A simple sampler for the horseshoe estimator

Enes Makalic, and Daniel F. Schmidt, *Member, IEEE*

***Abstract*—In this note we derive a simple Bayesian sampler for linear regression with the horseshoe hierarchy. A new interpretation of the horseshoe model is presented, and extensions to logistic regression and alternative hierarchies, such as horseshoe+, are discussed. Due to the conjugacy of the proposed hierarchy, Chib's algorithm may be used to easily compute the marginal likelihood of the model.**

***Index Terms*—Bayesian regression, horseshoe estimator, Markov Chain Monte Carlo sampling.**

I. Introduction

CONSIDER the following Bayesian linear regression model for data $y \in \mathbb{R}^n$

$$\begin{aligned} \mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n), && (1)\\ \beta_j|\lambda_j^2, \tau^2, \sigma^2 &\sim \mathcal{N}(0, \lambda_j^2\tau^2\sigma^2), && (2)\\ \sigma^2 &\sim \sigma^{-2}d\sigma^2, && (3)\\ \lambda_j &\sim \mathcal{C}^+(0,1), && (4)\\ \tau &\sim \mathcal{C}^+(0,1) && (5) \end{aligned}$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a matrix of predictor variables (not necessarily full rank), $\mathcal{N}_k(\cdot,\cdot)$ is the k-variate Gaussian distribution, $\mathcal{C}^+(0,1)$ is the standard half-Cauchy distribution with probability density function

$$p(z) = \frac{2}{\pi(1+z^2)}, \quad z > 0,$$

and $j = (1,\ldots,p)$. It is usual to require that the p predictors are standardized to have zero mean and unit length and that the data y is centred. This avoids the need to explicitly model a separate parameter for the intercept.

Equations (1–5) define the horseshoe regression hierarchy recently proposed in [1]. The horseshoe model is a global-local shrinkage procedure in which the local shrinkage for coefficient β_j is determined by $\lambda_j > 0$ and the overall level of shrinkage is determined by the hyperparameter $\tau > 0$. The particular choice of a half-Cauchy prior distribution over the global and local hyperparameters results in aggressive shrinkage of small coefficients (i.e., noise) and virtually no shrinkage of sufficiently large coefficients (i.e., signal). This is in contrast to the well known Bayesian lasso [2] and Bayesian ridge hierarchies where the shrinkage effect is uniform across all coefficients. Further favourable properties of the horseshoe model are discussed in [1], [3], [4].

E. Makalic and D. F. Schmidt are with the Centre of Epidemiology and Biostatistics, The University of Melbourne, Australia, e-mail: {emakalic,dschmidt}@unimelb.edu.au

Manuscript received September 1, 2015; revised September 17, 2015.

The original horseshoe paper does not provide details for efficient sampling from the posterior distribution of the regression coefficients. A standard Gibbs sampling approach is difficult to implement due to the non-standard form of the conditional posterior distributions for the hyperparameters $(\lambda_1,\ldots,\lambda_p)$ and τ. Subsequent papers have suggested the use of specialised algorithms, such as slice sampling [5], for the hyperparameters [6]. In this paper, we provide an alternative sampling scheme for all model parameters based on auxiliary variables that leads to conjugate conditional posterior distributions for all parameters, making the application of Gibbs sampling relatively straightforward.

II. Bayesian horseshoe with auxiliary variables

We make use of the following scale mixture representation of the half-Cauchy distribution. Let x and a be random variables such that

$$x^2|a \sim \mathcal{IG}(1/2, 1/a) \quad \text{and} \quad a \sim \mathcal{IG}(1/2, 1/A^2); \qquad (6)$$

then $x \sim \mathcal{C}^+(0, A)$ [7], where $\mathcal{IG}(\cdot,\cdot)$ is the inverse-gamma distribution with probability density function

$$p(z|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1}\exp\left(-\frac{\beta}{z}\right).$$

Using the decomposition (6) leads to the revised horseshoe hierarchy

$$\begin{aligned} \mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n),\\ \beta_j|\lambda_j^2, \tau^2, \sigma^2 &\sim \mathcal{N}(0, \lambda_j^2\tau^2\sigma^2),\\ \sigma^2 &\sim \sigma^{-2}d\sigma^2,\\ \lambda_j^2|\nu_j &\sim \mathcal{IG}(1/2, 1/\nu_j),\\ \tau^2|\xi &\sim \mathcal{IG}(1/2, 1/\xi),\\ \nu_1,\ldots,\nu_p,\xi &\sim \mathcal{IG}(1/2, 1). \end{aligned}$$

The above hierarchy makes Gibbs sampling from the posterior distribution straightforward. The conditional posterior distribution of the regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ [8] is

$$\begin{aligned} \boldsymbol{\beta}|\cdot &\sim \mathcal{N}_p(\mathbf{A}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \sigma^2\mathbf{A}^{-1}) && (7)\\ \mathbf{A} &= (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\Lambda}_*^{-1})\\ \boldsymbol{\Lambda}_* &= \tau^2\boldsymbol{\Lambda} \end{aligned}$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1^2,\ldots,\lambda_p^2)$. The conditional posterior distribution of σ^2 is an inverse-gamma distribution given by

$$\sigma^2|\cdot \sim \mathcal{IG}\left((n+p)/2, (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/2 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Lambda}_*^{-1}\boldsymbol{\beta}/2\right).$$

Following the approach in [9], we may improve the convergence behaviour of the sampler by integrating out the regression coefficients $\boldsymbol{\beta}$ from the posterior distribution of σ^2. After integration, the conditional posterior density of σ^2 is

$$\sigma^2|\cdot \sim \mathcal{IG}\left(n/2, \mathbf{y}^{\mathrm{T}}\left(\mathbf{I}_n - \mathbf{X}\boldsymbol{\Lambda}_*^{-1}\mathbf{X}^{\mathrm{T}}\right)\mathbf{y}/2\right).$$

where $\mathbf{I}_n$ is the $(n \times n)$ identity matrix. The above formula corrects a minor typographical error in the original manuscript [9].

The conditional posterior distributions for the local and global hypervariances are also of inverse-gamma type

$$\begin{aligned} \lambda_j^2|\cdot &\sim \mathcal{IG}\left(1, \frac{1}{\nu_j} + \frac{\beta_j^2}{2\tau^2\sigma^2}\right), \quad (j = 1, 2, \ldots, p), \quad (8) \\ \tau^2|\cdot &\sim \mathcal{IG}\left(\frac{p+1}{2}, \frac{1}{\xi} + \frac{1}{2\sigma^2}\sum_{j=1}^{p}\frac{\beta_j^2}{\lambda_j^2}\right). \end{aligned}$$

Finally, the conditional posterior distributions for the auxiliary variables are:

$$\begin{aligned} \nu_j|\cdot &\sim \mathcal{IG}\left(1, 1 + \frac{1}{\lambda_j^2}\right), \quad (j = 1, 2, \ldots, p), \\ \xi|\cdot &\sim \mathcal{IG}\left(1, 1 + \frac{1}{\tau^2}\right). \end{aligned}$$

It is interesting to note that the conditional posterior distributions for all parameters, except the regression coefficients, are inverse-gamma. Generation of inverse-gamma variates may be done by noting that a random variable $Y = 1/X$ follows an inverse-gamma distribution $\mathcal{IG}(a, 1/b)$ if X follows a Gamma(a, b) distribution. Efficient generation of random variates from the gamma distribution may be done using the algorithm in, for example, [10] (pp. 45–46).

III. Extensions

A. Horseshoe logistic regression

The Gibbs sampling approach proposed in this paper can be extended to other models and other prior distributions. In logistic regression, equation (1) in the hierarchy becomes

$$y_i|\mathbf{x}_i, \boldsymbol{\beta}, \beta_0 \sim \text{Binom}(1, 1/(1 + e^{-\psi_i})),$$

where $y_i \in \{0, 1\}$, β_0 is the intercept parameter and $\psi_i = \beta_0 + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ are the log-odds of success. Unlike in the case of linear regression, the intercept parameter β_0 may now be explicitly modelled with a uniform prior $\pi(\beta_0) \propto 1$ to ensure that β_0 is not penalized.

Bayesian logistic regression with the horseshoe hierarchy may be implemented using a scale mixture based on z-distributions [11] or the Pólya-gamma data augmentation framework [12] for modelling the logistic function at the top level of the hierarchy. Using the Pólya-gamma approach, the conditional posterior distribution of $\boldsymbol{\beta}$ given auxiliary variables $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^{\mathrm{T}}$ and data y_i $(i = 1, \ldots, n)$ is a multivariate Gaussian proportional to

$$\pi_\beta(\boldsymbol{\beta}) \exp\left(-\frac{1}{2}(\mathbf{z} - \beta_0\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\Omega(\mathbf{z} - \beta_0\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta})\right),$$

where $\mathbf{1}_n$ is a n-vector of ones, $\mathbf{z} = (\kappa_1/\omega_1, \ldots, \kappa_n/\omega_n)^{\mathrm{T}}$, $\kappa_i = y_i - 1/2$, $\boldsymbol{\Omega} = \text{diag}(\boldsymbol{\omega})$ and $\pi_\beta(\boldsymbol{\beta})$ is a prior distribution for the regression coefficients.

The conditional posterior distribution for the auxiliary variables $\boldsymbol{\omega}$ is

$$\omega_i|\boldsymbol{\beta} \sim \text{PG}(1, \beta_0 + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}),$$

which is a Pólya-gamma distribution with shape parameter 1 and scale parameter $\beta_0 + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$. With the horseshoe prior, the conditional posterior distribution for the regression coefficients becomes

$$\begin{aligned} \boldsymbol{\beta}|\cdot &\sim \mathcal{N}_p(\mathbf{A}^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Omega}(\mathbf{z} - \beta_0\mathbf{1}_n), \mathbf{A}^{-1}) \quad (9) \\ \mathbf{A} &= (\mathbf{X}^{\mathrm{T}}\boldsymbol{\Omega}\mathbf{X} + \boldsymbol{\Lambda}_*^{-1}) \\ \boldsymbol{\Lambda}_* &= \tau^2\boldsymbol{\Lambda} \end{aligned}$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1^2, \ldots, \lambda_p^2)$. The conditional posterior distribution for the intercept parameter is

$$\beta_0|\boldsymbol{\omega}, \mathbf{y} \propto \mathcal{N}\left(\frac{1}{s}\sum_{i=1}^{n} v_i, \frac{1}{s}\right) \quad (10)$$

where

$$\begin{aligned} \mathbf{v} &= (\mathbf{z} - \mathbf{X}\boldsymbol{\beta}) \odot \boldsymbol{\omega} \\ s &= \sum_{i=1}^{n}\omega_i \end{aligned}$$

and $\odot$ denotes element-wise vector product. The conditional posterior distributions for the prior hyperparameters remain unchanged.

B. Horseshoe negative-binomial regression

The Pólya-gamma data augmentation strategy may also be used to derive a Bayesian horseshoe estimator of negative-binomial regression for count data [12]. Data y_i $(i = 1, 2, \ldots, n)$ is now assumed to be generated by the negative binomial distribution

$$p(y_i|h, \pi_i) \propto (1 - \pi_i)^h \pi_i^{y_i}, (h > 0) \quad (11)$$

where $\pi_i = \exp(\psi_i)/(1 + \exp(\psi_i))$ and $\psi_i = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$. This is equivalent to the sampling model

$$\mathbf{z}|\mathbf{X}, \boldsymbol{\beta} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Omega}^{-1}) \quad (12)$$

where $z_i = (y_i - h)/2$ and $\boldsymbol{\Omega} = \text{diag}(\omega_1, \omega_2, \ldots, \omega_n)$. The conditional posterior distribution for the auxiliary variables $\boldsymbol{\omega}$ is now

$$\omega_i|\boldsymbol{\beta} \sim \text{PG}(y_i + h, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}),$$

which is a Pólya-gamma distribution with shape parameter $(y_i + h)$ and scale parameter $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$. The conditional posterior distribution of the regression coefficients $\boldsymbol{\beta}$ with the horseshoe prior is equivalent to the horseshoe logistic regression model (9). As before, the intercept parameter must be explicitly modelled and should not be subject to any shrinkage.

C. Alternative Prior Distribution

Extending the hierarchy (1)–(5) to the horseshoe+ [13] estimator is straightforward and involves a direct application of the auxiliary variable representation. In the horseshoe+ hierarchy, the local shrinkage parameters λ_j are given the half-Cauchy prior distribution

$$\lambda_j \sim \mathcal{C}^+(0, \eta_j) \quad (13)$$

where $j = (1, 2, \ldots, p)$ and η_j is a further half-Cauchy mixing variable which is given the prior density

$$\eta_j \sim \mathcal{C}^+(0, 1). \tag{14}$$

The conditional posterior distribution of the local-shrinkage coefficients λ_j is equivalent to (8) in the case of the horseshoe+ estimator. The conditional posterior density for the auxiliary variables ν_j is now

$$\nu_j|\cdot \sim \mathcal{IG}\left(1, \frac{1}{\eta_j^2} + \frac{1}{\lambda_j^2}\right), \quad (j = 1, 2, \ldots, p), \tag{15}$$

Sampling of the auxiliary mixing variables η_j is then a straightforward application of the decomposition (6).

IV. Evaluation

We have conducted a simulation experiment to compare run times between our implementation (`bhs`) and the R package `monomvn` implementation (available on the CRAN repository) of the Bayesian horseshoe estimator. The `monomvn` package uses the conventional approach to sampling from the posterior distribution of the horseshoe estimator by slice sampling the hyperparameters τ and $(\lambda_1, \ldots, \lambda_p)$. This comparison is expected to be somewhat biased as the two packages are implemented in different programming languages — our implementation is in pure MATLAB code, while the `monomvn` package is implemented as an R interface to compiled C code.

The execution times of both procedures were compared across a range of values for sample size and the number of predictors. For each timing run, both procedures were timed generating $1,000$ samples from the Bayesian horseshoe posterior distribution and all timing results were measured in seconds. The results of the timing experiment are given in Table I. For small sample sizes and small numbers of predictors, the `monomvn` implementation is faster than `bhs`, with both procedures finishing in less than one second of execution time. However, as the number of predictors grows our implementation is significantly faster than `monomvn`. As an example, for the experiment ($n = 1,000$, $p = 1,000$), our implementation is approximately 40 times faster than `monomvn`.

We also compared our sampling procedure based on auxiliary variables (`bhs`) against the `monomvn` implementation using the diabetes data set examined in [2]. This dataset consists of 442 observations and 10 predictors, some of which are highly correlated. The sampling efficiency of the two procedures was compared using the effective sample size metric as discussed in [10] (pp. 499–500) and [14], and is determined from the autocorrelation in the sampling chain. The effective sample size for each regression coefficient was estimated, under varying levels of thinning, from a chain of 50 million posterior samples generated by both `bhs` and `monomvn`. For all regression coefficients and under all levels of thinning, the `bhs` procedure was found to have a higher effective sample size. As an example, Figure 1 depicts the effective sample size for predictors `S2` and `S3` as a function of the level of thinning. These two predictors had the smallest observed effective sample size of the ten predictors in the diabetes data set. The `bhs` procedure is clearly more efficient than `monomvn` and the difference appears greatest for thinning levels of four to eight. In order to achieve an effective sample size of 80% for all ten predictors, the `monomvn` sampler required a thinning level of 16; in contrast, the `bhs` sampler required a thinning level of 11 to achieve the same sampling efficiency. Despite the introduction of latent variables, the `bhs` algorithm appears to be more efficient in terms of effective sample size than the `monomvn` procedure which is based on slice sampling with no auxiliary variables.

V. Discussion

The decomposition (6) reveals an interesting novel interpretation of the horseshoe hierarchy. Integrating out the hypervariances $(\lambda_1^2, \ldots, \lambda_p^2)$ implies a Cauchy prior distribution over each regression coefficient β_j of the form

$$\beta_j|\tau^2, \sigma^2, \nu_j \sim \mathcal{C}(0, 2\tau\sigma/(2\nu_j)^{1/2}).$$

The horseshoe model can therefore be viewed as placing a Cauchy prior distribution over each regression coefficient β_j with the scale of each prior inversely proportional to ν_j.

Gibbs sampling may also be used in the case where the global scale parameter $\sigma > 0$ is given a half-Cauchy prior distribution, as recommended in [15], by using the decomposition (6). An interesting consequence of the conjugacy of all conditional posterior distributions is that Chib's algorithm [16] for computing the marginal likelihood from the output of a Gibbs sampler is readily applicable as the normalizing constants for all conditionals are known. This is in contrast to sampling methods that use algorithms such as slice sampling.

Our MATLAB™ implementation of Bayesian horseshoe linear regression is available from the MATLAB Central File Exchange repository (File ID #52479). The implementation uses Rue's algorithm [17] for efficient sampling from the multivariate Gaussian conditional posterior distribution of the regression coefficients when the sample size is greater than the number of predictors ($n > p$). This algorithm is based on Cholesky factorization with cubic order of complexity in terms of the number of predictors p. For the case where the number of predictors is large, this approach for sampling from multivariate Gaussian distributions is computationally inefficient. In this setting, our implementation employs the sampling algorithm given in [18] which has linear complexity in terms of the number of predictors.

References

[1] C. M. Carvalho, N. G. Polson, and J. G. Scott, "The horseshoe estimator for sparse signals," *Biometrika*, vol. 97, no. 2, pp. 465–480, 2010.

[2] T. Park and G. Casella, "The Bayesian lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, June 2008.

[3] N. G. Polson and J. G. Scott, "Local shrinkage rules, Lévy processes and regularized regression," *Journal of the Royal Statistical Society (Series B)*, vol. 74, no. 2, pp. 287–311, 2012.

[4] ——, "Shrink globally, act locally: Sparse Bayesian regularization and prediction," in *Bayesian Statistics*, vol. 9, 2010.

[5] R. M. Neal, "Slice sampling," *The Annals of Statistics*, vol. 31, no. 3, pp. 705–741, June 2003.

[6] N. G. Polson, J. G. Scott, and J. Windle, "The Bayesian bridge," *Journal of the Royal Statistical Society (Series B)*, vol. 76, no. 4, pp. 713–733, 2014.

	Sample size (n)	Number of predictors (p)				
		10	50	100	500	1000
bhs	10	0.44	0.34	0.43	1.94	5.96
	50	0.21	0.28	0.45	2.31	6.76
	100	0.22	0.29	0.57	2.99	7.50
	500	0.24	0.44	0.65	8.36	32.46
	1000	0.30	0.53	0.83	10.05	34.75
monomvn	10	0.01	0.23	1.49	165.21	1470.74
	50	0.02	0.25	1.51	165.06	1480.70
	100	0.02	0.24	1.48	165.29	1474.30
	500	0.05	0.31	1.57	165.49	1488.92
	1000	0.09	0.41	1.73	166.27	1481.17

TABLE I
COMPARISON OF RUN TIMES BETWEEN THE R PACKAGE monomvn AND OUR IMPLEMENTATION (bhs) OF THE BAYESIAN HORSESHOE SAMPLER. IN EACH TEST, THE SAMPLERS GENERATED $1,000$ SAMPLES FROM THE POSTERIOR DISTRIBUTION OF THE BAYESIAN HORSESHOE. ALL TIMINGS ARE GIVEN IN SECONDS.

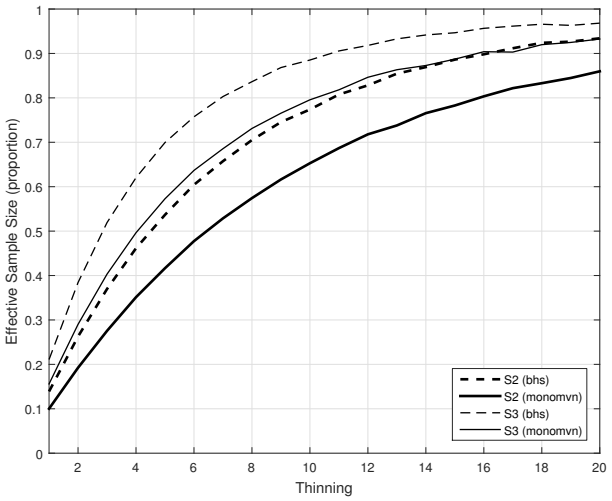

Fig. 1. Comparison of sampling efficiency between monomvn and bhs in terms of effective sample size (expressed as a proportion)

[7] M. P. Wand, J. T. Ormerod, S. A. Padoan, and R. Fruhwirth, "Mean field variational Bayes for elaborate distributions," *Bayesian Analysis*, vol. 6, no. 4, pp. 847–900, 2011.

[8] D. V. Lindley and A. F. M. Smith, "Bayes estimates for the linear model," *Journal of the Royal Statistical Society (Series B)*, vol. 34, no. 1, pp. 1–41, 1972.

[9] B. Rajaratnam and D. Sparks. (2015) Fast bayesian lasso for high-dimensional regression. ArXiv:1509.03697v1.

[10] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed., S. Fienberg, Ed. Springer, 2004.

[11] R. B. Gramacy and N. G. Polson, "Simulation-based regularized logistic regression," *Bayesian Analysis*, vol. 7, no. 3, pp. 567–590, 2012. [Online]. Available: arXiv:1005.3430v1

[12] N. G. Polson, J. G. Scott, and J. Windle, "Bayesian inference for logistic models using Pólya-gamma latent variables," *Journal of the American Statistical Association*, vol. 108, no. 504, pp. 1339–1349, 2013.

[13] A. Bhadra, J. Datta, N. G. Polson, and B. Willard, "The horseshoe+ estimator of ultra-sparse signals," 2015, arXiv:1502.00560.

[14] C. J. Geyer, "Practical markov chain monte carlo," *Statistical Science*, vol. 7, no. 4, pp. 473–483, 1992.

[15] N. G. Polson and J. G. Scott, "On the half-Cauchy prior for a global scale parameter," *Bayesian Analysis*, vol. 7, no. 4, pp. 887–902, 2012.

[16] S. Chib, "Marginal likelihood from the Gibbs output," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1313–1321, December 1995.

[17] H. Rue, "Fast sampling of gaussian markov random fields," *Journal of the Royal Statistical Society (Series B)*, vol. 63, no. 2, pp. 325–338, 2001.

[18] A. Bhattacharya, A. Chakraborty, and B. K. Mallick, "Fast sampling with Gaussian scale-mixture priors in high-dimensional regression," 2015, arXiv:1506.04778.