Scale Mixture Modeling of Priors for Sparse Signal Recovery

Bhaskar D Rao[1]
University of California, San Diego

[1]Thanks to David Wipf, Jason Palmer, Zhilin Zhang and Ritwik Giri

Outline

Outline

- Sparse Signal Recovery (SSR) Problem and some Extensions

Outline

- Sparse Signal Recovery (SSR) Problem and some Extensions
- Scale Mixture Priors
  - Gaussian Scale Mixture (GSM)
  - Laplacian Scale Mixture (LSM)
  - Power Exponential Scale Mixture (PESM)

Outline

- Sparse Signal Recovery (SSR) Problem and some Extensions
- Scale Mixture Priors
  - Gaussian Scale Mixture (GSM)
  - Laplacian Scale Mixture (LSM)
  - Power Exponential Scale Mixture (PESM)
- Bayesian Methods
  - MAP estimation (Type I)
  - Hierarchical Bayes (Type II)

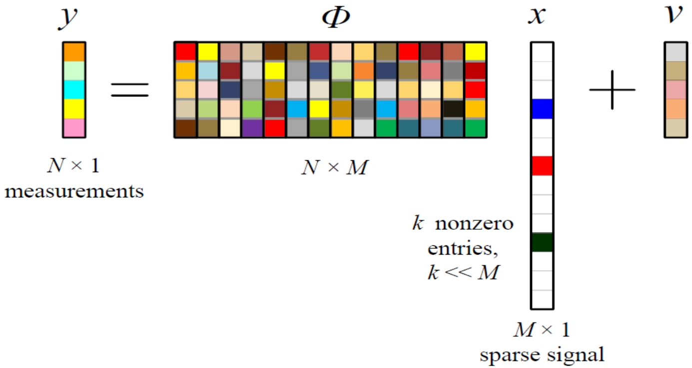
Outline

- Sparse Signal Recovery (SSR) Problem and some Extensions
- Scale Mixture Priors
  - Gaussian Scale Mixture (GSM)
  - Laplacian Scale Mixture (LSM)
  - Power Exponential Scale Mixture (PESM)
- Bayesian Methods
  - MAP estimation (Type I)
  - Hierarchical Bayes (Type II)
- Experimental Results

Outline

- Sparse Signal Recovery (SSR) Problem and some Extensions
- Scale Mixture Priors
  - Gaussian Scale Mixture (GSM)
  - Laplacian Scale Mixture (LSM)
  - Power Exponential Scale Mixture (PESM)
- Bayesian Methods
  - MAP estimation (Type I)
  - Hierarchical Bayes (Type II)
- Experimental Results
- Summary

Problem Description: Sparse Signal Recovery (SSR)



- y is a $N \times 1$ measurement vector.
- Φ is $N \times M$ dictionary matrix where $M >> N$.
- x is $M \times 1$ desired vector which is sparse with k non zero entries.
- v is the measurement noise.

Extensions

Extensions

- Block Sparsity

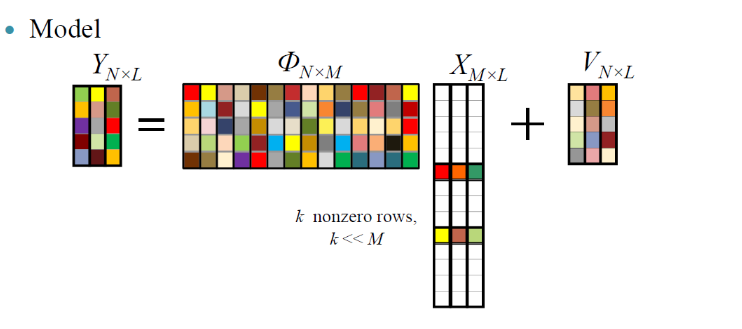
Extensions

- Block Sparsity
- Multiple Measurement Vectors (MMV)

Extensions

- Block Sparsity
- Multiple Measurement Vectors (MMV)
- Block MMV
- MMV with time varying sparsity

Multiple Measurement Vectors (MMV)



- Multiple measurements: L measurements
- Common Sparsity Profile: k nonzero rows

Applications

Applications

- Signal Representation (Mallat, Coifman, Donoho,..)
- EEG/MEG (Leahy, Gorodnitsky,Ioannides,..)
- Robust Linear Regression and Outlier Detection (Jin, Giannakis, ..)
- Speech Coding (Ozawa, Ono, Kroon,..)
- Compressed Sensing (Donoho, Candes, Tao,..)
- Magnetic Resonance Imaging (Lustig,..)
- Sparse Channel Equalization (Fevrier, Proakis,...)
- Face Recognition (Wright, Yang, ...)
- Cognitive Radio (Eldar, ..)

and many more.........

Potential Algorithmic Approaches

Finding the Optimal Solution is NP hard. So need low complexity algorithms with reasonable performance.

Potential Algorithmic Approaches

Finding the Optimal Solution is NP hard. So need low complexity algorithms with reasonable performance.

Greedy Search Techniques

Matching Pursuit (MP), Orthogonal Matching Pursuit (OMP), ...

Potential Algorithmic Approaches

Finding the Optimal Solution is NP hard. So need low complexity algorithms with reasonable performance.

Greedy Search Techniques

Matching Pursuit (MP), Orthogonal Matching Pursuit (OMP), ...

Minimizing Diversity Measures (Regularization Framework)

Tractable Surrogate Cost functions: e.g. ℓ_1 minimization, ...

Potential Algorithmic Approaches

Finding the Optimal Solution is NP hard. So need low complexity algorithms with reasonable performance.

Greedy Search Techniques

Matching Pursuit (MP), Orthogonal Matching Pursuit (OMP), ...

Minimizing Diversity Measures (Regularization Framework)

Tractable Surrogate Cost functions: e.g. ℓ_1 minimization, ...

Bayesian Methods

Make appropriate Statistical assumptions on the solution (sparsity): Choice of Prior

Bayesian Methods: Choice of Prior

- Super Gaussian Distributions: Heavy tailed and sharper peak at origin compared to Gaussian.
- Tractable representations using Scale Mixtures:
  - Gaussian Scale Mixture (GSM)
  - Laplacian Scale Mixture (LSM)
  - Power Exponential Scale Mixture (PESM)

Separability: $p(x) = \prod_i p(x_i)$

Gaussian Scale Mixtures

Separability: $p(x) = \prod_i p(x_i)$

$$p(x_i) = \int p(x_i|\gamma_i)p(\gamma_i)d\gamma_i = \int N(x_i; 0, \gamma_i)p(\gamma_i)d\gamma_i$$

Gaussian Scale Mixtures

Separability: $p(x) = \prod_i p(x_i)$

$$p(x_i) = \int p(x_i|\gamma_i)p(\gamma_i)d\gamma_i = \int N(x_i; 0, \gamma_i)p(\gamma_i)d\gamma_i$$

Theorem

A density $p(x)$ which is symmetric with respect origin, can be represented by a GSM iff $p(\sqrt{x})$ is completely monotonic on $(0, \infty)$.

Gaussian Scale Mixtures

Separability: $p(x) = \prod_i p(x_i)$

$$p(x_i) = \int p(x_i|\gamma_i)p(\gamma_i)d\gamma_i = \int N(x_i; 0, \gamma_i)p(\gamma_i)d\gamma_i$$

Theorem

A density $p(x)$ which is symmetric with respect origin, can be represented by a GSM iff $p(\sqrt{x})$ is completely monotonic on $(0, \infty)$.

Most of the sparse priors over x can be represented in this GSM form. [Palmer et al., 2006]

Examples of Gaussian Scale Mixture

Laplacian density

$$p(x; a) = \frac{a}{2} exp(-a|x|)$$

Scale mixing density: $p(\gamma) = \frac{a^2}{2} \exp(-\frac{a^2}{2}\gamma), \gamma \geq 0.$

Examples of Gaussian Scale Mixture

Laplacian density

$$p(x; a) = \frac{a}{2} exp(-a|x|)$$

Scale mixing density: $p(\gamma) = \frac{a^2}{2} \exp(-\frac{a^2}{2}\gamma), \gamma \geq 0.$

Student-t Distribution

$$p(x; a, b) = \frac{b^a \Gamma(a + 1/2)}{(2\pi)^{0.5}\Gamma(a)} \frac{1}{(b + x^2/2)^{a+1/2}}$$

Scale mixing density: Gamma Distribution.

Examples of Gaussian Scale Mixture

Laplacian density

$$p(x;a) = \frac{a}{2} exp(-a|x|)$$

Scale mixing density: $p(\gamma) = \frac{a^2}{2}\exp(-\frac{a^2}{2}\gamma), \gamma \geq 0.$

Student-t Distribution

$$p(x;a,b) = \frac{b^a\Gamma(a+1/2)}{(2\pi)^{0.5}\Gamma(a)}\frac{1}{(b+x^2/2)^{a+1/2}}$$

Scale mixing density: Gamma Distribution.

Generalized Gaussian

$$p(x;p) = \frac{1}{2\Gamma(1+\frac{1}{p})}e^{-|x|^p}$$

Scale mixing density: Positive alpha stable density of order $p/2$.

Generalized Scale Mixture Family

- GSM corresponds to ℓ_2 norm based SSR algorithm.
- LSM corresponds to ℓ_1 norm based SSR algorithm.
- Need a generalized scale mixture for a unified treatment of ℓ_1 and ℓ_2 minimization based SSR.

Power Exponential Distribution

Also known as Box and Tiao (BT) or Generalized Gaussian distribution (GGD).

$$p_{PE}(x; 0, \sigma, p) = K e^{-\frac{|x|^p}{\sigma^p}}$$

Power Exponential Scale Mixture Distributions (PESM)

Power Exponential Distribution

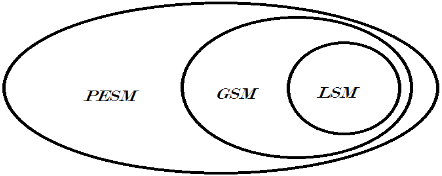
Also known as Box and Tiao (BT) or Generalized Gaussian distribution (GGD).

$$p_{PE}(x; 0, \sigma, p) = Ke^{-\frac{|x|^p}{\sigma^p}}$$

Scale Mixture of Power Exponential :

$$p(x_i) = \int p(x_i|\gamma_i)p(\gamma_i)d\gamma_i = \int p_{PE}(x_i; 0, \gamma_i, p)p(\gamma_i)d\gamma_i$$

Power Exponential Scale Mixture Distributions (PESM)



Choice of p=2

Gaussian Scale Mixtures (GSM): ℓ_2 norm minimization based algorithms.

Choice of p=1

Laplacian Scale Mixtures (LSM): ℓ_1 norm minimization based algorithms.

PESM

Unified treatment of both ℓ_1 and ℓ_2 based algorithms.

PESM Example: Generalized t distribution

PESM Example: Generalized t distribution

Inverse Generalized Gamma (GG) for scaling density:

$$p(\gamma_i) = p_{GG}(\gamma_i; -p, \sigma, q) = \eta(\sigma/\gamma_i)^{pq+1} e^{-(\sigma/\gamma_i)^p}$$

PESM Example: Generalized t distribution

Inverse Generalized Gamma (GG) for scaling density:

$$p(\gamma_i) = p_{GG}(\gamma_i; -p, \sigma, q) = \eta(\sigma/\gamma_i)^{pq+1} e^{-(\sigma/\gamma_i)^p}$$

$$p_{GT}(x; \sigma, p, q) = K(1 + \frac{|x|^p}{q\sigma^p})^{-(q+1/p)}$$

PESM Example: Generalized t distribution

Inverse Generalized Gamma (GG) for scaling density:

$$p(\gamma_i) = p_{GG}(\gamma_i; -p, \sigma, q) = \eta(\sigma/\gamma_i)^{pq+1} e^{-(\sigma/\gamma_i)^p}$$

$$p_{GT}(x; \sigma, p, q) = K(1 + \frac{|x|^p}{q\sigma^p})^{-(q+1/p)}$$

A wide class of heavy tailed super gaussian densities can be represented by GT using suitable shape parameters p and q.

Variants of GT

Table: Variants of Generalized t Distribution

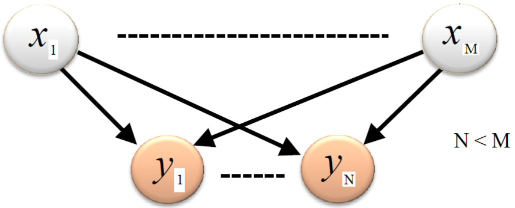
q	p	Distribution
$q \to \infty$	2	Normal
$q \to \infty$	1	Laplacian (Double Exponential)
$q \geq 0$ (degrees of freedom)	2	Student t distribution
$q \geq 0$ (shape parameter)	1	Generalized Double Pareto (GDP)

MAP Estimation (Type I)

MAP Estimation (Type I)

Hierarchical Bayes (Type II)

MAP Estimation Framework (Type I)

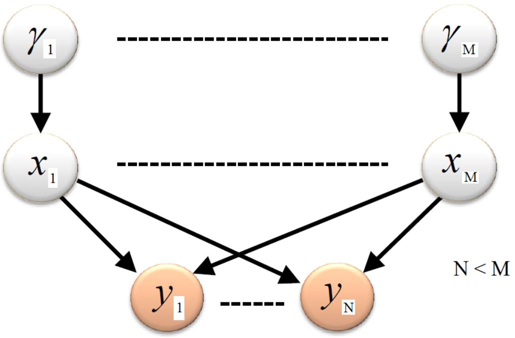


Problem Statement

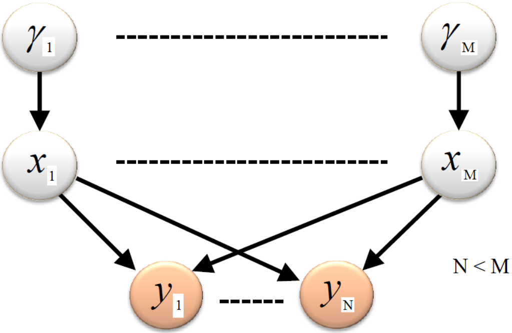
$$\hat{x} = \arg\max_x p(x|y) = \arg\max_x p(y|x)p(x)$$

Choice of $p(x) = \frac{a}{2}e^{-a|x|}$ as Laplacian and Gaussian Likelihood assumption will lead to the familiar LASSO framework.

Hierarchical Bayesian Framework (Type II)



Hierarchical Bayesian Framework (Type II)



Problem Statement

$$\hat{\gamma} = \arg\max_{\gamma} p(\gamma|y) = \arg\max_{\gamma} p(y|\gamma)p(\gamma)$$

Using this estimate of γ we can compute our concerned posterior $p(x|y;\hat{\gamma})$.

Hierarchical Bayesian Framework (Type II)

Potential Advantages

- Averaging over x leads to fewer minima in $p(\gamma|y)$.
- γ can tie several parameters, leading to fewer parameters.
- Maximizing the true posterior mass over the subspaces spanned by non zero indexes instead of looking for the mode.

Hierarchical Bayesian Framework (Type II)

Potential Advantages

- Averaging over x leads to fewer minima in $p(\gamma|y)$.
- γ can tie several parameters, leading to fewer parameters.
- Maximizing the true posterior mass over the subspaces spanned by non zero indexes instead of looking for the mode.

Bayesian LASSO

Laplacian $p(x)$ as GSM[a]:

$$\begin{aligned} p(x) &= \int p(x|\gamma)p(\gamma)d\gamma \\ &= \int \underbrace{\frac{1}{\sqrt{2\pi\gamma}} exp(-\frac{x^2}{2\gamma})}_{p(x|\gamma)} \times \underbrace{\frac{a^2}{2} exp(-\frac{a^2}{2}\gamma)}_{p(\gamma)} d\gamma \\ &= \frac{a}{2} exp(-a|x|) \end{aligned}$$

[a] "Bayesian Compressive Sensing Using Laplace Priors", Babacan et al

MAP Estimation (Type I) Framework

Problem Statement

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y}) = \arg\max_{\mathbf{x}} \log p(\mathbf{y}|\mathbf{x}) + \log p(\mathbf{x})$$

MAP Estimation (Type I) Framework

Problem Statement

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y}) = \arg\max_{\mathbf{x}} \log p(\mathbf{y}|\mathbf{x}) + \log p(\mathbf{x})$$

Examples:

Prior Distribution	Penalty Function	SSR Algorithm
Normal	$\|\|x\|\|_2$	Ridge Regression
Laplacian	$\|\|x\|\|_1$	LASSO
Student t distribution	$\log(\epsilon + x^2)$	Reweighted ℓ_2 (Chartrand's)
Generalized Double Pareto	$\log(\epsilon + \|x\|)$	Reweighted ℓ_1(Candes's)

MAP Estimation (Type I) Framework

Problem Statement

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y}) = \arg\max_{\mathbf{x}} \log p(\mathbf{y}|\mathbf{x}) + \log p(\mathbf{x})$$

Examples:

Prior Distribution	Penalty Function	SSR Algorithm
Normal	$\|\|x\|\|_2$	Ridge Regression
Laplacian	$\|\|x\|\|_1$	LASSO
Student t distribution	$\log(\epsilon + x^2)$	Reweighted ℓ_2 (Chartrand's)
Generalized Double Pareto	$\log(\epsilon + \|x\|)$	Reweighted ℓ_1(Candes's)

PESM as sparsity promoting prior $p(x)$: Unified Type I Framework

Unified Type I Framework

Choice of Prior: p(x)

Any distribution in PESM class.

Unified Type I Framework

Choice of Prior: p(x)

Any distribution in PESM class.

EM Algorithm

- Complete Data Log-Likelihood:

$$\log p(\mathbf{y}, \mathbf{x}, \gamma) = \log p(\mathbf{y}|\mathbf{x}) + \log p(\mathbf{x}|\gamma) + \log p(\gamma)$$

- Hidden Variable: γ
- Concerned Posterior: $p(\gamma|\mathbf{x}, \mathbf{y}) \sim p(\gamma|\mathbf{x})$ (From Markov chain).

Unified Type I: E step

$$Q(\mathbf{x}) = \mathbb{E}_{\gamma|\mathbf{x}}\left[\log p(\mathbf{y}|\mathbf{x}) + \log p(\mathbf{x}|\gamma) + \log p(\gamma)\right]$$

E Step

- Only second term has dependencies on both **x** and γ.
- Compute $E_{\gamma_i|x_i}\left[\frac{1}{\gamma_i^p}\right]$

Unified Type I: E step

$$
\begin{aligned}
p'(x_i) &= \frac{d}{dx_i}\int_0^\infty p(x_i|\gamma_i)p(\gamma_i)d\gamma_i \\
&= -p \times |x_i|^{p-1}\mathrm{sign}(x_i)p(x_i)\int_0^\infty \frac{1}{\gamma_i^p}p(\gamma_i|x_i)d\gamma_i \\
&= -p \times |x_i|^{p-1}\mathrm{sign}(x_i)p(x_i)E_{\gamma_i|x_i}[\frac{1}{\gamma_i^p}]
\end{aligned}
$$

E step

$$
E_{\gamma_i|x_i}[\frac{1}{\gamma_i^p}] = -\frac{p'(x_i)}{p \times |x_i|^{p-1}\mathrm{sign}(x_i)p(x_i)}
$$

Unified Type I: E step

$$p'(x_i) = \frac{d}{dx_i}\int_0^\infty p(x_i|\gamma_i)p(\gamma_i)d\gamma_i$$
$$= -p \times |x_i|^{p-1}\text{sign}(x_i)p(x_i)\int_0^\infty \frac{1}{\gamma_i^p}p(\gamma_i|x_i)d\gamma_i$$
$$= -p \times |x_i|^{p-1}\text{sign}(x_i)p(x_i)E_{\gamma_i|x_i}[\frac{1}{\gamma_i^p}]$$

E step

$$E_{\gamma_i|x_i}[\frac{1}{\gamma_i^p}] = -\frac{p'(x_i)}{p \times |x_i|^{p-1}\text{sign}(x_i)p(x_i)}$$

Note: No need to know $p(\gamma)$, as long as $p(x)$ is known and has a PESM representation.

Unified Type I: M step

M step

$$\hat{\mathbf{x}}^{(k+1)} = \arg\min_{\mathbf{x}} \frac{1}{2\lambda}||\mathbf{y} - \Phi\mathbf{x}||^2 + \sum_i w_i^{(k)}|x_i|^p$$

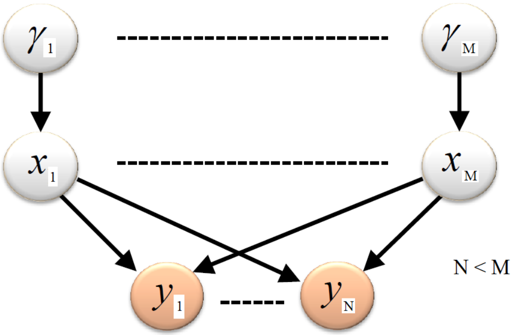
Where,

$$w_i^{(k)} = E_{\gamma_i|x_i^{(k)}}\left[\frac{1}{\gamma_i^p}\right]$$

Special Case: Generalized t distribution

$$w_i^{(k)} = \frac{q + 1/p}{q\sigma^p + |x_i^{(k)}|^p}$$

Hierarchical Bayesian Framework (Type II)



Estimate of the posterior distribution for x using estimated $\hat{\gamma}$; i.e.$p(x|y;\hat{\gamma})$.

Choice of GSM as $p(x)$ leads to Sparse Bayesian Learning

Sparse Bayesian Learning (Type II)

$$\mathbf{y} = \Phi\mathbf{x} + \mathbf{v}$$

Sparse Bayesian Learning (Type II)

$$\mathbf{y} = \Phi\mathbf{x} + \mathbf{v}$$

Solving for MAP estimate of γ

$$\hat{\gamma} = \arg\max_{\gamma} p(\gamma|y) = \arg\max_{\gamma} p(y|\gamma)p(\gamma)$$

$$\mathbf{y} = \Phi\mathbf{x} + \mathbf{v}$$

Solving for MAP estimate of γ

$$\hat{\gamma} = \arg\max_{\gamma} p(\gamma|y) = \arg\max_{\gamma} p(y|\gamma)p(\gamma)$$

What is $p(y|\gamma)$

Sparse Bayesian Learning (Type II)

$$\mathbf{y} = \Phi\mathbf{x} + \mathbf{v}$$

Solving for MAP estimate of γ

$$\hat{\gamma} = \arg\max_{\gamma} p(\gamma|y) = \arg\max_{\gamma} p(y|\gamma)p(\gamma)$$

What is $p(y|\gamma)$

Given γ, $\mathbf{x}$ is Gaussian with mean zero and Covariance matrix Γ with $\Gamma = \text{diag}(\gamma)$, i.e. $p(x|\gamma) = N(x; 0, \Gamma) = \Pi N(x_i; 0, \gamma_i)$.

Sparse Bayesian Learning (Type II)

$$\mathbf{y} = \Phi\mathbf{x} + \mathbf{v}$$

Solving for MAP estimate of γ

$$\hat{\gamma} = \arg\max_{\gamma} p(\gamma|y) = \arg\max_{\gamma} p(y|\gamma)p(\gamma)$$

What is $p(y|\gamma)$

Given γ, $\mathbf{x}$ is Gaussian with mean zero and Covariance matrix Γ with $\Gamma = \text{diag}(\gamma)$, i.e. $p(x|\gamma) = N(x; 0, \Gamma) = \Pi N(x_i; 0, \gamma_i)$.

Then $p(y|\gamma) = N(y; 0, \Sigma_y)$, where $\Sigma_y = \sigma^2 I + \Phi\Gamma\Phi^T$,

$$p(y|\gamma) = \frac{1}{\sqrt{(2\pi)^N|\Sigma_y|}} e^{-\frac{1}{2}y^T\Sigma_y^{-1}y}$$

Sparse Bayesian Learning (Tipping)

$$y = \Phi x + v$$

Solving for the optimal γ

$$\hat{\gamma} = \arg\max_{\gamma} p(\gamma|y) = \arg\max_{\gamma} p(y|\gamma)p(\gamma)$$

$$= \arg\min_{\gamma} log|\Sigma_y| + y^T \Sigma_y^{-1} y - 2\sum_i logp(\gamma_i)$$

where, $\Sigma_y = \sigma^2 I + \Phi \Gamma \Phi^T$ and $\Gamma = diag(\gamma)$

Sparse Bayesian Learning (Tipping)

$$y = \Phi x + v$$

Solving for the optimal γ

$$\hat{\gamma} = \arg\max_{\gamma} p(\gamma|y) = \arg\max_{\gamma} p(y|\gamma)p(\gamma)$$
$$= \arg\min_{\gamma} log|\Sigma_y| + y^T\Sigma_y^{-1}y - 2\sum_i logp(\gamma_i)$$

where, $\Sigma_y = \sigma^2 I + \Phi\Gamma\Phi^T$ and $\Gamma = diag(\gamma)$

Computational Methods

Many options for solving the above optimization problem, e.g. Majorization Minimization, Expectation-Maximization (EM).

$$\mathbf{y} = \Phi\mathbf{x} + \mathbf{v}$$

Computing Posterior

Now because of our convenient GSM choice, posterior can be easily computed, i.e, $p(x|y;\hat{\gamma}) = N(\mu_x, \Sigma_x)$ where,

$$\mu_x = E[x|y;\hat{\gamma}] = \hat{\Gamma}\Phi^T(\sigma^2 I + \Phi\hat{\Gamma}\Phi^T)^{-1}\mathbf{y}$$

$$\Sigma_x = Cov[x|y;\hat{\gamma}] = \hat{\Gamma} - \hat{\Gamma}\Phi^T(\sigma^2 I + \Phi\hat{\Gamma}\Phi^T)^{-1}\Phi\hat{\Gamma}$$

Sparse Bayesian Learning

$$\mathbf{y} = \Phi\mathbf{x} + \mathbf{v}$$

Computing Posterior

Now because of our convenient GSM choice, posterior can be easily computed, i.e, $p(x|y;\hat{\gamma}) = N(\mu_x, \Sigma_x)$ where,

$$\mu_x = E[x|y;\hat{\gamma}] = \hat{\Gamma}\Phi^T(\sigma^2 I + \Phi\hat{\Gamma}\Phi^T)^{-1}\mathbf{y}$$

$$\Sigma_x = Cov[x|y;\hat{\gamma}] = \hat{\Gamma} - \hat{\Gamma}\Phi^T(\sigma^2 I + \Phi\hat{\Gamma}\Phi^T)^{-1}\Phi\hat{\Gamma}$$

μ_x can be used as a point estimate.

$$\mathbf{y} = \Phi\mathbf{x} + \mathbf{v}$$

Computing Posterior

Now because of our convenient GSM choice, posterior can be easily computed, i.e, $p(x|y;\hat{\gamma}) = N(\mu_x, \Sigma_x)$ where,

$$\mu_x = E[x|y;\hat{\gamma}] = \hat{\Gamma}\Phi^T(\sigma^2 I + \Phi\hat{\Gamma}\Phi^T)^{-1}\mathbf{y}$$

$$\Sigma_x = Cov[x|y;\hat{\gamma}] = \hat{\Gamma} - \hat{\Gamma}\Phi^T(\sigma^2 I + \Phi\hat{\Gamma}\Phi^T)^{-1}\Phi\hat{\Gamma}$$

μ_x can be used as a point estimate.

Sparsity of μ_x is achieved through sparsity in γ.

$$\mathbf{y} = \Phi\mathbf{x} + \mathbf{v}$$

Computing Posterior

Now because of our convenient GSM choice, posterior can be easily computed, i.e, $p(x|y;\hat{\gamma}) = N(\mu_x, \Sigma_x)$ where,

$$\mu_x = E[x|y;\hat{\gamma}] = \hat{\Gamma}\Phi^T(\sigma^2 I + \Phi\hat{\Gamma}\Phi^T)^{-1}\mathbf{y}$$

$$\Sigma_x = Cov[x|y;\hat{\gamma}] = \hat{\Gamma} - \hat{\Gamma}\Phi^T(\sigma^2 I + \Phi\hat{\Gamma}\Phi^T)^{-1}\Phi\hat{\Gamma}$$

μ_x can be used as a point estimate.

Sparsity of μ_x is achieved through sparsity in γ.

Another parameter of interest for the EM algorithm

$$E(x_i^2|\mathbf{y},\hat{\gamma}) = \mu_x^2(i) + \Sigma_x(i,i)$$

EM algorithm: Updating γ

EM algorithm: Updating γ

Treating $(\mathbf{y}, \mathbf{x})$ as complete data and vector $\mathbf{x}$ as hidden variable.

$$\log p(y, x, \gamma) = \log p(y|x) + \log p(x|\gamma) + \log p(\gamma)$$

EM algorithm: Updating γ

Treating $(\mathbf{y}, \mathbf{x})$ as complete data and vector $\mathbf{x}$ as hidden variable.

$$\log p(y, x, \gamma) = \log p(y|x) + \log p(x|\gamma) + \log p(\gamma)$$

E step

$$Q(\gamma|\gamma^k) = \mathbb{E}_{x|y;\gamma^k}[\log p(y|x) + \log p(x|\gamma) + \log p(\gamma)]$$

EM algorithm: Updating γ

Treating $(\mathbf{y}, \mathbf{x})$ as complete data and vector $\mathbf{x}$ as hidden variable.

$$\log p(y, x, \gamma) = \log p(y|x) + \log p(x|\gamma) + \log p(\gamma)$$

E step

$$Q(\gamma|\gamma^k) = \mathbb{E}_{x|y;\gamma^k}[\log p(y|x) + \log p(x|\gamma) + \log p(\gamma)]$$

M step

$$\begin{aligned}\gamma^{k+1} &= \text{argmax}_\gamma Q(\gamma|\gamma^k) = \text{argmax}_\gamma \mathbb{E}_{x|y;\gamma^k}[\log p(x|\gamma) + \log p(\gamma)] \\ &= \text{argmin}_\gamma \mathbb{E}_{x|y;\gamma^k} \sum_{i=1}^{M} \left[\left(\frac{x_i^2}{2\gamma_i} + \frac{1}{2} \log \gamma_i \right) - \log p(\gamma_i) \right]\end{aligned}$$

EM algorithm: Updating γ

Treating $(\mathbf{y}, \mathbf{x})$ as complete data and vector $\mathbf{x}$ as hidden variable.

$$\log p(y, x, \gamma) = \log p(y|x) + \log p(x|\gamma) + \log p(\gamma)$$

E step

$$Q(\gamma|\gamma^k) = \mathbb{E}_{x|y;\gamma^k}[\log p(y|x) + \log p(x|\gamma) + \log p(\gamma)]$$

M step

$$\gamma^{k+1} = \text{argmax}_\gamma Q(\gamma|\gamma^k) = \text{argmax}_\gamma \mathbb{E}_{x|y;\gamma^k}[\log p(x|\gamma) + \log p(\gamma)]$$

$$= \text{argmin}_\gamma \mathbb{E}_{x|y;\gamma^k} \sum_{i=1}^{M} \left[\left(\frac{x_i^2}{2\gamma_i} + \frac{1}{2} \log \gamma_i \right) - \log p(\gamma_i) \right]$$

Solving this optimization problem with a non-informative prior $p(\gamma)$,

$$\gamma_i^{k+1} = E(x_i^2|\mathbf{y}, \gamma^k) = \mu_x(i)^2 + \Sigma_x(i,i)$$

Type II (SBL) properties

- Local minima are sparse, i.e. have at most N nonzero γ_i

Type II (SBL) properties

- Local minima are sparse, i.e. have at most N nonzero γ_i
- Cost function $p(\gamma|y)$ is generally much smoother than the associated MAP estimation objective $p(x|y)$. Fewer local minima.

Type II (SBL) properties

- Local minima are sparse, i.e. have at most N nonzero γ_i
- Cost function $p(\gamma|y)$ is generally much smoother than the associated MAP estimation objective $p(x|y)$. Fewer local minima.
- In high signal to noise ratio, the global minima is the sparsest solution. No structural problems.

Type II (SBL) properties

- Local minima are sparse, i.e. have at most N nonzero γ_i
- Cost function $p(\gamma|y)$ is generally much smoother than the associated MAP estimation objective $p(x|y)$. Fewer local minima.
- In high signal to noise ratio, the global minima is the sparsest solution. No structural problems.
- Attempts to approximate the posterior distribution $p(x|y)$ in the area with significant mass.

Algorithmic Variants

Algorithmic Variants

- Fixed Point iteration based on setting the derivative of the objective function to zero (Tipping)

Algorithmic Variants

- Fixed Point iteration based on setting the derivative of the objective function to zero (Tipping)
- Sequential search for the significant γ's (Tipping and Faul)

Algorithmic Variants

- Fixed Point iteration based on setting the derivative of the objective function to zero (Tipping)
- Sequential search for the significant γ's (Tipping and Faul)
- Majorization-Minimization based approach (Wipf and Nagarajan)

Algorithmic Variants

- Fixed Point iteration based on setting the derivative of the objective function to zero (Tipping)
- Sequential search for the significant γ's (Tipping and Faul)
- Majorization-Minimization based approach (Wipf and Nagarajan)
- Reweighted ℓ_1 and ℓ_2 algorithms (Wipf and Nagarajan)

Algorithmic Variants

- Fixed Point iteration based on setting the derivative of the objective function to zero (Tipping)
- Sequential search for the significant γ's (Tipping and Faul)
- Majorization-Minimization based approach (Wipf and Nagarajan)
- Reweighted ℓ_1 and ℓ_2 algorithms (Wipf and Nagarajan)
- Approximate Message Passing (AlShoukairi and Rao)

Type II using PESM

- In E step we need to compute the conditional expectation.
- Closed form may not be available depending on the choice of p (distributional parameter of PESM).
- Alternative: MCMC technique.

Type II using PESM

- In E step we need to compute the conditional expectation.
- Closed form may not be available depending on the choice of p (distributional parameter of PESM).
- Alternative: MCMC technique.

LSM

Using the fact that a Laplacian density has a GSM representation, a tractable 3 layer hierarchical model can be developed.

Simulation Results

Parameters

1. N = 50, M = 250.
2. Dictionary Elements: Normal Distribution with mean = 0 and standard deviation = 1.
3. Distribution of non zero elements
   - (I) Zero mean unit variance Gaussian.
   - (II) Student t distribution with degrees of freedom $\nu = 3$. (Super-Gaussian)
   - (III) Uniform ± 1 random spikes.

Simulation Results: Gaussian

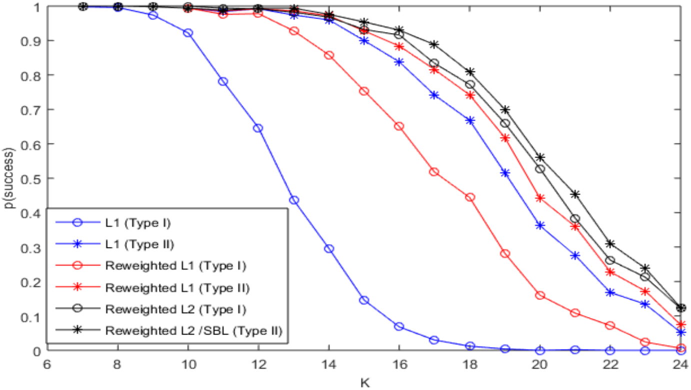


Figure: Recovery performance with Gaussian distributed non zero coefficients

Simulation Results: Super Gaussian

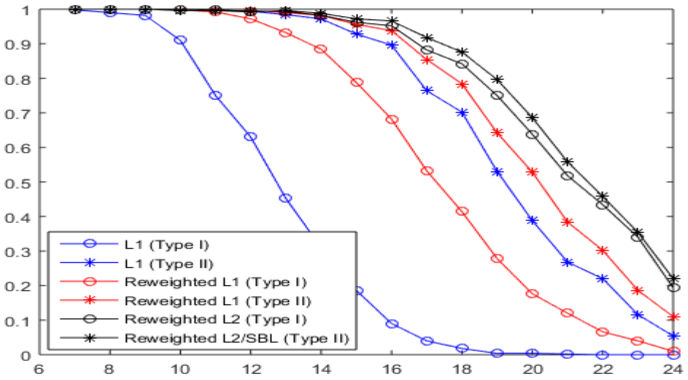


Figure: Recovery performance with Super Gaussian (Student t) distributed non zero coefficients

Simulation Results: Uniform

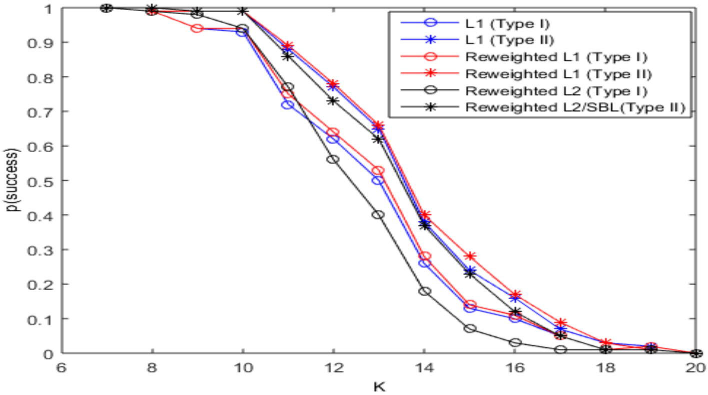
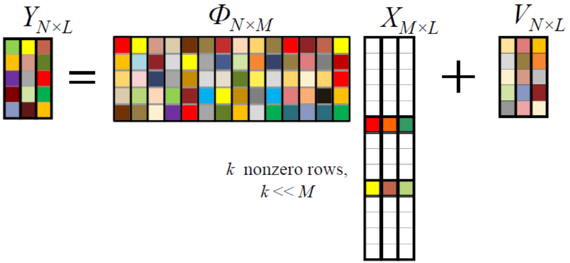


Figure: Recovery performance with uniform spikes as non zero coefficients

Special Case: MMV

- Model



- Multiple measurements: L measurements
- Common Sparsity Profile: k nonzero rows

Representation for Random Vectors (Rows for MMV)

$$\mathbf{X} = \gamma \mathbf{G} \text{ where, } \mathbf{G} \sim N(g; 0, \mathbf{B})$$

γ is a positive random variable, which is independent of $\mathbf{G}$.

Representation for Random Vectors (Rows for MMV)

$$\mathbf{X} = \gamma \mathbf{G} \text{ where, } \mathbf{G} \sim N(g; 0, \mathbf{B})$$

γ is a positive random variable, which is independent of $\mathbf{G}$.

$$p(\mathbf{x}) = \int p(\mathbf{x}|\gamma)p(\gamma)d\gamma = \int N(\mathbf{x}; 0, \gamma\mathbf{B})p(\gamma)d\gamma$$

Representation for Random Vectors (Rows for MMV)

$$\mathbf{X} = \gamma \mathbf{G} \text{ where, } \mathbf{G} \sim N(g; 0, \mathbf{B})$$

γ is a positive random variable, which is independent of $\mathbf{G}$.

$$p(\mathbf{x}) = \int p(\mathbf{x}|\gamma)p(\gamma)d\gamma = \int N(\mathbf{x}; 0, \gamma\mathbf{B})p(\gamma)d\gamma$$

- $\mathbf{B} = \mathbf{I}$ if the row entries are assumed independent.

Representation for Random Vectors (Rows for MMV)

$$\mathbf{X} = \gamma \mathbf{G} \text{ where, } \mathbf{G} \sim N(g; 0, \mathbf{B})$$

γ is a positive random variable, which is independent of $\mathbf{G}$.

$$p(\mathbf{x}) = \int p(\mathbf{x}|\gamma)p(\gamma)d\gamma = \int N(\mathbf{x}; 0, \gamma\mathbf{B})p(\gamma)d\gamma$$

- $\mathbf{B} = \mathbf{I}$ if the row entries are assumed independent.
- One γ per row vector. Complexity of estimating γ does not grow with L.

Bayesian Methods: GSM Extension

Representation for Random Vectors (Rows for MMV)

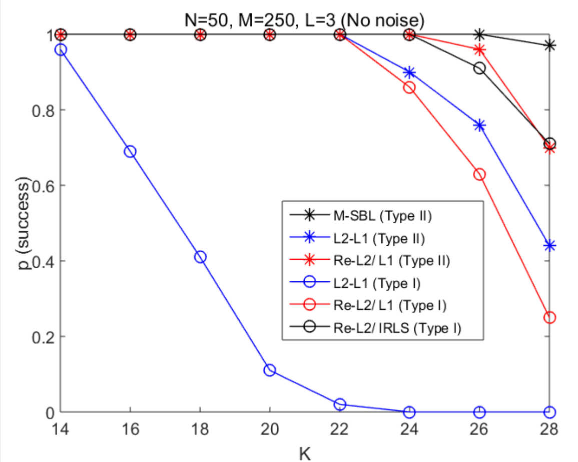
$$\mathbf{X} = \gamma \mathbf{G} \text{ where, } \mathbf{G} \sim N(g; 0, \mathbf{B})$$

γ is a positive random variable, which is independent of $\mathbf{G}$.

$$p(\mathbf{x}) = \int p(\mathbf{x}|\gamma)p(\gamma)d\gamma = \int N(\mathbf{x}; 0, \gamma\mathbf{B})p(\gamma)d\gamma$$

- $\mathbf{B} = \mathbf{I}$ if the row entries are assumed independent.
- One γ per row vector. Complexity of estimating γ does not grow with L.
- The EM algorithm is also very tractable.

MMV Empirical Comparison: 1000 trials



Summary

Summary

- Sparse Signal Recovery (SSR) and Compressed Sensing (CS) are interesting new signal processing tools with many potential applications.

Summary

- Sparse Signal Recovery (SSR) and Compressed Sensing (CS) are interesting new signal processing tools with many potential applications.
- Many algorithmic options exist to solve the underlying sparse signal recovery problem; Greedy Search Techniques, regularization methods, Bayesian methods, among others.

Summary

- Sparse Signal Recovery (SSR) and Compressed Sensing (CS) are interesting new signal processing tools with many potential applications.
- Many algorithmic options exist to solve the underlying sparse signal recovery problem; Greedy Search Techniques, regularization methods, Bayesian methods, among others.
- Bayesian methods offer interesting algorithmic options to the Sparse Signal Recovery problem

Summary

- Sparse Signal Recovery (SSR) and Compressed Sensing (CS) are interesting new signal processing tools with many potential applications.
- Many algorithmic options exist to solve the underlying sparse signal recovery problem; Greedy Search Techniques, regularization methods, Bayesian methods, among others.
- Bayesian methods offer interesting algorithmic options to the Sparse Signal Recovery problem
  - MAP methods (reweighted ℓ_1 and ℓ_2 methods)

Summary

- Sparse Signal Recovery (SSR) and Compressed Sensing (CS) are interesting new signal processing tools with many potential applications.
- Many algorithmic options exist to solve the underlying sparse signal recovery problem; Greedy Search Techniques, regularization methods, Bayesian methods, among others.
- Bayesian methods offer interesting algorithmic options to the Sparse Signal Recovery problem
  - MAP methods (reweighted ℓ_1 and ℓ_2 methods)
  - Hierarchical Bayesian Methods (Sparse Bayesian Learning)

Summary

- Sparse Signal Recovery (SSR) and Compressed Sensing (CS) are interesting new signal processing tools with many potential applications.
- Many algorithmic options exist to solve the underlying sparse signal recovery problem; Greedy Search Techniques, regularization methods, Bayesian methods, among others.
- Bayesian methods offer interesting algorithmic options to the Sparse Signal Recovery problem
  - MAP methods (reweighted ℓ_1 and ℓ_2 methods)
  - Hierarchical Bayesian Methods (Sparse Bayesian Learning)
  - Versatile and can be more easily employed in problems with structure

Summary

- Sparse Signal Recovery (SSR) and Compressed Sensing (CS) are interesting new signal processing tools with many potential applications.
- Many algorithmic options exist to solve the underlying sparse signal recovery problem; Greedy Search Techniques, regularization methods, Bayesian methods, among others.
- Bayesian methods offer interesting algorithmic options to the Sparse Signal Recovery problem
  - MAP methods (reweighted ℓ_1 and ℓ_2 methods)
  - Hierarchical Bayesian Methods (Sparse Bayesian Learning)
  - Versatile and can be more easily employed in problems with structure
  - Algorithms can often be justified by studying the resulting objective functions.