Bayesian Robust Regression with the Horseshoe+ Estimator

Enes Makalic, Daniel F. Schmidt, and John L. Hopper

The University of Melbourne
Centre for Epidemiology and Biostatistics
Carlton VIC 3053, Australia
{emakalic,dschmidt,j.hopper}@unimelb.edu.au

Abstract. The horseshoe+ estimator for Gaussian linear regression models is a novel extension of the horseshoe estimator that enjoys many favourable theoretical properties. We develop the first efficient Gibbs sampling algorithm for the horseshoe+ estimator for linear and logistic regression models. Importantly, our sampling algorithm incorporates robust data models that naturally handle non-Gaussian data and are less sensitive to outliers. The resulting software implementation provides a powerful, flexible and robust tool for building prediction and classification models from potentially high-dimensional data and represents the state-of-the-art in Bayesian machine learning techniques.

1 Introduction

Bayesian regression models are becoming increasingly common in the big data domain. Consider the following Bayesian regression hierarchy for data $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$:

$$
\begin{aligned}
y_i | \mathbf{X}, \boldsymbol{\beta}, \beta_0, \omega_i^2, \sigma^2 &\sim \mathcal{N}(\mathbf{x}_i^T \boldsymbol{\beta}, \omega_i^2 \sigma^2), && (1)\\
\omega_i^2 &\sim \pi_\omega(\omega_i^2) d\omega_i^2, && (2)\\
\sigma^2 &\sim \sigma^{-2} d\sigma^2, && (3)\\
\beta_j | \lambda_j^2, \tau^2, \sigma^2 &\sim \mathcal{N}(0, \lambda_j^2 \tau^2 \sigma^2), && (4)\\
\beta_0 &\sim d\beta_0, && (5)\\
\lambda_j &\sim \mathcal{C}^+(0, \eta_j), && (6)\\
\eta_j &\sim \mathcal{C}^+(0, 1) && (7)\\
\tau &\sim \mathcal{C}^+(0, 1) && (8)
\end{aligned}
$$

where $i = (1, \ldots, n)$, $j = (1, \ldots, p)$, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a matrix of predictor variables (not necessarily full rank), $\beta_0 \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ are the unknown regression coefficients, $\mathcal{N}(\cdot, \cdot)$ denotes the Gaussian distribution and $\mathcal{C}^+(0, 1)$ is the standard half-Cauchy distribution with probability density function

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

The hierarchy introduced in (1)–(8) comprises two groups: (i) the sampling distribution of the data $\mathbf{y} \in \mathbb{R}^n$ given by (1)–(3) and (ii) the prior distributions of the regression coefficients β_0 and $\boldsymbol{\beta}$ given by (4)–(8). The data model is built from a scale mixture of normal distributions, which is a standard technique for representing a wide range of statistical distributions [1]. For example, by appropriate choice of the prior distribution $\pi_\omega(\cdot)$, one may model the data $\mathbf{y}$ as arising from a heavy-tailed distribution, such as the Cauchy or the Student-t distribution, or even a distribution for categorical data, such as the logistic model.

In most regression problems, the main task is to estimate the unknown regression parameters as well as determine which predictors should be included in the model. It is becoming more common that the dimensionality of the predictors p is large, often much larger than the sample size n (i.e., $p \gg n$), and the performance of the aforementioned Bayesian estimator depends crucially on the particular choice of prior distributions (6)–(8). In big data problems, it is common to assume that the number of predictor variables associated with the outcome $\mathbf{y}$ is small relative to the overall dimensionality p. In other words, the majority of the regression coefficients $\boldsymbol{\beta}$ are assumed to be exactly equal to zero (i.e., the sparse model assumption).

Bhadra et al. [2] recently introduced the Bayesian horseshoe+ hierarchy for ultra-sparse regression problems, which enjoys many favourable theoretical properties. The usual implementation of Bayesian hierarchical regression models is by using standard Markov Chain Monte Carlo (MCMC) approaches, such as the Gibbs sampler [3]. However, due to the difficulties related to sampling from the half-Cauchy distribution, there is currently no efficient and simple MCMC approach for the horseshoe+. Recently, Makalic and Schmidt [4] identified a novel scale mixture representation of the half-Cauchy distribution that was applied successfully to a specific form of Bayesian regression with half-Cauchy prior distributions called the horseshoe estimator [5, 6]. This method outperformed the existing state-of-the-art implementations of the horseshoe, and given the favourable properties of the horseshoe+ relative to the horseshoe, it would be of great benefit if there was an efficient sampling algorithm for the horseshoe+.

This paper extends the work in [4] in two important ways: (i) we derive the first efficient Gibbs sampling scheme for the horseshoe+ estimator (see Section 2.1), and (ii) we exploit scale mixture representations to incorporate robust data models that naturally handle non-Gaussian data and are less sensitive to outliers (see Section 3). The resulting software implementation provides a powerful, flexible and robust tool for building prediction and classification models from potentially high-dimensional data and represents the state-of-the-art in Bayesian machine learning techniques.

2 Horseshoe+ estimator

Following [4], we model the half-Cauchy distribution as a scale mixture of inverse gamma distributions. Specifically, let x and a be random variables such that

$$x^2|a \sim \mathcal{IG}(1/2, 1/a) \quad \text{and} \quad a \sim \mathcal{IG}(1/2, 1/A^2); \tag{9}$$

then $x \sim \mathcal{C}^{+}(0, A)$ [7], where $\mathcal{IG}(\cdot, \cdot)$ is the inverse gamma distribution (see Appendix A). The decomposition (9) may be used to represent the horseshoe+ prior distributions (6)–(8) through the following latent variable representation:

$$\begin{aligned}
\lambda_j^2 | \nu_j &\sim \mathcal{IG}(1/2, 1/\nu_j), \\
\tau^2 | \xi &\sim \mathcal{IG}(1/2, 1/\xi), \\
\nu_j | \eta_j^2 &\sim \mathcal{IG}(1/2, 1/\eta_j^2), \\
\eta_j^2 | \phi_j &\sim \mathcal{IG}(1/2, 1/\phi_j), \\
\phi_1, \dots, \phi_p, \xi &\sim \mathcal{IG}(1/2, 1).
\end{aligned}$$

where $(j = 1, \dots, p)$. Although the above hierarchy introduces a number of additional latent variables, it leads to conjugate conditional posterior distributions for all parameters, greatly simplifying the resulting Gibbs sampling procedure.

2.1 Gibbs sampling for the horseshoe+

This section details the conditional posterior distributions for the horseshoe+ parameters required for the Gibbs sampler [3]. An advantage of the hierarchy (1)–(8) is that the conditional posterior distributions of the horseshoe+ do not depend on the choice of the sampling distribution of the data (1)–(3). Let

$$z_i = \begin{cases} (y_i^* - \frac{1}{2})/\omega_i^2 & \text{if } y_i \in \{0, 1\} \\ y_i & \text{otherwise} \end{cases}, \tag{10}$$

$$e_i = z_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} - \beta_0, \tag{11}$$

$$\boldsymbol{\Omega} = \sigma^2 \operatorname{diag}(\omega_1^2, \dots, \omega_n^2),$$

where $(y_1^*, \dots, y_n^*)$ are latent variables defined in Section 3.3. The conditional posterior distribution for the intercept term $\beta_0 \in \mathbb{R}$ is the Gaussian distribution $\mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$ where:

$$\tilde{\mu} = \left(\sum_{i=1}^{n} \frac{z_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}}{\omega_i^2} \right) \left(\sum_{i=1}^{n} \frac{1}{\omega_i^2} \right)^{-1}, \qquad \tilde{\sigma}^2 = \sigma^2 \left(\sum_{i=1}^{n} \frac{1}{\omega_i^2} \right)^{-1}$$

From the seminal paper by Lindley and Smith [8], the conditional posterior distribution for the regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ is the p-variate Gaussian distribution $\mathcal{N}_p\left(\tilde{\boldsymbol{\mu}}, \mathbf{A}^{-1}\right)$ where:

$$\begin{aligned}
\tilde{\boldsymbol{\mu}} &= \mathbf{A}^{-1} \mathbf{X}^{\mathrm{T}} \boldsymbol{\Omega}^{-1} (\mathbf{z} - \beta_0 \mathbf{1}_n) \\
\mathbf{A} &= \left(\mathbf{X}^{\mathrm{T}} \boldsymbol{\Omega}^{-1} \mathbf{X} + \boldsymbol{\Lambda}^{-1} \right) \\
\boldsymbol{\Lambda} &= \tau^2 \sigma^2 \operatorname{diag}(\lambda_1^2, \dots, \lambda_p^2)
\end{aligned}$$

This paper uses Rue's algorithm [9] 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 $(n > p)$. Rue's algorithm is based on Cholesky factorisation of the conditional posterior variance matrix and has cubic complexity in terms of the number of predictors p. An alternative approach to sampling multivariate Gaussian densities of this form was recently introduced by Cong et al. [10]. When the number of predictors is greater than the sample size ($p > n$), we instead use the sampling algorithm by Bhattacharya et al. [11], which has linear complexity in p.

In the case of continuous data $y_i \in \mathbb{R}$, the conditional posterior distribution for ($\sigma^2 > 0$) is the inverse gamma distribution $\mathcal{IG}(\tilde{\alpha}, \tilde{\beta})$ where:

$$\tilde{\alpha} = \frac{n+p}{2}, \qquad \tilde{\beta} = \frac{1}{2}\left(\sum_{i=1}^{n} \frac{e_i^2}{\omega_i^2} + \sum_{j=1}^{p} \frac{\beta_j^2}{\tau^2 \lambda_j^2} \right).$$

In the case of binary data, the noise variance parameter is simply set to ($\sigma^2 = 1$) and does not require sampling. The conditional posterior densities for the remaining hyperparameters are:

$$\begin{aligned}
\lambda_j^2 | \beta_j, \nu_j, \tau^2, \sigma^2 &\sim \mathcal{IG}\left(1, \frac{1}{\nu_j} + \frac{\beta_j^2}{2\tau^2\sigma^2}\right), \\
\tau^2 | \boldsymbol{\beta}, \boldsymbol{\lambda}, \xi, \sigma^2 &\sim \mathcal{IG}\left(\frac{p+1}{2}, \frac{1}{\xi} + \frac{1}{2\sigma^2} \sum_{j=1}^{p} \frac{\beta_j^2}{\lambda_j^2}\right), \\
\nu_j | \eta_j, \lambda_j &\sim \mathcal{IG}\left(1, \frac{1}{\eta_j^2} + \frac{1}{\lambda_j^2}\right), \\
\xi | \tau^2 &\sim \mathcal{IG}\left(1, 1 + \frac{1}{\tau^2}\right), \\
\eta_j^2 | \nu_j, \phi_j &\sim \mathcal{IG}\left(1, \frac{1}{\nu_j} + \frac{1}{\phi_j}\right), \\
\phi_j | \eta_j &\sim \mathcal{IG}\left(1, 1 + \frac{1}{\eta_j^2}\right).
\end{aligned}$$

Importantly, the conditional posterior distributions for all horseshoe+ hyperparameters are inverse gamma distributions for which computationally efficient sampling algorithms are readily available.

3 Robust data models

The decision to represent the sampling distribution of the data as a Gaussian scale mixture distribution naturally extends the data model (1)–(3) to a wide range of non-Gaussian distributions. For example, the Student-t and Laplace distributions can be represented in a scale mixture form. These distributions can be used to form hierarchical Bayesian estimators that are robust to data outliers and heavy-tailed errors. Additionally, Polson et al. [12] have recently extended this scale mixture representation to model discrete data through logistic regression (see Section 3.3) and negative binomial regression.

For the following robust linear regression models, we let $(z_i = y_i)$ as in equation (10). In the case of standard linear regression with Gaussian errors, the latent variables $(\omega_i^2 = 1)$ for all $i = (1, \dots, n)$ and sampling of ω_i^2 is not required. The prior distributions and corresponding conditional posterior densities for regression with Laplace noise (see Section 3.1) and regression with Student-t noise (see Section 3.2) are discussed below.

3.1 Regression with Laplace noise

It is well known that estimators using the Gaussian distribution to model errors can be negatively influenced by even a single outlying data point. This is due to the light tails of the Gaussian distribution, which are not able to capture large departures from the distribution mean. A popular alternative to the Gaussian distribution for modelling outliers is the Laplace distribution, which has heavier tails while still possessing finite mean and variance (see Appendix A.4). The Laplace distribution may be represented as a Gaussian-exponential scale mixture distribution where the scale parameters follow

$$\omega_i^2 \sim \text{Exp}(1), \quad (i = 1, \dots, n),$$

and Exp(1) denotes the exponential distribution with a mean of one. This choice of mixing distribution ensures that the scale parameter σ^2 is equal to the variance of the residuals $(e_i \in \mathbb{R})$ (see (11)), as in standard Gaussian regression.

Given the residuals $(e_1, \dots, e_n)$ and the variance parameter σ^2, the conditional posterior distribution of $1/\omega_i^2$ is

$$\frac{1}{\omega_i^2} \mid y_i, \mathbf{x}_i, \boldsymbol{\beta}, \beta_0, \sigma^2 \sim \text{IGauss}\left(\left(\frac{2\sigma^2}{e_i^2} \right)^{\frac{1}{2}}, 2 \right)$$

where IGauss(μ, λ) denotes the inverse Gaussian density with mean $(\mu > 0)$ and shape parameter $(\lambda > 0)$ (see Appendix A.2).

3.2 Regression with student-*t* noise

The Student-t distribution is often used in Bayesian robust regression as an alternative to both the Gaussian and Laplace distributions. The Student-t distribution is parameterised by a location, a scale and a degrees of freedom parameter $(\delta > 0)$ (see Appendix A.3). When $(\delta = 1)$, the Student-t distribution is equal to the Cauchy distribution. For all finite values of $(\delta > 0)$ the Student-t distribution has heavier tails than the Gaussian distribution. For $(\delta < 6)$, the Student-t distribution has heavier tails than the Laplace distribution and results in estimators that are significantly more resistant to outliers in the data. One potential drawback of modelling the errors with the Student-t distribution is that it has infinite variance when ($\delta \leq 2$).

The Student-t distribution with degrees of freedom ($\delta > 0$) may be represented as a Gaussian-inverse gamma scale mixture distribution where

$$\omega_i^2 \sim \mathcal{IG}\left(\frac{\delta}{2}, \frac{\delta}{2}\right).$$

Given the residuals $(e_1, \ldots, e_n)$ and the scale parameter σ^2, the conditional posterior distribution of ω_i^2 is

$$\omega_i^2 \mid y_i, \mathbf{x}_i, \boldsymbol{\beta}, \beta_0, \sigma^2, \delta \sim \mathcal{IG}\left(\frac{\delta+1}{2}, \frac{1}{2}\left(\frac{e_i^2}{\sigma^2} + \delta\right)\right).$$

When the degrees of freedom parameter ($\delta > 2$), the variance of the residuals is related to σ^2 by

$$\mathrm{Var}(e_i) = \sigma^2\left(\frac{\delta}{\delta - 2}\right).$$

When ($\delta \leq 2$), the parameter σ^2 can only be interpreted as a scale parameter.

3.3 Binary data

When dealing with binary data, we model the relationship between the predictors ($\mathbf{x}_i \in \mathbb{R}^p$) and the outcome variable $y_i \in \{0, 1\}$ $(i = 1, \ldots, n)$ using logistic regression. Directly sampling from the logistic regression model is difficult due to the analytic form of the likelihood function. Recently, Polson et al. [12] introduced a latent variable representation of the logistic likelihood function that is easily integrated into the hierarchy (1)–(8). Here, the logistic likelihood function is represented as a Gaussian scale mixture with a Pólya-gamma mixing density (see Appendix A.5).

Implementation of logistic regression within the hierarchy (1)–(8) requires only a minor change in the way the latent variables $(\omega_1, \ldots, \omega_n)$ and the scale parameter σ^2 are handled. In the case of logistic regression, we set ($y_i^* = y_i$) for all $i = (1, \ldots, n)$ in (10), ($\sigma^2 = 1$) and sample the latent variables $(\omega_1, \ldots, \omega_n)$ from

$$\frac{1}{\omega_i^2} \mid \mathbf{x}_i, \boldsymbol{\beta}, \beta_0 \sim \mathrm{PG}(1, \beta_0 + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}),$$

where $\mathrm{PG}(\cdot, \cdot)$ is a Pólya-gamma distribution (see Appendix A.5). An algorithm for efficient sampling from the Pólya-gamma distribution was recently introduced by Windle et al. [13] and is employed in this paper.

Robust logistic regression In real data, misrecording of the labels of the outcome variable or the values of the predictors is relatively common. For continuous outcomes, a standard approach to handle these data errors is to employ heavy-tailed, non-Gaussian error models to moderate the effects of erroneous data points. In the case binary outcome variables, the only possible mislabelling of the outcome is a transposition, a change from $(y_i = 0)$ to $(y_i = 1)$ or from $(y_i = 1)$ to $(y_i = 0)$. The effect of mislabelled binary data is to reduce the magnitude of the observed association, in contrast with the continuous case where data contamination usually results in artificially inflated estimates. For binary data, an interesting symmetry exists between the labels y_i and the predictors $\mathbf{x}_i$ in the sense that mislabelling an outcome variable is essentially equivalent to moving a predictor from one side of the decision plane to the other.

In this paper, we handle contaminated binary data by adapting the misclassification model introduced in [14] and later analysed by Copas [15] and Carrol and Pederson [16]. Let

$$\begin{aligned}
&\mathbb{P}(y_i = 1 | y_i^* = 0, \gamma_0) = \gamma_0, \\
&\mathbb{P}(y_i = 0 | y_i^* = 0, \gamma_0) = (1 - \gamma_0), \\
&\mathbb{P}(y_i = 0 | y_i^* = 1, \gamma_1) = \gamma_1, \\
&\mathbb{P}(y_i = 1 | y_i^* = 1, \gamma_1) = (1 - \gamma_1),
\end{aligned}$$

where $(y_1^*, \ldots, y_n^*)$ are latent variables representing the unobserved, correctly specified values of the observed outcomes $(y_1, \ldots, y_n)$, γ_0 is the probability of a $(y_i^* = 0)$ being misrecorded as a $(y_i = 1)$, and γ_1 is the probability of a $(y_i^* = 1)$ being misrecorded as a $(y_i = 0)$.

Rather than fix the values of the hyperparameters γ_0 and γ_1 *a priori*, they are integrated into the hierarchy (1)–(8) and sampled along with all other parameters. To simplify sampling, γ_0 and γ_1 are assigned independent Beta(1/2,1/2) prior distributions. Bayesian estimation of this misspecification model involves sampling from the conditional posterior distributions of the latent variables $(y_1^*, \ldots, y_n^*)$

$$\mathbb{P}(y_i^* = j | y_i, \beta_0, \boldsymbol{\beta}) \propto \mathbb{P}(y_i | y_i^* = j, \gamma_j)\ \mathbb{P}(y_i = j | \beta_0, \boldsymbol{\beta}), \tag{12}$$

where

$$\mathbb{P}(y_i = 1 | \beta_0, \boldsymbol{\beta}) = \frac{1}{1 + \exp(-\beta_0 - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta})}$$

is the logistic function. Normalising (12) is straightforward given that the conditional posterior distributions of $(y_1^*, \ldots, y_n^*)$ are independent and y_i^* is a binary variable $(y_i^* \in \{0, 1\})$. The conditional posterior distributions of the hyperparameters $(\gamma_0,\ \gamma_1)$ are

$$\gamma_k \mid (y_1^*, \ldots, y_n^*), (y_1, \ldots, y_n) \sim \mathrm{Beta}(n_{0k} + 1/2, n_{1k} + 1/2),$$

where $(k = 1, 2)$ and

$$n_{jk} = \sum_{i=1}^{n} \mathbb{I}(y_i = j, y_i^* = k)$$

is the count of how many times we observe $(y_i = j)$ together with $(y_i^* = k)$.

p	SNR	Gaussian		Laplace		Student-t ($\delta = 5$)	
		HS	HS+	HS	HS+	HS	HS+
50	1	3.308 (0.18)	**3.273** (0.17)	3.197 (0.15)	**3.175** (0.14)	3.261 (0.18)	**3.229** (0.17)
	4	0.816 (0.03)	**0.810** (0.03)	0.792 (0.03)	**0.789** (0.02)	0.806 (0.04)	**0.801** (0.03)
	8	0.407 (0.02)	**0.405** (0.02)	0.396 (0.01)	**0.394** (0.01)	0.403 (0.02)	**0.400** (0.02)
100	1	6.393 (0.71)	**6.324** (0.75)	5.787 (0.50)	**5.704** (0.50)	5.951 (0.51)	**5.862** (0.53)
	4	1.447 (0.09)	**1.427** (0.09)	1.367 (0.05)	**1.354** (0.05)	1.398 (0.07)	**1.381** (0.07)
	8	0.719 (0.04)	**0.711** (0.04)	0.682 (0.03)	**0.676** (0.03)	0.697 (0.03)	**0.689** (0.03)
200	1	**16.61** (1.49)	16.75 (1.59)	**16.28** (1.36)	16.36 (1.45)	**16.22** (1.50)	16.37 (1.63)
	4	3.705 (0.60)	**3.498** (0.59)	3.379 (0.52)	**3.181** (0.42)	3.417 (0.73)	**3.272** (0.67)
	8	1.712 (0.22)	**1.631** (0.20)	1.565 (0.15)	**1.512** (0.13)	1.592 (0.21)	**1.540** (0.19)

Table 1. Mean squared prediction errors for the horseshoe (HS) and horseshoe+ estimators computed over 100 test iterations (standard errors shown in parenthesis). For each test, the number of non-zero coefficients was set to 5% of the total number of predictors p. The sample sizes of the training and test data were set to $n = 100$ and $n = 10^5$ respectively.

The advantages of the Bayesian misspecification model over the frequentist approaches [15, 16] is that the Bayesian model automatically provides posterior estimates and credible intervals for both hyperparameters (γ_0, γ_1). These hyperparameters can be interpreted as the estimated rate of transposition errors in a given data set. Furthermore, the posterior expectation

$$\mathbb{E}\left(\mathbb{I}(y_i \neq y_i^*)|y_1, \ldots, y_n\right)$$

can be interpreted as the estimated posterior probability of an individual data point being misrecorded which is a useful diagnostic when analysing real data.

4 Results and discussion

This section compares our novel implementation of the horseshoe+ estimator against the horseshoe estimator implemented in [4]. As the key difference between these estimators is an extra level of Gibbs sampling required for the horseshoe+, the computational speed and memory usage of both implementations is expected to be approximately equivalent. Further, based on the theoretical analysis presented in [2] for the multiple means problem, the horseshoe+ estimator is expected to outperform the horseshoe estimator in terms of asymptotic prediction error (i.e., as the sample size $n \to \infty$). However, non-asymptotic comparison of the two estimators in Gaussian and non-Gaussian regression models has not been attempted to date. Consequently, we have chosen to compare the horseshoe and horseshoe+ estimators on their empirical prediction performance in finite sample regression tests.

4.1 Robust linear regression

The test procedure for comparing the two estimators is now described. For each test, we first generated a training set of predictors $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ where each $\mathbf{x}_i$ $(i = 1, \ldots, n)$ was generated from a zero mean p-variate Gaussian distribution

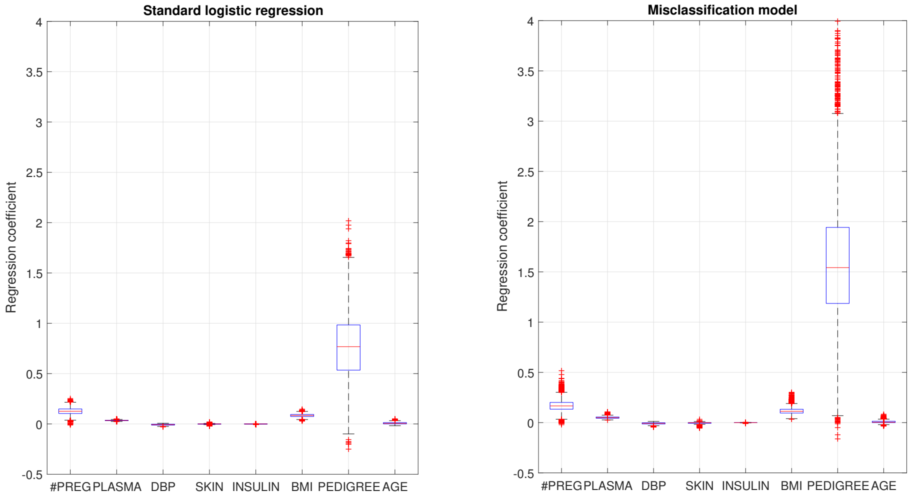

Fig. 1. Analysis of the Pima Indian diabetes data set with the Bayesian horseshoe+ estimator. The left Figure depicts the posterior estimates of the regression coefficients from the non-robust logistic regression model. The right Figure shows the regression coefficients estimated from the misclassification model.

$\mathbf{x}_i \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p)$ with the variance–covariance matrix equal to the identity matrix $\mathbf{I}_p$. The sample size of the training data was set to $(n = 100)$ in each test iteration. The training data $(\mathbf{y} \in \mathbb{R}^n)$ was generated from the model

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}^* + \epsilon_i, \quad (i = 1, \ldots, n),$$

where $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is the vector of true regression coefficients and the noise variable ϵ_i follows a zero-mean Gaussian, Laplace or Student-t distribution. The variance of the noise variables was chosen to control a pre-specified signal-to-noise (SNR) ratio for each test. The regression coefficients $\boldsymbol{\beta}^*$ were chosen to represent the ultra-sparse data model for which the horseshoe and horseshoe+ estimators were originally developed. In particular,

$$\boldsymbol{\beta}^* = (1, 1, \ldots, 1, 0, 0, \ldots, 0)^{\mathrm{T}},$$

where 5% of the entries of $\boldsymbol{\beta}^*$ were set to one, with the remaining 95% of the entries set to zero. The complete test procedure was repeated for 100 iterations for $p = \{50, 100, 200\}$ and SNR $= \{1, 4, 8\}$. The performance of the two estimators was measured using the mean squared prediction error metric computed on independently generated test data at the Rao-Blackwellised posterior mean. For each estimator, the posterior mean was computed from $1,000$ samples from the corresponding posterior distribution with a 'burnin' period of $1,000$ samples and a thinning level of 5. The sample size of the test data sets was set to $n = 10^5$. In the spirit of reproducible research, MATLAB code implementing both estimators and all simulations will be made available on the authors' web pages[1].

Table 1 presents the mean squared predictions errors for the horseshoe and horseshoe+ estimators. In the case of ($p = 200$) and (SNR $= 1$), the horseshoe estimator resulted in smaller prediction error in contrast to the new horseshoe+ estimator. However, the horseshoe+ estimator obtained improved prediction error for all other combinations of SNR and p that were examined. Importantly, the horseshoe+ estimator performed well even when the noise distribution was non-Gaussian. Consequently, the finite sample performance of the horseshoe+ estimator appears to be at least as good as the original horseshoe procedure in the robust linear regression setting.

4.2 Robust logistic regression

This section demonstrates the utility of the logistic regression misclassification model, discussed in Section 3.3, when applied to the analysis of a real data set. We consider the Pima Indians diabetes data set [17] collected by the National Institute of Diabetes and Digestive and Kidney Diseases and available for download from the UCI Machine Learning Repository. The data set comprises eight ($p = 8$) predictor variables and ($n = 768$) observations collected from female patients of Pima Indian heritage who are at least 21 years of age. The aim is to build a prediction model of diabetes from the eight predictor variables. The original donor of the data set indicated that there are no missing values in the data. However, this appears to be erroneous because the data contains predictors with values that are biologically implausible. For example, the variable diastolic blood pressure contains entries with zero blood pressure, which is clearly incorrect.

The Pima Indians diabetes data set was analysed using the following two models: (i) standard (non-robust) logistic regression and (ii) the misclassification model (see Section 3.3). For both analyses, the Bayesian horseshoe+ estimator was used to estimate all unknown parameters. Box and whisker plots of the posterior estimates of the regression coefficients are shown in Figure 1. It is clear that the observed associations under the misclassification model (Figure 1, right) are inflated in contrast with standard logistic regression (Figure 1, left). This is expected because the effect of ignoring data contamination in binary logistic regression is a reduction in the absolute magnitude of the regression coefficients. The posterior estimates of the rate of transposition errors in the data were found to be relatively large, ($\gamma_0 = 0.03$) and ($\gamma_1 = 0.13$). Further analysis of the unobserved latent variables ($y_1^*, \ldots, y_n^*$) was then used to discover which particular observations are likely to be contaminated. A manual examination of the top 10 observations identified as highly likely to contain contaminated data was then performed. These top observations were found to be clearly erroneous because they contained biologically impossible predictor values, highlighting the usefulness of the robust logistic regression model.

[1] `www.emakalic.org/blog` and `www.dschmidt.org`

A Appendix

A.1 Inverse gamma distribution

The inverse gamma probability density function is given by

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right), \quad (x > 0),$$

with shape parameter $(\alpha > 0)$ and scale parameter $(\beta > 0)$. The first two moments are

$$\mathrm{E}\left(x\right) = \frac{\beta}{\alpha - 1}, \quad \mathrm{Var}\left(x\right) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)},$$

where the mean and variance only exist for $(\alpha > 1)$ and $(\alpha > 2)$ respectively.

A.2 Inverse Gaussian distribution

The inverse Gaussian probability density function is given by

$$p(x|\mu, \lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right),$$

for $(x > 0)$, where $(\mu > 0)$ is the mean and $(\lambda > 0)$ is the shape parameter. The first two moments are

$$\mathrm{E}\left(x\right) = \mu, \quad \mathrm{Var}\left(x\right) = \frac{\mu^3}{\lambda}.$$

A.3 Student-*t* distribution

The Student-*t* distribution probability density function is given by

$$p(x|\mu, \sigma^2, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma^2}} \left(1 + \frac{1}{\nu}\frac{(x-\mu)^2}{\sigma^2}\right)^{-\frac{\nu+1}{2}},$$

where $(x \in \mathbb{R})$, $(\mu \in \mathbb{R})$, $(\sigma^2 > 0)$ and the degrees of freedom $(\nu > 0)$. The first two moments are

$$\mathrm{E}\left(x\right) = \mu, \quad (\nu > 1), \quad \mathrm{Var}\left(x\right) = \sigma^2\left(\frac{\nu}{\nu-2}\right), \tag{13}$$

where the mean and variance only exist for $(\nu > 1)$ and $(\nu > 2)$ respectively.

A.4 Laplace distribution

The probability density function of the Laplace distribution is

$$p(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right),$$

where $(x \in \mathbb{R})$, $(\mu \in \mathbb{R})$ is the location parameter and $(b > 0)$ is the scale parameter. The first two moments are

$$\mathrm{E}\left(x\right) = \mu, \quad \mathrm{Var}\left(x\right) = 2b^2.$$

A.5 Pólya-gamma distribution

A random variable x follows a Pólya-gamma distribution [12], $x \sim \mathrm{PG}(b, c)$, if

$$x \overset{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k-1/2)^2 + c^2/(4\pi^2)},$$

where $g_k \sim \mathrm{Ga}(b, 1)$ are independent gamma random variables, $(b > 0)$ and $(c \in \mathbb{R})$ are the parameters and $\overset{D}{=}$ denotes equality in distribution. The first two moments of x are

$$\mathrm{E}\left(x\right) = \frac{b}{2c} \tanh\left(\frac{c}{2}\right), \quad \mathrm{Var}\left(x\right) = \frac{b}{4c^3}\left(\sinh(c) - c\right) \mathrm{sech}^2\left(\frac{c}{2}\right).$$

References

1. Andrews, D.F., Mallows, C.L.: Scale mixtures of normal distributions. Journal of the Royal Statistical Society (Series B) **36**(1) (1974) 99–102
2. Bhadra, A., Datta, J., Polson, N.G., Willard, B.: The horseshoe+ estimator of ultra-sparse signals (2015) arXiv:1502.00560.
3. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration images. IEEE Tran. Pat. Anal. Mach. **6** (1984) 721–741
4. Makalic, E., Schmidt, D.F.: A simple sampler for the horseshoe estimator. IEEE Signal Processing Letters **23**(1) (2016) 179–182
5. Carvalho, C.M., Polson, N.G., Scott, J.G.: The horseshoe estimator for sparse signals. Biometrika **97**(2) (2010) 465–480
6. Polson, N.G., Scott, J.G.: Shrink globally, act locally: Sparse Bayesian regularization and prediction. In: Bayesian Statistics. Volume 9. (2010)
7. Wand, M.P., Ormerod, J.T., Padoan, S.A., Fruhwirth, R.: Mean field variational Bayes for elaborate distributions. Bayesian Analysis **6**(4) (2011) 847–900
8. Lindley, D.V., Smith, A.F.M.: Bayes estimates for the linear model. Journal of the Royal Statistical Society (Series B) **34**(1) (1972) 1–41
9. Rue, H.: Fast sampling of Gaussian markov random fields. Journal of the Royal Statistical Society (Series B) **63**(2) (2001) 325–338
10. Cong, Y., Chen, B., Zhou, M.: Fast simulation of hyperplane-truncated multivariate normal distributions (2016)
11. Bhattacharya, A., Pati, D., Pillai, N.S., Dunson, D.B.: Dirichlet—Laplace priors for optimal shrinkage. Journal of the American Statistical Association **110** (2015) 1479–1490
12. Polson, N.G., Scott, J.G., Windle, J.: Bayesian inference for logistic models using Pólya-gamma latent variables. **108**(504) (2013) 1339–1349
13. Windle, J., Polson, N.G., Scott, J.G.: Sampling Pólya-gamma random variates: alternate and approximate techniques (2014)
14. Ekholm, A., Palmgren, J.: Correction for misclassification using doubly sampled data. (1987) 419–429
15. Copas, J.B.: Binary regression models for contaminated data. Journal of the Royal Statistical Society, Series B (Methodological) **50**(2) (1988) 225–265
16. Carroll, R.J., Pederson, S.: On robustness in the logistic regression model. **55**(3) (1993) 693–706
17. Lichman, M.: UCI machine learning repository (2013)