Sparse Bayesian Nonparametric Regression

François Caron CARONFR@CS.UBC.CA
Arnaud Doucet ARNAUD@CS.UBC.CA
Departments of Computer Science and Statistics, University of British Columbia, Vancouver, Canada

Abstract

One of the most common problems in machine learning and statistics consists of estimating the mean response $X\beta$ from a vector of observations y assuming $y = X\beta + \varepsilon$ where X is known, β is a vector of parameters of interest and ε a vector of stochastic errors. We are particularly interested here in the case where the dimension K of β is much higher than the dimension of y. We propose some flexible Bayesian models which can yield sparse estimates of β. We show that as $K \to \infty$ these models are closely related to a class of Lévy processes. Simulations demonstrate that our models outperform significantly a range of popular alternatives.

1. Introduction

Consider the following linear regression model

$$y = X\beta + \varepsilon \tag{1}$$

where $y \in \mathbb{R}^L$ is the observation, $\beta = (\beta_1, \ldots, \beta_K) \in \mathbb{R}^K$ is the vector of unknown parameters, X is an known $L \times K$ matrix. We will assume that ε follows a zero-mean normal distribution $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_L\right)$ where I_L is the identity matrix of dimension L.

We do not impose here any restriction on L and K but we are particularly interested in the case where $K >> L$. This scenario is very common in many application domains. In such cases, we are interested in obtaining a sparse estimate of β; that is an estimate $\widehat{\beta} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_K)$ such that only a subset of the components $\widehat{\beta}_k$ differ from zero. This might be for sake of variable selection (Tibshirani, 1996; Figueiredo, 2003; Griffin & Brown, 2007) or to decompose a signal over an overcomplete basis (Lewicki & Sejnowski, 2000; Chen et al., 2001).

Numerous models and algorithms have been proposed in the machine learning and statistics literature to address this problem including Bayesian stochastic search methods based on the 'spike and slab' prior (West, 2003), Lasso (Tibshirani, 1996), projection pursuit or the Relevance Vector Machine (RVM) (Tipping, 2001). We follow here a Bayesian approach where we set a prior distribution on β and we will primarily focus on the case where $\widehat{\beta}$ is the resulting Maximum a Posteriori (MAP) estimate or equivalently the Penalized Maximum Likelihood (PML) estimate. Such MAP/PML approaches have been discussed many times in the literature and include the Lasso (the corresponding prior being the Laplace distribution) (Tibshirani, 1996; Lewicki & Sejnowski, 2000; Girolami, 2001), the normal-Jeffreys (NJ) prior (Figueiredo, 2003) or the normal-exponential gamma prior (Griffin & Brown, 2007). Asymptotic theoretical properties of such PML estimates are discussed in (Fan & Li, 2001).

We propose here a class of prior distributions based on scale mixture of Gaussians for β. For a finite K, our prior models correspond to normal-gamma (NG) and normal-inverse Gaussian (NIG) models. This class of models includes as limiting cases both the popular Laplace and normal-Jeffreys priors but is more flexible. As $K \to \infty$, we show that the proposed priors are closely related to the variance gamma and normal-inverse Gaussian processes which are Lévy processes (Applebaum, 2004). In this respect, our models are somehow complementary to two recently proposed Bayesian nonparametric models: the Indian buffet process (Ghahramani et al., 2006) and the infinite gamma-Poisson process (Titsias, 2007). Under given conditions, the normal-gamma prior yields sparse MAP estimates $\widehat{\beta}$. The log-posterior distributions associated to these prior distributions are not convex but we propose an Expectation-Maximization (EM) algorithm to find modes of the posteriors and

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

a Markov Chain Monte Carlo (MCMC) algorithm to sample from them. We demonstrate through simulations that these Bayesian models outperform significantly a range of established procedures on a variety of applications.

The rest of the paper is organized as follows. In Section 2, we propose the NG and NIG models for β. We establish some properties of these models for K finite and in the asymptotic case where $K \to \infty$. We also relate our model to the Indian buffet process (Ghahramani et al., 2006) and the infinite gamma-Poisson process (Titsias, 2007). In Section 3, we establish conditions under which the MAP/PML estimate $\widehat{\beta}$ can enjoy sparsity properties. Section 4 presents an EM algorithm to find modes of the posterior distributions and a Gibbs sampling algorithm to sample from them. We demonstrate the performance of our models and algorithms in Section 5. Finally we discuss some extensions in Section 6.

2. Sparse Bayesian Nonparametric Models

We will consider models where the components β are independent and identically distributed

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

and $p(\beta_k)$ is a scale mixture of Gaussians; that is

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

where $\mathcal{N}(x; \mu, \sigma^2)$ denotes the Gaussian distribution of argument x, mean μ and variance σ^2. We propose two conjugate distributions for σ_k^2; namely the gamma and the inverse Gaussian distributions. The resulting marginal distribution for β_k belongs in both cases to the class of generalized hyperbolic distributions.

In the models presented here, the unknown scale parameters are random and integrated out so that the marginal priors on the regression coefficients are not Gaussian. This differs from the RVM (Tipping, 2001) where these parameters are unknown and estimated through maximum likelihood.

2.1. Normal-Gamma Model

2.1.1. Definition

Consider the following gamma prior distribution

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

whose probability density function (pdf) $\mathcal{G}(\sigma_k^2; \frac{\alpha}{K}, \frac{\gamma^2}{2})$ is given by

$$\frac{(\frac{\gamma^2}{2})^{\frac{\alpha}{K}}}{\Gamma(\frac{\alpha}{K})} (\sigma_k^2)^{\frac{\alpha}{K}-1} \exp(-\frac{\gamma^2}{2}\sigma_k^2).$$

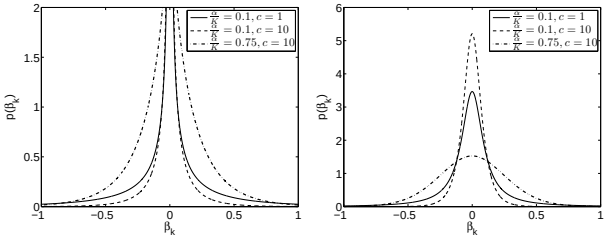
Following Eq. (2), the marginal pdf of β_k is given for $\beta_k \neq 0$ by

$$p(\beta_k) = \frac{\gamma^{\alpha/K+1/2}}{\sqrt{\pi} 2^{\alpha/K-1/2} \Gamma(\frac{\alpha}{K})} |\beta_k|^{\frac{\alpha}{K}-\frac{1}{2}} \mathcal{K}_{\frac{\alpha}{K}-\frac{1}{2}}(\gamma|\beta_k|) \qquad (3)$$

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

$$\lim_{\beta_k \to 0} p(\beta_k) = \begin{cases} \frac{\gamma}{2\sqrt{\pi}} \frac{\Gamma(\frac{\alpha}{K}-\frac{1}{2})}{\Gamma(\frac{\alpha}{K})} & \text{if } \frac{\alpha}{K} > \frac{1}{2} \\ \infty & \text{otherwise} \end{cases}$$

and the tails of this distribution decrease in $|\beta_k|^{\frac{\alpha}{K}-1} \exp(-\gamma|\beta_k|)$, see Figure 1(a). The parameters α and γ resp. control the shape and scale of the distribution. When $\alpha \to 0$, there is a high discrepancy between the values of σ_k^2, while when $\alpha \to \infty$, most of the values are equal.



(a) Normal-gamma (b) Normal-inverse Gaussian

Figure 1. Probability density functions of the NG and NIG for different values of the parameters.

This class of priors includes many standard priors. Indeed, Eq. (3) reduces to the Laplace prior when $\frac{\alpha}{K} = 1$ and we obtain the NJ prior when $\frac{\alpha}{K} \to 0$ and $\gamma \to 0$.

In Figure 2 some realizations of the process are given for different values $\alpha = 1, 5, 100$ and $\gamma^2/2 = \alpha$.

2.1.2. Properties

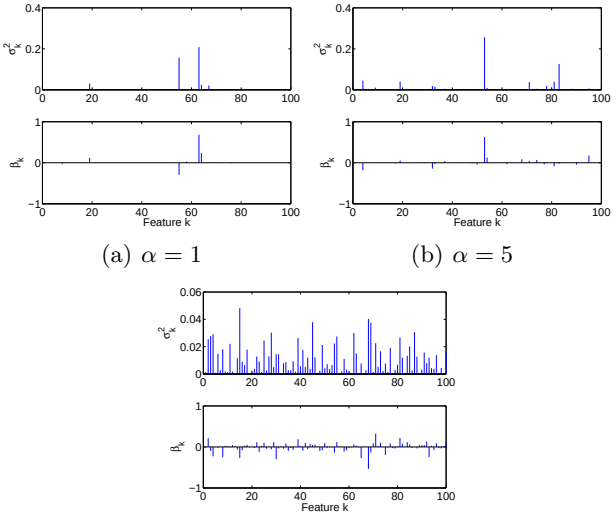
It follows from Eq. (3) that

$$\mathbb{E}[|\beta_k|] = \sqrt{\frac{4}{\pi\gamma^2}} \frac{\Gamma(\frac{\alpha}{K}+\frac{1}{2})}{\Gamma(\frac{\alpha}{K})}, \mathbb{E}[\beta_k^2] = \frac{2\alpha}{\gamma^2 K}$$

and we obtain

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

Hence the sum of the terms remains bounded whatever being K.



(a) $\alpha = 1$ (b) $\alpha = 5$

(c) $\alpha = 100$

Figure 2. Realizations (top) $\left\{\sigma_k^2\right\}_{k=1,\dots,K}$ and (bottom) $\{\beta_k\}_{k=1,\dots,K}$ from the NG model for $\alpha = 1, 5, 100$.

Using properties of the gamma distribution, it is possible to relate β to a Lévy process known as the variance gamma process as $K \to \infty$. First consider a finite K. Let $\sigma^2_{(1)} \geq \sigma^2_{(2)} \geq \ldots \geq \sigma^2_{(K)}$ be the order statistics of the sequence $\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2$ and let $\pi_1, \ldots, \pi_K$ be random variables verifying the following (finite) stick-breaking construction

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

where $\mathcal{B}$ is the Beta distribution. Finally if $g \sim \mathcal{G}(\alpha, \frac{\gamma^2}{2})$ then we can check that the order statistics $\left(\sigma^2_{(k)}\right)$ follow the same distribution as the order statistics of $(g\pi_k)$. The characteristic function of β_k is given by

$$\Phi_{\beta_k}(u) = \frac{1}{\left(1 - \frac{iu}{\gamma}\right)^{\frac{\alpha}{K}}} \frac{1}{\left(1 + \frac{iu}{\gamma}\right)^{\frac{\alpha}{K}}}$$

and therefore

$$\beta_k \stackrel{d}{=} w_1 - w_2 \text{ where } w_1 \sim \mathcal{G}(\frac{\alpha}{K},\gamma) \text{ and } w_2 \sim \mathcal{G}(\frac{\alpha}{K},\gamma)$$

It follows that β_k can be written as the difference of two variables following a gamma distribution.

As $K \to \infty$, the order statistics $\left(\sigma^2_{(k)}\right)$ are the conic part of a gamma process with shape parameter α and scale parameter $\gamma^2/2$; see (Tsilevich et al., 2000) for details. In particular $\overline{\sigma}^2 = \left(\frac{\sigma^2_{(1)}}{\sum_k \sigma^2_{(k)}}, \frac{\sigma^2_{(2)}}{\sum_k \sigma^2_{(k)}}, \ldots\right)$ and $\sum_k \sigma^2_{(k)}$ are independent and respectively distributed according to $PD(\alpha)$ and $\mathcal{G}(\alpha, \gamma^2/2)$ where $PD(\alpha)$ is the Poisson-Dirichlet distribution of scale parameter α. It is well-known that this distribution can be recovered by the following (infinite) stick-breaking construction (Tsilevich et al., 2000) as if we set

$$\pi_k = \zeta_k \prod_{j=1}^{k-1}(1-\zeta_j) \text{ with } \zeta_j \sim \mathcal{B}(1, \alpha) \quad (5)$$

for any k then the order statistics $\left(\pi_{(k)}\right)$ are distributed from the Poisson-Dirichlet distribution.

The coefficients (β_k) are thus nothing but the weights (jumps) of the so-called variance gamma process which is a Brownian motion evaluated at times given by a gamma process (Applebaum, 2004; Madan & Seneta, 1990).

2.2. Normal-Inverse Gaussian Model

2.2.1. Definition

Consider the following inverse Gaussian prior distribution

$$\sigma_k^2 \sim \mathcal{IG}(\frac{\alpha}{K}, \gamma) \quad (6)$$

whose pdf $\mathcal{IG}(\sigma_k^2; \frac{\alpha}{K}, \gamma)$ is given by (Barndorff-Nielsen, 1997)

$$\frac{1}{\sqrt{2\pi}} \frac{\alpha}{K} \exp(\gamma \frac{\alpha}{K})(\sigma_k^2)^{-3/2} \exp(-\frac{1}{2}(\frac{\alpha^2}{K^2\sigma_k^2} + \gamma^2\sigma_k^2)) \quad (7)$$

Following Eq. (2), the marginal pdf of β_k is given

$$\frac{\alpha\gamma}{\pi K} \exp(\frac{\alpha\gamma}{K}) \left(\frac{\alpha^2}{K^2} + \beta_k^2\right)^{-\frac{1}{2}} \mathcal{K}_1\left(\gamma\sqrt{\frac{\alpha^2}{K^2} + \beta_k^2}\right) \quad (8)$$

and the tails of this distribution decrease in $|\beta_k|^{-3/2} \exp(-\gamma|\beta_k|)$. It is displayed in Figure 1(b). The parameters α and γ resp. control the shape and scale of the distribution. When $\alpha \to 0$, there is a high discrepancy between the values of σ_k^2, while when $\alpha \to \infty$, most of the values are equal. Some realizations of the model, for different values of α are represented in Figure 3.

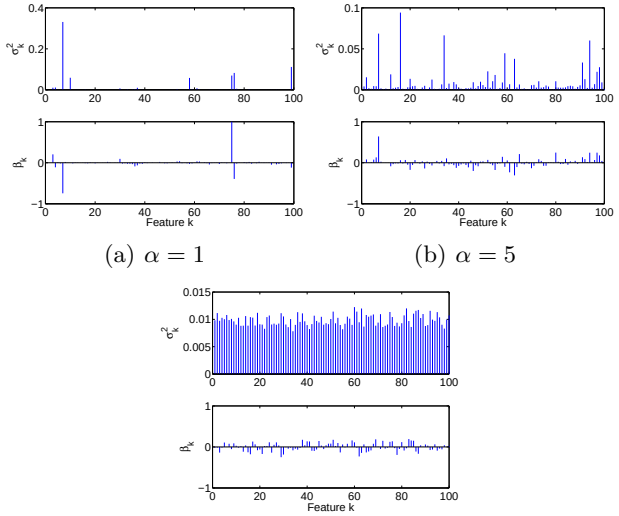
2.2.2. Properties

The moments are given

$$\mathbb{E}[|\beta_k|] = \frac{2\alpha}{K\pi} \exp(\frac{\gamma\alpha}{K}) \mathcal{K}_0(\frac{\alpha\gamma}{K}), \; \mathbb{E}[\beta_k^2] = \frac{\alpha}{K\gamma}$$

Therefore, as $K \to \infty$, the mean of sum of the absolute values is infinite while the sum of the square is $\frac{\alpha}{\gamma}$.

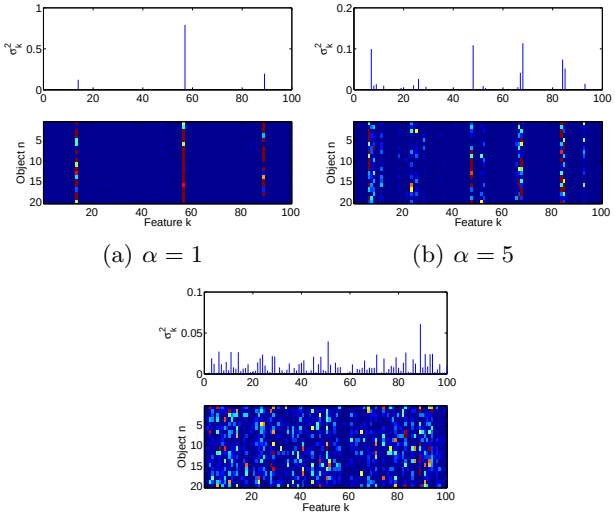
We can also establish in this case that the coefficients (β_k) tend to weights (jumps) of a normal-inverse Gaussian process (Barndorff-Nielsen, 1997).



(a) $\alpha = 1$ (b) $\alpha = 5$

(c) $\alpha = 100$

Figure 3. Realizations (top) $(\sigma_k^2)_{k=1,\dots,K}$ and (bottom) $(\beta_k)_{k=1,\dots,K}$ from the NIG model for $K = 100$, $N = 20$, $\alpha = 1, 10, 100$ and $\gamma = \alpha$.



(a) $\alpha = 1$ (b) $\alpha = 5$

(c) $\alpha = 100$

Figure 4. Realizations (top) $(\sigma_k^2)_{k=1,\dots,K}$ and (bottom) $(\beta_k^n)_{n=1,\dots,N,k=1,\dots,K}$ from the normal-gamma model for $K = 100$, $N = 20$, $\alpha = 1, 10, 100$ and $\gamma^2/2 = \alpha$. The lighter the colour, the larger $|\beta_k^n|$.

2.3. Extension

Consider now the case where we have N vectors of observations $\{y_n\}_{n=1}^N$ where $y_n \in \mathbb{R}^L$. We would like to model the fact that for a given k the random variables $\{\beta_k^n\}_{n=1}^N$ are statistically dependent and exchangeable. We consider the following 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 $k = 1, \dots, K$ and

$$\beta_k^n \sim \mathcal{N}(0, \sigma_k^2)$$

for $n = 1, \dots, N$. Some realizations of the process for different values $\alpha = 1, 5, 100$ are represented in Figure 4.

In this respect, this work is complementary to two recently proposed Bayesian nonparametric models: the Indian buffet process (Ghahramani et al., 2006) and the infinite gamma-Poisson process (Titsias, 2007). In these two contributions, prior distributions over infinite matrices with integer-valued entries are defined. These models are constructed as the limits of finite-dimensional models based respectively on the beta-binomial and gamma-Poisson models. They enjoy the following property: while the number of non-zero entries of an (infinite) row is potentially infinite, the expected number of these entries is finite. These models are also closely related to the beta and gamma processes which are Lévy processes (Applebaum, 2004; Teh et al., 2007; Thibaux & Jordan, 2007). Our models could be interpreted as prior distributions over infinite matrices with real-valued entries. In our case, the number of non-zero entries of an (infinite) row is always infinite but we can have

$$\lim_{K\to\infty} \mathbb{E}\left[\sum_{k=1}^{K} |\beta_k^n|^\rho\right] < \infty \tag{9}$$

for $\rho = 1$ or $\rho = 2$. Morever for some values of $\frac{\alpha}{K}$ and γ we can also ensure that for any $x > 0$

$$\lim_{K\to\infty} \Pr\left(\exists k : |\beta_k^n| > x\right) > 0; \tag{10}$$

that is there is still a non-vanishing probability of having coefficients with large values as $K \to \infty$ despite Eq. (9).

The joint distribution is given by $p(\beta_{1:K}^{1:N}) = \prod_{k=1}^K p(\beta_k^{1:N})$ where for the NG model

$$p(\beta_k^{1:N}) \propto u_k^{\frac{\alpha}{K}-\frac{N}{2}} \mathcal{K}_{\frac{\alpha}{K}-\frac{N}{2}}(\gamma u_k)$$

and for the NIG model

$$p(\beta_k^{1:N}) \propto (q_k)^{-(N+1)/2} \mathcal{K}_{\frac{N+1}{2}}(\gamma q_k)$$

where

$$u_k = \sqrt{\sum\nolimits_{n=1}^{N} (\beta_k^n)^2}, \; q_k = \sqrt{\frac{\alpha^2}{K^2} + u_k^2} \tag{11}$$

3. Sparsity Properties

Further on we will also use the following notation for any random variable u

$$pen(u) \equiv \log(p(u))$$

Table 1. Penalizations and their derivatives for different prior distributions

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

where '$\equiv$' denotes equal up to an additive constant independent of u. When computing the MAP/PML estimate for N data, we select

$$\widehat{\beta}^{1:N} = \arg\min_{\beta^{1:N}} \sum_{n=1}^{N} \frac{\|y_n - X\beta^n\|_2^2}{2\sigma^2} - \sum_{k=1}^{K} pen(\beta_k^{1:N}). \tag{12}$$

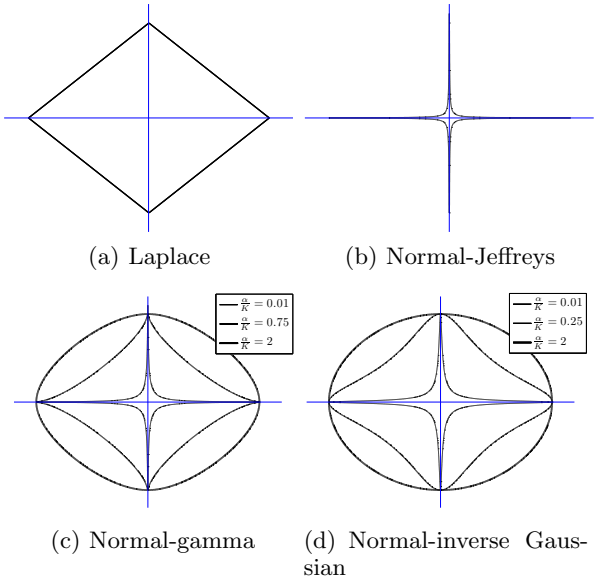
We give in Table 1 the penalizations $pen(\beta_k^{1:N})$ and their derivatives for different prior distributions as a function of u_k and q_k defined in Eq. (11).

When $\alpha/K = 1$, the NG prior is equal to the Laplace prior so its penalization reduces to the ℓ_1 penalization used in Lasso and basis pursuit (Tibshirani, 1996; Chen et al., 2001). When $\alpha/K \to 0$ and $c \to 0$ the prior is the NJ prior and the penalization reduces to $\log(|\beta_k|)$ which has been used in (Figueiredo, 2003). We display in Figure 5 the contours of constant value for various prior distributions when $N = 1$ and $K = 2$. For $\alpha/K < 1/2$, the MAP estimate (12) does not exist as the pdf (3) is unbounded. For other values of the parameters, a mode can dominate at zero whereas we are interested in the data driven turning point/local minimum (Griffin & Brown, 2007).

Consider now the case where the matrix X is orthogonal, $\sigma = 1$ and $N = 1$. The turning point and/or MAP/PML estimate is obtained by minimizing Eq. (12) which is equivalent to minimize componentwise

$$\frac{1}{2}(z_k - \beta_k)^2 + pen(\beta_k) \tag{13}$$

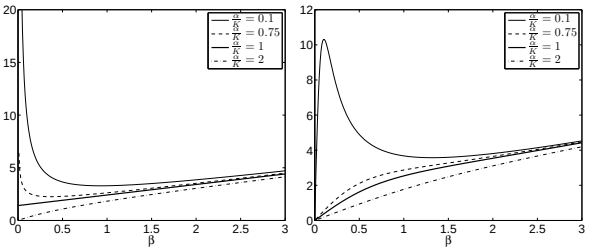
where $z = X^T y$. The first derivative of (13) is $sign(\beta_k)\,(|\beta_k| + pen'(|\beta_k|)) - z_k$. As stated in (Fan & Li, 2001, p. 1350), a sufficient condition for the estimate to be a thresholding rule is that the minimum of the function $|\beta_k| + pen'(|\beta_k|)$ is strictly positive. Plots of the function $|\beta_k| + pen'(|\beta_k|)$ are given in Figure 6 and the resulting thresholds corresponding to the argument minimizing (13) are presented in Figure 7. It follows that 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 as the derivative of the penalization is 0 when $\beta_k = 0$ whatever being the values of the parameters. However, from Figure 7(d), it is clear that it can yield "almost sparse" estimates; that is most components are such that $\left|\widehat{\beta}_k\right| \simeq 0$.



(a) Laplace (b) Normal-Jeffreys

(c) Normal-gamma (d) Normal-inverse Gaussian

Figure 5. Contour of constant value of $pen(\beta_1) + pen(\beta_2)$ for different prior distributions.



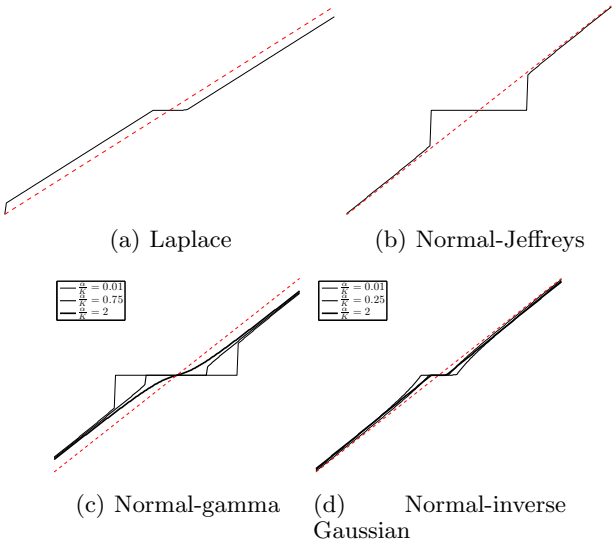
(a) Normal-gamma (b) Normal-inverse Gaussian

Figure 6. Plots of $|\beta_k| + pen'(|\beta_k|)$.

4. Algorithms

4.1. EM

The log-posterior in Eq. (12) is not concave but we can use the EM algorithm to find modes of it. The EM algorithm relies on the introduction of the missing data $\sigma_{1:K} = (\sigma_1, ..., \sigma_K)$. Conditional upon these missing data, the regression model is linear Gaussian and all the EM quantities can be easily computed in closed form; see for example (Figueiredo, 2003; Griffin & Brown, 2007). We have at iteration $i+1$ of the EM



(a) Laplace (b) Normal-Jeffreys

(c) Normal-gamma (d) Normal-inverse Gaussian

Figure 7. Thresholds for the different prior distributions.

$$\widehat{\beta}^{1:N}_{(i+1)} = \arg\max_{\beta_{1:N}} Q(\beta^{1:N}; \widehat{\beta}^{1:N}_{(i)})$$

where $Q(\beta^{1:N}; \widehat{\beta}^{1:N}_{(i)})$ is given by

$$\int \log(p(\beta^{1:N}|y_{1:N}, \sigma_{1:K})).p(\sigma_{1:K}|\widehat{\beta}^{1:N}_{(i)}, y_{1:N})d\sigma_{1:K}.$$

After a few calculations, we obtain

$$\widehat{\beta}^{n}_{(i+1)} = \left(\sigma^2 V_{(i)} + X^T X\right)^{-1} X^T y_n$$

with $V_{(i)} = diag(m_{1,(i)}, \ldots, m_{K,(i)})$ and $m_{k,(i)} = \left(\widehat{u}_{k,(i)}\right)^{-1} pen'(\widehat{u}_{k,(i)})$ where $\widehat{u}_{k,(i)} = \sqrt{\sum_{n=1}^{N}\left(\widehat{\beta}^{n}_{k,(i)}\right)^2}$, $pen'(\widehat{u}_{k,(i)}) = \frac{\partial pen(u_k)}{\partial u_k}\Big|_{\widehat{u}_{k,(i)}}$ (see Table 1).

4.2. MCMC

We can also easily sample from the posterior distribution $p(\beta^{1:N}|y_{1:N})$ by sampling from $p(\beta^{1:N}, \sigma^2_{1:K}|y_{1:N})$ using the Gibbs sampler. Indeed the full conditional distributions $p(\beta^{1:N}|\sigma_{1:K}, y_{1:N})$ and $p(\sigma^2_{1:K}|\beta^{1:N}, y_{1:N})$ are available in closed-form. The distribution $p(\beta^{1:N}|\sigma_{1:K}, y_{1:N})$ is a multivariate normal whereas we have $p(\sigma^2_{1:K}|\beta^{1:N}, y_{1:N}) = \prod_{k=1}^{K} p(\sigma^2_k|\beta^{1:N}_k)$. For the NG prior, we obtain

$$p(\sigma^2_k|\beta^{1:N}_k) = \frac{(\sigma^2_k)^{\frac{\alpha}{K}-\frac{N}{2}-1}\exp\left(-\frac{1}{2}\frac{u^2_k}{\sigma^2_k} - \gamma\sigma^2_k\right)}{2\left(\frac{u_k}{\gamma}\right)^{\frac{\alpha}{K}-\frac{N}{2}}\mathcal{K}_{\frac{\alpha}{K}-\frac{N}{2}}(\gamma u_k)}$$

which is a generalized inverse Gaussian distribution from which we can sample exactly. For the NIG distribution, we also obtain a generalized inverse Gaussian distribution.

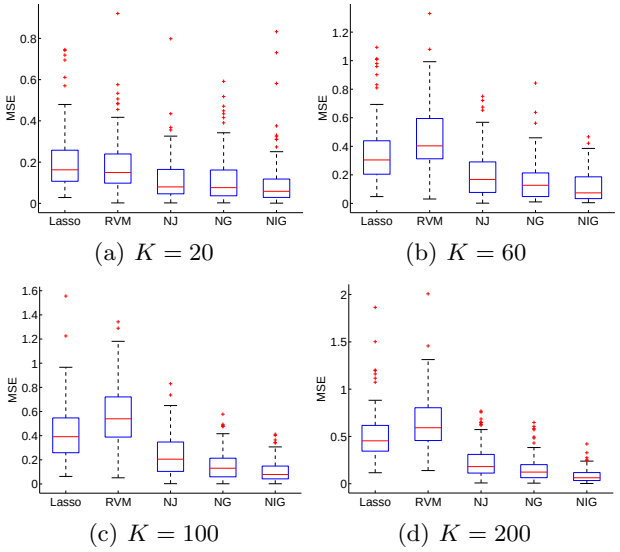
5. Applications

5.1. Simulated Data

In the following, we provide numerical comparisons between the Laplace (that is Lasso), the RVM, NJ, NG and NIG models. We simulate 100 datasets from (1) with $L = 50$ and $\sigma = 1$. The correlation between $X_{k,i}$ and $X_{k,j}$ is $\rho^{|i-j|}$ with $\rho = 0.5$. We set $\beta = (3\ 1.5\ 0\ 0\ 2\ 0\ 0 \ldots)^T \in \mathbb{R}^K$ where the remaining components of the vector are set to zero. We consider the cases where $K = 20, 60, 100, 200$. Parameters of the Lasso, NG and NIG are estimated by 5-fold cross-validation, as described in (Tibshirani, 1996). The Lasso estimate is obtained with the Matlab implementation of the interior point method downloadable at http://www.stanford.edu/~boyd/l1_ls/. For the other priors, the estimate is obtained via 100 iterations of the EM algorithm. Box plots of the mean square error (MSE) are reported in Figure 8. These plots show that the performance of the estimators based on the NG and NIG priors outperform those of classical models in that case. In Figure 9 are represented the box plots of the number of estimated coefficients whose absolute value is below T, $T = 10^{-10}$ (the precision tuned for the Lasso estimate) and $T = 10^{-3}$, for $K = 200$. The true number of zeros in that case is 197. The NG outperforms the other models in identifying the zeros of the model. On the contrary, as the NIG estimate is not a thresholding rule, the median number of coefficients whose absolute value is below 10^{-10} for this model is zero. However, most of the coefficients have a very low absolute value, as the median of the coefficients with absolute value below 10^{-3} is equal to the true value 197 (see Figure 9(b)). Moreover, the estimator obtained by thresholding the coefficients whose absolute value is below 10^{-3} to zero yields very minor differences in terms of MSE.

5.2. Biscuit NIR Dataset

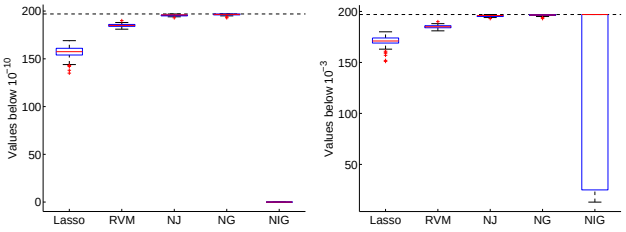
We consider the biscuits data which have been studied in (Griffin & Brown, 2007; West, 2003). The matrix X is composed of 300 (centered) NIR reflectance measurements from 70 biscuit dough pieces. The observations y are the percentage of fat, sucrose, flour and water associated to each piece. The objective here is to predict the level of each of the ingredients from the NIR reflectance measurements. The data are divided into a training dataset (39 measurements) and a test dataset (31 measurements). The fitted coefficients of fat and flour, using 5-fold cross-validation, are represented in Figure 10. The estimated spikes are consistent with the results obtained in (West, 2003; Griffin & Brown, 2007). In particular, both models detect



(a) $K = 20$ (b) $K = 60$

(c) $K = 100$ (d) $K = 200$

Figure 8. Box plots of the MSE associated to the simulated data.



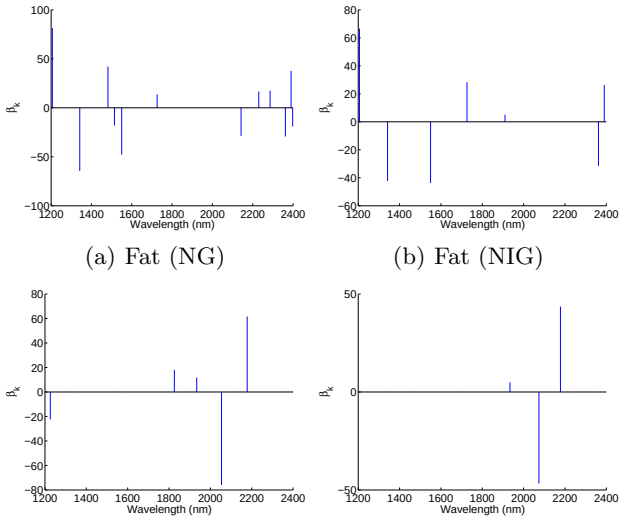
(a) $T = 10^{-10}$ (b) $T = 10^{-3}$

Figure 9. Box plots of the number of estimated coefficients whose absolute value is below a threshold T. Dash line represents the true value of zero coefficients (197).

a spike at 1726nm, which lies in a region known for fat absorbance. The predicted observations versus the true observations are given in Figure 11 for the training and test datasets. The test data are well fitted by the estimated coefficients. MSE errors for the test dataset are reported in Table 2. The proposed models show better performances for flour and similar performances for fat.

6. Discussion

We have presented some flexible priors for linear regression based on the NG and NIG models. The NG prior yields sparse local maxima of the posterior distribution whereas the NIG prior yields "almost sparse" estimates; that is most of the coefficients are extremely close to zero. We have shown that asymptotically these models are closely related to the variance gamma process and the normal-inverse Gaussian process. Contrary to the NJ model or the RVM,



(a) Fat (NG) (b) Fat (NIG)

(c) Flour (NG) (d) Flour (NIG)

Figure 10. Coefficients estimated with a normal-gamma (left) and normal-inverse Gaussian (right) prior for fat (top) and flour (bottom) ingredients.

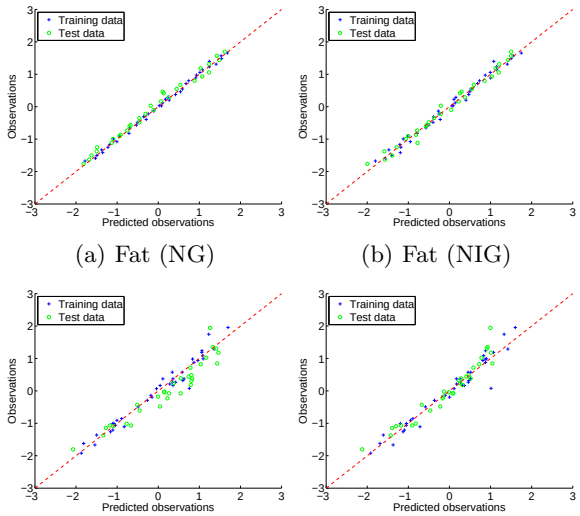
Table 2. MSE for biscuits NIR data

	Flour	Fat
NJ	9.93	0.56
RVM	6.48	0.56
NG	3.44	0.55
NIG	1.94	0.49

these models require specifying two hyperparameters. However, using a simple cross-validation procedure we have demonstrated that these models can perform significantly better that well-established procedures. In particular, the experimental performance of the NIG model are surprisingly good and deserve being further studied. The NG prior has been discussed in (Griffin & Brown, 2007). It was discarded because of its spike at zero and the flatness of the penalty for large values but no simulations were provided. They favour another model which relies on a cylinder parabolic function[1]. The NG prior has nonetheless interesting asymptotic properties in terms of Lévy processes and we have demonstrated its empirical performances. The NG, NIG and Laplace priors can also be considered as particular cases of generalized hyperbolic distributions. This class of distributions has been used in (Snoussi & Idier, 2006) for blind source separation.

The extension to (probit) classification is straightfor-

[1]The authors provide a link to a program to compute this function. Unfortunately, it is extremely slow. The resulting algorithm is at least one order of magnitude slower than our algorithms which rely on Bessel functions.



(a) Fat (NG) (b) Fat (NIG)

(c) Flour (NG) (d) Flour (NIG)

Figure 11. Observations versus predicted observations estimated with a normal-gamma (left) and normal-inverse Gaussian (right) prior for fat (top) and flour (bottom) ingredients.

ward by adding latent variables corresponding to the regression function plus some normal noise. Computationally it only requires adding one line in the EM algorithm and one simulation step in the Gibbs sampler.

References

Applebaum, D. (2004). *Lévy processes and stochastic calculus*. Cambridge University Press.

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

Chen, S., Donoho, D., & Saunders, M. (2001). Atomic decomposition by basis pursuit. *SIAM review*, *43*, 129–159.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

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

Ghahramani, Z., Griffiths, T., & Sollich, P. (2006). Bayesian nonparametric latent feature models. *Proceedings Valencia/ISBA World meeting on Bayesian statistics*.

Girolami, M. (2001). A variational method for learning sparse and overcomplete representations. *Neural computation*, *13*, 2517–2532.

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

Lewicki, M. S., & Sejnowski, T. (2000). Learning overcomplete representations. *Neural computation*, *12*, 337–365.

Madan, D., & Seneta, E. (1990). The variance-gamma model for share market returns. *Journal of Business*, *63*, 511–524.

Snoussi, H., & Idier, J. (2006). Bayesian blind separation of generalized hyperbolic processes in noisy and underdeterminate mixtures. *IEEE Transactions on Signal Processing*, *54*, 3257–3269.

Teh, Y., Gorur, D., & Ghahramani, Z. (2007). Stick-breaking construction for the Indian buffet process. *International Conference on Artificial Intelligence and Statistics*.

Thibaux, R., & Jordan, M. (2007). Hierarchical beta processes and the Indian buffet process. *International Conference on Artificial Intelligence and Statistics*.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society B*, *58*, 267–288.

Tipping, M. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, *211-244*, 211–244.

Titsias, M. (2007). The infinite gamma-Poisson feature model. *International Conference on Neural Information Processing Systems*.

Tsilevich, N., Vershik, A., & Yor, M. (2000). *Distinguished properties of the gamma process, and related topics* (Technical Report). Laboratoire de Probabilités et Modèles aléatoires, Paris.

West, M. (2003). Bayesian factor regression models in the "Large p, Small n" paradigm. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith and M. West (Eds.), *Bayesian statistics 7*, 723–732. Oxford University Press.