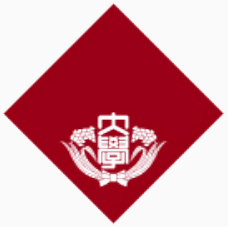


Bayesian Sparse-Smooth Modeling and Variational Inference

Shunsuke Horii

November 13, 2017

Waseda University

Introduction

Background

- Learning from high-dimensional data is one of the emerging tasks in machine learning
  - In some cases, dimension p is larger than sample size n
- Sparsity is an important concept
- Optimization based approach such as LASSO has been well studied
  - LASSO can be considered as a method of finding MAP estimate assuming the Laplace prior
- Full information of the posterior distribution is useful for some problems
  - Ex: Bayesian experimental design
- Some methods for sparse Bayesian modeling
  - Relevance Vector Machine, Gibbs sampler (Bayesian LASSO), Variational Bayes

Introduction

Past works

- Some extensions of LASSO:
  - Group LASSO: Group Sparsity
  - Fused LASSO: Sparsity + Smoothness
    - Parameter is sparse and locally constant
- Some Bayesian studies for extended LASSO models:
  - Kyung et al, 2010: Gibbs sampler for Group LASSO, Fused LASSO, Elastic net
  - Babacan et al, 2014: Variational Bayes for Group LASSO

Our work

- Variational Bayes for sparse and smooth model (Fused LASSO model)
  - Variational Bayes is more efficient than Gibbs sampler

Introduction

Relationship with past works:

	Sparsity only	Group Sparsity	Sparsity + Smoothness
Optimization based	Tibshirani, 1996	Yuan and Lin, 2006	Tibshirani, 2005
MCMC	Park and Casella, 2008	Kyung et al., 2010	Kyung et al., 2010
Variational Bayes	Babacan et al., 2014	Babacan et al., 2014	**Our work**

Strength of our work:

- Considering sparsity and smoothness: effective for sparse and smooth data
  - Ex: Denoising for sparse image
- Variational Bayes: more efficient than Gibbs sampler
  - Applicable for very high dimensional data analysis

Problem setup

Linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\boldsymbol{y} \in \mathbb{R}^n$: Observation vector
- $\boldsymbol{X} \in \mathbb{R}^{n \times p}$: Design matrix
- $\boldsymbol{\beta} \in \mathbb{R}^p$: Unknown parameter
- $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, s^{-1}\boldsymbol{I}_p)$: Gaussian noise vector (s: precision parameter)

Assumption:

Sparsity: $\boldsymbol{\beta}$ is sparse (number of nonzero elements is small)

Smoothness: β_i and β_j take similar values for $(i, j) \in E$

- $G = (V, E)$ is a predefined graph

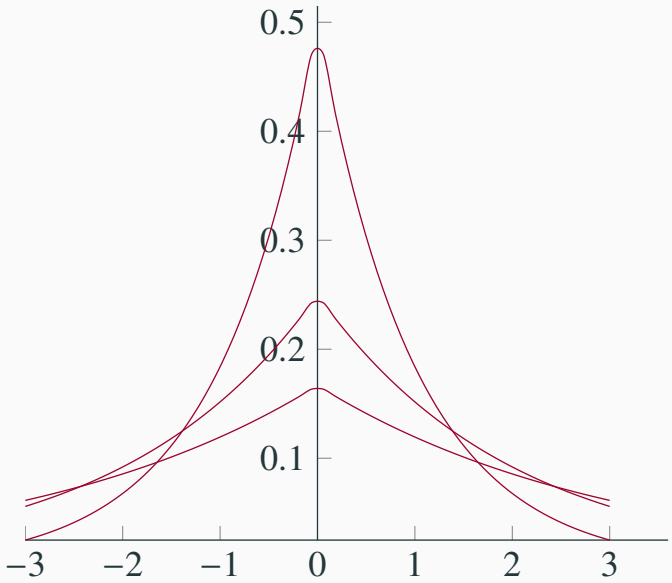
How to statistically model sparsity and smoothness?

Straightforward approach

Assume Laplace priors for $\{\beta_j\}$ and $\{\beta_i - \beta_j\}$

$$p(\boldsymbol{\beta}) \propto \prod_{j=1}^{p} \exp\left(-\frac{|\beta_j|}{\sqrt{a_{\tau,j}}}\right) \prod_{(j,k)\in E} \exp\left(-\frac{|\beta_j - \beta_k|}{\sqrt{a_{\nu,jk}}}\right)$$

Sparsity Smoothness



- MAP estimator corresponds to the Fused LASSO estimator
- Difficult to calculate the posterior distribution
- What values should be set for α?

$\Rightarrow$ needs another approach

Hierarchical representation of Laplace distribution

Laplace distribution can be expressed as a scale mixture of Gaussian with exponential mixture:

$$\frac{\sqrt{\alpha}}{2}\exp\left(-\sqrt{\alpha}|\beta|\right) = \int_0^\infty p(\beta|\tau)p(\tau|\gamma)d\tau$$

$$p(\beta|\tau) = \mathcal{N}(\beta|0,\tau^{-1})$$

$$p(\tau|\gamma) = \frac{\gamma}{2}\exp\left(-\frac{\gamma}{2}\tau\right)$$

- This property is used in the past Bayesian studies on the sparse modeling
- If we use inverse gamma distribution for the scale mixture, the marginal distribution becomes Student's t distribution

Sparse-Smooth hierarchical modeling

Gaussian distribution for $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta}|\boldsymbol{\tau},\boldsymbol{\nu}) \propto \prod_{j=1}^{p} \exp\left(-\tau_j \beta_j^2\right) \prod_{(j,k)\in E} \exp\left(-\nu_{jk}(\beta_j - \beta_k)^2\right)$$

Exponential (or IG) distribution for $\boldsymbol{\tau}$ and $\boldsymbol{\nu}$:

$$p(\boldsymbol{\tau}|\boldsymbol{a_\tau}) = \prod_{j=1}^{p} \text{Exp}(\tau_j|a_{\tau,j}),\ p(\boldsymbol{\nu}|\boldsymbol{a_\nu}) = \prod_{(j,k)\in E} \text{Exp}(\nu_{jk}|a_{\nu,jk})$$

Gamma distribution for $\boldsymbol{a_\tau}, \boldsymbol{a_\nu}$:

$$p(\boldsymbol{a_\tau}) = \prod_{j=1}^{p} \text{Gam}(a_{\tau,j}|\theta_\tau, k_\tau),\ p(\boldsymbol{a_\nu}) = \prod_{(j,k)\in E} \text{Gam}(a_{\nu,jk}|\theta_\nu, k_\nu)$$

Variational Bayes

Joint distribution:

$$p(\boldsymbol{y}, \boldsymbol{\beta}, s, \boldsymbol{\tau}, \boldsymbol{\nu}, \boldsymbol{a_\tau}, \boldsymbol{a_\nu}) = p(\boldsymbol{y}|\boldsymbol{\beta}, s)p(\boldsymbol{\beta}|\boldsymbol{\tau}, \boldsymbol{\nu})p(\boldsymbol{\tau}|\boldsymbol{a_\tau})p(\boldsymbol{\nu}|\boldsymbol{a_\nu})p(\boldsymbol{a_\tau}, \boldsymbol{a_\nu})p(s)$$

We want to calculate posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{y})$

$\Rightarrow$ complex integral calculation is required

Mean field approximation:

For $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\nu}, \boldsymbol{a_\tau}, \boldsymbol{a_\nu}, s)$, find an approximate distribution $q(\boldsymbol{\theta})$ which minimizes Kullback-Leibler divergence:

$$q^*(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta})}{\operatorname{argmin}} \int q(\boldsymbol{\theta}) \ln \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})} d\boldsymbol{\theta}$$

w.r.t. factorized distribution $q(\boldsymbol{\theta}) = q(\boldsymbol{\beta})q(\boldsymbol{\tau}, \boldsymbol{\nu})q(\boldsymbol{a_\tau}, \boldsymbol{a_\nu})q(s)$

Variational Bayes

$\langle \cdot \rangle$: Expectation w.r.t. the corresponding distribution

Update equation for $q(\boldsymbol{\beta})$:

$$q^*(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}|\bar{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}),$$
$$\bar{\boldsymbol{\beta}} = \langle s \rangle \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \boldsymbol{X}^T \boldsymbol{y},$$
$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \left(\langle s \rangle \boldsymbol{X}^T \boldsymbol{X} + \langle \boldsymbol{S}_{\boldsymbol{\tau},\boldsymbol{\nu}} \rangle \right)^{-1}$$

Update equation for $q(\boldsymbol{\tau}, \boldsymbol{\nu})$:

$$q^*(\boldsymbol{\tau}, \boldsymbol{\nu}) = \prod_{j=1}^{p} \mathrm{Gig}\left(\tau_j \left| \left\langle a_{\tau,j} \right\rangle, \left\langle \beta_j^2 \right\rangle, \frac{1}{2} \right. \right) \prod_{(j,k) \in E} \mathrm{Gig}\left(\nu_{jk} \left| \left\langle a_{\nu,jk} \right\rangle, \left\langle \left(\beta_j - \beta_k \right)^2 \right\rangle, \frac{1}{2} \right. \right)$$

$\mathrm{Gig}(x|a, b, \rho)$ is the generalized inverse Gaussian:

$$\mathrm{Gig}(x|a, b, \rho) \propto x^{\rho - 1} \exp\left(-\frac{1}{2}(ax + bx^{-1}) \right)$$

Variational Bayes

Update equation for $q(\boldsymbol{a_\tau}, \boldsymbol{a_\nu})$:

$$q^*(\boldsymbol{a_\tau}, \boldsymbol{a_\nu}) = \prod_{j=1}^{p} \text{Gam}\left(a_{\tau,j} \middle| k_\tau + 1, \theta_\tau + \left\langle \tau_j \right\rangle /2\right) \cdot \prod_{(j,k)\in E} \text{Gam}\left(a_{\nu,jk} \middle| k_\nu + 1, \theta_\nu + \left\langle \nu_{jk} \right\rangle /2\right)$$

Update equation for $q(s)$:

$$q^*(s) = \text{Gam}\left(s \middle| k_s + n/2, \theta_s + \left\langle (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^2 \right\rangle /2\right)$$

Assuming that $p(s)$ is $\text{Gam}(s|k_s, \theta_s)$

Experiment on synthetic data

- We considered two cases:

  Case 1: True parameter is $\beta_1^* = (\mathbf{0}_{10}, \mathbf{2}_{10}, \mathbf{0}_{10}, \mathbf{2}_{10})$ (sparse and smooth)

  Case 2: True parameter is $\beta_2^* = (0, 2, 0, 2, \ldots) \in \mathbb{R}^{40}$ (sparse, but not smooth)

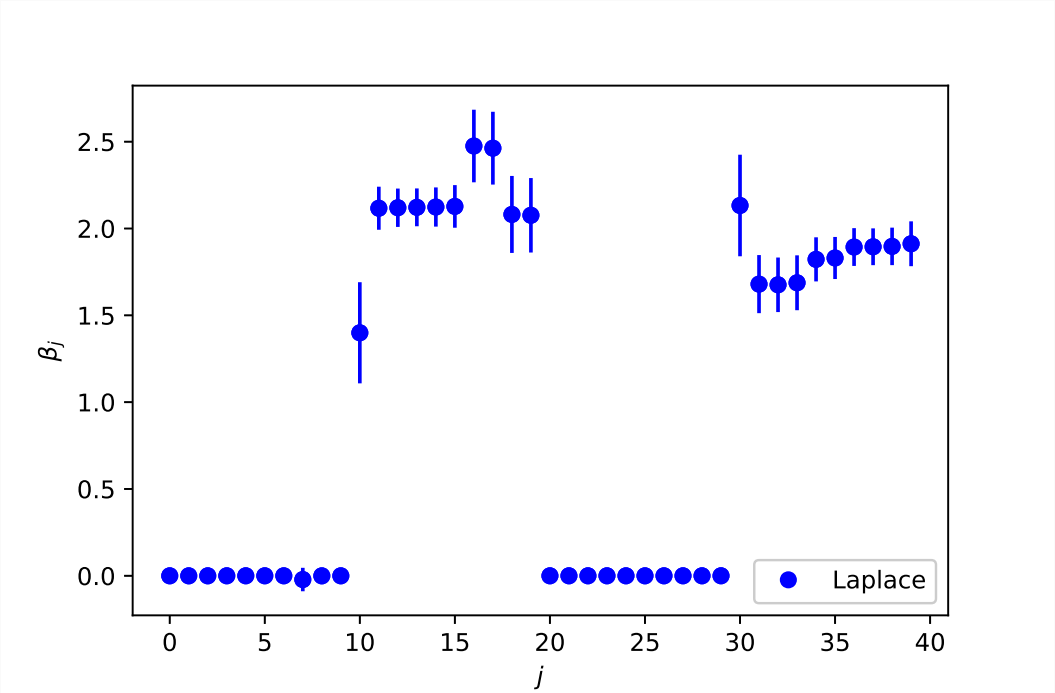
- Randomly generate sample (size = 40)
- Number of experiments is 100
- For $p(\tau, \nu)$, besides exponential distribution, we also consider inverse gamma distribution
  - Referred as Laplace when exponential distribution is assumed and Student when inverse gamma distribution is assumed

Experiment on synthetic data

Mean squared error:

	Lasso	Laplace	Student
β_1^*	5.75×10^{-1}	8.07×10^{-3}	7.98×10^{-3}
β_2^*	5.53×10^{-1}	1.37	1.11

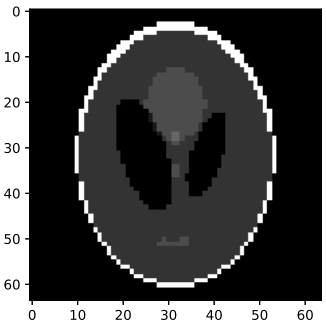
Example of estimation result:



Application for denoising of sparse image

$\beta \in \mathbb{R}^{60\times 60}$:

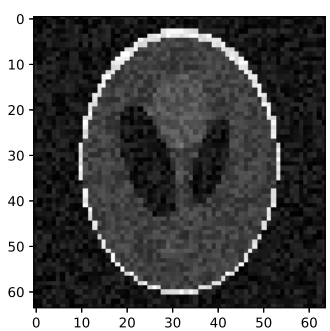
Original sparse image



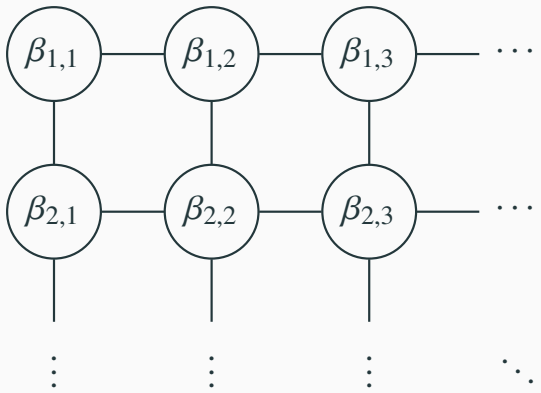
$y \in \mathbb{R}^{60\times 60}$:

Noisy image

$y = \beta + \epsilon$



Graph G: 2D-grid



Problem: Estimate the original image from the noisy image

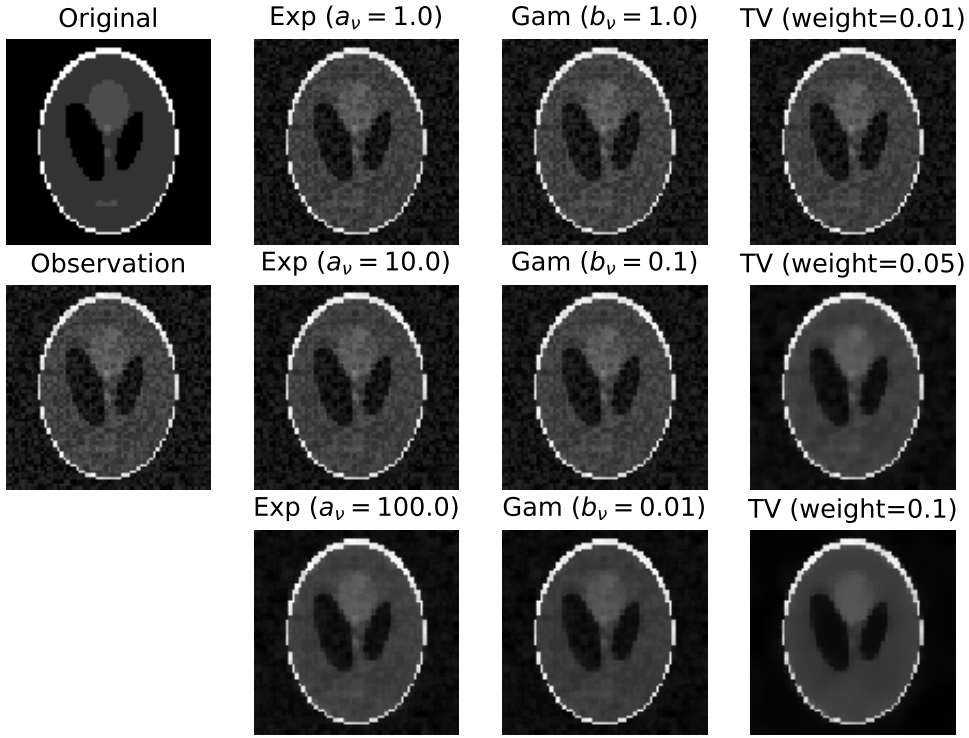
- In this experiment, τ, that controls sparsity, is fixed

Application for denoising of sparse image

PSNR [DB] and SSIM of restored images:

method	Exponential			Gamma			TV		
parameter	$a_\nu = 1.0$	$a_\nu = 10.0$	$a_\nu = 100.0$	$b_\nu = 1.0$	$b_\nu = 0.1$	$b_\nu = 0.01$	λ=0.01	λ=0.05	λ=0.1
PSNR	24.89	24.96	**25.90**	24.18	24.46	25.04	24.51	24.85	24.49
SSIM	0.5541	0.5680	**0.6257**	0.5407	0.5620	0.6138	0.5670	0.6119	0.6028

Restored images:



Summary

- We proposed hierarchical model for the problem of estimating parameter which is sparse and smooth.
- We also proposed an approximate estimation algorithm based on the variational Bayes method.
- The effectiveness of the proposed method was proved by experiments on synthetic data and real image data