Bayesian Methods for Sparse Signal Recovery

Bhaskar D Rao
University of California, San Diego

Thanks to David Wipf, Jason Palmer, Zhilin Zhang and Ritwik Giri
Motivation

Sparse Signal Recovery is an interesting area with many potential applications. Methods developed for solving sparse signal recovery problems can be a valuable tool for signal processing practitioners. Many interesting developments in recent past that make the subject timely.

Bayesian Framework offers some interesting options.
Motivation

Sparse Signal Recovery is an interesting area with many potential applications.
Motivation

Sparse Signal Recovery is an interesting area with many potential applications.

Methods developed for solving sparse signal recovery problem can be a valuable tool for signal processing practitioners.

Bhaskar D Rao  University of California, San Diego
Motivation

Sparse Signal Recovery is an interesting area with many potential applications.

Methods developed for solving sparse signal recovery problem can be a valuable tool for signal processing practitioners.

Many interesting developments in recent past that make the subject timely.
Motivation

Sparse Signal Recovery is an interesting area with many potential applications.

Methods developed for solving sparse signal recovery problem can be a valuable tool for signal processing practitioners.

Many interesting developments in recent past that make the subject timely.

Bayesian Framework offers some interesting options.
Outline
Outline

- Sparse Signal Recovery (SSR) Problem and some Extensions
Outline

- Sparse Signal Recovery (SSR) Problem and some Extensions
- Applications
Outline

- Sparse Signal Recovery (SSR) Problem and some Extensions
- Applications
- Bayesian Methods
  - MAP estimation
  - Empirical Bayes
Outline

- Sparse Signal Recovery (SSR) Problem and some Extensions
- Applications
- Bayesian Methods
  - MAP estimation
  - Empirical Bayes
- Summary
Problem Description: Sparse Signal Recovery (SSR)

\[ y = \Phi x + v \]

- \( y \) is a \( N \times 1 \) measurement vector.
- \( \Phi \) is \( N \times M \) dictionary matrix where \( M \gg N \).
- \( x \) is \( M \times 1 \) desired vector which is sparse with \( k \) non zero entries.
- \( v \) is the measurement noise.
Problem Statement: SSR

Noise Free Case

Given a target signal $y$ and dictionary $\mathcal{D}$, find the weights $x$ that solve,

$$\min_x \sum_i I(x_i \neq 0) \text{ subject to } y = \mathcal{D} x$$

$I(.)$ is the indicator function.

Noisy case

Given a target signal $y$ and dictionary $\mathcal{D}$, find the weights $x$ that solve,

$$\min_x \sum_i I(x_i \neq 0) \text{ subject to } k y \preceq x_k^2$$
Problem Statement: SSR

Noise Free Case

Given a target signal $y$ and dictionary $\Phi$, find the weights $x$ that solve,

$$\min_x \sum_i I(x_i \neq 0) \text{ subject to } y = \Phi x$$

$I(.)$ is the indicator function.
Problem Statement: SSR

**Noise Free Case**

Given a target signal $y$ and dictionary $\Phi$, find the weights $x$ that solve,

$$
\min_x \sum_i I(x_i \neq 0) \text{ subject to } y = \Phi x
$$

$I(.)$ is the indicator function.

**Noisy case**

Given a target signal $y$ and dictionary $\Phi$, find the weights $x$ that solve,

$$
\min_x \sum_i I(x_i \neq 0) \text{ subject to } \|y - \Phi x\|_2 < \beta
$$
Useful Extensions

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

- Block Sparsity
Useful Extensions

- Block Sparsity
- Multiple Measurement Vectors (MMV)
Useful Extensions

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

\[ y = \Phi_{N \times M} x + v \]

Variations include equal blocks, unequal blocks, block boundary known or unknown.
Multiple Measurement Vectors (MMV)

- Model

\[
Y_{N \times L} = \Phi_{N \times M} \Phi_{M \times L}^T + V_{N \times L}
\]

- Multiple measurements: \( L \) measurements
- Common Sparsity Profile: \( k \) nonzero rows
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........
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........
Forward model dictionary \( \Phi \) can be computed using Maxwell’s equations [Sarvas, 1987].

In many situations the active brain regions may be relatively sparse, and so solving a sparse inverse problem is required.

[Baillet et al., 2001]
Sparse Channel Estimation

Potential Application: Underwater Acoustics
Speech Modeling and Deconvolution

Speech specific assumptions: Voiced excitation is block sparse and the filter is an all pole filter $\frac{1}{A(z)}$
Compressive Sampling (CS)
Compressive Sampling (CS)

\[ \Psi \mathbf{x} = \mathbf{b} \]

\( \Psi \) is the transform and \( \mathbf{b} \) is the original data/image.
Compressive Sampling (CS)

Computation:
Solve for $x$ such that $x = y$.

Reconstruction:
$b = x$.

Issues:
Need to recover sparse signal $x$ with constraint $x = y$.

Need to design sampling matrix $A$. 

Bhaskar D Rao  University of California, San Diego
Compressive Sampling (CS)

Computation:

- Solve for $x$ such that $\Phi x = y$.
- Reconstruction: $b = \Psi x$
Compressive Sampling (CS)

Computation:
- Solve for $x$ such that $\Phi x = y$.
- Reconstruction: $b = \Psi x$

Issues:
- Need to recover sparse signal $x$ with constraint $\Phi x = y$.
- Need to design sampling matrix $A$. 
Robust Linear Regression

\[
X, y: \text{data; } \\
c: \text{regression coeffs.; } \\
n: \text{model noise; }
\]

\[
Y = Xc + \Phi w + \varepsilon, \text{ where } \Phi = I,
\]

or

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

Indicator function is not continuous. Define Surrogate Cost functions that are more tractable and whose minimization leads to sparse solutions, e.g. \( \ell_1 \) minimization.

**Bayesian Methods**

Make appropriate Statistical assumptions on the solution and apply estimation techniques to identify the desired sparse solution.
Bayesian Methods

1. MAP Estimation Framework (Type I)

2. Hierarchical Bayesian Framework (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 \( P(y|x) \) as Gaussian will lead to the familiar LASSO framework.

Bhaskar D Rao  University of California, San Diego
Hierarchical Bayesian Framework (Type II)
Hierarchical Bayesian Framework (Type II)

Problem Statement

$$\hat{\gamma} = \arg \max_{\gamma} P(\gamma | y) = \arg \max_{\gamma} \int P(y | x) P(x | \gamma) P(\gamma) dx$$

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

Example: Bayesian LASSO

Laplacian prior $P(x)$ can be represented as a Gaussian Scale Mixture in this fashion,

$$P(x) = \int P(x | \gamma) P(\gamma) d\gamma = \int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \times \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{a^2}{2}\right) d\gamma = a^2 \exp\left(-\frac{a |x|}{x}\right)$$
Hierarchical Bayesian Framework (Type II)

**Problem Statement**

\[ \hat{\gamma} = \arg \max_{\gamma} P(\gamma | y) = \arg \max_{\gamma} \int P(y|x)P(x|\gamma)P(\gamma)dx \]

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

**Example: Bayesian LASSO**

Laplacian prior \( P(x) \) can be represented as a Gaussian Scale Mixture in this fashion,

\[
P(x) = \int P(x|\gamma)P(\gamma)d\gamma \\
= \int \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{x^2}{2\gamma}\right) \times \frac{a^2}{2} \exp\left(-\frac{a^2}{2\gamma}\right)d\gamma \\
= \frac{a}{2} \exp(-a|x|)
\]
MAP Estimation

Problem Statement

\[ \hat{x} = \arg \max_x P(x|y) = \arg \max_x P(y|x)P(x) \]

Advantages

- Many options to promote sparsity, i.e. choose some sparse prior over \( x \).
- Growing options for solving the underlying optimization problem.
- Can be related to LASSO and other \( \ell_1 \) minimization techniques by using suitable \( P(x) \).
Assumption: Gaussian Noise

\[
\hat{x} = \arg \max_x P(y|x)P(x) \\
= \arg \min_x -\log P(y|x) - \log P(x) \\
= \arg \min_x \|y - \Phi x\|_2^2 + \lambda \sum_{i=1}^{m} g(|x_i|)
\]
MAP Estimation

Assumption: Gaussian Noise

\[ \hat{x} = \arg \max_x P(y|x)P(x) \]
\[ = \arg \min_x -\log P(y|x) - \log P(x) \]
\[ = \arg \min_x \|y - \Phi x\|_2^2 + \lambda \sum_{i=1}^{m} g(|x_i|) \]

Theorem

If \( g \) is non decreasing and strictly concave function for \( x \in \mathbb{R}^+ \), the local minima of the above optimization problem will be the extreme points, i.e. have max of \( N \) non-zero entries.
Special cases of MAP estimation

**Gaussian Prior**

Gaussian assumption of $P(x)$ leads to $\ell_2$ norm regularized problem

$$\hat{x} = \arg \min_x \| y - \Phi x \|^2_2 + \lambda \| x \|^2_2$$
Special cases of MAP estimation

**Gaussian Prior**

Gaussian assumption of $P(x)$ leads to $\ell_2$ norm regularized problem

$$\hat{x} = \arg\min_x \|y - \Phi x\|_2^2 + \lambda \|x\|_2^2$$

**Laplacian Prior**

Laplacian assumption of $P(x)$ leads to standard $\ell_1$ norm regularized problem i.e. LASSO.

$$\hat{x} = \arg\min_x \|y - \Phi x\|_2^2 + \lambda \|x\|_1$$
Examples of Sparse Distributions

Sparse distributions can be viewed using a general framework of supergaussian distribution.

\[ P(x; \beta, p) = \frac{p}{2^{\beta/2} \Gamma(\frac{1}{p})} e^{-\frac{|x|^p}{2 \beta}} , \quad p \leq 1 \]

If a unit variance distribution is desired, \( \beta \) becomes a function of \( p \).
Example of Sparsity Penalties

Practical Selections

\[ g(x_i) = \log(x_i^2 + \epsilon), \quad \text{[Chartrand and Yin, 2008]} \]
\[ g(x_i) = \log(|x_i| + \epsilon), \quad \text{[Candes et al., 2008]} \]
\[ g(x_i) = |x_i|^p, \quad \text{[Rao et al., 1999]} \]

Different choices favor different levels of sparsity.
Which Sparse prior to choose?

\[ \hat{x} = \arg \min_x \|y - \Phi x\|_2^2 + \lambda \sum_{i=1}^M |x_i|^p \]

Two issues:

- If the prior is too sparse, i.e. \( p \sim 0 \), then we may get stuck at a local minima which results in convergence error.

- If the prior is not sparse enough, i.e. \( p \sim 1 \), then though global minima can be found, it may not be the sparsest solution, which results in a structural error.
MAP Estimation

Underlying Optimization problem is

\[ \hat{x} = \arg \min_x \| y - \Phi x \|_2^2 + \lambda \sum_{i=1}^{m} g(|x_i|) \]
MAP Estimation

Underlying Optimization problem is

\[ \hat{x} = \arg \min_x \|y - \Phi x\|_2^2 + \lambda \sum_{i=1}^{m} g(|x_i|) \]

- Useful algorithms exist to minimize the cost function with a strictly concave penalty function \( g \) on \( R^+ \) (Reweighted \( \ell_2/\ell_1 \) algorithms).
- The essence of this algorithm is to create a bound for the concave penalty function and follow the steps of a Majorize-Minimization (MM) algorithm.
MAP Estimation

Underlying Optimization problem is

\[ \hat{x} = \arg \min_x \| y - \Phi x \|_2^2 + \lambda \sum_{i=1}^{m} g(|x_i|) \]

- Useful algorithms exist to minimize the cost function with a strictly concave penalty function \( g \) on \( R^+ \) (Reweighted \( \ell_2/\ell_1 \) algorithms).
- The essence of this algorithm is to create a bound for the concave penalty function and follow the steps of a Majorize-Minimization (MM) algorithm.
Reweighted $\ell_1$ optimization

**Assume:** $g(x_i) = h(|x_i|)$ with $h$ concave.

Now we have to bound this concave penalty function.
Reweighted $\ell_1$ optimization

Assume: $g(x_i) = h(|x_i|)$ with $h$ concave.
Reweighted $\ell_1$ optimization

Assume: $g(x_i) = h(|x_i|)$ with $h$ concave.

Updates

$$x^{(k+1)} \rightarrow \text{argmin}_x \|y - \Phi x\|_2^2 + \lambda \sum_i w_{i}^{(k)} |x_i|$$

$$w_i^{k+1} \rightarrow \frac{\partial g(x_i)}{\partial |x_i|} \bigg|_{x_i=x_i^{(k+1)}}$$
Reweighted $\ell_1$ optimization

Candes et al., 2008

- Penalty: $g(x_i) = \log(|x_i| + \epsilon), \ 0\epsilon \geq 0$
- Weight Update: $w_{i}^{(k+1)} \rightarrow [|x_{i}^{(k+1)} + \epsilon]^{-1}$
Reweighted $\ell_2$ optimization

- **Assume:** $g(x_i) = h(x_i^2)$ with $h$ concave
- Upper bound $h(.)$ as before.
- Bound will be quadratic in the variables leading to a weighted 2-norm optimization problem
Reweighted $\ell_2$ optimization

- **Assume:** $g(x_i) = h(x_i^2)$ with $h$ concave
- Upper bound $h(.)$ as before.
- Bound will be quadratic in the variables leading to a weighted 2-norm optimization problem

**Updates**

\[
x^{(k+1)} \rightarrow \arg\min_x \|y - \Phi x\|_2^2 + \lambda \sum_i w_i^{(k)} x_i^2
\]

\[
= \tilde{W}^{(k)} \Phi^T (\lambda I + \Phi \tilde{W}^{(k)} \Phi^T)^{-1} y
\]

\[
w_i^{k+1} \rightarrow \frac{\partial g(x_i)}{\partial x_i^2} \bigg|_{x_i = x_i^{(k+1)}} , \quad \tilde{W}^{(k+1)} \rightarrow \text{diag}[w^{(k+1)}]^{-1}
\]
Reweighted $\ell_2$ optimization: Examples

**FOCUSS Algorithm [Rao et al., 2003]**

- **Penalty:** $g(x_i) = |x_i|^p$, $0 \leq p \leq 2$
- **Weight Update:** $w_i^{(k+1)} \rightarrow |x_i^{(k+1)}|^{p-2}$
- **Properties:** Well-characterized convergence rates; very susceptible to local minima when $p$ is small.
Reweighted $\ell_2$ optimization: Examples

**FOCUSS Algorithm [Rao et al., 2003]**

- **Penalty:** $g(x_i) = |x_i|^p$, $0 \leq p \leq 2$
- **Weight Update:** $w_i^{(k+1)} \rightarrow |x_i^{(k+1)}|^{p-2}$
- **Properties:** Well-characterized convergence rates; very susceptible to local minima when $p$ is small.

**Chartrand and Yin (2008) Algorithm**

- **Penalty:** $g(x_i) = \log(x_i^2 + \epsilon)$, $\epsilon \geq 0$
- **Weight Update:** $w_i^{(k+1)} \rightarrow [(x_i^{(k+1)})^2 + \epsilon]^{-1}$
- **Properties:** Slowly reducing $\epsilon$ to zero smoothes out local minima initially allowing better solutions to be found;
Empirical Comparison

1. **Φ** matrix with 50 rows and 250 columns generated from zero mean Gaussian with normalized columns.

2. Select the support of true sparse coefficient vector $x_0$ randomly.

3. Generate the non-zero components of true sparse coefficient vector $x_0$ from normal distribution.

4. Compute measurement $y = \Phi x_0$.

5. Measurements and the **Φ** matrix are presented to our estimation algorithms to generate the estimate of the coefficient vector $\hat{x}$.

6. Repeat for 1000 trials.
Empirical Comparison

**Figure:** Probability of Successful recovery vs Number of non zero coefficients

Sparse Algorithms' performance in terms of Probability of Successful Recovery

Bhaskar D Rao  University of California, San Diego
Limitation of MAP based methods

To retain the same maximally sparse global solution as the $\ell_0$ norm in general conditions, then any possible MAP algorithm will possess $O\left(\binom{M}{N}\right)$ local minima.
Hierarchical Bayes: Sparse Bayesian Learning (SBL)

MAP estimation is just a penalized regression, hence Bayesian interpretation has not contributed much as of now. MAP methods were interested in the mode of the posterior but SBL uses posterior information beyond the mode, i.e. posterior distribution.

Problem
For all sparse priors it is not possible to compute the normalized posterior $P(x|y)$, hence some approximations are needed.
Hierarchical Bayes: Sparse Bayesian Learning (SBL)

MAP estimation is just a penalized regression, hence Bayesian Interpretation has not contributed much as of now.
Hierarchical Bayes: Sparse Bayesian Learning (SBL)

MAP estimation is just a penalized regression, hence Bayesian Interpretation has not contributed much as of now.

MAP methods were interested in the mode of the posterior but SBL uses posterior information beyond the mode, i.e. posterior distribution.
MAP estimation is just a penalized regression, hence Bayesian Interpretation has not contributed much as of now.

MAP methods were interested in the mode of the posterior but SBL uses posterior information beyond the mode, i.e. posterior distribution.

Problem
For all sparse priors it is not possible to compute the normalized posterior $P(x|y)$, hence some approximations are needed.

Bhaskar D Rao  University of California, San Diego
Hierarchical Bayesian Framework (Type II)

In order for this framework to be useful, we need tractable representations: Gaussian Scaled Mixtures.
Hierarchical Bayesian Framework (Type II)

In order for this framework to be useful, we need tractable representations: Gaussian Scaled Mixtures
Construction of Sparse priors

Separability: \( P(x) = \prod_i P(x_i) \)

Most of the sparse priors over \( x \) (including those with concave \( g \)) can be represented in this GSM form, and different scale mixing density \( i \), \( P(i) \) will lead to different sparse priors. [Palmer et al., 2006]

Instead of solving a MAP problem in \( x \), in the Bayesian framework one estimates the hyperparameters leading to an estimate of the posterior distribution for \( x \), i.e. \( P(x|y; \hat{\theta}) \). (Sparse Bayesian Learning)

Bhaskar D Rao  University of California, San Diego
Construction of Sparse priors

**Separability:** \( P(x) = \prod_i P(x_i) \)

**Gaussian Scale Mixture:**

\[
P(x_i) = \int P(x_i | \gamma_i) P(\gamma_i) d\gamma_i = \int N(x_i; 0, \gamma_i) P(\gamma_i) d\gamma_i
\]
Construction of Sparse priors

**Separability:** \( P(x) = \prod_i P(x_i) \)

**Gaussian Scale Mixture:**

\[
P(x_i) = \int P(x_i|\gamma_i) P(\gamma_i) d\gamma_i = \int N(x_i; 0, \gamma_i) P(\gamma_i) d\gamma_i
\]

Most of the sparse priors over \( x \) (including those with concave \( g \)) can be represented in this GSM form, and different scale mixing density i.e, \( P(\gamma_i) \) will lead to different sparse priors. [Palmer et al., 2006]
Construction of Sparse priors

**Separability:** \( P(x) = \prod_i P(x_i) \)

**Gaussian Scale Mixture:**
\[
P(x_i) = \int P(x_i | \gamma_i) P(\gamma_i) d\gamma_i = \int N(x_i; 0, \gamma_i) P(\gamma_i) d\gamma_i
\]

Most of the sparse priors over \( x \) (including those with concave \( g \)) can be represented in this GSM form, and different scale mixing density i.e, \( P(\gamma_i) \) will lead to different sparse priors. [Palmer et al., 2006]

Instead of solving a MAP problem in \( x \), in the Bayesian framework one estimates the hyperparameters \( \gamma \) leading to an estimate of the posterior distribution for \( x \), i.e. \( P(x | y; \hat{\gamma}) \). (Sparse Bayesian Learning)
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 \( p/2. \)
Sparse Bayesian Learning (Tipping)

\[ y = \Phi x + v \]

Solving for the optimal \( \gamma \)

\[
\hat{\gamma} = \arg \max_{\gamma} P(\gamma|y) = \arg \max_{\gamma} P(y|\gamma)P(\gamma) \\
= \arg \min_{\gamma} \log|\Sigma_y| + y^T \Sigma_y^{-1} y - 2 \sum_i \log P(\gamma_i)
\]

where, \( \Sigma_y = \sigma^2 I + \Phi \Gamma \Phi^T \) and \( \Gamma = \text{diag}(\gamma) \)
y = \Phi x + v

Solving for the optimal $\gamma$

$\hat{\gamma} = \arg\max \ P(\gamma|y) = \arg\max \ P(y|\gamma)P(\gamma)$

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

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

Empirical Bayes

Choose $P(\gamma_i)$ to be a non-informative prior
Sparse Bayesian Learning

**Computing Posterior**

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

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

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

Computing Posterior

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

\[
\begin{align*}
\mu_x &= E[x|y; \hat{\gamma}] = \hat{\gamma} \Phi^T(\sigma^2 I + \Phi \hat{\gamma} \Phi^T)^{-1} y \\
\Sigma_x &= Cov[x|y; \hat{\gamma}] = \hat{\gamma} - \hat{\gamma} \Phi^T(\sigma^2 I + \Phi \hat{\gamma} \Phi^T)^{-1} \Phi \hat{\gamma}
\end{align*}
\]

Updating \( \gamma \)

Using EM algorithm with a non informative prior over \( \gamma \), the update rule becomes:

\[
\gamma_i \leftarrow \mu_x(i)^2 + \Sigma_x(i, i)
\]
SBL properties

Local minima are sparse, i.e. have at most $N$ nonzero. Bayesian inference cost is generally much smoother than associated MAP estimation. Fewer local minima.

In high signal to noise ratio, the global minima is the sparsest solution. No structural problems.

Bhaskar D Rao  University of California, San Diego
SBL properties

- Local minima are sparse, i.e. have at most $N$ nonzero $\gamma_i$
SBL properties

- Local minima are sparse, i.e. have at most $N$ nonzero $\gamma_i$
- Bayesian inference cost is generally much smoother than associated MAP estimation. Fewer local minima.
SBL properties

- Local minima are sparse, i.e. have at most $N$ nonzero $\gamma_i$.
- Bayesian inference cost is generally much smoother than associated MAP estimation. Fewer local minima.
- In high signal to noise ratio, the global minima is the sparsest solution. No structural problems.
Empirical Comparison

For each test case

1. Generate a random dictionary $\Phi$ with 50 rows and 250 columns from the normal distribution and normalize each column to have 2-norm of 1.

2. Select the support for the true sparse coefficient vector $x_0$ randomly.

3. Generate the non-zero components of $x_0$ from the normal distribution.

4. Compute signal, $y = \Phi x_0$ (Noiseless case).

5. Compare SBL with previous methods with regard to estimating $x_0$.

6. Average over 1000 independent trials.
Empirical Comparison: 1000 trials

Figure: Probability of Successful recovery vs Number of non zero coefficients
Empirical Comparison: Multiple Measurement Vectors (MMV)

Generate data matrix via $Y = \Phi X_0$ (noiseless), where:

1. $X_0$ is 100-by-5 with random non-zero rows.
2. $\Phi$ is 50-by-100 with Gaussian iid entries.
Empirical Comparison: 1000 trials

![Graph showing probability of success vs. row sparsity for different methods: SBL, Candès et al. (2008), Chartrand and Yin (2008), L₁ solution.](image)
Bayesian methods offer interesting and useful options to the Sparse Signal Recovery problem. MAP estimation (Reweighted $\ell_2$/$\ell_1$ algorithms) is a common approach in sparse signal recovery. Sparse Bayesian Learning is versatile and can be more easily employed in problems with structure. Algorithms can often be justified by studying the resulting objective functions.
Bayesian methods offer interesting and useful options to the Sparse Signal Recovery problem

- MAP estimation (Reweighted $\ell_2/\ell_1$ algorithms)
- Sparse Bayesian Learning
Summary

- Bayesian methods offer interesting and useful options to the Sparse Signal Recovery problem
  - MAP estimation (Reweighted $\ell_2/\ell_1$ algorithms)
  - Sparse Bayesian Learning
- Versatile and can be more easily employed in problems with structure
Summary

- Bayesian methods offer interesting and useful options to the Sparse Signal Recovery problem
  - MAP estimation (Rewighted $\ell_2/\ell_1$ algorithms)
  - 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.