An Approximate Bayesian Computational Approach to Stable Priors for Linear Regression Model

Genya Kobayashi and Hideo Kozumi *

August 1, 2012

Abstract

The symmetric alpha stable (SAS) prior has appeared in the image processing field as a wavelet shrinker. The SAS prior includes the normal and Cauchy priors as special cases and has some desirable properties for the shrinkage problem such as spike at zero and heavy tail. However, the use of the SAS prior is currently quite limited, because it does not allow a closed form density. The non-existence of a closed form density causes to difficulty in estimating the hyperparameter, which plays an important role for shrinkage to be data adaptive. We propose a convenient way to estimate the regression model under the prior distribution with intractable density with the hyperparameter estimated from the data. We develop an Markov chain Monte Carlo algorithms coupled with the technique of approximate Bayesian computation. Our method can be applied to the case of the generalised Linnik prior, which can be regarded as an extension of the SAS prior. The proposed method is demonstrated using the simulated datasets and the real dataset.

Keywords: α stable distribution; Approximate Bayesian computation; Likelihood-free method; Linnik distribution; Markov chain Monte Carlo; Shrinkage prior; Symmetric α stable distribution;

Address for correspondence: Graduate School of Business Administration, Kobe University, 2-1 Rokkodai, Nada-ku, Kobe 657-8501, Japan (E-mail:kozumi@kobe-u.ac.jp).

1 Introduction

Consider the linear regression model given by

$$y_i = \mu + \mathbf{x}_i'\boldsymbol{\beta} + e_i, \quad i = 1, \ldots, n, \tag{1}$$

where y_i is the dependent variable, μ is the intercept parameter, $\mathbf{x}_i$ is the $k \times 1$ vector of independent variables, $\boldsymbol{\beta}$ is the coefficient parameter, and e_i is the error term which independently follows $\mathcal{N}(0, \sigma^2)$. We do not impose any restriction on p and n. In many fields of application, there is a growing interest in the case where $p >> n$. For example, in the economic growth model, the number of nations n can be no larger than approximately 200 while the independent variables include geographical, demographic, and policy variables and p is not restricted. In the gene microarray analysis, there are thousands of candidate genes to explain the relationship with a certain phenotype, but the number of subjects is quite limited. In such a case, our goal is to obtain a sparse estimate $\hat{\boldsymbol{\beta}}$ such that only a subset of its components differs from zero.

In the Bayesian framework, extracting the nonzero components is achieved through placing a prior distribution on $\boldsymbol{\beta}$. A shrinkage prior ought to have a certain property, as studied by Polson and Scott (2010), such as tail robustness and predictive efficiency. Park and Casella (2008) and Hans (2009) proposed the Bayesian Lasso based on the double exponential (Laplace) distribution. It is known that the relatively light tails of the double exponential prior causes overshrinking of large coefficients. Caron and Doucet (2008) and Griffin and Brown (2010) considered the normal-gamma prior which includes the Bayesian Lasso as a special case. The double Pareto prior considered by Armagan *et al*. (2012) permits a similar mixture representation to the normal gamma prior. Li and Lin (2009) and Hans (2011) proposed the Bayesian elastic net based on the scale mixture of normal representation. The bridge estimator based on the exponential power prior was proposed by Polson and Scott (2011a) using the scale mixture of normal representation and later was represented using the mixture of Bartlett-Fejer kernel by Polson and Scott (2011b). The horseshoe prior of Carvalho *et al*. (2010) used the Cauchy mixing distribution for the

scale mixture of normal and resulted in some desirable property as a shrinkage prior.

In this paper, we consider another class of prior distribution for $\boldsymbol{\beta}$, namely the symmetric α stable (SAS) prior and the symmetric generalised Linnik prior (SGL) distributions. The SAS prior has appeared in the image processing field as a wavelet shrinker (Achim *et al*., 2003; Boubchir and Fadili, 2006). Although SAS includes some interesting special cases, such as normal and Cauchy and presumably has a desirable property as a shrinkage prior, it has not gather much attention in the literature of shrinkage prior. This is particularly because that SAS does not have a closed form density. The non-existance of the prior density cases serious computational difficulty when we try to estimate the hyperparameter from the data in order that the shrinkage is data adaptive. The SGL prior can be regarded as an extension of the SAS prior. The flexible SGL has many known distributions as special cases, including the double exponential distribution. However, just as SAS, the density for SGL does not generally exist.

To overcome the computational difficulty due to the intractable prior density, we develop an Markov chain Monte Carlo (MCMC) algorithm coupled with the technique of approximate Bayesian computation (ABC). While ABC is usually used to approximate the intractable likelihood function (see Sisson and Fan, 2011; Marin *et al*. 2011). The ABC algorithm is intended to sample from the posterior distribution by finding parameter values under which the simulated values are close to the observed values. Since it is easy to simulate from SAS and SGL, our MCMC algorithm with ABC is expected to be computationally efficient. To our knowledge, this work is the first to apply the ABC method to approximate the intractable prior distribution. Although we use ABC in a irregular way approximating the intractable prior density, the simulation study shows that our method performs well under an appropriate choice of tuning parameter.

The rest of this paper is organised as follows. Section 2 introduces the SAS and SGL prior distributions and briefly look at the shrinkage pattern of those priors. In Section 3, we develop the MCMC algorithms coupled with ABC for the SAS and SGL priors. We conduct some simulation study in Section 4 to examine how ABC approximation to the posterior distribution of the hyperparameter influences

the performance of the shrinkage prior. In Section 5, we apply the priors to a high dimensional dataset and compare the performance with other existing prior distributions. Finally, we have some concluding remark in Section 6.

2 Stable priors

2.1 Symmetric α stable prior

The SAS distribution does not generally allow an analytical form of density function. Instead, it is best characterised by the characteristic function given by

$$\phi_{SAS}(t) = \exp\{-\tau|t|^{\alpha}\}, \tag{2}$$

where $\tau > 0$ is the scale parameter, and $\alpha \in (0, 2]$ is the parameter called characteristic exponent. The parameter α controls the heaviness of the tails. As it approaches to 2, the tails behave like normal distribution. Under small value of α, particularly for $\alpha < 1$, the tails decay more slowly and the distribution is more peaked around the centre (see *e.g.*, Burnecki *et al.*, 2008). In fact, SAS includes two important special cases, normal distribution when $\alpha = 2$ and Cauchy distribution when $\alpha = 1$.

SAS allows a convenient mixture of normal representation. If a random variable $X \sim \mathcal{S}\alpha S(\alpha, \tau)$, then

$$X = \sqrt{\tau^2\lambda}Z, \quad \lambda \sim \mathcal{S}^{+}\left(\frac{\alpha}{2}\right), \quad \mathcal{N}(0, 1), \tag{3}$$

where $\mathcal{S}^{+}(a)$ denotes the positive stable distribution with $a \in (0, 1)$.

2.2 Symmetric generalised Linnik prior

We consider another prior distribution, SGL prior. It is known that the density of Linnik distribution is more peaked at zero and heavier tail than SAS (Burnecki *et al.*, 2008). Thus, it is expected that SGL has better performance on controlling the noise. SGL is also defined only by the characteristic function,

$$\phi_{SGL}(t) = \frac{1}{(1 + \tau^{\alpha}|t|^{\alpha})^{\gamma}}, \tag{4}$$

where τ is the scale parameter, α is the characteristic exponent, and γ is the shape parameter. The extra γ parameter is expected to introduce further flexibility into the prior. SGL also includes some important special cases. This class of prior nests double Pareto, meridian, Cauchy, and double exponential (Polson and Scott, 2011b). Therefore, stable priors provide an alternative approach to an extension of the mixture of Bartlett-Fejer kernel for the Bayesian bridge.

SGL can be also represented using the scale mixture of normal. If $Y \sim \mathcal{SGL}(\alpha, \tau, \gamma)$, then

$$Y = \xi^{1/\alpha} \sqrt{\tau^2 \lambda} Z, \quad \xi \sim \mathcal{G}a(\gamma, 1), \quad \lambda \sim \mathcal{S}^+\left(\frac{\alpha}{2}\right), \quad Z \sim \mathcal{N}(0, 1), \tag{5}$$

see Devroye (1996). Therefore, it is possible to sample $\boldsymbol{\beta}$ from its full conditional distribution using Gibbs sampling.

2.3 Shrinkage coefficient

Apart from being tail robust (see Choy and Smith, 1997 for SAS), little is known about the shrinkage rules of the SAS and SGL priors. Here we have a look at the densities of the shrinkage coefficients $\kappa \in (0, 1)$ to get some intuitive understanding of their shrinkage rules. The behaviour of the density $f(\kappa)$ near $\kappa = 1$ will control the shrinkage of the noise while the behaviour near $\kappa = 0$ will control the tail robustness. As stated in Carvalho *et al*. (2010), a prior distribution with the U-shaped $f(\kappa)$ is desirable, since it can both keep the noise near zero and extract the signal without overshrinking.

Figure 1 shows $f(\kappa)$ for the SAS prior. For large values of α, for example, $\alpha = 1.5$ and 1, $f(\kappa)$ has little mass around $\kappa = 1$. This implies that under large α the prior may not sufficiently suppress the noise. On the other hand, when $\alpha < 1$, the shape of the density becomes close to U-shape. Figure 2 shows $f(\kappa)$ for the SGL prior with $\gamma = 0.5, 1$, and 2. Similar to the case of SAS, large value of α for SGL would lead to poor performance of the prior. When $\alpha > 1$, the density has either no mass or finite mass at $\kappa = 0$. Thus, under large α, the prior may overly shrink the signal. On the other hand, the SGL prior when $\alpha < 1$ results in the U-shaped density. Therefore, from our observations, for both SAS and SGL we may restrict the support of α, example to $(0, 1)$, for a complex problem as in our real data example.

3 Posterior inference

3.1 SAS prior

To proceed a posterior inference, we assume the independent SAS prior for the coefficient:

$$\beta_j \sim \mathcal{S}\alpha S(\alpha, \sigma/\nu), \quad j = 1, \ldots, p, \tag{6}$$

where the scale of the prior is also controlled by the error standard deviation. Furthermore, we assume

$$\mu \sim \mathcal{N}(0, M_0), \quad \sigma^2 \sim \mathcal{IG}\left(\frac{n_0}{2}, \frac{s_0}{2}\right), \quad \nu^2 \sim \mathcal{G}a\left(\frac{m_0}{2}, \frac{t_0}{2}\right). \tag{7}$$

Exploiting the scale of normal representation (3) for SAS, our MCMC algorithm samples form the full conditional distributions of $\tilde{\boldsymbol{\beta}} = (\mu, \beta_1, \ldots, \beta_p)'$, σ^2, ν^2, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$, and α. Except for α, it is not difficult to sample the values from the full conditional distributions as demonstrated by Tsionas (1999).

The full conditional distribution of $\tilde{\boldsymbol{\beta}}$ is given by

$$\tilde{\boldsymbol{\beta}} | \sigma^2, \nu^2, \boldsymbol{\lambda}, \alpha, \mathbf{y} \sim \mathcal{N}(\mathbf{b}_1, \mathbf{B}_1), \tag{8}$$

where $\mathbf{B}_1 = \left(\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{X}} + \boldsymbol{\Lambda}^{-1}\right)^{-1}$ and $\mathbf{b}_1 = \mathbf{B}_1\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\mathbf{y}$ with $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n)'$, $\tilde{\mathbf{x}}_i = (1, x_1, \ldots, x_p)'$, $\boldsymbol{\Sigma} = \text{diag}(\sigma^2, \ldots, \sigma^2)$, and $\boldsymbol{\Lambda} = \text{diag}(M_0, \sigma^2\lambda_1/\nu^2, \ldots, \sigma^2\lambda_p/\nu^2)$. The full conditional distribution of σ^2 is given by

$$\sigma^2 | \tilde{\boldsymbol{\beta}}, \nu^2, \boldsymbol{\lambda}, \alpha, \mathbf{y} \sim \mathcal{IG}\left(\frac{n_1}{2}, \frac{s_1}{2}\right), \tag{9}$$

where $n_1 = n_0 + n + p$ and $s_1 = s_0 + \sum_{i=1}^{n}(y_i - \tilde{\mathbf{x}}_i'\tilde{\boldsymbol{\beta}})^2 + \nu^2 \sum_{j=1}^{p} \beta_j^2/\lambda_j$. Similarly, the full conditional distribution of ν^2 is given by

$$\nu^2 | \tilde{\boldsymbol{\beta}}, \sigma^2, \boldsymbol{\lambda}, \alpha, \mathbf{y} \sim \mathcal{G}a\left(\frac{m_1}{2}, \frac{t_1}{2}\right), \tag{10}$$

where $m_1 = m_0 + p$ and $t_1 = t_0 + \frac{1}{\sigma^2}\sum_{j=1}^{p} \beta_j^2/\lambda_j$. The full conditional density of the local variance λ_j is given by

$$\pi(\lambda_j | \tilde{\boldsymbol{\beta}}, \sigma^2, \nu^2, \alpha, \mathbf{y}) \propto \lambda_j^{-1/2} \exp\left\{-\frac{\nu^2\beta_j}{2\sigma^2\lambda_j}\right\} f_{\alpha/2}(\lambda_j), \quad j = 1, \ldots, p, \tag{11}$$

where $f_a(\cdot)$ is the density of $\mathcal{S}^+(a)$. Although the density of positive stable distribution is not analytically tractable, it is possible to sample the values of λ_j. Following Tsionas (1999), we use independence Metropolis-Hastings (MH) algorithm with the proposal density $f_{\alpha/2}(\cdot)$.

After integrating out the local variances $\boldsymbol{\lambda}$, the full conditional density of α is given by

$$\pi(\alpha|\tilde{\boldsymbol{\beta}}, \sigma, \nu, \mathbf{y}) \propto \prod_{j=1}^{p} \pi(\beta_j|\alpha, \sigma, \nu)\pi(\alpha), \tag{12}$$

where $\pi(\beta_j|\alpha, \sigma, \nu)$ is the density of $\mathcal{S}\alpha S(\alpha, \sigma/\nu)$ and $\pi(\alpha)$ is the prior density for α. As mentioned above, this density is not analytically tractable. One can use numerical integration of Nolan (1998) or fast Fourier transform used by Tsionas (1999) and Lombardi (2007). However, since the density of SAS must be numerically evaluated at each iteration of MCMC, those methods can be computationally intensive and unstable for small α, in which we would be interested for shrinkage problem. Therefore, it is desirable to avoid the direct evaluation of the prior density.

We apply the ABC technique to approximate the posterior distribution of α. We try to find the α values under which the current value of $\boldsymbol{\beta}$ and the simulated value from SAS are close. Specifically, consider the augmented posterior of α given by

$$\pi_\epsilon(\alpha, \mathbf{b}|\tilde{\boldsymbol{\beta}}, \sigma^2, \nu^2, \mathbf{y}) \propto g_\epsilon(\boldsymbol{\beta}|\mathbf{b}, \alpha) \prod_{j=1}^{p} \pi(b_j|\alpha, \sigma, \nu)\pi(\alpha), \tag{13}$$

where $\mathbf{b} = (b_1, \ldots, b_p)'$ is the simulation from SAS on the same space as $\boldsymbol{\beta}$, $g_\epsilon(\boldsymbol{\beta}|\mathbf{b}, \alpha)$ is the weighting function which places greater weight to the point under which $\boldsymbol{\beta}$ and $\mathbf{b}$ are closer and becomes constant when $\boldsymbol{\beta} = \mathbf{b}$, and ϵ is the tuning parameter which determines the precision of approximation. Popular choice of the weighting function includes, for example,

$$g_\epsilon(\boldsymbol{\beta}|\mathbf{b}, \alpha) \propto I(\rho(\boldsymbol{\beta}, \mathbf{b}) < \epsilon), \tag{14}$$

where $\rho(\boldsymbol{\beta}, \mathbf{b})$ measures the distance between $\boldsymbol{\beta}$ and $\mathbf{b}$. We use the following particular form of distance metric given by

$$\rho(\boldsymbol{\beta}, \mathbf{b}) = \sqrt{\frac{1}{p}\sum_{j=1}^{p}(\beta_{(j)} - b_{(j)})^2}, \tag{15}$$

where $\beta_{(j)}$ and $b_{(j)}$, $j = 1, \ldots, p$, are the order statistics of $\boldsymbol{\beta}$ and $\mathbf{b}$, respectively. As $\epsilon \to 0$, the augmented posterior approaches to the true posterior (see, *e.g.*, Sisson and Fan, 2011).

Then, we implement the following MH algorithm to sample from the augmented posterior (13).

1. Suppose the current state is $(\alpha, \mathbf{b})$.

2. Draw α^* from the proposal density $q(\alpha^*|\alpha)$.

3. Draw $\mathbf{b}^*$ from $\mathcal{S}\alpha S(\alpha^*, \sigma/\nu)$.

4. Accept α^* and $\mathbf{b}^*$ with probability

$$\min\left\{1, \frac{g_\epsilon(\boldsymbol{\beta}|\mathbf{b}^*, \alpha^*)\pi(\alpha^*)q(\alpha|\alpha^*)}{g_\epsilon(\boldsymbol{\beta}|\mathbf{b}, \alpha)\pi(\alpha)q(\alpha^*|\alpha)}\right\}, \tag{16}$$

Notice that the acceptance probability (16) involves no intractable prior densities (see Marjoram *et al.*, 2003).

Since it is easy to simulate from SAS, our approach is expected to be computationally efficient. Under a feasible choice of ϵ, we may approximate the posterior distribution of α well.

3.2 SGL prior

We proceed similarly to the case of SAS prior. The independent SGL prior for the coefficient is assumed:

$$\beta_j \sim \mathcal{SGL}(\alpha, \sigma/\nu, \gamma), \quad j = 1, \ldots, p. \tag{17}$$

In addition to (7), we assume

$$\gamma \sim \mathcal{G}a(r_0, R_0). \tag{18}$$

Then, our MCMC scheme samples from the full conditional distribution of $\tilde{\boldsymbol{\beta}}$, σ^2, ν^2, $\boldsymbol{\lambda}$, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p)$, γ, and α.

The full conditional distribution of $\tilde{\boldsymbol{\beta}}$ is the normal distribution (8) with $\boldsymbol{\Lambda}$ in $\mathbf{B}_1$ replaced with $\boldsymbol{\Omega} = \text{diag}(M_0, \sigma^2\lambda_1\xi^{2/\alpha}/\nu^2, \ldots, \sigma^2\lambda_p\xi_p^{2/\alpha}/\nu^2)$. The full conditional distribution of σ^2 is $\mathcal{IG}(n_1/2, s_2/2)$ where $s_2 = s_o + \sum_{i=1}^n (y_i - \tilde{x}_i'\tilde{\boldsymbol{\beta}})^2 + \nu^2 \sum_{j=1}^p \frac{\beta_j^2}{\lambda_j\xi_j^{2/\alpha}}$. The full conditional distribution of ν^2 is $\mathcal{G}a(m_1/2, t_2/2)$ where

$t_2 = t_o + \frac{1}{\sigma^2}\sum_{j=1}^{p}\frac{\beta_j^2}{\lambda_j\xi_j^{2/\alpha}}$. The full conditional density of λ_j is proportional to $\lambda_j^{-1/2}\exp\left\{-\frac{\nu^2\beta_j^2}{2\sigma^2\lambda_j\xi^{2/\alpha}}\right\} f_{\alpha/2}(\lambda_j)$. We may use the same independence MH algorithm as in the case of SAS prior. The full conditional density of the other local variance ξ_j is given by

$$\pi(\xi_j|\tilde{\boldsymbol{\beta}}, \sigma^2, \nu^2, \boldsymbol{\lambda}, \alpha, \gamma, \mathbf{y}) \propto \xi^{\gamma-1/\alpha-1}\exp\left\{-\frac{1}{2}\left(\frac{\nu^2\beta_j^2}{\sigma^2\lambda_j}\xi_j^{-2/\alpha} + 2\xi_j\right)\right\}, \quad j = 1, \ldots, p. \tag{19}$$

To sample from (19), we employ the independence MH algorithm with generalised inverse Gaussian proposal distribution whose density kernel is given by

$$q(\xi_j) \propto \xi_j^{c-1}\exp\left\{-\frac{1}{2}\left(d^2\xi_j^{-1} + e^2\xi_j\right)\right\}, \tag{20}$$

where $c = \gamma - 1/\alpha$, $d^2 = \frac{\nu^2\beta_j^2}{\sigma^2\lambda_j}$, and $e^2 = 2$. The full conditional density of γ is given by

$$\pi(\gamma|\tilde{\boldsymbol{\beta}}, \sigma^2, \nu^2, \boldsymbol{\lambda}, \boldsymbol{\xi}, \alpha, \mathbf{y}) \propto \prod_{j=1}^{p}\xi_j^{\gamma-1}\left(\frac{1}{\Gamma(\gamma)}\right)^p \gamma^{r_0-1}\exp\{-R_0\gamma\}. \tag{21}$$

We use random walk MH with Gaussian proposal distribution to sample γ values. Finally, the full conditional density of α after integrating out the local variances $\boldsymbol{\lambda}$ and $\boldsymbol{\xi}$ is given by

$$\pi(\alpha|\tilde{\boldsymbol{\beta}}, \sigma^2, \nu^2, \boldsymbol{\lambda}, \boldsymbol{\xi}, \gamma, \mathbf{y}) \propto \prod_{j=1}^{p}\pi(\beta_j|\alpha, \sigma, \nu, \gamma)\pi(\alpha), \tag{22}$$

where $\pi(\beta_j|\alpha, \sigma, \nu, \gamma)$ is the density of SGL prior. Our ABC method can be applied to sample α in the same manner as in the case of SAS.

4 Simulation study

Since the value of α plays an important role in the shrinkage problem as seen in Section 2.3, we are interested in how the ABC approximation to the posterior of α under SAS prior influence the performance of the prior. To this end, we try different ϵ values under various situations. In this simulation study, we assume the following prior distributions for the parameters: $\mu \sim \mathcal{N}(0, 100), \sigma^2 \sim \mathcal{IG}(5/2, 0.1/2), \nu^2 \sim \mathcal{G}a(5/2, 0.1/2)$, and $\alpha \sim \mathcal{U}(0, 2)$.

4.1 Simulation 1

We simulated the data from (1) with $\mu = 0$, $\mathbf{x}_i$ generated from $\mathcal{N}(\mathbf{0}, \mathbf{V})$, $\boldsymbol{\beta}$ generated from $\mathcal{S}\alpha S(\alpha, 1)$, and $\sigma^2 = 1$. We consider uncorrelated and correlated independent variables. The uncorrelated independent variables were generated under $\mathbf{V} = \mathbf{I}$ while the correlated independent variables were generated under $\mathbf{V}_{kl} = 0.5$, if $k \neq l$ and 1, otherwise. In this simulation, we set $\alpha = 1.3, 0.7$, $n = 100$, and $p = 50, 200$.

For ABC, we use (14) and (15) with different ϵ values. MCMC is run for 30,000 iterations with the initial burn-in period of 10,000 iterations. The performance of the prior is measured through the root mean squared errors (RMSE) given by

$$\text{RMSE} = \sqrt{\frac{1}{p}\sum_{j=1}^{p}(\beta_j - \hat{\beta}_j)^2}, \tag{23}$$

where $\hat{\beta}_j$ denotes the posterior mean of β_j.

Table 1 and 2 show the posterior means, 95% credible intervals for α and σ/ν, and RMSE under the different values of ϵ for the uncorrelated and correlated independent variables, respectively. It can be seen from the tables that as ϵ increases the posterior mean of α become larger deviating from the true value. This can be explained as follows. For SAS, matching of the values around the centre is easier than that of the values in the tails and SAS does not produce b_j with large magnitude under some large α. When ϵ is large, α can be accepted even though the values in the tails are dissimilar. Therefore, under large ϵ, the matching of the tail values can less influence the acceptance of α and the posterior distribution of α approximated by the ABC algorithm tends to be in the region where matching of the values around the centre is easy. On the other hand, when ϵ is small, the ABC algorithm would reject the α value under which the values in the tail are dissimilar. Therefore, it was possible for the ABC algorithm to locate the true value of α.

It is found from Table 1 that as long as $n > p$ and the independent variables are not correlated, the choice of ϵ does not seem to affect RMSE. However, when $p > n$, the choice of ϵ does affect RMSE. Thus, the performance of the prior is more sensitive to the approximation quality of the posterior distribution of α when $p > n$. Even when $n > p$, when the independent variables are correlated, the choice of ϵ affects

the performance of the prior as seen in Table 2. Therefore, the results in Table 1 and 2 suggest that we use smaller ϵ for more accurate estimation.

Although we do not show the result here, we also implemented the random walk MH sampler using numerical integration of Nolan (1997) to estimate α. However, the algorithm was unstable and the acceptance rate was too low to be practical.

4.2 Simulation 2

Next, we consider a sparse situation where $\boldsymbol{\beta}$ is designed as

$$\beta_j = \begin{cases} 2 & \mod (j-1, p/10) = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

We set $p = 100, 200$. The dataset of 70 observations was generated using the same setting as Section 4.1 and then we split the data into the training and testing set of 50 and 20 observations, respectively. Here, we focus on predicting the values in the training set and performance is measured through the prediction RMSE given by

$$RMSE = \sqrt{\frac{1}{20}\sum_{i=1}^{20}(y_{test,i} - \hat{y}_i)^2}, \tag{25}$$

where $y_{test,i}$ is the observation in the testing set and $\hat{y}_i$ is the mean of the corresponding predictive distribution.

Although now β_j does not follow SAS, it would be assuring if the ABC approximation to the posterior distribution of α works well. Therefore, it is interesting to study the impact of choice of ϵ on the performance of the prior.

Table 3 and 4 show the posterior means and 95% credible intervals for α and σ/ν and prediction RMSE under the different values of ϵ for the uncorrelated and correlated independent variables. Similarly to the results in Section 4.1, it is found that as ϵ becomes large, the posterior mean of α shifts towards larger values. Interestingly, RMSE becomes worse when larger ϵ is used. The impact of choice of ϵ on the performance is profound when $p = 200$ and the independent variables are correlated. Combining the

observations from Table 1 to 4, we may conclude that we should use smaller ϵ for better performance of the prior.

5 Application: NIR ternary mixture data

The data consists of 95 near-infrared (NIR) spectra of ternary mixture of ethanol, water and isopropanol at five different temperatures (Wulfuer *et al*., 1998; Wehrens, 2012). We consider the content of ethanol and isopropanol measured in percentage as the dependent variables. The independent variables include the temperature measured in Celsius and 200-channel absorbance spectrum in the wavelength range 850-1050 nm. Average pairwise absolute correlation between the spectrum is 0.726, thus the exists some strong collinearity in the independent variables. We split the data into the training and testing set of 60 and 35 observations and consider predicting ethanol and isopropanol contents on the basis of their infrared spectrum.

Since it is of interest to study how the SAS and SGL priors perform against some existing priors. We apply the double exponential (DE), Caushy, normal gamma (NG), horseshoe (HS), SAS, and SGL priors for the coefficients of the spectrum. The performance of the priors are compared in terms of prediction RMSE. We assume the following prior distributions for the rest of the parameters. It is assumed the intercept and the coefficient for temperature independently follow $\mathcal{N}(0, 100)$, and $\sigma^2 \sim \mathcal{IG}(5/2, 0.1/2), \nu^2 \sim \mathcal{G}a(5/2, 0.1/2), \gamma \sim \mathcal{G}a(5, 0.1)$. We assume $\alpha \sim \mathcal{U}(0, 1)$ in order that we can avoid the posterior distribution of α to concentrate on a region with large α values and fully exploit the good shrinkage pattern of the priors.

Table 5 shows RMSE under the different prior distributions. As expected, DE performed the worst followed by NG. The performance of Cauchy, HS, SAS, and SGL seems to be quite comparable. Actually, SGL performed the best among the six priors for both ethanol and isopropanol cases. Figure 3 shows the plots of posterior means under each prior distribution. It can be seen that DE failed to distinguish the signals from the noise. NG managed to extract the signals, although it still picks up the noise as signal.

Similarly to the result of Table 5, the posterior means under Cauchy, HS, SAS, and SGL appears to be quite similar. They have peaks in the similar wavelength regions. Under the SAS prior, the posterior mean of α was 0.792 with 95% credible interval $(0.589, 0.986)$ for ethanol, and 0.835 with $(0.502, 0.994)$ for isopropanol. Under the SGL prior, the posterior mean of α was 0.903 with 95% credible interval $(0.741, 0.996)$ for ethanol, and 0.917 with $(0.785, 0.997)$ for isopropanol.

6 Conclusion

We have considered two classes of shrinkage priors based on the stable distribution. To estimate the hyperparameters of those priors for data driven shrinkage, we proposed an MH algorithm based on the ABC method which avoids the direct evaluation of the intractable prior densities. Since our method only needs simulation from the prior distribution, it is computationally efficient. It was found from our simulation study that under small ϵ our method performs well. From the application to NIR data, we found that SAS performs comparable with other existing priors and SGL performs the best among the six prior distributions we considered.

A direction for the future study is as follows. In our ABC algorithm to estimate α, we compare the order statistics of $\boldsymbol{\beta}$ and the simulated $\mathbf{b}$. In the ABC literature, it is often recommended to use summary statistics with lower dimension to avoid the dimensionality problem. However, from our experience, it was difficult to find a set of appropriate summary statistics in our model setting probably due to severe loss of information. For example, in the setting of Section 4.1, we considered using the estimators of McCulloch (1986, 1998) which performed well in Peters *et al*. (2011). The method did not work well especially when $p > n$ and α is small. To our knowledge, the ABC method has never been applied to the SGL distribution, it would be interesting to study which kind of statistics can be used for SGL. Therefore, we need to develop a set of effective summary statistics for SAS and SGL in the ABC algorithm in order that our method is more efficient. Furthermore, some analytical results on the shrinkage rule of the priors would be required. Since both priors are derived from the infinitely divisible preocesses, we may employ

the Lévy process approach of Polson and Scott (2012).

References

[1] Achim, A., Tsakalides, P., and Bezerianos, A. (2003). "SAR image demonising via Bayesian wavelet shrinkage based on heavy-tailed modeling," *IEEE Transactions on Geoscience and Remote Sensing*, **41**, 1773–1784.

[2] Armagan, A. Dinson, D, B. and Lee, J. (2012). "Generalized double Pareto shrinkage," arXiv: 1104.0861.

[3] Boubchir, L. and Fadili, J. M. (2006). "A closed-form nonparametric Bayesian estimator in the wavelet domain of images using an approximate α-stable prior," *Pattern Recognition Letters*, **27**, 1370–1382.

[4] Burnecki, K., Janczura, J., Magdziarz, M., and Weron, A. (2008). "Can one see a competition between sub diffusion and Léve flights? A case of geometric-stable noise," *Acta Physica Polonica B*, **39**, 1043–1054.

[5] Caron, F. and Doucet, A. (2008). "Sparse Bayesian nonparametric regression," In *Proceedings of the 25th International Conference on Machine Learning*, 88-95. New York: Association for Computing Machinery Press.

[6] Carvalho, C. M., Polson, N. G., and Scott, S. L. (2010). "The horseshoe estimator for sparse signals," *Biometrika*, **97**, 465–480.

[7] Choi, S. T. B. and Smith, A. F. M. (1997). "On robust analysis of a normal location parameter," *Journal of the Royal Statistical Society Series B*, **59**, 463–474.

[8] Devroye. L. (1996). "Random variate generation in one line of code," In *Proceedings of the 1996 Winter Simulation Conference*, 265–272.

[9] Griffin, J. E. and Brown, P. J. (2010). "Inference with normal-gamma prior distributions in regression problems," *Bayesian Analysis*, **5**, 171–188.

[10] Hans, C. (2009). "Bayesian lasso regression," *Biometrika*, **96**, 835–845.

[11] Hans, C. (2011). "Elastic net regression modeling with the outhunt normal prior," *Journal of the American Statistical Association*, **106**, 1383–1393.

[12] Li, Q. and Lin, N. (2010). "The Bayesian elastic net," *Bayesian Analysis*, **5**, 151–170.

[13] Lombardi, M. (2007). "Bayesian inference for α-stable distributions: a random walk MCMC approach," *Computational Statistics & Data Analysis*, **51**, 2688–2700.

[14] Marin, J., Pudlo, P., Robert, C. P. and Ryder, R. (2011). "Approximate Bayesian computational methods," arXiv: 1101.0955.

[15] McCulloch, J. H. (1986). "Simple consistent estimators of stable distribution parameters," *Communications in Statistics - Simulation and Computation*, **15**, 1109–1136.

[16] McCulloch, J. H. (1998). "Numerical approximation of the symmetric stable distribution and density," In *Practical Guide to Heavy Tails: Statistical Techniques and Applications*, Eds. Adler, R., Feldman, R., and Taqqu, M. S. Birkhäuser, Boston.

[17] Nolan, J. P. (1997). "Numerical computation of stable densities and distributions," *Communications in Statistics - Simulation and Computation* **13**, 759?774.

[18] Park, T. and Casella, G. (2008). "The Bayesian Lasso," *Journal of the American Statistical Association*, **103**, 681–686.

[19] Peters, G. W., Sisson, S. A., and Fan, Y. "Likelihood-free Bayesian inference for α-stable models," *Computational Statistics & Data Analysis*, **56**, 3743–3756.

[20] Polson, N. G. and Scott, S. L. (2011a). "Data augmentation for support vector machines," *Bayesian Analysis*, **6**, 1–24.

[21] Polson, N. G. and Scott, S. L. (2011b). "The Bayesian bridge," arXiv:1109.2279.

[22] Polson, N. G. and Scott, S. L. (2012). "Local shrinkage rules, Lévy processes and reguralized regression," *Journal of the Royal Statistical Society Series B*, **74**, 287–311.

[23] Sisson, A. S. and Fan, Y. (2011). "Likelihood-free Markov chain Monte Carlo," In *Handbook of Markov Chain Monte Carlo*, Eds. S. P. Brooks, A. Gelman, G. Jones and X . L. Meng. Chapman and Hall/CRC Press.

[24] Tsionas, E. G. (1999). "Monte Carlo inference in econometric models with symmetric stable disturbances," *Journal of Econometrics*, **88**, 365–401.

[25] Wehrens, R. (2012). kohonen: *Supervised and unsupervised self-organising maps*. R package version 2.0.9; available at `http://cran.r-project.org/web/packages/kohonen/`

[26] Wulfert, F., Kok, W. T., and Smilde A. K. (1998). "Influence of temperature on vibrational spectra and sonsequences for the predictive ability of multivariate models," *Analytical Chemistry*, **70**, 1761–1767.

Table 1: Effect of ϵ on posterior means of $\boldsymbol{\beta}$ with uncorrelated independent variables

			Posterior of α			Posterior of σ/ν			
p	α	ϵ^*	Mean	95% CI		Mean	95% CI		RMSE
50	1.3	0.612	1.266	(0.735,	1.733)	0.874	(0.531,	1.229)	0.134
		0.724	1.410	(0.895,	1.908)	0.881	(0.598,	1.205)	0.134
		0.835	1.518	(0.905,	1.981)	0.871	(0.573,	1.199)	0.134
	0.7	0.913	0.775	(0.449,	1.419)	0.574	(0.248,	1.180)	0.131
		0.981	0.916	(0.485,	1.515)	0.678	(0.485,	1.515)	0.130
		1.027	1.361	(0.608,	1.968)	1.102	(0.625,	1.651)	0.131
200	1.3	0.517	1.172	(0.861,	1.528)	0.757	(0.429,	1.128)	1.339
		0.718	1.279	(0.904,	1.724)	0.825	(0.570,	1.120)	1.370
		1.005	1.518	(0.947,	1.976)	0.926	(0.659,	1.206)	1.414
	0.7	0.811	0.850	(0.599,	1.228)	1.136	(0.599,	1.228)	1.928
		0.965	1.046	(0.634,	1.595)	1.498	(0.920,	1.595)	1.982
		1.110	1.388	(0.714,	1.976)	1.838	(1.327,	2.443)	2.138

$\epsilon^* = \epsilon/$(standard deviation of $\boldsymbol{\beta}$)

Table 2: Effect of ϵ on posterior means of $\boldsymbol{\beta}$ with correlated independent variables

			Posterior of α			Posterior of σ/ν			
p	α	ϵ^*	Mean	95% CI		Mean	95% CI		RMSE
50	1.3	0.612	1.261	(0.818,	1.716)	0.830	(0.514,	1.164)	0.182
		0.724	1.385	(0.864,	1.907)	0.842	(0.560,	1.173)	0.181
		0.835	1.499	(0.884,	1.972)	0.838	(0.568,	1.151)	0.180
	0.7	0.913	0.823	(0.493,	1.575)	0.729	(0.328,	1.432)	0.197
		0.981	0.958	(0.510,	1.589)	0.796	(0.397,	1.589)	0.204
		1.027	1.353	(0.617,	1.964)	1.160	(0.676,	1.741)	0.223
200	1.3	0.517	1.216	(0.866,	1.682)	0.811	(0.469,	1.185)	1.443
		0.718	1.298	(0.917,	1.724)	0.873	(0.605,	1.151)	1.448
		1.005	1.523	(0.955,	1.978)	0.934	(0.674,	1.190)	1.500
	0.7	0.811	0.839	(0.579,	1.199)	0.947	(0.499,	1.630)	2.282
		0.965	1.002	(0.616,	1.528)	1.181	(0.755,	1.690)	2.284
		1.110	1.357	(0.689,	1.967)	1.496	(1.074,	1.996)	2.367

$\epsilon^* = \epsilon/$(standard deviation of $\boldsymbol{\beta}$)

Table 3: Effect of ϵ under sparse $\boldsymbol{\beta}$ with uncorrelated independent variables

		Posterior of α			Posterior of σ/ν			
p	ϵ^*	Mean	95% CI		Mean	95% CI		RMSE
100	0.833	1.304	(0.874,	1.878)	0.092	(0.051,	0.187)	1.476
	0.917	1.453	(0.878,	1.966)	0.091	(0.057,	0.152)	1.475
	1.000	1.527	(0.923,	1.979)	0.090	(0.061,	0.136)	1.494
200	0.803	1.146	(0.854,	1.500)	0.038	(0.018,	0.066)	1.358
	0.895	1.384	(0.920,	1.903)	0.049	(0.025,	0.090)	1.449
	1.032	1.544	(0.963,	1.982)	0.058	(0.033,	0.147)	1.704

$\epsilon^* = \epsilon/$(standard deviation of $\boldsymbol{\beta}$)

Table 4: Effect of ϵ under sparse $\boldsymbol{\beta}$ with correlated independent variables

		Posterior of α			Posterior of σ/ν			
p	ϵ^*	Mean	95% CI		Mean	95% CI		RMSE
100	0.750	1.293	(0.742,	1.791)	0.121	(0.056,	0.199)	1.458
	0.833	1.414	(0.903,	1.916)	0.124	(0.071,	0.196)	1.484
	0.917	1.550	(1.007,	0.272)	0.135	(0.080,	0.224)	1.557
200	0.780	1.366	(0.976,	1.995)	0.078	(0.033,	0.210)	1.902
	0.849	1.584	(1.074,	1.982)	0.125	(0.045,	0.211)	3.566
	1.032	1.615	(1.094,	1.980)	0.129	(0.056,	0.199)	3.901

$\epsilon^* = \epsilon/$(standard deviation of $\boldsymbol{\beta}$)

Table 5: RMSE for ternary mixture data

	DE	Cauchy	NG	HS	SAS	SGL
Ethanol	1.083	0.637	1.067	0.646	0.665	0.624
Isopropanol	0.948	0.656	0.882	0.638	0.649	0.625

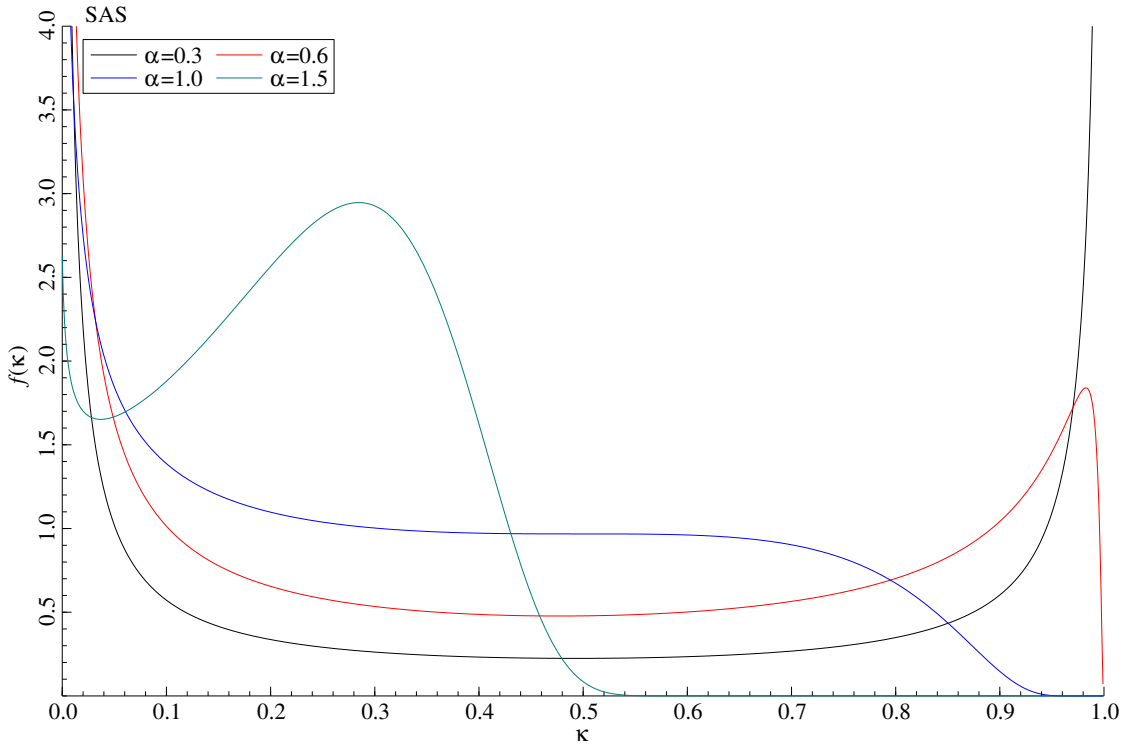

Figure 1: Density of κ for SAS prior

Figure 2: Densities of κ for SGL prior with $\gamma = 0.5, 1$, and 2 (from top to bottom)

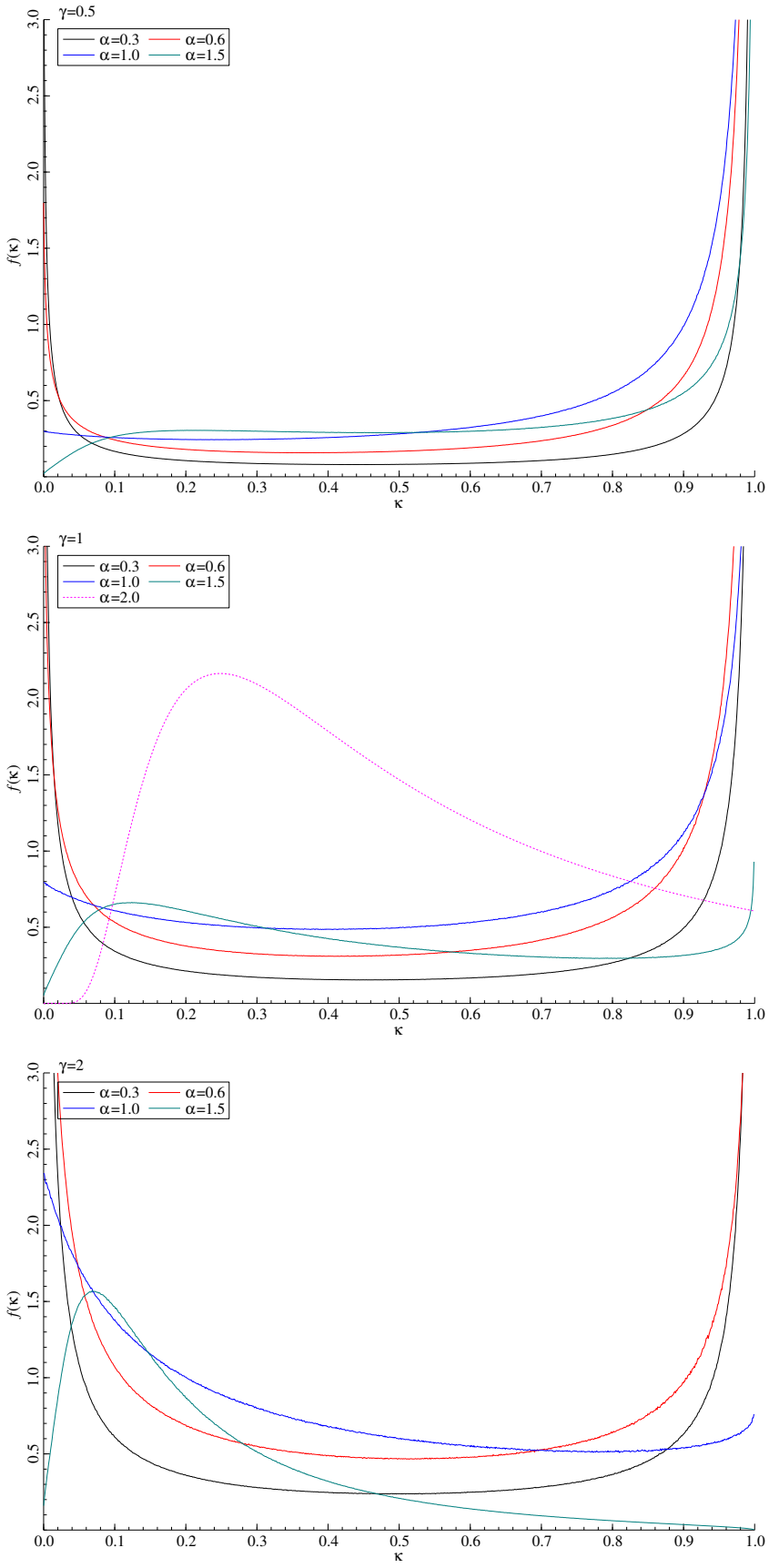

Figure 3: Posterior means for ternary data