Sparse Bayesian nonparametric regression

François Caron and Arnaud Doucet

Depts of Computer Science & Statistics, UBC

July 7, 2008

Introduction

- Linear regression model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_L\right) \tag{1}$$

Introduction

- Linear regression model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_L\right) \tag{1}$$

- $y \in \mathbb{R}^L$, X design matrix of size $L \times K$, $\beta \in \mathbb{R}^K$

Introduction

- Linear regression model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_L\right) \tag{1}$$

- $y \in \mathbb{R}^L$, X design matrix of size $L \times K$, $\beta \in \mathbb{R}^K$
- Sparse estimate of β

Introduction

- Linear regression model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_L\right) \tag{1}$$

- $y \in \mathbb{R}^L$, X design matrix of size $L \times K$, $\beta \in \mathbb{R}^K$
- Sparse estimate of β
- Variable selection, decomposition of a signal over an overcomplete basis, etc.

Introduction

- Linear regression model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_L\right) \tag{1}$$

- $y \in \mathbb{R}^L$, X design matrix of size $L \times K$, $\beta \in \mathbb{R}^K$
- Sparse estimate of β
- Variable selection, decomposition of a signal over an overcomplete basis, etc.
- Numerous models/algorithms

Introduction

- Linear regression model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_L\right) \tag{1}$$

- $y \in \mathbb{R}^L$, X design matrix of size $L \times K$, $\beta \in \mathbb{R}^K$
- Sparse estimate of β
- Variable selection, decomposition of a signal over an overcomplete basis, etc.
- Numerous models/algorithms
  - Spike and Slab

Introduction

- Linear regression model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_L\right) \tag{1}$$

- $y \in \mathbb{R}^L$, X design matrix of size $L \times K$, $\beta \in \mathbb{R}^K$
- Sparse estimate of β
- Variable selection, decomposition of a signal over an overcomplete basis, etc.
- Numerous models/algorithms
  - Spike and Slab
  - Lasso

Introduction

- Linear regression model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_L\right) \tag{1}$$

- $y \in \mathbb{R}^L$, X design matrix of size $L \times K$, $\beta \in \mathbb{R}^K$
- Sparse estimate of β
- Variable selection, decomposition of a signal over an overcomplete basis, etc.
- Numerous models/algorithms
  - Spike and Slab
  - Lasso
  - Relevance Vector Machine

Introduction

- Prior distribution $p(\beta) = \prod_{k=1}^{K} p(\beta_k)$

Introduction

- Prior distribution $p(\beta) = \prod_{k=1}^{K} p(\beta_k)$
- Local minima of the objective function

$$-\frac{1}{2\sigma^2}||y - X\beta||^2 - \log p(\beta) \tag{2}$$

Introduction

- Prior distribution $p(\beta) = \prod_{k=1}^{K} p(\beta_k)$
- Local minima of the objective function

$$-\frac{1}{2\sigma^2}||y - X\beta||^2 - \log p(\beta) \tag{2}$$

- Scale-mixture of Gaussians

$$p(\beta_k) = \int \mathcal{N}(\beta_k; 0, \sigma_k^2) p\left(\sigma_k^2\right) d\sigma_k^2$$

Introduction

- Prior distribution $p(\beta) = \prod_{k=1}^{K} p(\beta_k)$
- Local minima of the objective function

$$-\frac{1}{2\sigma^2}||y - X\beta||^2 - \log p(\beta) \tag{2}$$

- Scale-mixture of Gaussians

$$p(\beta_k) = \int \mathcal{N}(\beta_k; 0, \sigma_k^2) p\left(\sigma_k^2\right) d\sigma_k^2$$

  - Laplace prior $\longrightarrow$ Lasso objective function

Introduction

- Prior distribution $p(\beta) = \prod_{k=1}^{K} p(\beta_k)$
- Local minima of the objective function

$$-\frac{1}{2\sigma^2}||y - X\beta||^2 - \log p(\beta) \tag{2}$$

- Scale-mixture of Gaussians

$$p(\beta_k) = \int \mathcal{N}(\beta_k; 0, \sigma_k^2) p\left(\sigma_k^2\right) d\sigma_k^2$$

  - Laplace prior $\longrightarrow$ Lasso objective function
  - Normal-Jeffreys

Introduction

- Prior distribution $p(\beta) = \prod_{k=1}^{K} p(\beta_k)$
- Local minima of the objective function

$$-\frac{1}{2\sigma^2}||y - X\beta||^2 - \log p(\beta) \tag{2}$$

- Scale-mixture of Gaussians

$$p(\beta_k) = \int \mathcal{N}(\beta_k; 0, \sigma_k^2) p\left(\sigma_k^2\right) d\sigma_k^2$$

  - Laplace prior $\longrightarrow$ Lasso objective function
  - Normal-Jeffreys
  - Normal-exponential gamma

Introduction

- Prior distribution $p(\beta) = \prod_{k=1}^{K} p(\beta_k)$
- Local minima of the objective function

$$-\frac{1}{2\sigma^2}||y - X\beta||^2 - \log p(\beta) \tag{2}$$

- Scale-mixture of Gaussians

$$p(\beta_k) = \int \mathcal{N}(\beta_k; 0, \sigma_k^2) p\left(\sigma_k^2\right) d\sigma_k^2$$

  - Laplace prior $\longrightarrow$ Lasso objective function
  - Normal-Jeffreys
  - Normal-exponential gamma
- Find local minimum of Eq. (2) with EM algorithm

Why this title is too bad

- Sparse... but... one model does not lead to 'strictly' sparse estimates!

Why this title is too bad

- Sparse... but... one model does not lead to 'strictly' sparse estimates!
- Bayesian... but... EM algorithm

Why this title is too bad

- Sparse... but... one model does not lead to 'strictly' sparse estimates!
- Bayesian... but... EM algorithm
- nonparametric... but... the number of parameters is finite!

Why this title is too bad

- Sparse... but... one model does not lead to 'strictly' sparse estimates!
- Bayesian... but... EM algorithm
- nonparametric... but... the number of parameters is finite!
- regression... YES IT IS :-)

Overview

Normal-gamma model

- Gamma prior over σ_k^2

$$\sigma_k^2 \sim \mathcal{G}(\frac{\alpha}{K}, \frac{\gamma^2}{2})$$

Normal-gamma model

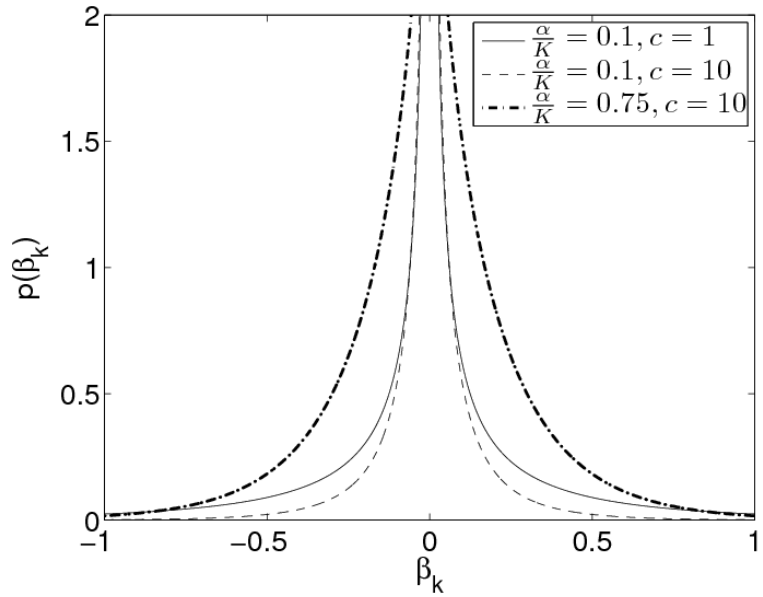
- Gamma prior over σ_k^2

$$\sigma_k^2 \sim \mathcal{G}(\frac{\alpha}{K}, \frac{\gamma^2}{2})$$

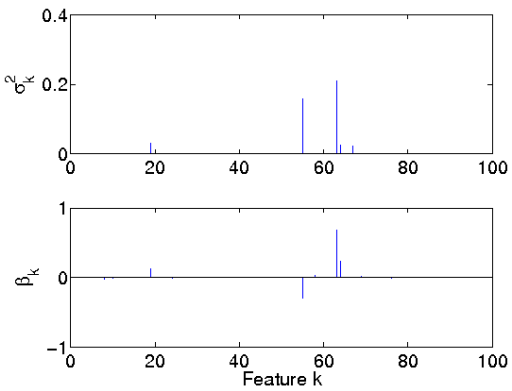
- Marginal distribution over β_k

$$p(\beta_k) \propto |\beta_k|^{\frac{\alpha}{K}-\frac{1}{2}} \mathcal{K}_{\frac{\alpha}{K}-\frac{1}{2}}(\gamma|\beta_k|) \tag{3}$$

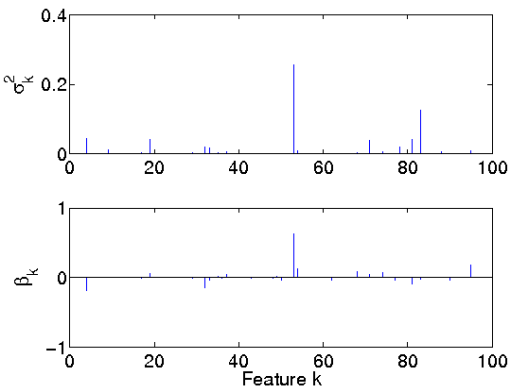
where $\mathcal{K}_\nu(\cdot)$ is the modified Bessel function of the second kind.



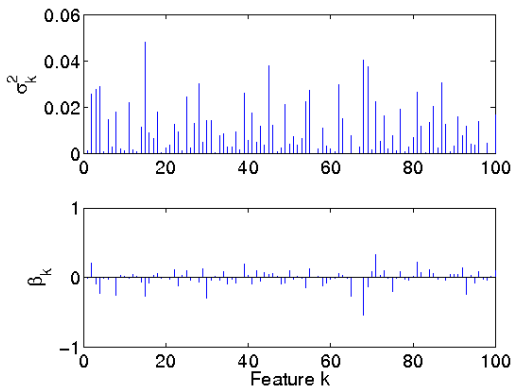
Normal-gamma model



(a) $\alpha = 1$



(b) $\alpha = 5$



(c) $\alpha = 100$

Asymptotic properties

- Bounded sum of the terms

$$\lim_{K\to\infty} \mathbb{E}[\sum_{k=1}^{K} |\beta_k|] = \frac{2\alpha}{\gamma} \qquad , \mathbb{E}[\sum_{k=1}^{K} \beta_k^2] = \frac{2\alpha}{\gamma^2}.$$

Asymptotic properties

- Bounded sum of the terms

$$\lim_{K\to\infty} \mathbb{E}[\sum_{k=1}^{K} |\beta_k|] = \frac{2\alpha}{\gamma} \qquad , \mathbb{E}[\sum_{k=1}^{K} \beta_k^2] = \frac{2\alpha}{\gamma^2}.$$

- Stick breaking construction for the weights

Asymptotic properties

- Bounded sum of the terms

$$\lim_{K\to\infty} \mathbb{E}[\sum_{k=1}^{K} |\beta_k|] = \frac{2\alpha}{\gamma} \qquad , \mathbb{E}[\sum_{k=1}^{K} \beta_k^2] = \frac{2\alpha}{\gamma^2}.$$

- Stick breaking construction for the weights
  - Let $\sigma^2_{(1)} \geq \sigma^2_{(2)} \geq \ldots \geq \sigma^2_{(K)}$ be the order statistics of the sequence $\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2$.

Asymptotic properties

- Bounded sum of the terms

$$\lim_{K\to\infty} \mathbb{E}[\sum_{k=1}^{K} |\beta_k|] = \frac{2\alpha}{\gamma} \qquad , \mathbb{E}[\sum_{k=1}^{K} \beta_k^2] = \frac{2\alpha}{\gamma^2}.$$

- Stick breaking construction for the weights
  - Let $\sigma^2_{(1)} \geq \sigma^2_{(2)} \geq \ldots \geq \sigma^2_{(K)}$ be the order statistics of the sequence $\sigma^2_1, \sigma^2_2, \ldots, \sigma^2_K$.
  - $\overline{\sigma}^2 = \left(\frac{\sigma^2_{(1)}}{\sum_k \sigma^2_{(k)}}, \frac{\sigma^2_{(2)}}{\sum_k \sigma^2_{(k)}}, \ldots \right)$ and $\sum_k \sigma^2_{(k)}$ are independent and respectively distributed according to $PD(\alpha)$ and $\mathcal{G}(\alpha, \gamma^2/2)$

Asymptotic properties

- Bounded sum of the terms

$$\lim_{K\to\infty} \mathbb{E}[\sum_{k=1}^{K} |\beta_k|] = \frac{2\alpha}{\gamma} \qquad , \mathbb{E}[\sum_{k=1}^{K} \beta_k^2] = \frac{2\alpha}{\gamma^2}.$$

- Stick breaking construction for the weights
  - Let $\sigma^2_{(1)} \geq \sigma^2_{(2)} \geq \ldots \geq \sigma^2_{(K)}$ be the order statistics of the sequence $\sigma^2_1, \sigma^2_2, \ldots, \sigma^2_K$.
  - $\overline{\sigma}^2 = \left(\frac{\sigma^2_{(1)}}{\sum_k \sigma^2_{(k)}}, \frac{\sigma^2_{(2)}}{\sum_k \sigma^2_{(k)}}, \ldots \right)$ and $\sum_k \sigma^2_{(k)}$ are independent and respectively distributed according to $PD(\alpha)$ and $\mathcal{G}(\alpha, \gamma^2/2)$
  - Can be recovered from the (Infinite) stick-breaking construction

$$\pi_k = \zeta_k \prod_{j=1}^{k-1} (1 - \zeta_j) \text{ with } \zeta_j \sim \mathcal{B}(1, \alpha) \tag{4}$$

Asymptotic properties

- Bounded sum of the terms

$$\lim_{K\to\infty} \mathbb{E}[\sum_{k=1}^{K} |\beta_k|] = \frac{2\alpha}{\gamma} \qquad , \mathbb{E}[\sum_{k=1}^{K} \beta_k^2] = \frac{2\alpha}{\gamma^2}.$$

- Stick breaking construction for the weights
  - Let $\sigma^2_{(1)} \geq \sigma^2_{(2)} \geq \ldots \geq \sigma^2_{(K)}$ be the order statistics of the sequence $\sigma^2_1, \sigma^2_2, \ldots, \sigma^2_K$.
  - $\overline{\sigma}^2 = \left(\frac{\sigma^2_{(1)}}{\sum_k \sigma^2_{(k)}}, \frac{\sigma^2_{(2)}}{\sum_k \sigma^2_{(k)}}, \ldots \right)$ and $\sum_k \sigma^2_{(k)}$ are independent and respectively distributed according to $PD(\alpha)$ and $\mathcal{G}(\alpha, \gamma^2/2)$
  - Can be recovered from the (Infinite) stick-breaking construction

$$\pi_k = \zeta_k \prod_{j=1}^{k-1} (1 - \zeta_j) \text{ with } \zeta_j \sim \mathcal{B}(1, \alpha) \tag{4}$$

  - Coefficients (β_k) are the weights (jumps) of the so-called variance gamma process (Brownian motion evaluated at times given by a gamma process)

Normal-inverse Gaussian model

- Inverse-Gaussian prior over σ_k^2

$$\sigma_k^2 \sim \mathcal{IG}(\frac{\alpha}{K}, \gamma) \tag{5}$$

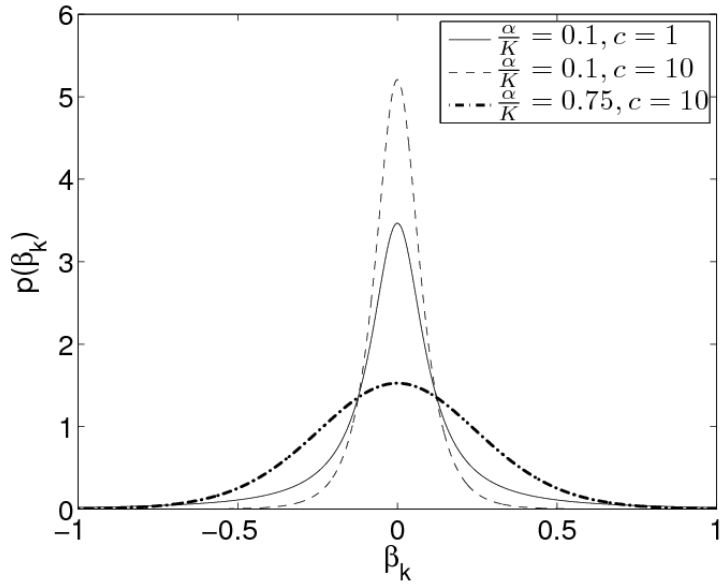
Normal-inverse Gaussian model

- Inverse-Gaussian prior over σ_k^2

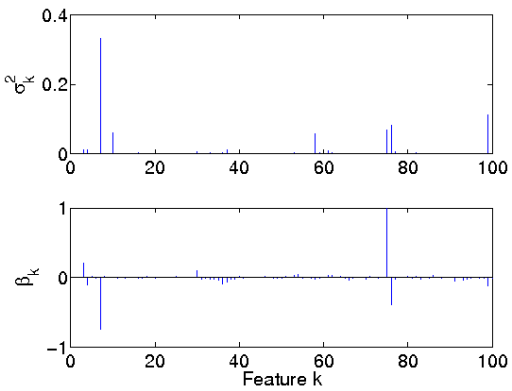
$$\sigma_k^2 \sim \mathcal{IG}(\frac{\alpha}{K}, \gamma) \tag{5}$$

- Marginal pdf of β_k

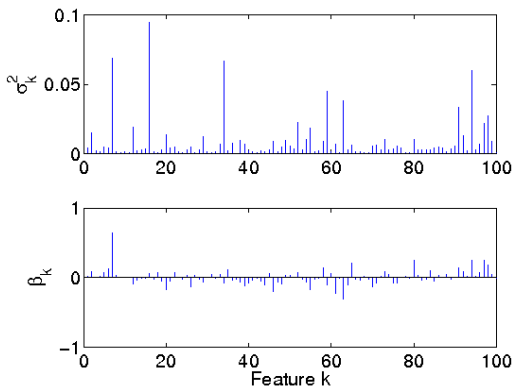
$$p(\beta_k) \propto \left(\frac{\alpha^2}{K^2} + \beta_k^2\right)^{-1/2} \mathcal{K}_1\left(\gamma\sqrt{\frac{\alpha^2}{K^2} + \beta_k^2}\right) \tag{6}$$



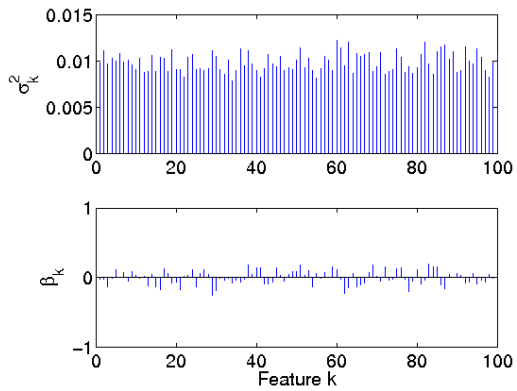
Normal-inverse Gaussian model



(d) $\alpha = 1$



(e) $\alpha = 5$



(f) $\alpha = 100$

Extension

- N vectors $\{y_n\}_{n=1}^N$ where $y_n \in \mathbb{R}^L$.

Extension

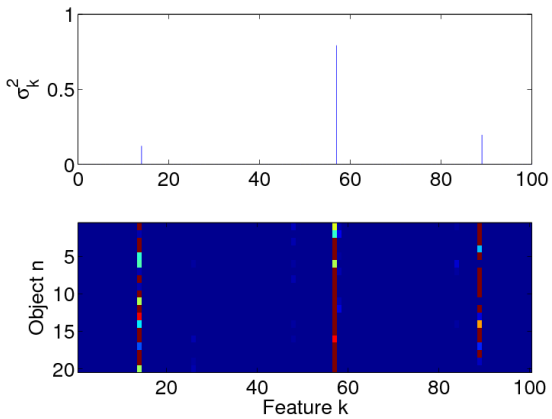
- N vectors $\{y_n\}_{n=1}^N$ where $y_n \in \mathbb{R}^L$.
- For a given k the random variables $\{\beta_k^n\}_{n=1}^N$ are statistically dependent and exchangeable

Extension

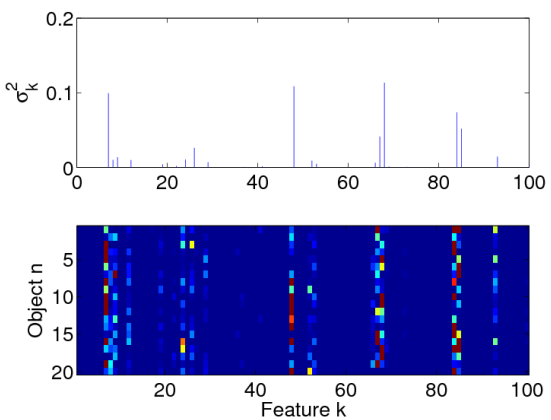
- N vectors $\{y_n\}_{n=1}^N$ where $y_n \in \mathbb{R}^L$.
- For a given k the random variables $\{\beta_k^n\}_{n=1}^N$ are statistically dependent and exchangeable
- Hierarchical model

$$\sigma_k^2 \sim \mathcal{G}(\frac{\alpha}{K}, \frac{\gamma^2}{2}) \text{ or } \sigma_k^2 \sim \mathcal{IG}(\frac{\alpha}{K}, \gamma)$$

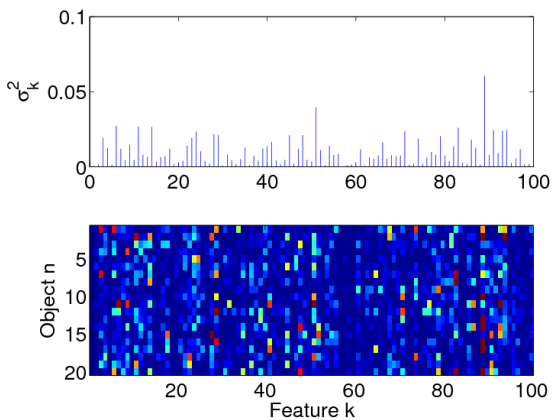
for $k = 1, \ldots, K$ and $\beta_k^n \sim \mathcal{N}(0, \sigma_k^2)$ for $n = 1, \ldots, N$.



(m) $\alpha = 1$



(n) $\alpha = 5$



(o) $\alpha = 100$

Extensions

- As $K \to \infty$, prior distributions over infinite matrices with real-valued entries

Extensions

- As $K \to \infty$, prior distributions over infinite matrices with real-valued entries
- Complementary to the Indian buffet process and the infinite gamma-Poisson process which are prior distributions over infinite matrices with integer-valued entries.

Overview

Sparsity properties

- Minimize

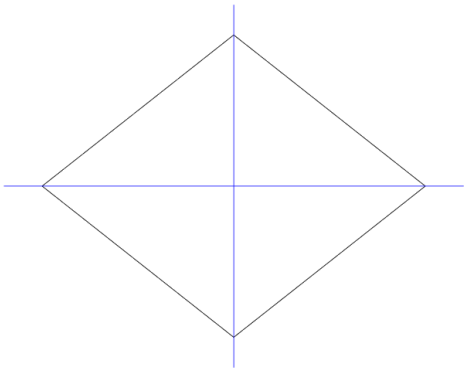
$$-\sum_{n=1}^{N}\frac{1}{2\sigma_n^2}||y^n - X\beta^n||^2 - \sum_{k=1}^{K} pen(\beta_k^{1:N})$$

where

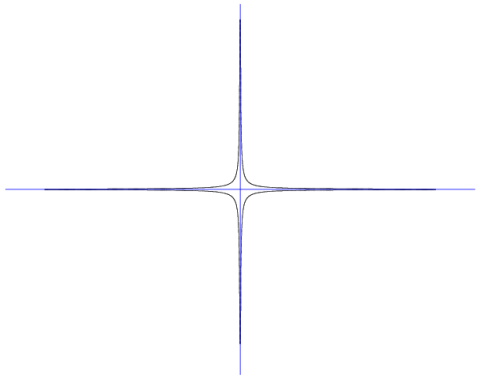
	$pen(\beta_k^{1:N})$
Lasso ($N=1$)	$\gamma\lvert\beta_k\rvert$
NJ	$N\log(u_k)$
NG	$(\frac{N}{2}-\frac{\alpha}{K})\log u_k - \log \mathcal{K}_{\frac{\alpha}{K}-\frac{N}{2}}(\gamma u_k)$
NIG	$\frac{N+1}{2}\log(q_k) - \log \mathcal{K}_{\frac{N+1}{2}}(\gamma q_k)$

where $u_k = \sqrt{\sum_{n=1}^{N}(\beta_k^n)^2}$, $q_k = \sqrt{\frac{\alpha^2}{K^2}+u_k^2}$

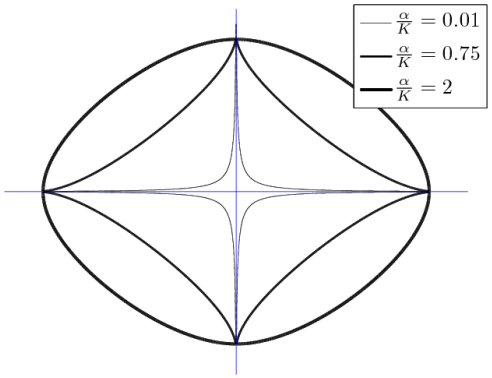
Sparsity properties



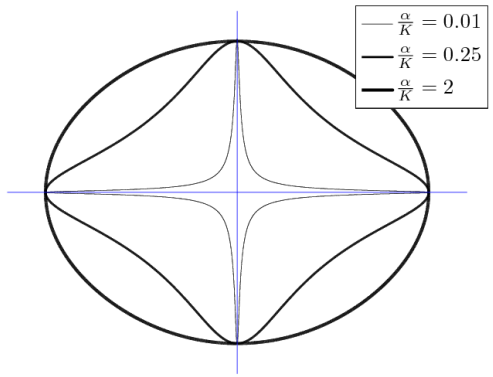
(p) Laplace



(q) Normal-Jeffreys



(r) Normal-gamma



(s) Normal-inverse Gaussian

Figure: Contour of constant value of $pen(\beta_1) + pen(\beta_2)$ for different priors

Sparsity properties

- The normal-gamma prior is a thresholding rule for $\alpha/K \leq 1$ and yields sparse estimates
- The normal-inverse Gaussian is not a thresholding rule but it can yield "almost sparse" estimates

Overview

Empirical results

- 100 datasets with $L = 50$ and $\sigma = 1$
- Correlation between $X_{k,i}$ and $X_{k,j}$ is $\rho^{|i-j|}$ with $\rho = 0.5$
- True $\beta = (3\ 1.5\ 0\ 0\ 2\ 0\ 0 \ldots)^T \in \mathbb{R}^K$, where $K = 20, 60, 100, 200$
- Parameters of the Lasso, NG and NIG are estimated by 5-fold cross-validation

Empirical results

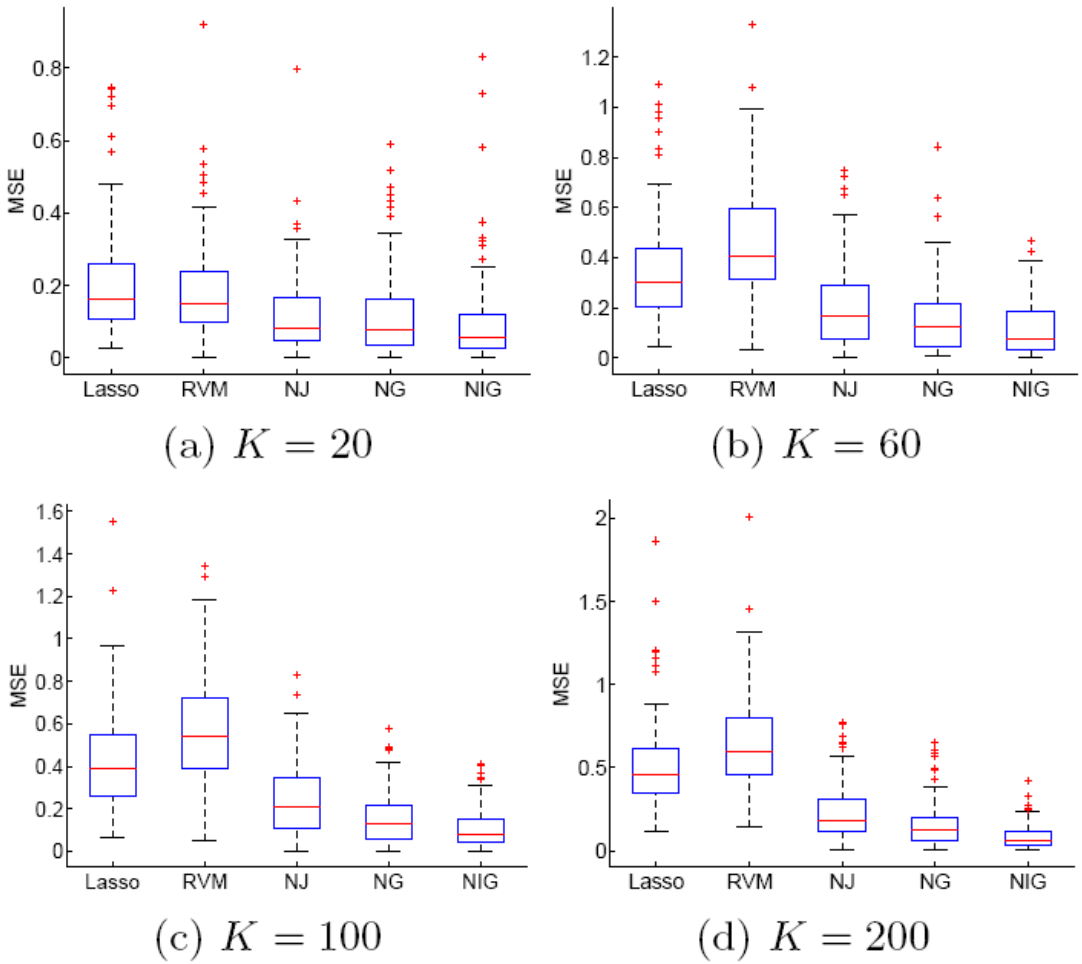


Figure: Box plots of the MSE associated to the simulated data.

Conclusion

Why this title is not so bad

- Sparse...

Conclusion

Why this title is not so bad

- Sparse...
  - Two classes of models which lead to sparser estimates

Conclusion

Why this title is not so bad

- Sparse...
  - Two classes of models which lead to sparser estimates
- Bayesian nonparametric...

Conclusion

Why this title is not so bad

- Sparse...
  - Two classes of models which lead to sparser estimates
- Bayesian nonparametric...
  - Related to a class of nonparametric Bayesian model when $K \to \infty$

Conclusion

Why this title is not so bad

- Sparse...
  - Two classes of models which lead to sparser estimates
- Bayesian nonparametric...
  - Related to a class of nonparametric Bayesian model when $K \to \infty$
- regression...

Conclusion

Why this title is not so bad

- Sparse...
  - Two classes of models which lead to sparser estimates
- Bayesian nonparametric...
  - Related to a class of nonparametric Bayesian model when $K \to \infty$
- regression...
  - Extension to probit regression

Conclusion

- Ongoing work with K. Murphy on graph learning with group sparsity

Conclusion

- Ongoing work with K. Murphy on graph learning with group sparsity
- How to use the stick-breaking construction?

Conclusion

- Ongoing work with K. Murphy on graph learning with group sparsity
- How to use the stick-breaking construction?
- Marginal distribution?

Bibliography


Barndorff-Nielsen, O. (1997).
Normal inverse Gaussian distributions and stochastic volatility modelling.
Scandinavian Journal of Statistics, *24*, 1–13.


Figueiredo, M. (2003).
Adaptive sparseness for supervised learning.
IEEE Transactions on Pattern Analysis and Machine Intelligence, *25*, 1150–1159.


Griffin, J., & Brown, P. (2007).
Bayesian adaptive lasso with non-convex penalization (Technical Report).
Dept of Statistics, University of Warwick.