Performance Analysis for Sparse Support Recovery

Gongguo Tang and Arye Nehorai

ESE, Washington University

April 21st 2009
1 Background and Motivation
2 Research Overview
3 Mathematical Model
4 Theoretical Analysis
5 Conclusions
6 Future Work
Sparse signals refer to a set of signals that have only a few nonzero components under a common basis/dictionary. The set of indices corresponding to the nonzero components are called the support for the signal. If several sparse signals share a common support, we call them jointly sparse. Sparse signal support recovery aims at identifying the true support of jointly sparse signals through its noisy linear measurements. Suppose that $S$ is an index set, then for $x \in \mathbb{F}^N$ a vector, $x_S$ denotes the vector formed by those components of $x$ indicated by $S$; for $A \in \mathbb{F}^{M \times N}$ a matrix, $A_S$ denotes the matrix formed by those columns indicated by $S$. 

Gongguo Tang and Arye Nehorai (Institute) Performance Analysis for Sparse Support Recovery April 21st 2009 4 / 41
Long-established paradigm for digital data acquisition

- sample data at Nyquist rate (2x bandwidth)
- compress data (signal-dependent, nonlinear)
- brick wall to resolution/performance
"Why go to so much effort to acquire all the data when most of what we get will be thrown away? Can't we just directly measure the part that won't end up being thrown away?"

— David L. Donoho
Directly acquire “compressed” data
Replace samples by more general “measurements”

\[ K < M \ll N \]

This slide is adapted from R. Baraniuk, J. Romberg and M. Wakin’s "Tutorial on Compressive Sensing".
When data is sparse/compressible, we can directly acquire a condensed representation with no/little information loss.

Random projection will work.

\[ M \times 1 \text{ measurements} = \Phi \]

\[ M = O(K \log(N/K)) \]

\[ N \times 1 \text{ sparse signal} \]

\[ K \text{ nonzero entries} \]

This slide is adapted from R. Baraniuk, J. Romberg and M. Wakin’s "Tutorial on Compressive Sensing".
Background

Previous Assumptions

- When there are measurement noises, there are different criteria for measuring the recovery performance
  - various $l_p$ norms $\mathbb{E} \| \hat{x} - x^* \|_p$, especially $l_2$ and $l_1$
  - predictive power (e.g., $\mathbb{E} \| y - \hat{y} \|^2_2$, where $\hat{y}$ is the estimate of $y$ based on $\hat{x}$
  - 0–1 loss associated with the event of recovering the correct support $S$

- Assumptions on noise
  - bounded noise
  - sparse noise
  - Gaussian noise
Background

Previous Assumptions

- Assumptions on sparse signal
  - deterministic with unknown support but known component values
  - deterministic with unknown support and unknown component values
  - random with unknown support

- Assumptions on measurement matrix
  - standard Gaussian ensemble
  - Bernoulli ensemble
  - random but with a structure such as Toeplitz
  - deterministic
Motivation

Why Support Recovery?

The support of a sparse signal has physical significance

- the timing of events
- the locations of objects or anomalies
  - Compressive Radar Imaging
  - Compressive Sensor Network
- the frequency components
  - Compressive Spectrum Analysis
- the existence of certain substances such as chemicals and mRNAs
  - Compressed Sensing DNA Microarrays
After the recovery of the support, the magnitudes of the nonzero components can be obtained by solving a least-square problem.
Consider parameter estimation problem associated with the following widely applied model,

\[ y(t) = A(\theta)x(t) + w(t), t = 1, \ldots, T, \]

where \( A(\theta) = \begin{bmatrix} \varphi(\theta_1) & \varphi(\theta_2) & \cdots & \varphi(\theta_K) \end{bmatrix} \) and \( \theta_1, \theta_2, \ldots, \theta_K \) are true parameters.

In order to solve this problem, we sample the parameter space to \( \{\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_N\} \) and form

\( \tilde{A}(\tilde{\theta}) = \begin{bmatrix} \varphi(\tilde{\theta}_1) & \varphi(\tilde{\theta}_2) & \cdots & \varphi(\tilde{\theta}_N) \end{bmatrix} \). Define vector \( \tilde{x}(t) \) by setting its components to those of \( x(t) \) when their locations correspond to true parameters and zero otherwise. Then we have transformed a traditional parameter estimation problem to one of support recovery.
Research Overview
Introduce hypothesis testing problems for sparse signal support recovery

Derive an upper bound for the probability of error (PoE) for general measurement matrix

Study the effect of different parameters

Analyze the PoE for multiple hypothesis testing and its implications for system design
Mathematical Model
We will focus on the following model:

\[ y(t) = Ax(t) + w(t), \quad t = 1, \ldots, T, \]  

(1)

or in matrix form

\[ Y = AX + W. \]

- Here we have \( x(t) \in \mathbb{F}^N, \ w(t) \in \mathbb{F}^M, \ y(t) \in \mathbb{F}^M \) with \( \mathbb{F} = \mathbb{R} \) or \( \mathbb{C} \).
- \( X, \ W, \ Y \) are matrices with columns formed by \( \{x(t)\}_{t=1}^T, \ \{w(t)\}_{t=1}^T, \ \{y(t)\}_{t=1}^T \) respectively.
- Our analysis involves a constant \( \kappa \) which is \( \frac{1}{2} \) for \( \mathbb{F} = \mathbb{R} \) and 1 for \( \mathbb{F} = \mathbb{C} \).
- Generally \( M \) is the dimension of hardware while \( T \) is the number of time samples. Hence increasing \( M \) is more expensive.
Mathematical Model
Assumptions on Signal and Noise

We have the following assumptions:

- \( \{x(t)\}_{t=1}^T \) are jointly sparse signals with a common support \( S = \text{supp}(X) \).
- \( \{x_S(t)\}_{t=1}^T \) follow i.i.d. \( \mathcal{N}(0, I_K) \).
- \( \{w(t)\}_{t=1}^T \) follow i.i.d. \( \mathcal{N}(0, \sigma^2 I_M) \) and are independent of \( \{x(t)\}_{t=1}^T \). Note that the noise variance \( \sigma^2 \) can be viewed as \( 1/\text{SNR} \).
We consider two types of measurement matrices:

1. **Non-degenerate measurement matrix**: we say that a general measurement matrix $A_{M \times N}$ is non-degenerate if every $M \times M$ submatrix of $A$ is nonsingular.

2. **Gaussian measurement matrix**: The element of $A$, say, $a_{ij}$ follows i.i.d. $\mathcal{N}(0, 1)$. 
We focus on two hypothesis testing problem:

1. **Binary hypothesis testing (BHT)** with $|S_0| = |S_1|:

   \[
   \begin{cases}
   H_0 : \text{supp} (X) = S_0 \\
   H_1 : \text{supp} (X) = S_1 
   \end{cases}
   \]

2. **Multiple hypothesis testing (MHT)**:

   \[
   \begin{cases}
   H_1 : \text{supp} (X) = S_1 \\
   \vdots \\
   H_L : \text{supp} (X) = S_L 
   \end{cases}
   \]

   where $S_i$’s are candidate supports with the same cardinality $|S_i| = K$. 
Our aim is to calculate an accurate upper bound for the PoE and analyze the effect of $M$, $T$, and noise variance $\sigma^2$.

\[
p_{\text{err}}(A) = \frac{1}{2} \int_{H_1} \Pr(Y|H_0) \, dY + \frac{1}{2} \int_{H_0} \Pr(Y|H_1) \, dY
\]

for BHT and

\[
p_{\text{err}}(A) = \sum_{i=1}^{L} \frac{1}{L} \int_{H_{j}:j\neq i} \Pr(Y|H_i) \, dY
\]

for MHT.
Theoretical Analysis
The BHT problem is equivalent to deciding between two distributions of $Y$:

$$Y|H_0 \sim \mathcal{F}\mathcal{N}_{M,T}(0, \Sigma_0 \otimes I_T) \quad \text{or} \quad Y|H_1 \sim \mathcal{F}\mathcal{N}_{M,T}(0, \Sigma_1 \otimes I_T),$$

where $\Sigma_i = \sigma^2 I_M + A_{S_i} A_{S_i}^\dagger$.

With equal prior probabilities of $S_0$ and $S_1$, the optimal decision rule is given by the likelihood ratio test:

$$\frac{f(Y|H_1)}{f(Y|H_0)} \overset{H_1}{\gtrless} \frac{H_1}{H_0} 1 \Leftrightarrow \text{tr} \left[ Y^\dagger \left( \Sigma_1^{-1} - \Sigma_0^{-1} \right) Y \right] \overset{H_1}{\gtrless} T \log \frac{\Sigma_0}{\Sigma_1}$$
Due to the symmetry of $H_0$ and $H_1$, we can just compute the probability of false alarm

\[
p_{FA} = \Pr\{H_1|H_0\} = \Pr\left\{\operatorname{tr}\left[\mathbf{Y}^\dagger \left(\Sigma_1^{-1} - \Sigma_0^{-1}\right) \mathbf{Y}\right] < T \log \frac{\left|\Sigma_0\right|}{\left|\Sigma_1\right|} H_0\right\} = \Pr\left\{\operatorname{tr}\left[\mathbf{Z}^\dagger \left(\Sigma_0^{1/2}\Sigma_1^{-1}\Sigma_0^{1/2} - I_M\right) \mathbf{Z}\right] < T \log \frac{\left|\Sigma_0\right|}{\left|\Sigma_1\right|} H_0\right\},
\]

where $\mathbf{Z} = \Sigma_0^{-1/2} \mathbf{Y} \sim \mathcal{F}\mathcal{N}(0, I_M \otimes I_T)$.

We define $H = \Sigma_0^{1/2}\Sigma_1^{-1}\Sigma_0^{1/2}$ with $\Sigma_i = A_{S_i}A_{S_i}^\dagger + \sigma^2I_M$, which is a fundamental matrix in our analysis.
Suppose the ordered eigenvalues of $H$ are 

$$
\sigma_1 < \sigma_2 < \cdots < \sigma_{k_1} < 1 = 1 = \cdots = 1 < \lambda_1 < \lambda_2 < \cdots < \lambda_{k_0},
$$

and $H$ can be diagonalized by an orthogonal/unitary matrix $Q$. Then the transformation of $Z = QN$ will give us

$$
p_{FA} = \Pr \left\{ \sum_{i=1}^{k_0} (\lambda_i - 1) \sum_{t=1}^{T} |N_{it}|^2 - \sum_{i=1}^{k_1} (1 - \sigma_i) \sum_{t=1}^{T} \left| N_{(i+k_0)t} \right|^2 \right. \\
< T \log \frac{|\Sigma_0|}{|\Sigma_1|} |H_0| \right\}
$$
The eigenvalue structure of $H$, especially the eigenvalues that are greater than 1, determines the performance of measurement matrix $A$ in distinguishing between different supports. We study the structure of $H$ in a slightly general setting where the sizes of the two candidate supports might not be equal.

### Problem

1. **How many eigenvalues of $H$ are less than 1, greater than 1 and equal to 1? Is there a general rule?**

2. **Can we give tight lower bounds on the eigenvalues that are greater than 1? The bounds should have a nice distribution that can be handled easily.**
$M = 200, |S_0 \cap S_1| = 20, |S_0 \setminus S_1| = 80, |S_1 \setminus S_0| = 60$ and the elements of $A$ are $i.i.d.$ real Gaussian.
Note that $|S_1 \setminus S_0| = 60$ eigenvalues of $H$ are less than 1, $|S_0 \setminus S_1| = 80$ greater than 1, and $M - (|S_0 \setminus S_1| + |S_1 \setminus S_0|) = 60$ identical to 1.
Theoretical Analysis
Eigenvalue Structure of $H$

**Theorem**

Suppose $k_i = |S_0 \cap S_1|$, $k_0 = |S_0 \setminus S_1|$, $k_1 = |S_1 \setminus S_0|$ and $M \geq k_0 + k_1$, for general non-degenerate measurement matrix, $k_0$ eigenvalues of matrix $H$ are greater than 1, $k_1$ less than 1 and $M - (k_0 + k_1)$ equal to 1.

Note that from the bound we present later, $\sqrt{\prod_{i=1}^{k_0} \lambda_i \prod_{i=1}^{k_1} (1/\sigma_i)}$ determines the performance of the optimal BHT decision rule. Hence, generally and quite intuitively, the larger the difference set $S_0 \Delta S_1$, the easier to distinguish between the two candidate supports.
Theoretical Analysis
Eigenvalue Structure of H

**Theorem**

For Gaussian measurement matrix, the sorted eigenvalues of $H$ that are greater than 1 are lower bounded by those of $I_{k_0} + \frac{1}{\sigma^2}V$ with probability one, where $V$ is a matrix obtained from measurement matrix $A$ and $V$ follows $\mathcal{N}_{k_0}(I_{k_0}, 2\kappa(M - k_1 - k_i))$.

We comment that generally the larger $M - k_1 - k_i = M - |S_1|$, the larger the eigenvalues of $I_{k_0} + \frac{1}{\sigma^2}V$, and hence the better we can distinguish the true support from the false one.
Theoretical Analysis
A Lower Bound on Eigenvalues

- $M = 200$, $|S_0 \cap S_1| = 20$, $|S_0 \setminus S_1| = 80$, $|S_1 \setminus S_0| = 60$, $\sigma^2 = 4$ and the element of $A$ are \textit{i.i.d.} real Gaussian.
- Blue line represents the true sorted eigenvalues of $H$ that are greater than 1 and red line represents the lower bound.
Theoretical Analysis
Bound on PoE

**Theorem**

The Probability of False Alarm can be bounded by

\[
p_{FA} = \Pr (S_1|H_0) \leq \left\{ \left[ \frac{\bar{\lambda}_g (S_0, S_1)}{4} \right]^{k_d/2} \left[ \frac{\bar{\lambda}_g (S_1, S_0)}{4} \right]^{k_d/2} \right\}^{-\kappa T},
\]

where \( k_d = |S_0 \setminus S_1| \), \( \bar{\lambda}_g (S_0, S_1) = k_d \sqrt{\prod_{j=1}^{k_d} \lambda_j} \) with \( \lambda_j \)'s the eigenvalues of \( H = \left( A_{S_0} A_{S_0}^\dagger + \sigma^2 I_M \right)^{1/2} \left( A_{S_1} A_{S_1}^\dagger + \sigma^2 I_M \right)^{-1} \left( A_{S_0} A_{S_0}^\dagger + \sigma^2 I_M \right)^{1/2} \) that are greater than one.
Theoretical Analysis
Implications of the Bound

- The bound can be equivalently written as

\[ \left( \frac{\sqrt{\prod_{i=1}^{kd} \lambda_i \prod_{i=1}^{kd} (1/\sigma_i)}}{4} \right)^{-\kappa k_d T} \]

with \( \lambda_i \)'s and \( \sigma_i \)'s eigenvalues of \( H \) that are greater and less than 1, respectively. Hence these eigenvalues determines the systems ability in distinguishing two supports.

- As we will see the minimum of all \( \tilde{\lambda}_g (S_i, S_j) \)'s determines the systems ability in distinguishing all candidate supports, and can be viewed as a measure of incoherence.

- The logarithm of the bound can be approximated by

\[ -\kappa k_d T \left\{ \frac{1}{2} \log \left[ \tilde{\lambda}_g (S_0, S_1) \tilde{\lambda}_g (S_1, S_0) \right] - \log 4 \right\}. \]

Hence, if we can guarantee that \( \tilde{\lambda}_g (S_0, S_1) \tilde{\lambda}_g (S_1, S_0) \) of our measurement matrix is greater than some constant, then we can make the \( p_{\text{FA}} \) arbitrarily small by taking more temporal samples.
Now we turn to the MHT problem

\[
\begin{cases}
H_1 : \text{supp} \ (X) = S_1 \\
\vdots \\
H_L : \text{supp} \ (X) = S_L
\end{cases}
\]

where \( S_i \)'s are candidate supports with the same cardinality \( |S_i| = K \) and \( L = C^K_N \), the total number of candidate supports with size \( K \).
Theoretical Analysis
PoE for MHT

Theorem

Denote by $\lambda_{\text{min}} = \min \{ \lambda_g(S_i, S_j) \}$, then the total PoE for MHT can be bounded by

$$p_{\text{err}} \leq C \exp \left\{ -\kappa T \left[ \log (\lambda_{\text{min}}) - \log (4K(N - K))^{\frac{1}{\kappa T}} \right] \right\}.$$
Theoretical Analysis
Multiple Hypothesis Testing

Theorem

For \( T = O \left( \frac{\log N}{\log \left[ K \log \frac{N}{K} \right]} \right) \) and \( M = O(K \log (N/K)) \), then

\[
\text{Pr} \left\{ \bar{\lambda}_{\text{min}} > 4 \left[ K (N - K) \right]^{\frac{1}{KT}} \right\} \longrightarrow 1,
\]

as \( N, K, M \longrightarrow \infty \).
Theoretical Analysis

Discussion

- \( M = O(K \log(N/K)) \) is the same as conventional compressive sensing. We need \( MT \) samples in total. When \( K \) is sufficiently small compared with \( N \), this value is still much smaller than \( N \).

- Actually the value of \( T \) is not very large. For example, for \( N = 10^{100}, K = 10^5 \), we have \( \frac{\log N}{\log[K \log \frac{N}{K}]} \approx 13 \); for \( N = 10^{100}, K = 10^{98} \), we have \( \frac{\log N}{\log[K \log \frac{N}{K}]} \approx 1 \);

- After we recover the support, we can get the component values by solving a least-square problem.
In practice, given $N, K$, we take $M = O(K \log (N/K))$, $T = O \left( \frac{\log N}{\log [K \log \frac{N}{K}]} \right)$ and generate measurement matrix $A$. Then with large probability, we will get $\bar{\lambda}_{\text{min}} > 4 \left[ K (N - K) \right]^{\frac{1}{\kappa T}}$. For safety, we can

- compute $\bar{\lambda}_{\text{min}}$
- find $T$ large enough such that $\bar{\lambda}_{\text{min}} > 4 \left[ K (N - K) \right]^{\frac{1}{\kappa T}}$
- continue to increase $T$ so that $p_{\text{err}} < \alpha$. 
Conclusions

- Hypothesis testing for sparse signal support recovery
  - BHT
  - MHT
- Bound for PoE non-degenerate measurement matrix
- The behavior of critical quantity
- Implications in system design
  - Another dimension of data collection gives us more flexibility
Future Work

- Design measurement system with optimal $\bar{\lambda}_{\text{min}}$.
- Establish a necessary condition imposed on $M$ and $T$.
- Analyze the behavior of $\bar{\lambda}(S_0, S_1)$ and $\bar{\lambda}_{\text{min}}$ for other measurement matrix structures.
- Devise an efficient algorithm for support recovery and compare its performance with the optimal one.
  - The performance of $l_1$ minimization algorithm.
- Develop an algorithm to compute $\bar{\lambda}_{\text{min}}$ for given measurement matrix.
- Explore the relationship between $\bar{\lambda}_{\text{min}}$ and Restricted Isometry Property (RIP).
- Apply this result to the design of transmitted signals in Compressive Radar Imaging.
Thank you!