ECE 285: Sparsity and Compressed Sensing

Bhaskar Rao
Department of Electrical and Computer Engineering
University of California, San Diego

Outline

- Course Outline
- Motivation for Course
- Sparse Signal Recovery Problem
- Applications
- Computational Algorithms
  - Greedy Search
  - ℓ_1 norm minimization
  - Bayesian Methods
- Performance Guarantees
- Simulations
- Conclusions

Topics

- Sparse Signal Recovery Problem and Compressed Sensing
- Uniqueness
  - Spark
- Greedy search techniques and their performance evaluation
  - Coherence condition
- ℓ_1 methods and their performance evaluation
  - Restricted isometry property (RIP)
- Bayesian methods
  - MAP (Reweighted ℓ_1 and Reweighted ℓ_2)
  - Hierarchical Bayesian Methods (Sparse Bayesian Learning)
- Extensions (Block Sparsity, Multiple Measurement vectors)
- Dictionary Learning
- Message passing algorithms

Reference Books

- Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing by Michael Elad
- Compressed Sensing: Theory and Applications, edited by Yonina C. Eldar and Gitta Kutyniok
- An Introduction to Compressive Sensing, Collection Editors: Richard Baraniuk, Mark A. Davenport, Marco F. Duarte, Chinmay Hegde
- A Mathematical Introduction to Compressive Sensing by Simon Foucart and Holger Rauhut

Administrative details

- Who should take this class and background?
  - ≥ Second year graduate students, recommend S/U
  - Optimization theory, Estimation theory
  - Recommend an application to motivate the work
- Grades
  - Homework (60%)
  - Project (40%)
- Office hours
  - Tuesday 1-2pm
  - Office: Jacobs Hall 6407
  - Email: brao@ucsd.edu
  - Class Website: dsp.ucsd.edu

Outline

- Motivation for Course
- Sparse Signal Recovery Problem
- Applications
- Computational Algorithms
  - Greedy Search
  - ℓ_1 norm minimization
  - Bayesian Methods
- Performance Guarantees
- Simulations
- Conclusions

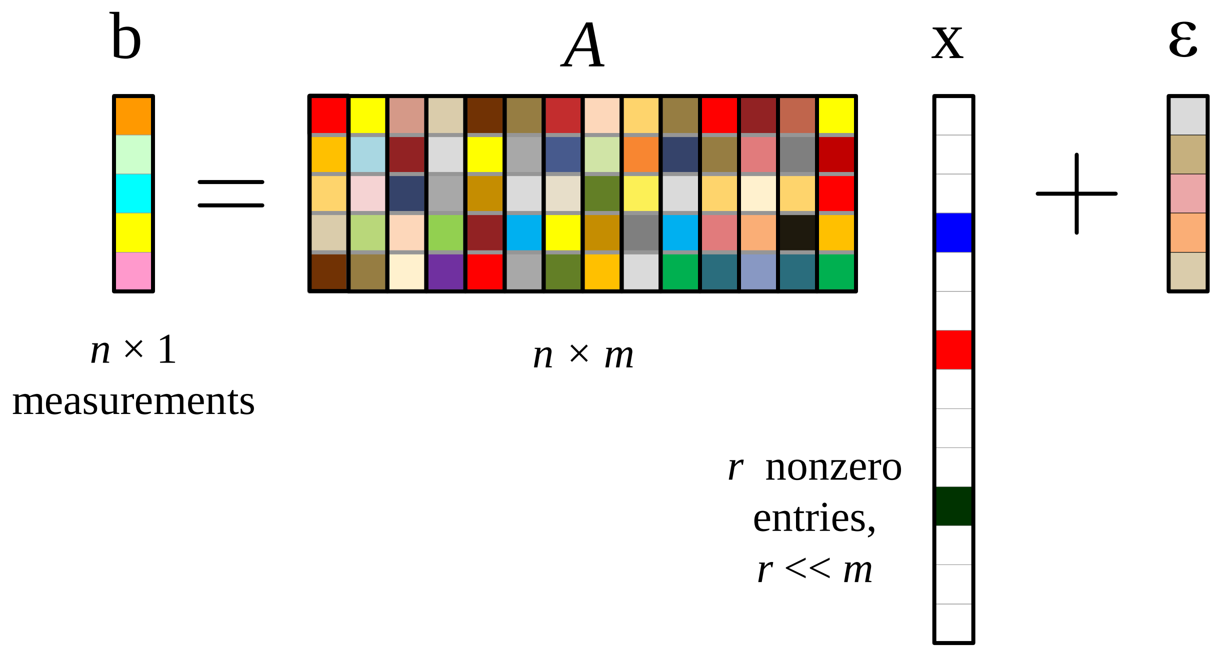
Motivation

- The concept of Sparsity has many potential applications. Unification of the theory will provide synergy.
- Methods developed for solving the Sparse Signal Recovery problem can be a valuable tool for signal processing practitioners.
- Many interesting developments in the recent past that make the subject timely.

Outline

- Motivation for Course
- Sparse Signal Recovery Problem
- Applications
- Computational Algorithms
  - Greedy Search
  - ℓ_1 norm minimization
  - Bayesian Methods
- Performance Guarantees
- Simulations
- Conclusions

Sparse Signal Recovery: Problem Description



$m \times 1$
sparse signal

- b is $n \times 1$ measurement vector
- A is $n \times m$ measurement/Dictionary matrix, $m >> n$
- x is $m \times 1$ desired vector which is sparse with r nonzero entries
- ε is the measurement noise

Early Works

- R. R. Hocking and R. N. Leslie , "Selection of the Best Subset in Regression Analysis," Technometrics, 1967.
- S. Singhal and B. S. Atal, "Amplitude Optimization and Pitch Estimation in Multipulse Coders," *IEEE Trans. Acoust., Speech, Signal Processing*, 1989
- S. D. Cabrera and T. W. Parks, "Extrapolation and spectral estimation with iterative weighted norm modification," *IEEE Trans. Acoust., Speech, Signal Processing, April 1991.*
- Many More works
- Our first work
  - I.F. Gorodnitsky, B. D. Rao and J. George, "Source Localization in Magnetoencephalography using an Iterative Weighted Minimum Norm Algorithm, IEEE Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, Pages: 167-171, Oct. 1992

Problem Statement

- **Noise Free Case**: Given a target signal *y* and a dictionary Φ, find the weights *x* that solve:

$$\min_x \sum_{i=1}^{m} I(x_i \neq 0) \quad \text{subject to} \quad b = Ax$$

where *I*(.) is the indicator function

- **Noisy Case**: Given a target signal *y* and a dictionary Φ, find the weights *x* that solve:

$$\min_x \sum_{i=1}^{m} I(x_i \neq 0) \quad \text{subject to} \quad \|b - Ax\|_2 \leq \beta$$

Complexity

- Search over all possible subsets, which would mean a search over a total of $({}^{m}C_{r})$ subsets. Combinatorial Complexity.

  With $m = 30$; $n = 20$; and $r = 10$ there are 3×10^7 subsets (Very Complex)

- A branch and bound algorithm can be used to find the optimal solution. The space of subsets searched is pruned but the search may still be very complex.

- Indicator function not continuous and so not amenable to standard optimization tools.

Challenge: Find low complexity methods with acceptable performance

Outline

- Motivation for Course
- Sparse Signal Recovery Problem
- Applications
- Computational Algorithms
  - Greedy Search
  - ℓ_1 norm minimization
  - Bayesian Methods
- Performance Guarantees
- Simulations
- Conclusions

Applications

- **Signal Representation** (Mallat, Coifman, Wickerhauser, Donoho, ...)
- **EEG/MEG** (Leahy, Gorodnitsky, Ioannides, ...)
- **Functional Approximation and Neural Networks** (Chen, Natarajan, Cun, Hassibi, ...)
- **Bandlimited extrapolations and spectral estimation** (Papoulis, Lee, Cabrera, Parks, ...)
- **Speech Coding** (Ozawa, Ono, Kroon, Atal, ...)
- **Sparse channel equalization** (Fevrier, Greenstein, Proakis, ...)
- **Compressive Sampling** (Donoho, Candes, Tao...)
- **Magnetic Resonance Imaging** (Lustig,..)
- **Cognitive Radio (**Eldar, ..)

DFT Example

- **Measurement *y***

$$y[l] = 2(\cos\omega_0 l + \cos\omega_1 l),\ l = 0,1,2,\ldots,n-1.\ n = 64.$$

$$\omega_0 = \frac{2\pi}{64}\frac{33}{2},\quad \omega_1 = \frac{2\pi}{64}\frac{34}{2}.$$

- **Dictionary Elements**:

$$a_l^{(m)} = [1, e^{-j\omega_l}, e^{-j2\omega_l}, \ldots, e^{-j(n-1)\omega_l}]^T, \omega_l = \frac{2\pi}{m} l$$

- Consider m = 64, 128, 256 and 512.

Questions:

- What is the result of a zero padded DFT?
- When viewed as problem of solving a linear system of equations (dictionary), what solution does the DFT give us?
- Are there more desirable solutions for this problem?

DFT Example

- Note that

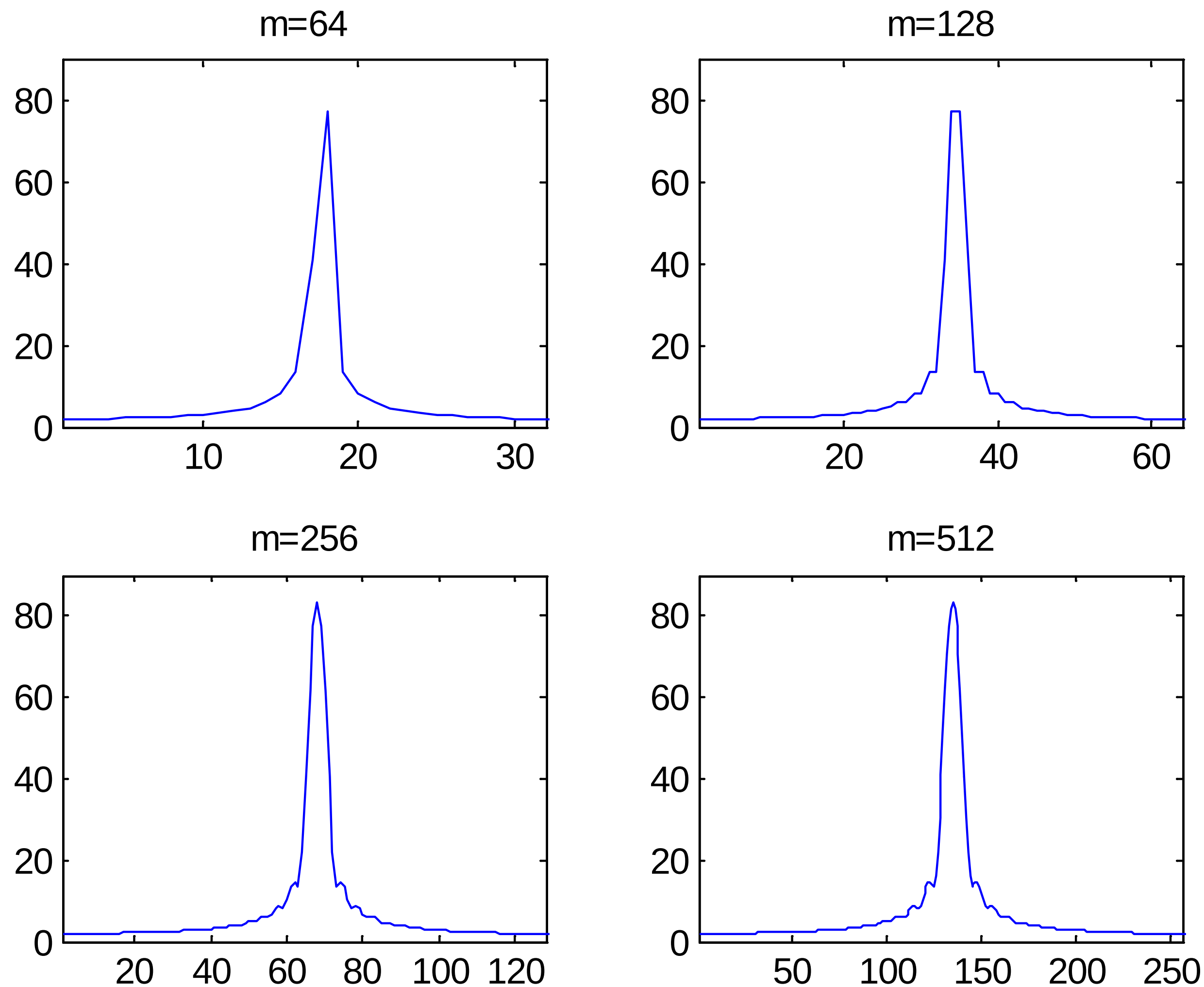
$$
\begin{aligned}
b &= a_{33}^{(128)} + a_{34}^{(128)} + a_{94}^{(128)} + a_{95}^{(128)} \\
&= a_{66}^{(256)} + a_{68}^{(256)} + a_{188}^{(256)} + a_{190}^{(256)} \\
&= a_{132}^{(512)} + a_{136}^{(512)} + a_{376}^{(512)} + a_{380}^{(512)}
\end{aligned}
$$

- Consider the linear system of equations

$$b = A^{(m)}x$$

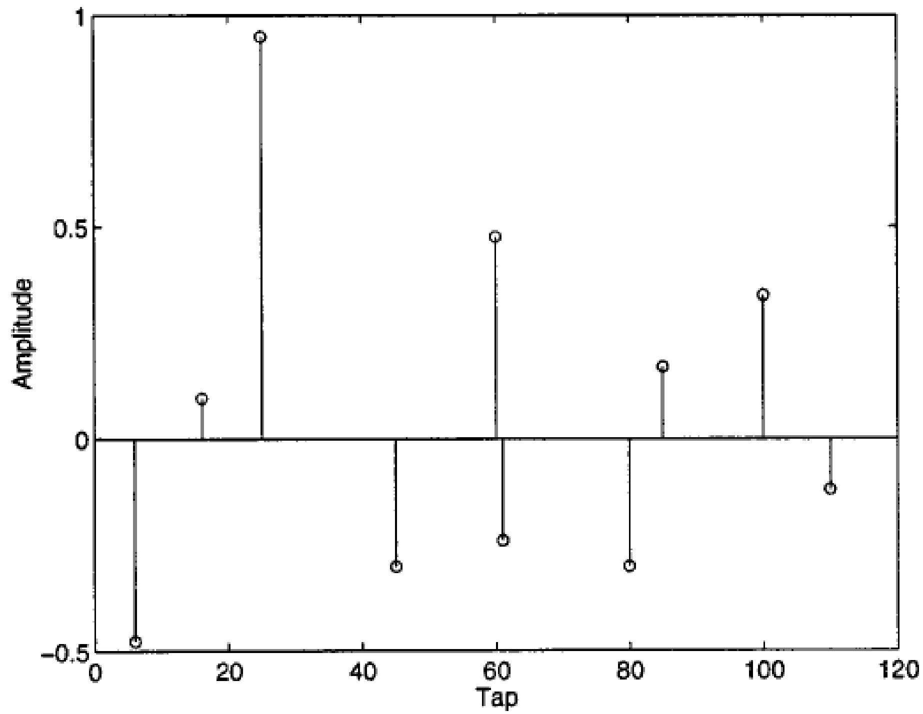
- The frequency components in the data are in the dictionaries $A^{(m)}$ for $m = 128, 256, 512$.
- What solution among all possible solutions does the DFT compute?

DFT Example



m=64
m=128
m=256
m=512
0
20
40
60
80
10
20
30
40
50
60
100
120
150
200
250

Sparse Channel Estimation



Noise

$$r(i) = \sum_{j=0}^{m-1} s(i-j)c(j) + \varepsilon(i), \quad i = 0,1,\ldots,n-1$$

Received seq. Training seq. Channel impulse response

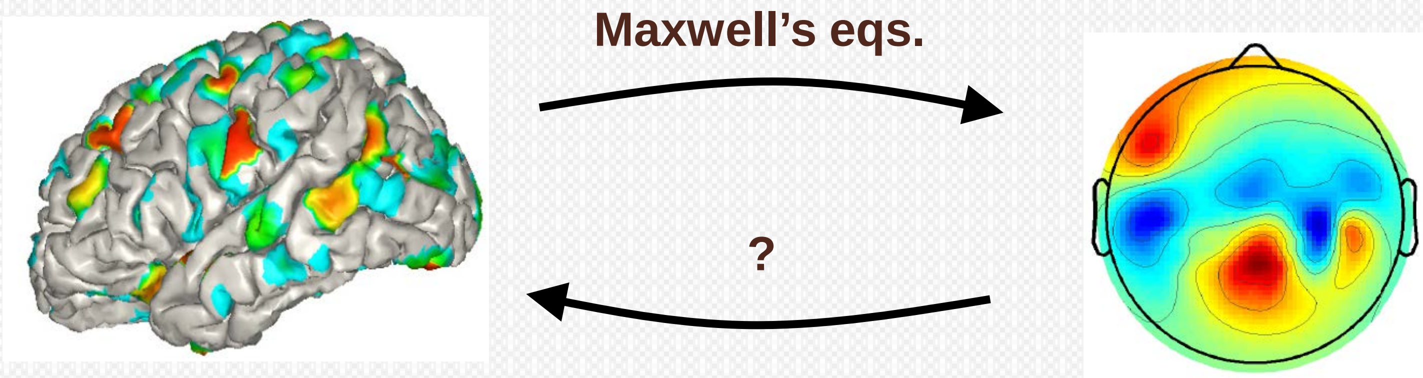
Example: Sparse Channel Estimation

- Formulated as a sparse signal recovery problem

$$\begin{bmatrix} r(0) \\ r(1) \\ \vdots \\ r(n-1) \end{bmatrix} = \begin{bmatrix} s(0) & s(-1) & \cdots & s(-m+1) \\ s(1) & s(0) & \cdots & s(-m+2) \\ \vdots & \vdots & \ddots & \vdots \\ s(n-1) & s(n-2) & \cdots & s(-m+n) \end{bmatrix} \begin{bmatrix} c(0) \\ c(1) \\ \vdots \\ c(m-1) \end{bmatrix} + \begin{bmatrix} \varepsilon(0) \\ \varepsilon(1) \\ \vdots \\ \varepsilon(n-1) \end{bmatrix}$$

- Can use any relevant algorithm to estimate the sparse channel coefficients

MEG/EEG Source Localization



source space (x)

sensor space (b)

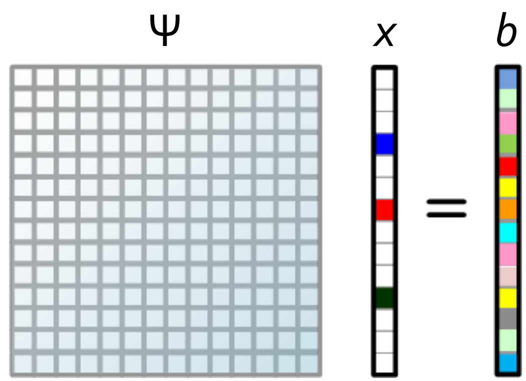
Compressive Sampling

- D. Donoho, “Compressed Sensing,” IEEE Trans. on Information Theory, 2006

- E. Candes and T. Tao, “Near Optimal Signal Recovery from random Projections: Universal Encoding Strategies,” IEEE Trans. on Information Theory, 2006

Compressive Sampling

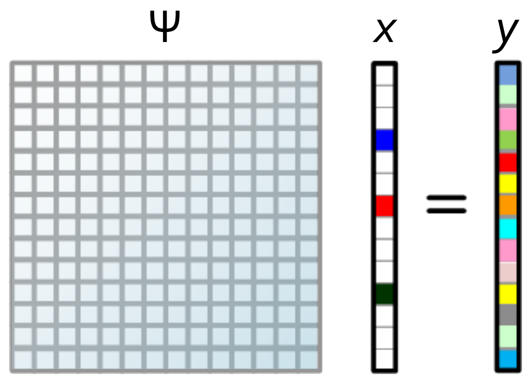
- Transform Coding



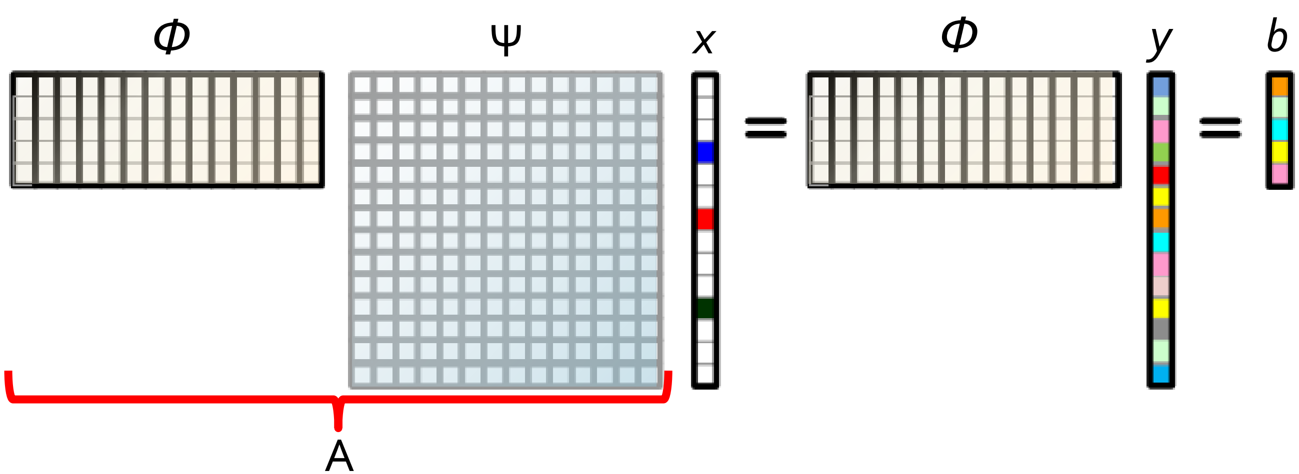
- What is the problem here?
  - Sampling at the Nyquist rate
  - Keeping only a small amount of nonzero coefficients
  - Can we directly acquire the signal below the Nyquist rate?

Compressive Sampling

- Transform Coding

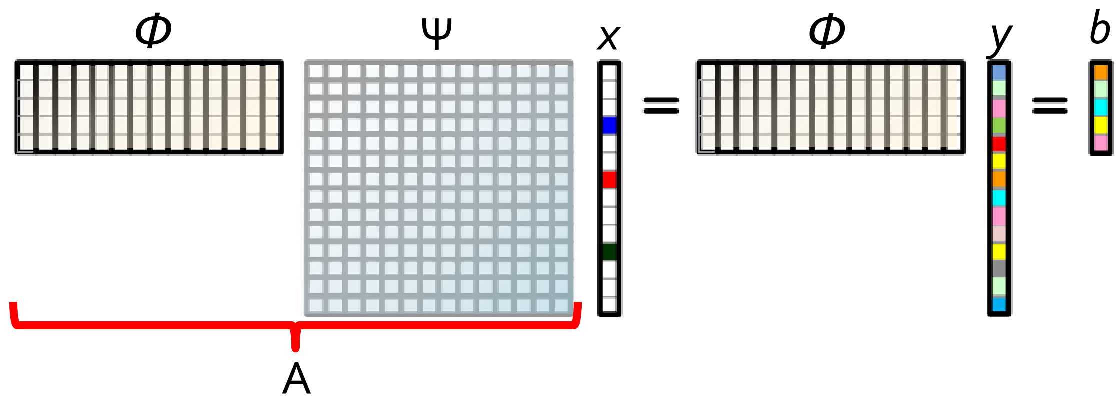


- Compressive Sampling



Compressive Sampling

- Compressive Sampling



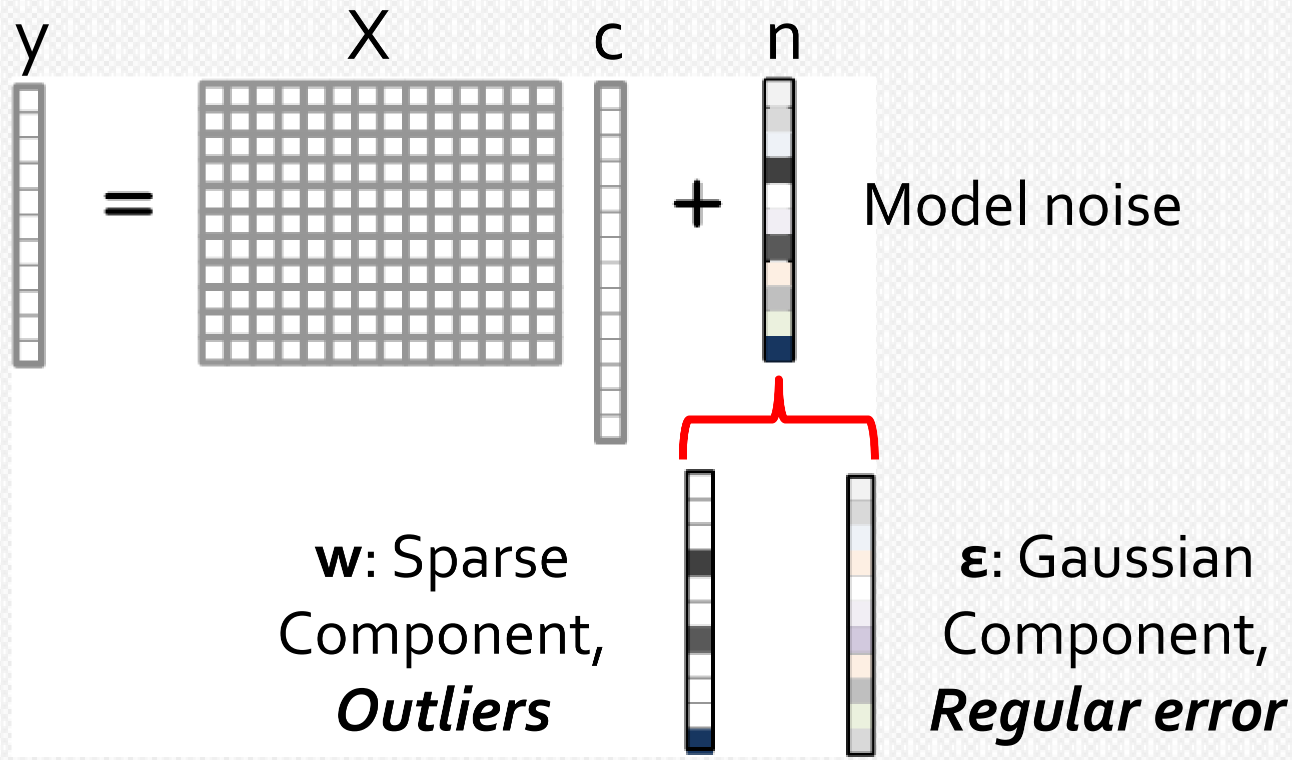
- Computation:
  1. Solve for x such that $Ax = b$
  2. Reconstruction: $y = \Psi x$
- Issues
  - Need to recover sparse signal x with constraint $Ax = b$
  - Need to design sampling matrix Φ

Robust Linear Regression

X, y: data;

c: regression coeffs.;

n: model noise;



Transform into overcomplete representation:

$$\mathbf{Y = X\,c + \Phi\,w + \varepsilon, \text{ where } \Phi = I,}$$

or

$$\mathbf{Y = [X\,,\,\Phi]}\begin{bmatrix} c \\ w \end{bmatrix} + \boldsymbol{\varepsilon}$$

Outline

- Motivation for Course
- Sparse Signal Recovery Problem
- Applications
- Computational Algorithms
  - Greedy Search
  - ℓ_1 norm minimization
  - Bayesian Methods
- Performance Guarantees
- Simulations
- Conclusions

Potential Approaches

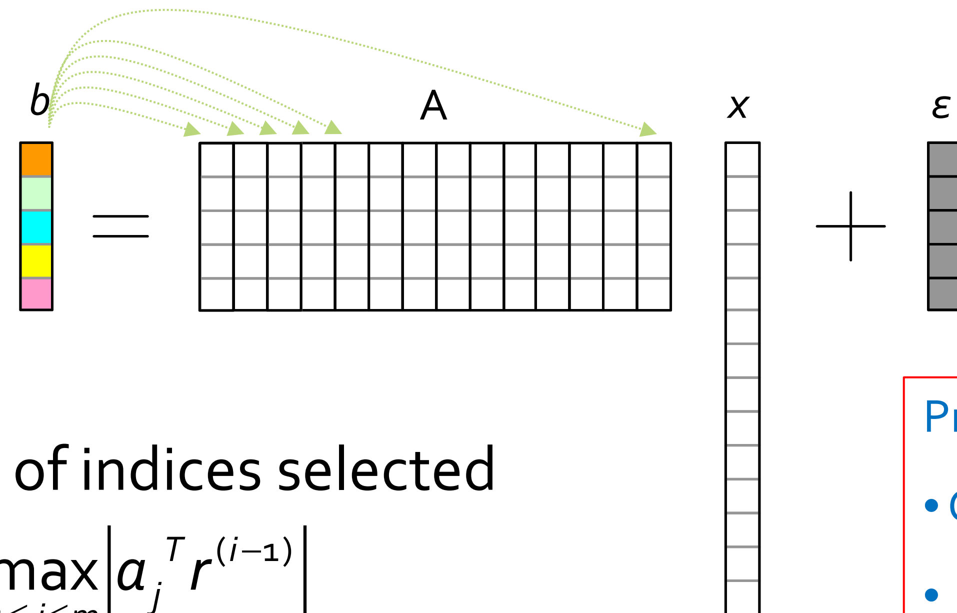
Combinatorial Complexity and so need alternate strategies

- Greedy Search Techniques: Matching Pursuit, Orthogonal Matching Pursuit
- Minimizing Diversity Measures: Indicator function not continuous. Define Surrogate Cost functions that are more tractable and whose minimization leads to sparse solutions, e.g. minimization ℓ_1
- Bayesian Methods:
  - MAP estimation (Reweighted ℓ_1 and reweighted $\ell 2$)
  - Hierarchical Bayesian Methods (Sparse Bayesian Learning)
- Message Passing Algorithms

GREEDY SEARCH TECHNIQUES

Greedy Search Method: Matching Pursuit

- Select a column that is most aligned with the current residual



  - $r^{(0)} = b$
  - $S^{(i)}$: set of indices selected
  - $l = \underset{1 \le j \le m}{\operatorname{argmax}} \left| {a_j}^T r^{(i-1)} \right|$

Practical stop criteria:

- Certain # iterations
- $\left\| r^{(i)} \right\|_2$ smaller than threshold

- Remove its contribution from the residual
  - Update $S^{(i)}$: If $l \notin S^{(i-1)}, S^{(i)} = S^{(i-1)} \bigcup \{l\}$. Or, keep $S^{(i)}$ the same
  - Update $r^{(i)}$: $r^{(i)} = \mathrm{P}_{a_l}^{\perp} r^{(i-1)} = r^{(i-1)} - a_l {a_l}^T r^{(i-1)}$

Question related to Matching Pursuit Type Algorithms

- Alternate search techniques
- Performance Guarantees

MINIMIZING DIVERSITY MEASURES

Inverse Techniques

- For the systems of equations Ax = b, the solution set is characterized by $\{x_s : x_s = A^+ y + v;\ v \in N(A)\}$, where N(A) denotes the null space of A and $A^+ = A^T(AA^T)^{-1}$.

- Minimum Norm solution: The minimum ℓ_2 norm solution

  $x_{mn} = A^+ b$ is a popular solution

- Noisy Case: regularized ℓ_2 norm solution often employed and is given by

$$x_{reg} = A^T(AA^T + \lambda I)^{-1} b$$

Minimum 2-Norm Solution

- Problem: Minimum ℓ_2 norm solution is not sparse

Example:

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$x_{mn} = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} & \frac{1}{3} \end{bmatrix}^T \quad \text{vs.} \quad x = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T$$

DFT: Also computes minimum 2-norm solution

Diversity Measures

$$\min_{x} \sum_{i=1}^{m} I(x_i \neq 0) \quad \text{subject to} \quad b = Ax$$

- Functionals whose minimization leads to sparse solutions
- Many examples are found in the fields of economics, social science and information theory
- These functionals are usually concave which leads to difficult optimization problems

Examples of Diversity Measures

- $\ell_{(p\leq 1)}$ Diversity Measure

$$E^{(p)}(x)=\sum_{i=1}^{m}|x_i|^p,\ p\leq 1$$

- As $p \rightarrow 0$,

$$\lim_{p\to 0}E^{(p)}(x)=\lim_{p\to 0}\sum_{i=1}^{m}|x_i|^p=\sum_{i=1}^{m}I(x_i\neq 0)$$

- Gaussian Entropy

$$E^{(G)}(x)=\sum_{i=1}^{m}\ln(\varepsilon+|x_i|^2)$$

ℓ_1 Diversity Measure

- Noiseless case

$$\min_x \sum_{i=1}^{m} |x_i| \text{ subject to } Ax = b$$

- Noisy case
  - ℓ_1 regularization [Candes, Romberg, Tao]

$$\min_x \sum_{i=1}^{m} |x_i| \text{ subject to } \|b - Ax\|_2 \le \beta$$

  - Lasso [Tibshirani], Basis Pursuit De-noising [Chen, Donoho, Saunders]

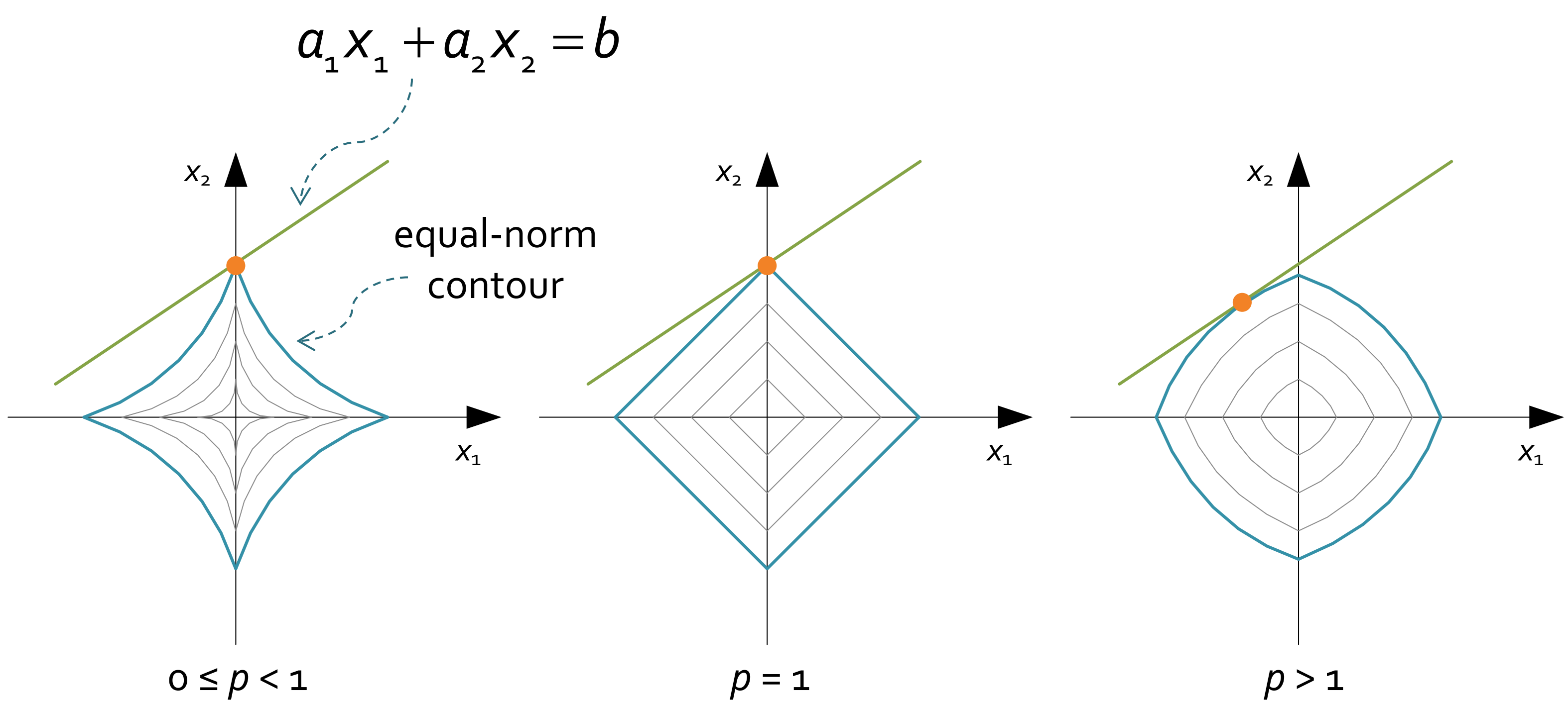
$$\min_x \|b - Ax\|_2^2 + \lambda \sum_{i=1}^{m} |x_i|$$

Attractiveness of ℓ_1 methods

- Convex Optimization and associated with rich class of optimization algorithms
  - Interior-point methods
  - Coordinate descent method
  - ....
- Question
  - What is the ability to find the sparse solution?

Why diversity measure encourages sparse solutions?

$$\min \left\| [x_1, x_2]^T \right\|_p^p \quad \text{subject to} \quad a_1 x_1 + a_2 x_2 = b$$



Example with ℓ_1 diversity measure

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

- Noiseless Case
  - $x_{BP} = [1, \ 0, \ 0]^T$ (machine precision)
- Noisy Case
  - Assume the measurement noise $\varepsilon = [0.01, \ -0.01]^T$
  - ℓ_1 regularization result: $x_{l1R} = [0.986, \ 0, \ 8.77 \times 10^{-6}]^T$
  - Lasso result ($\lambda = 0.05$): $x_{lasso} = [0.975, \ 0, \ 2.50 \times 10^{-5}]^T$

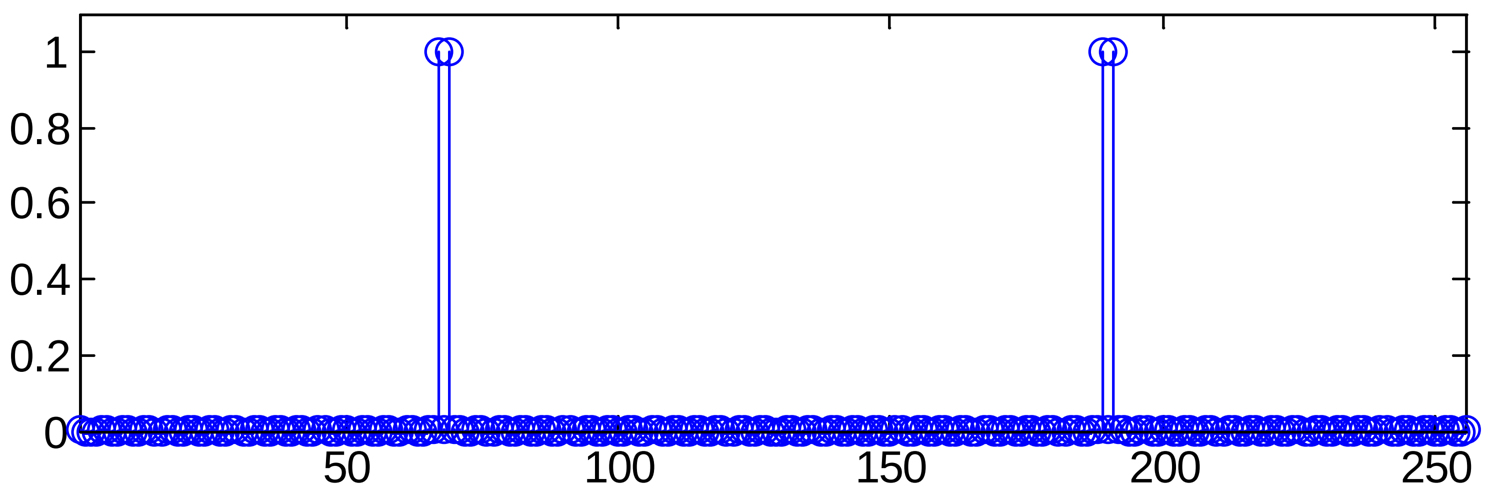
Example with ℓ_1 diversity measure

- Continue with the DFT example:

$$y[l] = 2(\cos\omega_0 l + \cos\omega_1 l),\ l = 0,1,2,\ldots,n-1.\ n = 64.$$

$$\omega_0 = \frac{2\pi}{64}\frac{33}{2},\quad \omega_1 = \frac{2\pi}{64}\frac{34}{2}.$$

- 64, 128,256,512 DFT cannot separate the adjacent frequency components
- Using ℓ_1 diversity measure minimization (m=256)



BAYESIAN METHODS

Bayesian Methods

- Maximum Aposteriori Approach (MAP)
  - Assume a sparsity inducing prior on the latent variable x
  - Developing an appropriate MAP estimation algorithm

$$\hat{x} = \arg\max_{x} \; p(x \mid b) = \arg\max_{x} \; p(b \mid x)\, p(x)$$

- Empirical Bayes
  - Assume a parameterized prior for the latent variable x (hyper-parameters)
  - Marginalize over the latent variable x and estimate the hyper-parameters
  - Determine the posterior distribution of x and obtain a point as the mean, mode or median of this density

Outline

- Motivation for Course
- Sparse Signal Recovery Problem
- Applications
- Computational Algorithms
  - Greedy Search
  - ℓ_1 norm minimization
  - Bayesian Methods
- Performance Guarantees
- Simulations
- Conclusions

Important Questions

- When is the ℓ_0 solution unique?
- When is the ℓ_1 solution equivalent to that of ℓ_0?
  - Noiseless Case
  - Noisy Measurements
- What are the limits of recovery in the presence of noise?
- How to design the dictionary matrix A?

Outline

- Motivation for Course
- Sparse Signal Recovery Problem
- Applications
- Computational Algorithms
  - Greedy Search
  - ℓ_1 norm minimization
  - Bayesian Methods
- Performance Guarantees
- Simulations
- Conclusions

Empirical Example

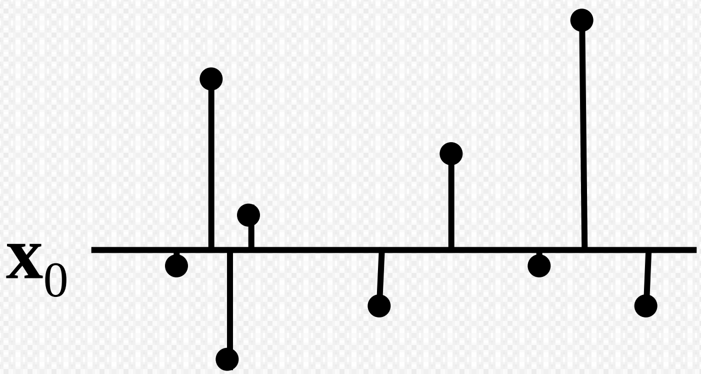
- For each test case:

1. Generate a random dictionary A with ***50 rows*** and ***100 columns***.
2. Generate a sparse coefficient vector $\mathbf{x}_0$.
3. Compute signal via $\mathbf{b} = A\,\mathbf{x}_0$ (noiseless).
4. Run ***BP*** and ***OMP***, as well as a competing Bayesian method called ***SBL*** (more on this later) to try and correctly estimate $\mathbf{x}_0$.
5. Average over1000 trials to compute empirical probability of failure.

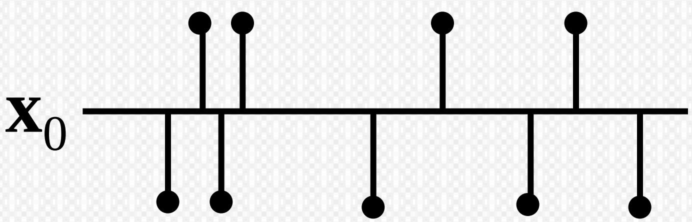
- Repeat with different sparsity values, i.e., $\|\mathbf{x}_0\|_0$ ranging from ***10*** to ***30***.

Amplitude Distribution

- If the magnitudes of the non-zero elements in $\mathbf{x}_0$ are highly scaled, then the canonical sparse recovery problem should be easier.



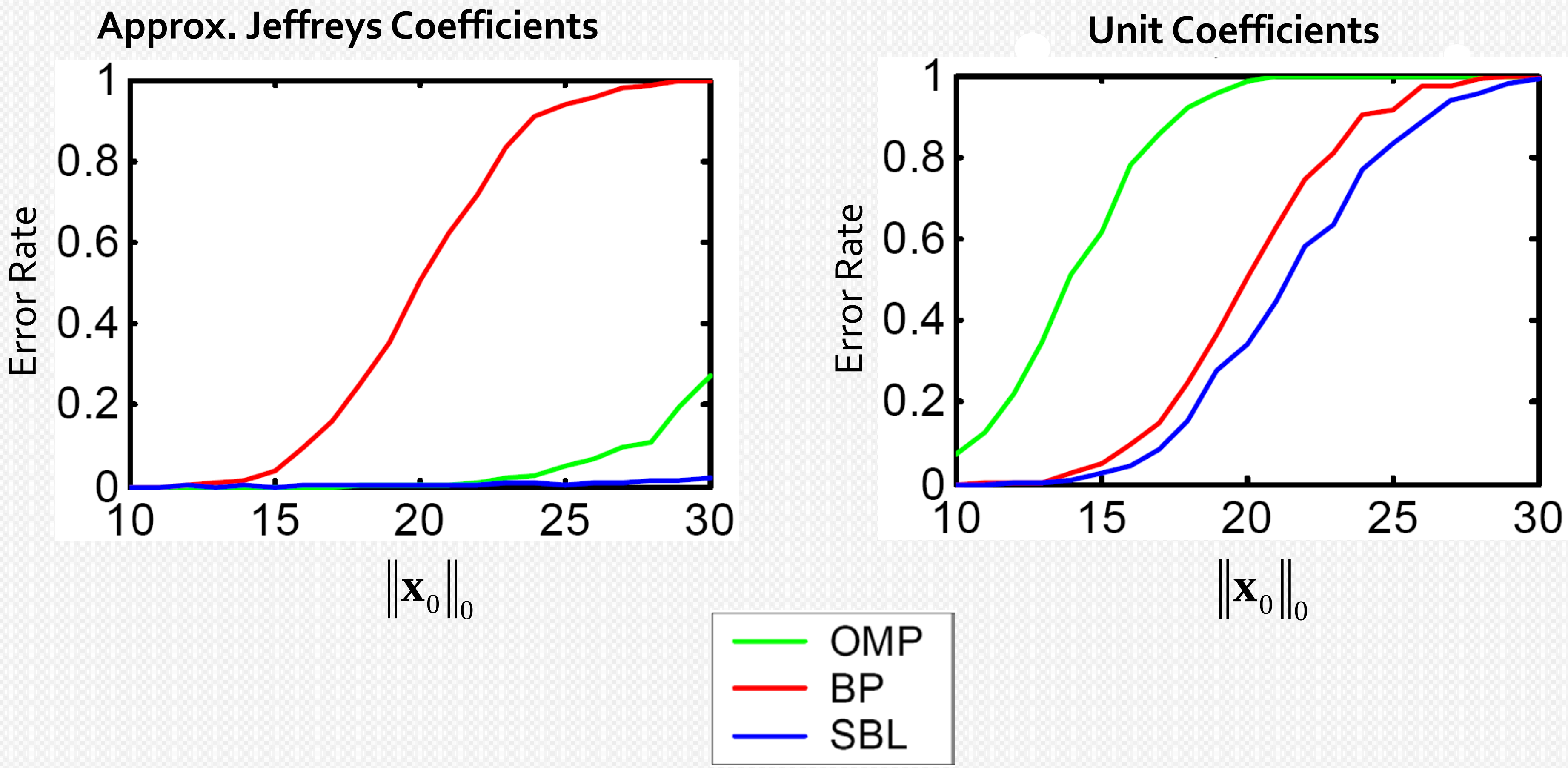
scaled coefficients (easy)



uniform coefficients (hard)

- The (approximate) Jeffreys distribution produces sufficiently scaled coefficients such that best solution can always be easily computed.

Sample Results ($n = 50$, $m = 100$)



Imaging Applications

1. Recovering fiber track geometry from diffusion weighted MR images [Ramirez-Manzanares et al. 2007].
2. Multivariate autoregressive modeling of fMRI time series for functional connectivity analyses [Harrison et al. 2003].
3. Compressive sensing for rapid MRI [Lustig et al. 2007].
4. MEG/EEG source localization [Sato et al. 2004; Friston et al. 2008].

Variants and Extensions

- Block Sparsity
- Multiple Measurement Vectors
- Dictionary Learning
- Scalable Algorithms
  - Message Passing Algorithms
- Sparsity for more general inverse problems
- More to come

Summary

- *Sparse Signal Recovery* is an interesting area with many potential applications.
- Methods developed for solving the Sparse Signal Recovery problem can be valuable tools for signal processing practitioners.
- Rich set of computational algorithms, e.g.,
  - Greedy search (OMP)
  - ℓ_1 norm minimization (Basis Pursuit, Lasso)
  - MAP methods (Reweighted ℓ_1 and ℓ_2 methods)
  - Bayesian Inference methods like SBL (show great promise)
- Potential for great theory in support of performance guarantees for algorithms.
- Expectation is that there will be continued growth in the application domain as well as in the algorithm development.