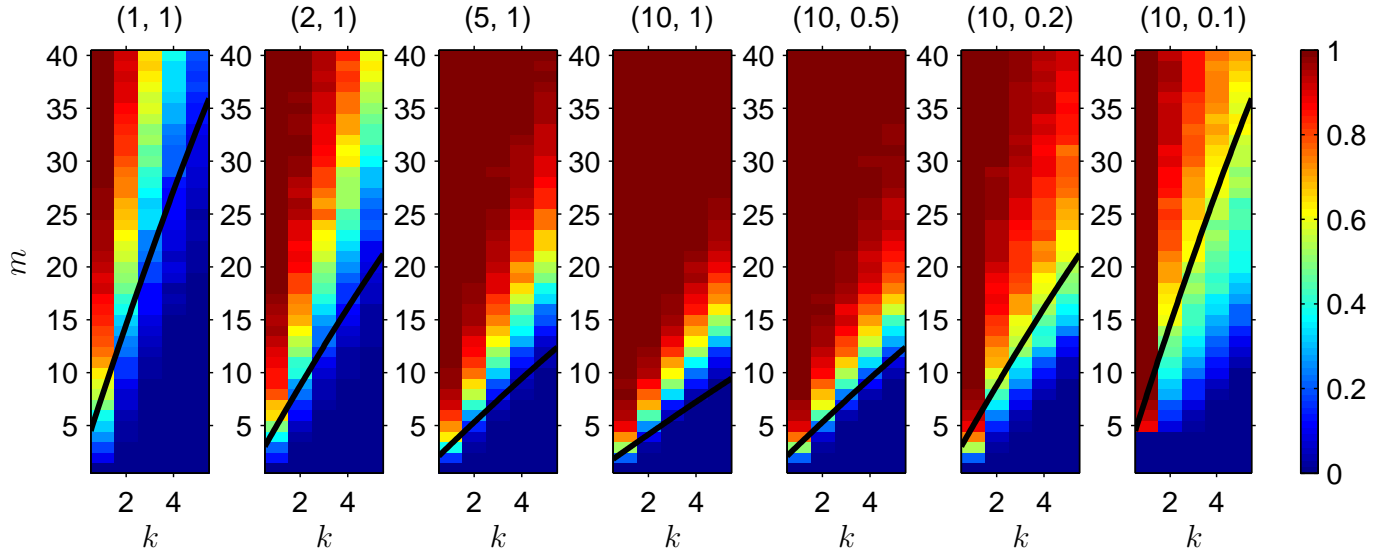
Fig. 1. Simulated success probability of ML detection for $n = 20$ and many values of k, m, SNR, and MAR. Each subfigure gives simulation results for $k \in \{1, 2, \ldots, 5\}$ and $m \in \{1, 2, \ldots, 40\}$ for one (SNR, MAR) pair. Each subfigure heading gives (SNR, MAR). Each point represents at least 500 independent trials. Overlaid on the color-intensity plots is a black curve representing (6).
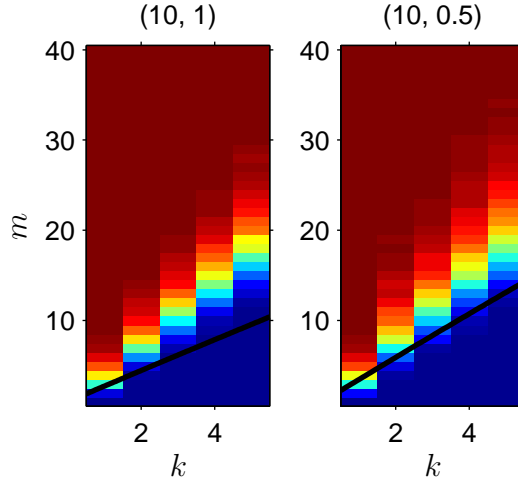

Fig. 2. Simulated success probability of ML detection for $n = 40$; SNR = 10; MAR = 1 (left) or MAR = 0.5 (right); and many values of k and m. Each subfigure gives simulation results for $k \in \{1, 2, \ldots, 5\}$ and $m \in \{1, 2, \ldots, 40\}$, with each point representing at least 1000 independent trials. Overlaid on the color-intensity plots (with scale as in Fig. 1) is a black curve representing (6).

This algorithm simply correlates the observed signal y with all the frame vectors a_j and selects the indices j with the highest correlation. It is significantly simpler than both lasso and matching pursuit and is not meant to be proposed as a competitive alternative. Rather, the MC method is introduced and analyzed to illustrate that a trivial method can obtain optimal scaling with respect to n and k.

Theorem 2: Let $k = k(n)$ and $m = m(n)$ vary with n such that $\lim_{n \to \infty} k = \infty$, $\limsup_{n \to \infty} k/n \leq 1/2$, and

$$m > \frac{(8 + \delta)(1 + \mathsf{SNR})}{\mathsf{MAR} \cdot \mathsf{SNR}} k \log(n - k) \tag{9}$$

for some $\delta > 0$. Then the maximum correlation estimator asymptotically detects the sparsity pattern, i.e.,

$$\lim_{n \to \infty} \Pr\left(\hat{I}_{\mathrm{MC}} = I_{\mathrm{true}}\right) = 1.$$

Proof: See Appendix C.   ∎

Remarks:

1) Comparing (6) and (9), we see that for a fixed SNR and minimum-to-average ratio, the simple MC estimator needs only a constant factor more measurements than the optimal ML estimator. In particular, the results show that the scaling of the minimum number of measurements $m = \Theta(k \log(n-k))$ is both necessary and sufficient. Moreover, the optimal scaling factor not only does not require ML estimation, it does not even require lasso or matching pursuit—it can be achieved with a remarkably simply method such as maximum correlation.

   There is, of course, a difference in the constant factors of the expressions (6) and (9). Specifically, the MC method requires a factor $4(1 + \mathsf{SNR})$ more measurements than ML detection. In particular, for low SNRs (i.e. $\mathsf{SNR} \ll 1$), the factor reduces to 4.

2) For high SNRs, the gap between the MC estimator and the ML estimator can be large. In particular, the lower bound on the number of measurements required by ML decreases to $k - 1$ as $\mathsf{SNR} \to \infty$.[3] In contrast, with the MC estimator increasing the SNR has diminishing returns: as $\mathsf{SNR} \to \infty$, the bound on the number of measurements in (9) approaches

$$m > \frac{8}{\mathsf{MAR}} k \log(n-k). \tag{10}$$

   Thus, even with $\mathsf{SNR} \to \infty$, the minimum number of measurements is not improved from $m = O(k \log(n-k))$.

   This diminishing returns for improved SNR exhibited by the MC method is also a problem for more sophisticated methods such as lasso. For example, the analysis of [15] shows that when the $\mathsf{SNR} = \Theta(n)$ (so $\mathsf{SNR} \to \infty$) and MAR is bounded strictly away from zero, lasso requires

$$m > 2k \log(n-k) + k + 1 \tag{11}$$

   for reliable recovery. Therefore, like the MC method, lasso does not achieve a scaling better than $m = O(k \log(n-k))$, even at infinite SNR.

3) There is certainly a gap between MC and lasso. Comparing (10) and (11), we see that, at high SNR, the simple MC method requires a factor of at most $4/\mathsf{MAR}$ more measurements than lasso. This factor is largest when MAR is small, which occurs when there are relatively small non-zero components. Thus, in the high SNR regime, the main benefit of lasso is not that it achieves an optimal scaling with respect to k and n (which can be achieved with the simpler MC), but rather that lasso is able to detect small coefficients, even when they are much below the average power.

Numerical validation: MC sparsity pattern detection is extremely simple and can thus be simulated easily for large problem sizes. Fig. 3 reports the results of a large number Monte Carlo simulations of the MC method with $n = 100$. The threshold predicted by (9) matches well to the parameter combinations where the probability of success drops below about 0.995.

V. Conclusions

We have considered the problem of detecting the sparsity pattern of a sparse vector from noisy random linear measurements. Our main conclusions are:

- *Necessary and sufficient scaling with respect to n and k.* For a given SNR and minimum-to-average ratio, the scaling of the number of measurements

$$m = O(k \log(n-k))$$

  is both necessary and sufficient for asymptotically reliable sparsity pattern detection. This scaling is significantly worse than predicted by previous information-theoretic bounds.

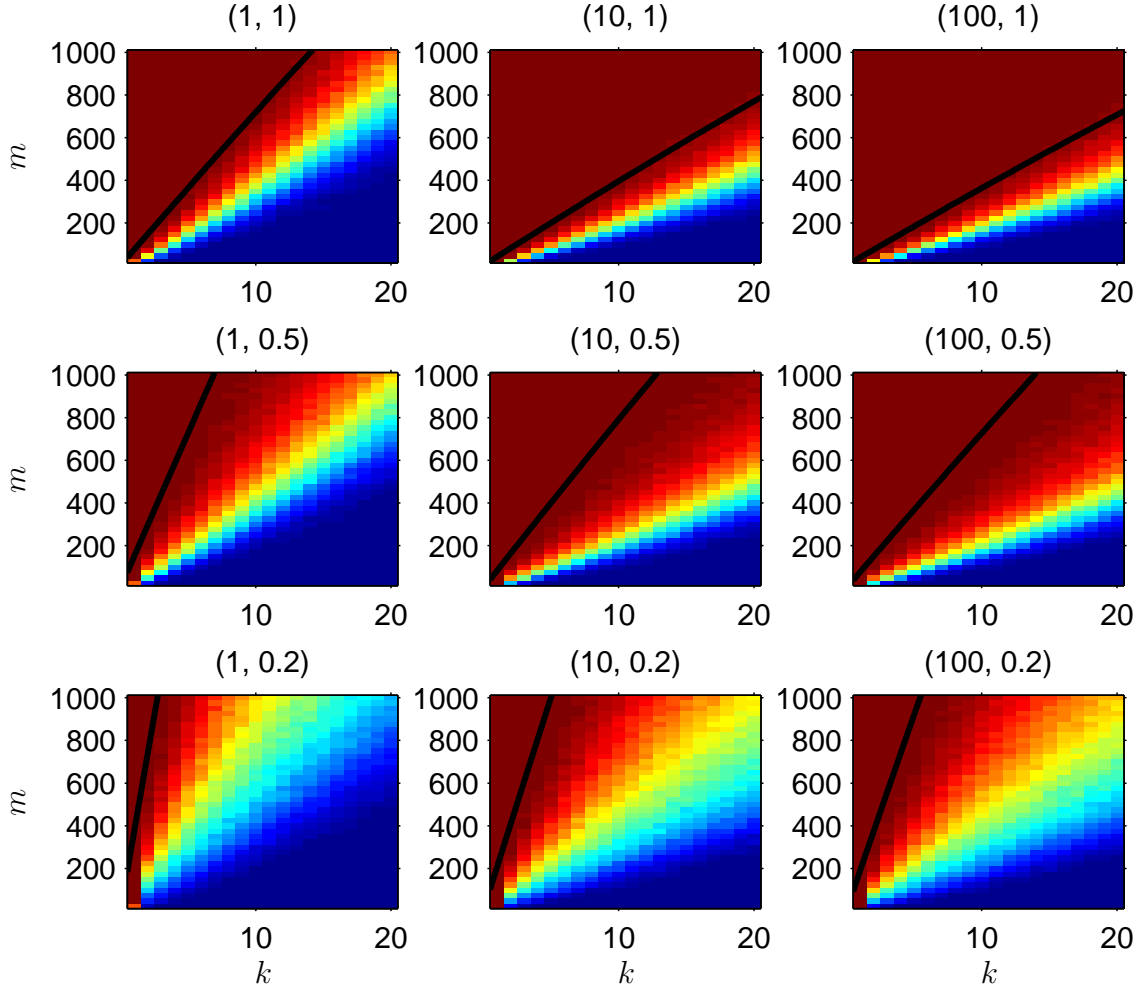[3]Of course, at least $k + 1$ measurements are necessary.

Fig. 3.   Simulated success probability of MC detection for $n = 100$ and many values of k, m, SNR, and MAR. Each subfigure gives simulation results for $k \in \{1, 2, \ldots, 20\}$ and $m \in \{25, 50, \ldots, 1000\}$ for one (SNR, MAR) pair. Each subfigure heading gives (SNR, MAR), so SNR $= 1, 10, 100$ for the three columns and MAR $= 1, 0.5, 0.2$ for the three rows. Each point represents 1000 independent trials. Overlaid on the color-intensity plots (with scale as in Fig. 1) is a black curve representing (9).

- *Scaling optimality of a trivial method.* The optimal scaling with respect to k and n can be achieved with a trivial maximum correlation (MC) method. In particular, both lasso and OMP, while likely to do better, are not necessary to achieve this scaling.
- *Dependence on SNR.* While the threshold number of measurements for ML and MC sparsity recovery to be successful have the same dependence on n and k, the dependence on SNR differs significantly. Specifically, the MC method requires a factor of up to $4(1 + \mathsf{SNR})$ more measurements than ML. Moreover, as SNR $\to \infty$, the number of measurements required by ML may be as low as $m = k + 1$. In contrast, even letting SNR $\to \infty$, the maximum correlation method still requires a scaling $m = O(k \log(n - k))$.
- *Lasso and dependence on MAR.* MC can also be compared to lasso, at least in the high SNR regime. There is a potential gap between MC and lasso, but the gap is smaller than the gap to ML. Specifically, in the high SNR regime, MC requires at most $4/\mathsf{MAR}$ more measurements than lasso, where MAR is the mean-to-average ratio defined in (5). Both lasso and MC scale as $m = O(k \log(n - k))$. Thus, the benefit of lasso is not in its scaling with respect to the problem dimensions, but rather its ability to detect the sparsity pattern even in the presence of relatively small nonzero coefficients (i.e. low MAR).

While our results settle the question of the optimal scaling of the number of measurements m in terms of k and n, there is clearly a gap in the necessary and sufficient conditions in terms of the scaling of the SNR. We have seen that full ML estimation could potentially have a scaling in SNR as small as $m = O(1/\mathsf{SNR}) + k - 1$. An open question is whether there is any practical algorithm that can achieve a similar scaling.

A second open issue is to determine conditions for partial sparsity recovery. The above results define conditions for recovering all the positions in the sparsity pattern. However, in many practical applications, obtaining some large fraction of these positions would be sufficient. Neither the limits of partial sparsity recovery nor the performance of practical algorithms are completely understood, though some results have been reported in [20]–[22], [24].

APPENDIX

A. Deterministic Necessary Condition

The proof of Theorem 1 is based on the following deterministic necessary condition for sparsity recovery. Recall the notation that for any $J \subseteq \{1, 2, \ldots, n\}$, P_J denotes the orthogonal projection onto the span of the vectors $\{a_j\}_{j \in J}$. Additionally, let $P_J^\perp = I - P_J$ denote the orthogonal projection onto the orthogonal complement of $\mathrm{span}(\{a_j\}_{j \in J})$.

Lemma 1: A necessary condition for ML detection to succeed (i.e. $\hat{I}_{\mathrm{ML}} = I_{\mathrm{true}}$) is:

$$\text{for all } i \in I_{\mathrm{true}} \text{ and } j \notin I_{\mathrm{true}}, \quad \frac{|a_i' P_K^\perp y|^2}{a_i' P_K^\perp a_i} \geq \frac{|a_j' P_K^\perp y|^2}{a_j' P_K^\perp a_j} \tag{12}$$

where $K = I_{\mathrm{true}} \setminus \{i\}$.

Proof: Note that $y = P_K y + P_K^\perp y$ is an orthogonal decomposition of y into the portions inside and outside the subspace $S = \mathrm{span}(\{a_j\}_{j \in K})$. An approximation of y in subspace S leaves residual $P_K^\perp y$. Intuitively, the condition (12) requires that the residual be at least as highly correlated with the remaining "correct" vector a_i as it is with any of the "incorrect" vectors $\{a_j\}_{j \notin I_{\mathrm{true}}}$.

Fix any $i \in I_{\mathrm{true}}$, $j \notin I_{\mathrm{true}}$ and let

$$J = K \cup \{j\} = (I_{\mathrm{true}} \setminus \{i\}) \cup \{j\}.$$

That is, J is equal to the true sparsity pattern I_{true}, except that a single "correct" index i has been replaced by an "incorrect" index j. If the ML estimator is to select $\hat{I}_{\mathrm{ML}} = I_{\mathrm{true}}$ then the energy of the noisy vector y must be larger on the true subspace I_{true}, than the incorrect subspace J. Therefore,

$$\|P_{I_{\mathrm{true}}} y\|^2 \geq \|P_J y\|^2. \tag{13}$$

Now, a simple application of the matrix inversion lemma shows that since $I_{\mathrm{true}} = K \cup \{i\}$,

$$\|P_{I_{\mathrm{true}}} y\|^2 = \|P_K y\|^2 + \frac{|a_i' P_K^\perp y|^2}{a_i' P_K^\perp a_i}. \tag{14a}$$

Also, since $J = K \cup \{j\}$,

$$\|P_J y\|^2 = \|P_K y\|^2 + \frac{|a_j' P_K^\perp y|^2}{a_j' P_K^\perp a_j}. \tag{14b}$$

Substituting (14a)–(14b) into (13) and cancelling $\|P_K y\|^2$ shows (12). ∎

B. Proof of Theorem 1

To simplify notation, assume without loss of generality that $I_{\text{true}} = \{1, 2, \ldots, k\}$. Also, assume that the minimization in (5) occurs at $j = 1$ with

$$|x_1|^2 = \frac{m}{k} \text{SNR} \cdot \text{MAR}. \tag{15}$$

Finally, since adding measurements (i.e. increasing m) can only improve the chances that ML detection will work, we can assume that in addition to satisfying (6), the numbers of measurements satisfy the lower bound

$$m > \epsilon k \log(n - k) + k - 1, \tag{16}$$

for some $\epsilon > 0$. This assumption implies that

$$\lim \frac{\log(n - k)}{m - k + 1} = \lim \frac{1}{\epsilon k} = 0. \tag{17}$$

Here and in the remainder of the proof the limits are as m, n and $k \to \infty$ subject to (6) and (16). With these requirements on m, we need to show $\lim \Pr(\hat{I}_{\text{ML}} = I_{\text{true}}) = 0$.

From Lemma 1, $\hat{I}_{\text{ML}} = I_{\text{true}}$ implies (12). Thus

$$
\begin{aligned}
&\Pr\left(\hat{I}_{\text{ML}} = I_{\text{true}}\right) \\
&\leq \ \Pr\left(\frac{|a_i' P_{\overline{K}}^{\perp} y|^2}{a_i' P_{\overline{K}}^{\perp} a_i} \geq \frac{|a_j' P_{\overline{K}}^{\perp} y|^2}{a_j' P_{\overline{K}}^{\perp} a_j} \quad \forall \, i \in I_{\text{true}}, \, j \notin I_{\text{true}}\right) \\
&\leq \ \Pr\left(\frac{|a_1' P_{\overline{K}}^{\perp} y|^2}{a_1' P_{\overline{K}}^{\perp} a_1} \geq \frac{|a_j' P_{\overline{K}}^{\perp} y|^2}{a_j' P_{\overline{K}}^{\perp} a_j} \quad \forall \, j \notin I_{\text{true}}\right) \\
&= \ \Pr(\Delta^- \geq \Delta^+),
\end{aligned}
$$

where

$$
\begin{aligned}
\Delta^- &= \ \frac{1}{\log(n - k)} \frac{|a_1' P_{\overline{K}}^{\perp} y|^2}{a_1' P_{\overline{K}}^{\perp} a_1}, \\
\Delta^+ &= \ \frac{1}{\log(n - k)} \max_{j \in \{k+1, \ldots, n\}} \frac{|a_j' P_{\overline{K}}^{\perp} y|^2}{a_j' P_{\overline{K}}^{\perp} a_j},
\end{aligned}
$$

and $K = I_{\text{true}} \setminus \{1\} = \{2, \ldots, k\}$. The $-$ and $+$ superscripts are used to reflect that Δ^- is the energy lost from removing "correct" index 1, and Δ^+ is the energy added from adding the worst "incorrect" index. The theorem will be proven if we can show that

$$\limsup \Delta^- < \liminf \Delta^+ \tag{18}$$

with probability approaching one. We will consider the two limits separately.

1) Limit of Δ^+: Let V_K be the $k - 1$ dimensional space spanned by the vectors $\{a_j\}_{j \in K}$. For each $j \notin I_{\text{true}}$, let u_j be the unit vector

$$u_j = P_{\overline{K}}^{\perp} a_j / \|P_{\overline{K}}^{\perp} a_j\|.$$

Since a_j has i.i.d. Gaussian components, it is spherically symmetric. Also, if $j \notin K$, a_j is independent of the subspace V_K. Hence, in this case, u_j will be a unit vector uniformly distributed on the unit sphere in $V_K^{\perp}$. Since $V_K^{\perp}$ is an $m - k + 1$ dimensional subspace, it follows from Lemma 4 (see Appendix D) that if we define

$$z_j = |u_j' P_{\overline{K}}^{\perp} y|^2 / \|P_{\overline{K}}^{\perp} y\|^2,$$

then z_j follows a Beta$(1, m - k + 1)$ distribution. See Appendix D for a review of the chi-squared and beta distributions and some simple results on these variables that will be used in the proofs below.

By the definition of u_j,

$$\frac{|a_j' P_K^\perp y|^2}{a_j' P_K^\perp a_j} = |u_j' P_K^\perp y|^2 = z_j \|P_K^\perp y\|^2,$$

and therefore

$$\Delta^+ = \frac{1}{\log(n-k)} \|P_K^\perp y\|^2 \max_{j \in \{k+1,\dots,n\}} z_j. \tag{19}$$

Now the vectors a_j are independent of one another, and for $j \notin I_{\text{true}}$, each a_j is independent of $P_K^\perp y$. Therefore, the variables z_j will be i.i.d. Hence, using Lemma 5 (see Appendix D) and (17),

$$\lim \frac{m-k+1}{\log(n-k)} \max_{j=k+1,\dots,n} z_j = 2 \tag{20}$$

in distribution. Also,

$$\liminf \frac{1}{m-k+1} \|P_K^\perp y\|^2 \overset{(a)}{\geq} \lim \frac{1}{m-k+1} \|P_{I_{\text{true}}}^\perp y\|^2$$

$$\overset{(b)}{=} \lim \frac{1}{m-k+1} \|P_{I_{\text{true}}}^\perp d\|^2$$

$$\overset{(c)}{=} \lim \frac{m-k}{m-k+1} = 1 \tag{21}$$

where (a) follows from the fact that $K \subset I_{\text{true}}$ and hence $\|P_K^\perp y\| \geq \|P_{I_{\text{true}}}^\perp y\|$; (b) is valid since $P_{I_{\text{true}}}^\perp a_j = 0$ for all $j \in I_{\text{true}}$ and therefore $P_{I_{\text{true}}} x = 0$; and (c) follows from the fact that $P_{I_{\text{true}}}^\perp d$ is a unit-variance white random vector in an $m-k$ dimensional space. Combining (19), (20) and (21) shows that

$$\liminf \Delta^+ \geq 2. \tag{22}$$

2) Limit of Δ^-: For any $j \in K$, $P_K^\perp a_j = 0$. Therefore,

$$P_K^\perp y = P_K^\perp \left(\sum_{j=1}^k a_j x_j + d \right) = x_1 P_K^\perp a_1 + P_K^\perp d.$$

Hence,

$$\frac{|a_1' P_K^\perp y|^2}{a_1' P_K^\perp a_1} = \left| \|P_K^\perp a_1\| x_1 + v \right|^2,$$

where v is given by

$$v = a_1' P_K^\perp d / \|P_K^\perp a_1\|.$$

Since $P_K^\perp a_1 / \|P_K^\perp a_1\|$ is a random unit vector independent of d, and d is a zero-mean, unit-variance Gaussian random vector, $v \sim \mathcal{N}(0,1)$. Therefore,

$$\lim \Delta^- = \lim \left| \frac{\|P_K^\perp a_1\| x_1}{\log^{1/2}(n-k)} + \frac{1}{\log^{1/2}(n-k)} v \right|^2$$

$$= \lim \frac{\|P_K^\perp a_1\|^2 |x_1|^2}{\log(n-k)}, \tag{23}$$

where, in the last step, we used the fact that

$$v / \log^{1/2}(n-k) \to 0.$$

Now, a_1 is a Gaussian vector with variance $1/m$ in each component and $P_K^\perp$ is a projection onto an $m-k+1$ dimensional space. Hence,

$$\lim \frac{m \|P_K^\perp a_1\|^2}{m-k+1} = 1. \tag{24}$$

Starting with a combination of (23) and (24),

$$
\begin{aligned}
\limsup \Delta^- &= \limsup \frac{|x_1|^2 (m - k + 1)}{m \log(n - k)} \\
&\overset{(a)}{=} \limsup \frac{(\mathsf{SNR} \cdot \mathsf{MAR})(m - k + 1)}{k \log(n - k)} \\
&\overset{(b)}{<} 2 - \delta
\end{aligned}
\tag{25}
$$

where (a) uses (15); and (b) uses (6).

Comparing (22) and (25) proves (18), thus completing the proof.

C. Proof of Theorem 2

We will show that there exists a $\mu > 0$ such that, with high probability,

$$
\begin{aligned}
|a_i' y|^2 &> \mu \quad \text{for all } i \in I_{\text{true}}; \\
|a_j' y|^2 &< \mu \quad \text{for all } j \notin I_{\text{true}}.
\end{aligned}
\tag{26}
$$

When (26) is satisfied,

$$
|a_i' y| > |a_j' y| \quad \text{for all indices } i \in I_{\text{true}} \text{ and } j \notin I_{\text{true}}.
$$

Thus, (26) implies that the maximum correlation estimator $\hat{I}_{\text{MC}}$ in (8) will select $\hat{I}_{\text{MC}} = I_{\text{true}}$. Consequently, the theorem will be proven if can find a μ such that (26) holds with high probability.

Since $\delta > 0$, we can find an $\epsilon > 0$ such that

$$
\sqrt{8 + \delta} - \sqrt{2 + \epsilon} > \sqrt{2}.
\tag{27}
$$

Define

$$
\mu = (2 + \epsilon)(1 + \mathsf{SNR}) \log(n - k).
\tag{28}
$$

Define two probabilities corresponding to the two conditions in (26):

$$
\begin{aligned}
p_{\text{MD}} &= \Pr\left(|a_i' y|^2 < \mu \quad \text{for some } i \in I_{\text{true}}\right) \tag{29} \\
p_{\text{FA}} &= \Pr\left(|a_j' y|^2 > \mu \quad \text{for some } j \notin I_{\text{true}}\right). \tag{30}
\end{aligned}
$$

The first probability p_{MD} is the probability of missed detection, i.e., the probability that the energy on one of the "true" vectors, a_i with $i \in I_{\text{true}}$, is below the threshold μ. The second probability p_{FA} is the false alarm probability, i.e., the probability that the energy on one of the "incorrect" vectors, a_j with $j \notin I_{\text{true}}$, is above the threshold μ. Since the correlation estimator detects the correct sparsity pattern when there are no missed vectors or false alarms, we have the bound

$$
\Pr\left(\hat{I}_{\text{MC}} \neq I_{\text{true}}\right) \leq p_{\text{MD}} + p_{\text{FA}}.
$$

So the result will be proven if we can show that p_{MD} and p_{FA} approach zero as m, n and $k \to \infty$ satisfying (9). We analyze these two probabilities separately.

1) Limit of p_{FA}: Consider any index $j \notin I_{\text{true}}$. Since y is a linear combination of vectors $\{a_i\}_{i \in I_{\text{true}}}$ and the noise vector d, a_j is independent of y. Also, recall that the components of a_j are $\mathcal{N}(0, 1/m)$. Therefore, conditional on $\|y\|^2$, the inner product $a_j'y$ is Gaussian with mean zero and variance $\|y\|^2/m$. For large m, $\|y\|^2/m \to 1 + \text{SNR}$. Hence, we can write

$$|a_j'y|^2 = (1 + \text{SNR})u_j^2,$$

where u_j is a random variable that converges in distribution to a zero mean Gaussian with unit variance. Using the definitions of p_{FA} in (30) and μ in (28), we see that

$$
\begin{aligned}
p_{\text{FA}} &= \Pr\left(\max_{j \notin I_{\text{true}}} |a_j'y|^2 > \mu \right) \\
&= \Pr\left(\max_{j \notin I_{\text{true}}} (1 + \text{SNR})u_j^2 > \mu \right) \\
&= \Pr\left(\max_{j \notin I_{\text{true}}} u_j^2 > \mu/(1 + \text{SNR}) \right) \\
&= \Pr\left(\max_{j \notin I_{\text{true}}} u_j^2 > (2 + \epsilon)\log(n - k) \right) \to 0
\end{aligned}
$$

where the last limit uses Lemma 3 (see Appendix D) on the maxima of chi-squared random variables.

2) Limit of p_{MD}: Consider any index $i \in I_{\text{true}}$. Observe that

$$a_i'y = \|a_i\|^2|x_i|^2 + a_i'e_i,$$

where

$$e_i = y - a_ix_i = \sum_{\ell \in I_{\text{true}}, \ell \neq i} a_\ell x_\ell + d.$$

It is easily verified that $\|a_i\|^2 \to 1$ and $\|e_i\|^2/m \to 1 + \text{SNR}$. Using a similar argument as above, one can show that

$$a_i'y = x_i + (1 + \text{SNR})^{1/2}u_i, \tag{31}$$

where u_i approaches a zero-mean, unit-variance Gaussian in distribution.

Now, using (4), (5) and (9),

$$
\begin{aligned}
|x_i|^2 &\geq \frac{\text{MAR}\|x\|^2}{k} = \frac{m\,\text{MAR} \cdot \text{SNR}}{k} \\
&> (8 + \delta)(1 + \text{SNR})\log(n - k). \tag{32}
\end{aligned}
$$

Combining (27), (28), (31) and (32)

$$
\begin{aligned}
|a_i'y|^2 \leq \mu &\iff \big| \; x_i + (1 + \text{SNR})^{1/2}u_i \big| \leq \mu^{1/2} \\
&\implies (1 + \text{SNR})u_i^2 \geq \left(|x_i| - \mu^{1/2} \right)^2 \\
&\iff u_i^2 > 2\log(n - k) \\
&\implies u_i^2 > 2\log(k)
\end{aligned}
$$

where, in the last step, we have used the fact that since $k/n < 1/2$, $n - k > k$. Therefore, using Lemma 3

$$
\begin{aligned}
p_{\text{MD}} &= \Pr\left(\min_{i \in I_{\text{true}}} |a_i'y|^2 \leq \mu \right) \\
&\leq \Pr\left(\max_{i \in I_{\text{true}}} u_i^2 > 2\log(k) \right) \to 0. \tag{33}
\end{aligned}
$$

Hence, we have shown both $p_{\text{FA}} \to 0$ and $p_{\text{MD}} \to 0$ as $n \to \infty$, and the theorem is proven.

D. Maxima of Chi-Squared and Beta Random Variables

The proofs of the main results above require a few simple results on the maxima of large numbers of chi-squared and beta random variables. A complete description of chi-squared and beta random variables can be found in [25].

A random variable U has a *chi-squared* distribution with r degrees of freedom if it can be written as

$$U = \sum_{i=1}^{r} Z_i^2,$$

where Z_i are i.i.d. $\mathcal{N}(0,1)$. For every n and r define the random variables

$$\overline{M}_{n,r} = \max_{i \in \{1,\dots,n\}} U_i,$$
$$\underline{M}_{n,r} = \min_{i \in \{1,\dots,n\}} U_i,$$

where the U_i's are i.i.d. chi-squared with r degrees of freedom.

Lemma 2: For $\overline{M}_{n,r}$ defined as above,

$$\lim_{n \to \infty} \frac{1}{\log(n)} \overline{M}_{n,1} = 2,$$

where the convergence is in distribution.

Proof: We can write $\overline{M}_{n,1} = \max_{i \in \{1,\dots,n\}} Z_i^2$ where Z_i are i.i.d. $\mathcal{N}(0,1)$. Then, for any $a > 0$,

$$
\begin{aligned}
&\Pr\left(\frac{1}{\log(n)} \overline{M}_{n,1} < a \right) \\
&= \Pr\left(|Z_1|^2 < a\log(n) \right)^n \\
&= \operatorname{erf}\left(\sqrt{a\log(n)/2} \right)^n \\
&\approx \left[1 - \sqrt{\frac{2}{\pi a \log(n)}} \exp(-a\log(n)/2) \right]^n \\
&= \left[1 - \sqrt{\frac{2}{\pi a \log(n)}} \frac{1}{n^{a/2}} \right]^n
\end{aligned}
$$

where the approximation is valid for large n. Taking the limit as $n \to \infty$, one can now easily show that

$$
\lim_{n \to \infty} \Pr\left(\frac{1}{\log(n)} \overline{M}_n < a \right) = \begin{cases} 0, & \text{for } a < 2; \\ 1, & \text{for } a > 2 \end{cases}
$$

and therefore $\overline{M}_n / \log(n) \to 2$ in distribution. ∎

Lemma 3: In any limit where $r \to \infty$ and $\log(n)/r \to 0$,

$$\lim_{r \to \infty} \frac{1}{r} \overline{M}_{n,r} = \lim_{r \to \infty} \frac{1}{r} \underline{M}_{n,r} = 1,$$

where the convergence is in distribution.

Proof: It suffices to show

$$
\begin{aligned}
\limsup_{r \to \infty} \frac{1}{r} \overline{M}_{n,r} &\leq 1, \\
\liminf_{r \to \infty} \frac{1}{r} \underline{M}_{n,r} &\geq 1.
\end{aligned}
$$

We will just prove the first inequality since the proof of the second is similar. We can write

$$\frac{1}{r}\overline{M}_{n,r} = \max_{i=1,\ldots,n} V_i,$$

where each $V_i = U_i/r$ and the U_i's are i.i.d. chi-squared random variables with r degree of freedom. Using the characteristic function of U_i and Chebyshev's inequality, one can show that for all $\epsilon > 0$,

$$\begin{aligned} \Pr(V_i > (1+\epsilon)) &= \Pr(U_i > (1+\epsilon)r) \\ &\leq (1+\epsilon)e^{-\epsilon r/2}. \end{aligned}$$

Therefore,

$$\begin{aligned} \Pr\left(\overline{M}_{n,r} \leq 1+\epsilon\right) &= [\Pr(V_i \leq 1+\epsilon)]^n \\ &\geq \left[1-(1+\epsilon)e^{-\epsilon r/2}\right]^n \\ &\geq 1-(1+\epsilon)ne^{-\epsilon r/2} \\ &= 1-(1+\epsilon)\exp\left[\log(n)-\epsilon r/2\right] \\ &\to 1, \end{aligned}$$

where the limit in the last step follows from the fact that $\log(n)/r \to 0$. Since this is true for all ϵ it follows that $\limsup r^{-1}\overline{M}_{n,r} \leq 1$. Similarly, one can show $\liminf r^{-1}\overline{M}_{n,r} \geq 1$ and this proves the lemma. ∎

The next two lemmas concern certain beta distributed random variables. A real-valued scalar random variable W follows a $\mathrm{Beta}(r,s)$ distribution if it can be written as

$$W = U_r/(U_r + V_s),$$

where the variables U_r and V_s are independent chi-squared random variables with r and s degrees of freedom, respectively. The importance of the beta distribution is given by the following lemma.

Lemma 4: Let x and $u \in \mathbb{R}^s$ be any two independent random vectors, with u being uniformly distributed on the unit sphere. Let $w = |u'x|^2/\|x\|^2$ be the energy of w projected onto u. Then w is independent of x and follows a $\mathrm{Beta}(1, s-1)$ distribution.

Proof: This can be proven along the lines of the arguments in [9]. ∎

The following lemma provides a simple expression for the maxima of certain beta distributed variables.

Lemma 5: Given any s and n, let $w_{j,s}$, $j = 1,\ldots,n$, be i.i.d. $\mathrm{Beta}(1, s-1)$ random variables and define

$$T_{n,s} = \max_{j=1,\ldots,n} w_{j,s}.$$

Then for any limit with n and $s \to \infty$ and $\log(n)/s \to 0$,

$$\lim_{n,s\to\infty} \frac{s}{\log(n)}T_{n,s} = 2,$$

where the convergence is in distribution.

Proof: We can write $w_{j,s} = u_j/(u_j+v_{j,s-1})$ where u_j and $v_{j,s-1}$ are independent chi-squared random variables with 1 and $s-1$ degrees of freedom, respectively. Let

$$\begin{aligned} \overline{M}_n &= \max_{j\in\{1,\ldots,n\}} u_j \\ \overline{M}_{n,s-1} &= \max_{j\in\{1,\ldots,n\}} v_{j,s-1} \\ \underline{M}_{n,s-1} &= \min_{j\in\{1,\ldots,n\}} v_{j,s-1}. \end{aligned}$$

Using the definition of $T_{n,s}$,

$$T_{n,s} \leq \frac{\overline{M}_n}{\overline{M}_n + \underline{M}_{n,s-1}}.$$

Now Lemmas 2 and 3 and the hypothesis of this lemma show that $\overline{M}_n/\log(n) \to 2$, $\underline{M}_{n,s-1}/(s-1) \to 1$, and $\log(n)/s \to 0$. One can combine these limits to show that

$$\limsup_{n,s\to\infty} \frac{s}{\log(n)} T_{n,s} \leq 2.$$

Similarly, one can show that

$$\liminf_{n,s\to\infty} \frac{s}{\log(n)} T_{n,s} \geq 2,$$

and therefore $sT_{n,s}/\log(n) \to 2$.   ∎

ACKNOWLEDGMENT

REFERENCES

[1] A. Miller, *Subset Selection in Regression*, 2nd ed., ser. Monographs on Statistics and Applied Probability.   New York: Chapman & Hall/CRC, 2002, no. 95.

[2] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[3] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Computing*, vol. 24, no. 2, pp. 227–234, Apr. 1995.

[4] 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, 1999.

[5] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc., Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[6] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inform. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.

[7] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[8] ——, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 1030–1051, Mar. 2006.

[9] 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. Sig. Process.*, vol. 2006, pp. 1–19, Mar. 2006.

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

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

[12] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[13] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Sparse approximation, denoising, and large random frames," in *Proc. Wavelets XI, part of SPIE Optics & Photonics*, vol. 5914, San Diego, CA, Jul.–Aug. 2005, pp. 172–181.

[14] D. L. Donoho and J. Tanner, "Counting faces of randomly-projected polytopes when the projection radically lowers dimension," *J. Amer. Math. Soc.*, submitted.

[15] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy recovery of sparsity," Univ. of California, Berkeley, Dept. of Statistics, Tech. Rep., May 2006, arXiv:math.ST/0605740 v1 30 May 2006.

[16] S. Sarvotham, D. Baron, and R. G. Baraniuk, "Measurements vs. bits: Compressed sensing meets information theory," in *Proc. 44th Ann. Allerton Conf. on Commun., Control and Comp.*, Monticello, IL, Sep. 2006.

[17] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Rate-distortion bounds for sparse approximation," in *IEEE Statist. Sig. Process. Workshop*, Madison, WI, Aug. 2007, pp. 254–258.

[18] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," Univ. of California, Berkeley, Dept. of Statistics, Tech. Rep. 725, Jan. 2007.

[19] V. K. Goyal, A. K. Fletcher, and S. Rangan, "Compressive sampling and lossy compression," *IEEE Sig. Process. Mag.*, vol. 25, no. 2, pp. 48–56, Mar. 2008.

[20] G. Reeves, "Sparse signal sampling using noisy linear projections," Univ. of California, Berkeley, Dept. of Elec. Eng. and Comp. Sci., Tech. Rep. UCB/EECS-2008-3, Jan. 2008.

[21] M. Akçakaya and V. Tarokh, "Shannon theoretic limits on noisy compressive sampling," arXiv:0711.0366v1 [cs.IT]., Nov. 2007.

[22] ——, "Noisy compressive sampling limits in linear and sublinear regimes," in *Proc. Conf. on Inform. Sci. & Sys.*, Princeton, NJ, Mar. 2008.

[23] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Trans. Inform. Theory*, vol. 52, no. 9, pp. 4036–4048, Sep. 2006.

[24] S. Aeron, M. Zhao, and V. Saligrama, "On sensing capacity of sensor networks for the class of linear observation, fixed SNR models," arXiv:0704.3434v3 [cs.IT]., Jun. 2007.

[25] M. Evans, N. Hastings, and J. B. Peacock, *Statistical Distributions*, 3rd ed.   New York: John Wiley & Sons, 2000.