Bayesian Grouped Horseshoe Regression with Application to Additive Models

Zemei Xu, Daniel F. Schmidt, Enes Makalic, Guoqi Qian, and John L. Hopper

Centre for Epidemiology and Biostatistics,
Melbourne School of Population and Global Health,
School of Mathematics and Statistics,
The University of Melbourne, VIC 3010
zemeix@student.unimelb.edu.au
{dschmidt,emakalic,qguoqi,j.hopper}@unimelb.edu.au

Abstract. The Bayesian horseshoe estimator is known for its robustness at handling noisy and sparse big data problems. This paper presents two extensions of the regular Bayesian horseshoe: (i) the grouped Bayesian horseshoe, and (ii) the hierarchical Bayesian grouped horseshoe. The advantages of the proposed methods are their flexibility of handling grouped variables through extra shrinkage parameters at the group and within-group levels. We apply the proposed methods to the important class of additive models where group structures naturally exist and demonstrate that the grouped hierarchical Bayesian horseshoe has promising performance on both simulated and real data.

Keywords: Bayesian regression, grouped variables, horseshoe, additive models.

1 Introduction

Statistical variable selection, also known as feature selection, has become an indispensable tool in many different research areas involving machine learning and data mining. The object is to select the best subset of predictors for fitting or predicting the response variable from a collection of candidate predictors of possibly very large size. It is particularly important in high dimensional problems, where there are potentially millions of predictors and only a few of them are associated with the outcome.

Consider the linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y}$ is an n by 1 observation vector of the response variable, $\mathbf{X}$ is an n by p observation or design matrix of the regressors or predictors, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is a p by 1 vector of regression coefficients to be estimated, and $\boldsymbol{\epsilon}$ is an n by 1 vector of i.i.d. $\mathcal{N}(0, \sigma^2)$ random errors with σ^2 unknown. Here, the vector $\boldsymbol{\beta}$

is assumed to be sparse in the sense that most of its components equal zero in truth. Therefore, dimensionality reduction is necessary, especially for large-p problems.

Recent approaches for variable selection in ultra-high dimensions are based on penalised likelihood methods, which select a model by minimising a loss function that is usually proportional to the negative log likelihood plus a penalty term:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \cdot q(\boldsymbol{\beta}) \right\}, \tag{2}$$

where $\lambda > 0$ is the regularisation parameter and $q(\cdot)$ is a penalty function. Many penalised regression approaches have been proposed in the literature, such as the non-negative garrote [3], ridge regression [8], the adaptive lasso [14], the elastic net [15], smoothly clipped absolute deviation (SCAD) [7], etc. One of the most widely used penalised approaches is the lasso [11] which shrinks coefficients while setting other coefficients to zero, effectively producing a simpler interpretable model. However, penalised regression methods such as the lasso, are designed for selecting individual explanatory variables.

1.1 Grouped Variables

Group structures naturally exist in predictor variables, and in this situation, variable selection should select groups of variables, rather than individual variables. For example, a multi-level categorical predictor in the regression model can be represented by a group of dummy variables; a continuous predictor in the additive model can be expressed as a composition of basis functions; and prior knowledge such as genes in the same biological pathway also can be used to form a natural group.

In order to find sparse solutions at group level, several group selection methods have been proposed, such as the group lasso [12], the group lasso for logistic regression [4], group selection with general composite absolute penalty [13], group bridge method [9], and bi-level selection in generalised linear models [2].

1.2 Bayesian Regression

An important class of alternative variable selection methods are the Bayesian penalised regression approaches. These are motivated by the fact that a good solution for β in Equation (1) can be interpreted as the posterior mode of β in the Bayesian model when β follows a certain prior distribution. For example, the assumption of a Laplace prior distribution for β leads to the Bayesian interpretation of the lasso [11].

A natural way of estimating β in the Bayesian approaches is by generating sparsity through a ‘spike and slab’ prior for each element of β, the spike concentrating near zero and the slab spreading away from zero:

$$\beta_j | \gamma_j \sim (1 - \gamma_j)\mathcal{N}(0, \tau_j^2) + \gamma\mathcal{N}(0, c_j \tau_j^2), \; j = 1, \cdots, p, \tag{3}$$

where γ_j are binary variables, τ_j^2 is small and $c_j > 0$ is large. However, a fully Bayesian approach with a spike and slab mixture for each β component requires exploration of a model space of size 2^p which becomes difficult when p is large.

A competitive Bayesian alternative to the lasso is the Bayesian model resulting from the horseshoe prior, which is a one-component prior [5]. The horseshoe arises from the same class of multivariate scale mixtures of normals as the lasso does, but it is almost universally superior to the double-exponential prior at handling sparsity [6]. Furthermore, the horseshoe prior is known for its robustness at handling large outlying signals. The estimator of the horseshoe model does not face the computational issues of the point mass mixture models. However, variable selection methods for grouped variables based on the Bayesian horseshoe models have not been analysed in the literature to date.

In this paper, we extend the Bayesian horseshoe model to handle the grouped variables by introducing shrinkage parameters at group level as well as within each group. We apply the proposed methods to the important class of additive models where group structures naturally exist. Both simulated and real data experiments demonstrate the promising performance of the proposed methods.

2 Bayesian Grouped Horseshoe models

In this section, the Bayesian horseshoe model is briefly introduced and two extensions of the Bayesian horseshoe model for grouped variables are proposed.

2.1 Bayesian Horseshoe Model

The horseshoe prior assumes that β_j are conditionally independent, and each of them has a density function that can be represented as a scale mixture of normals. It leaves strong signals unshrunk and penalises noise variables severely. Therefore, the horseshoe prior has the ability of adapting to different sparsity patterns, while simultaneously avoiding over-shrinkage of large coefficients.

Without loss of generality, $\mathbf{y}$ and $\mathbf{X}$ are assumed to be standardised for all models. The response $\mathbf{y}$ is centered and the covariates $\mathbf{X}$ are column-standardised to have mean zero and unit length. The Bayesian horseshoe estimator is defined as follows:

$$\begin{aligned}
\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2 &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_n),\\
\boldsymbol{\beta}|\sigma^2,\tau^2,\delta_1,\cdots,\delta_p &\sim \mathcal{N}(\mathbf{0},\sigma^2\tau^2\mathbf{D}_{\boldsymbol{\delta}}),\ \text{where } \mathbf{D}_{\boldsymbol{\delta}} = \mathrm{diag}(\delta_1^2,\cdots,\delta_p^2),\\
\delta_j &\sim C^+(0,1),\ j=1,\cdots,p,\\
\tau &\sim C^+(0,1),\\
\sigma^2 &\sim \frac{1}{\sigma^2}d\sigma^2,
\end{aligned} \tag{4}$$

where $C^+(0,1)$ is a standard half-Cauchy distribution with the probability density function:

$$f(x) = \frac{2}{\pi(1+x^2)},\ x>0. \tag{5}$$

The scale parameters δ_j are local shrinkage parameters, and τ is the global shrinkage parameter. A simple sampler proposed for the Bayesian horseshoe hierarchy [10] enables straightforward sampling of the full conditional posterior distributions.

2.2 Bayesian Grouped Horseshoe Model

The original Bayesian horseshoe model does not allow for grouped structure. Therefore, the Bayesian grouped horseshoe is proposed. Suppose there are $G \in \{1, \cdots, p\}$ groups of predictors in the data and the gth group has size s_g, where $g = 1, \cdots, G$ (i.e. there are s_g variables in group g). The horseshoe hierarchical representation of the full model for grouped variables can be constructed as:

$$\begin{aligned} \mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n), \\ \boldsymbol{\beta}|\sigma^2, \tau^2, \lambda_1^2, \cdots, \lambda_G^2 &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \tau^2 \mathbf{D}_{\boldsymbol{\lambda}}), \text{ where } \mathbf{D}_{\boldsymbol{\lambda}} = \operatorname{diag}(\lambda_1^2 \mathbf{I}_{s_1}, \cdots, \lambda_G^2 \mathbf{I}_{s_G}), \\ \lambda_g &\sim C^+(0,1), \; g = 1, \cdots, G, \\ \tau &\sim C^+(0,1), \\ \sigma^2 &\sim \frac{1}{\sigma^2} d\sigma^2, \end{aligned} \tag{6}$$

where λ_g are the shrinkage parameters at group level. Instead of having shrinkage parameters for all predictors, we have shrinkage parameters λ_g for each group. Hence, the model either shrinks all variables in the same group towards zero or leaves them untouched.

2.3 Hierarchical Bayesian Grouped Horseshoe Model

By combining the regular horseshoe with the grouped horseshoe, we develop a hierarchical Bayesian grouped horseshoe model that does selection and shrinking at group level as well as within groups.

Suppose the total number of groups G is assumed to be greater than one, since $G = 1$ implies that all predictors are in the same group which results in the regular Bayesian horseshoe model. The full hierarchical Bayesian grouped horseshoe model is:

$$\begin{aligned} \mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n), \\ \boldsymbol{\beta}|\sigma^2, \tau^2, \lambda_1^2, \cdots, \lambda_G^2, \delta_1, \cdots, \delta_p &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \tau^2 \mathbf{D}_{\boldsymbol{\lambda}} \mathbf{D}_{\boldsymbol{\delta}}), \\ \text{where } \mathbf{D}_{\boldsymbol{\lambda}} = \operatorname{diag}(\lambda_1^2 I_{s_1}, \cdots, \lambda_G^2 I_{s_G}) \; \mathbf{D}_{\boldsymbol{\delta}} &= \operatorname{diag}(\lambda_1^2 \mathbf{I}_{s_1}, \cdots, \lambda_G^2 \mathbf{I}_{s_G}), \\ \lambda_g &\sim C^+(0,1), \; g = 1, \cdots, G, \\ \delta_j &\sim C^+(0,1), \; j = 1, \cdots, p, \\ \tau &\sim C^+(0,1), \\ \sigma^2 &\sim \frac{1}{\sigma^2} d\sigma^2, \end{aligned} \tag{7}$$

where $\delta_1, \cdots, \delta_p$ are the shrinkage parameters for each predictor variable and $\lambda_1, \cdots, \lambda_G$ are the shrinkage parameters for group variables. Therefore, the model has a global shrinkage parameter τ, local shrinkage parameters λ and δ which control shrinkage at and within group levels, respectively. The full conditional distributions required to sample from (7) for HBHSG can be found in the Appendix.

3 Additive Models

Nonparametric regression methods, such as the additive model are widely used in statistics. In an additive model, each predictor can be expressed as a set of basis functions which forms a group structure. Given a data set $\{y_i, x_{i1}, \cdots, x_{ip}\}_{i=1}^n$, the additive model has the form:

$$y = \mu_0 + \sum_{j=1}^{p} f_j(X_j) + \epsilon, \tag{8}$$

where μ_0 is an intercept term and $f_j(\cdot)$ are unknown smooth functions. In the ideal situation, estimation of the selected smooth functions is expected to be as close to the corresponding true underlying functions or target functions as possible.

There are various classes of basis functions that can be used to approximate the target functions. The basis functions include polynomials, spline functions, etc. Let $g_j(x), j = 1, \cdots, p$, be a set of basis functions. Each smooth function component in the additive model can be represented as:

$$f(x) = a_0 + a_1 g_1(x) + a_2 g_2(x) + \cdots + a_p g_p(x). \tag{9}$$

One limitation of the regular polynomial basis functions (i.e. $g_j(x) = x^{j-1}$) is that there potentially exists correlation between those data generated by the basis functions. As a result, orthogonal polynomials are widely used. In this paper, one of the orthogonal polynomials, the Legendre polynomials, are used for all polynomial expansions. The Legendre polynomials are defined on the interval $[-1, 1]$. Each Legendre polynomial $g_j(x)$ is a pth-degree polynomial, and it can be expressed as:

$$g_j(x) = \frac{1}{2^j j!} \frac{d^j}{dx^j} [(x^2 - 1)^j], \ p = 0, 1, \cdots . \tag{10}$$

The additive models allow for nonlinear effects and grouped structures, and therefore, form perfect test functions for our proposed methods.

4 Simulation Studies

In this section, we compared the prediction performance of four methods: (i) the regular Bayesian horseshoe method (BHS), (ii) the hierarchical Bayesian grouped horseshoe (HBHSG), (iii) lasso with regularisation parameter selected by the Bayesian information criterion (lasso-BIC), and (iv) the regular Bayesian horseshoe without expansions of the predictors (BHS-NE). The performance of the four methods were compared on three test functions (see Section 4.2).

To compare the performance of the four methods, we generated $\mathbf{X}$ with a large sample size n from a uniform distribution $U(-1, 1)$, and computed the mean squared prediction error (MSPE) as the comparison metric:

$$\frac{1}{n}\sum_{i=1}^{n}[\mathrm{E}(y_i|x_i) - \hat{y}_i]^2, \tag{11}$$

where y_i are the responses of the test functions and $\hat{y}_i$ are the fitted values based on the estimates from the four methods.

4.1 Simulation Procedures

For simulated data, we first generated $t = 100$ data sets including $p = 10$ candidate predictors $\mathbf{X}_{n\times p}$ from the uniform distribution $U(-1, 1)$. For each realisation, we calculated $\mathbf{y}$ based on the true test functions. Then, we expanded $\mathbf{X}$ using Legendre polynomials with degree K in. Three methods, BHS, HBHSG, and lasso-BIC, were applied to t samples of expanded data $\tilde{\mathbf{X}}$. The benchmark method, BHS-NE, was applied to non-expanded covariates $\mathbf{X}$. We obtained t posterior samples $\boldsymbol{\beta}$ for each method, and computed the posterior median as the point estimate of each methods.

4.2 Test functions

Simulations were performed on data generate from the three true test functions shown below. The first two true functions are linear and non-linear functions and the third one consists of Legendre polynomials.

1. Function 1 (simple linear function):

$$y = X_1 + X_2 - X_3 - X_4 \tag{12}$$

2. Function 2 (nonlinear function):

$$y = \cos(8X_1) + X_2^2 + \mathrm{sign}(X_3) + |X_4| + X_5 + X_5^2 - X_5^3 \tag{13}$$

3. Function 3:

$$y = f_1(X_1) + f_2(X_2) + f_3(X_3), \tag{14}$$

   where $f_j = \beta_{j1}P_1(X_j) + \beta_{j2}P_2(X_j) + \beta_{j3}P_3(X_j),\ j = 1, 2, 3$ that consists of the Legendre polynomials of order up to three and the standardised true coefficients are: $\boldsymbol{\beta} = (2, 1, 1/2, 1, 1, 1, -1, -4, 1)'_{9\times1}$.

Function 1 is a linear function where BHS-NE is expected to benefit the most. We expect all methods except BHS-NE to do well in Function 2 where there is a highly nonlinear structure in the model. In Function 3, we expect HBHSG to perform well as the true functions are from basis functions.

To generate three test functions, we first generated $\mathbf{X}_{n\times m} = (\mathbf{X}_1, \cdots, \mathbf{X}_m)$ with a large sample size n from $U(-1,1)$, where m is the number of non-zero components. We then scaled each component to ensure that all components have same variance and are contributed equally to the final function. The variance of the noise variables σ^2 can be computed to achieve the designated variance of y.

For all three tests functions, we generated $p = 10$ predictors and varied the number of samples $n = \{100, 200\}$, signal-to-noise ratio SNR= $\{1, 5, 10\}$ and maximum degree of Legendre polynomial expansions $K = \{3, 6, 9, 12\}$.

4.3 Simulation Results

The grouped Bayesian horseshoe results were not presented, as they performed uniformly worse than the BHS and HBHSG in all experiments.

The average mean squared prediction error and the corresponding standard deviation for Function 1 are shown in Table 1. In the simple linear test, the Bayesian horseshoe model without expansions of covariates unsurprisingly produces the smallest MSPE values consistently. From the remaining three methods, HBHSG and BHS are competitive when sample size $n = 100$. As the sample size grows, HBHSG shows significant improvement in terms of MSPE compared to BHS, and the MSPE of HBHSG is smaller in most cases when $n = 200$. The prediction performance of BHS, HBHSG and lasso-BIC drop as the degrees of expansions increases.

For Function 2 where there exists a highly nonlinear relationship in the model, the HBHSG improves performance in almost all scenarios according to Table 2. The only scenario where the BHS slightly outperforms HBHSG is when $n = 100$ and the signal-to-noise ratio is low (SNR $= 1$). The BHS-NE performs poorly in this case because it is unable to capture nonlinear effects.

For Function 3, the HBHSG is expected to achieve good performance because the true underlying additive model consists of a set of polynomial expansions. Indeed, referring to Table 3, the HBHSG gives the smallest mean MSPE for all scenarios. The performance of BHS is better than lasso-BIC and BHS-NE unsurprisingly performs poorly for these polynomial test functions.

In general, the hierarchical Bayesian grouped horseshoe method outperforms the regular Bayesian horseshoe. As an illustration, boxplots of component-wise squared prediction error for BHS and HBHSG when $p = 10$, $n = 100$, SNR $= 5$, $K = 3$ are shown in Figure 1. The first three components are associated with the response variable and the rest are noise variables. From the figure, we see the HBHSG penalises noise variables heavier than the BHS does. When all the variables within a group have small effects, the HBHSG tends to put extra shrinkage for the whole group and shrink them together.

Table 1. Mean and standard deviation (in parentheses) of squared prediction error for BHS, HBHSG, BHS-NE and lasso-BIC of Function 1, when sample size $n = \{100, 200\}$, signal-to-noise-ratio SNR $= \{1, 5, 10\}$, and the highest degree of Legendre polynomial expansions $K = \{3, 6, 9, 12\}$.

n	SNR	Degree	BHS	HBHSG	BHS-NE	lasso-BIC
100	1	3	0.1137(0.0786)	0.1107(0.0861)	0.0856(0.0513)	0.1941(0.1090)
		6	0.1258(0.0918)	0.1322(0.1034)	0.0855(0.0520)	0.2543(0.1312)
		9	0.1536(0.1274)	0.1635(0.1423)	0.0852(0.0511)	0.5293(2.3280)
		12	0.2247(0.3075)	0.2200(0.2793)	0.0857(0.0519)	17.810(18.670)
	5	3	0.0178(0.0125)	0.0173(0.0125)	0.0154(0.0086)	0.0375(0.0190)
		6	0.0173(0.0115)	0.0176(0.0123)	0.0153(0.0088)	0.0491(0.0225)
		9	0.0194(0.0144)	0.0196(0.0143)	0.0153(0.0086)	0.1029(0.4565)
		12	0.0377(0.0797)	0.0346(0.0759)	0.0154(0.0087)	3.1820(3.6550)
	10	3	0.0088(0.0062)	0.0086(0.0062)	0.0077(0.0043)	0.0185(0.0090)
		6	0.0085(0.0056)	0.0087(0.0060)	0.0077(0.0044)	0.0251(0.0113)
		9	0.0095(0.0070)	0.0096(0.0070)	0.0077(0.0043)	0.0516(0.2269)
		12	0.0186(0.0394)	0.0171(0.0380)	0.0077(0.0044)	1.4160(1.7350)
200	1	3	0.0442(0.0258)	0.0422(0.0256)	0.0378(0.0200)	0.0954(0.0461)
		6	0.0461(0.0293)	0.0468(0.0294)	0.0377(0.0198)	0.1224(0.0457)
		9	0.0473(0.0316)	0.0467(0.0318)	0.0378(0.0202)	0.1354(0.0516)
		12	0.0497(0.0352)	0.0479(0.0317)	0.0376(0.0202)	0.1487(0.0544)
	5	3	0.0084(0.0051)	0.0082(0.0052)	0.0076(0.0040)	0.0194(0.0100)
		6	0.0087(0.0058)	0.0089(0.0059)	0.0076(0.0039)	0.0242(0.0100)
		9	0.0089(0.0062)	0.0088(0.0062)	0.0076(0.0040)	0.0269(0.0100)
		12	0.0094(0.0071)	0.0091(0.0063)	0.0076(0.0040)	0.0297(0.0109)
	10	3	0.0042(0.0026)	0.0041(0.0026)	0.0038(0.0020)	0.0094(0.0047)
		6	0.0043(0.0030)	0.0044(0.0030)	0.0038(0.0020)	0.0120(0.0044)
		9	0.0044(0.0031)	0.0044(0.0031)	0.0039(0.0020)	0.0133(0.0053)
		12	0.0047(0.0037)	0.0045(0.0032)	0.0038(0.0020)	0.0148(0.0056)

5 Real Data

To evaluate the performance of the grouped Bayesian horseshoe method and the hierachichal Bayesian horseshoe method, we applied them to the Electrical-Maintenance (ELE) data set [1]. There are $n = 1056$ samples and $p = 4$ input variables in the ELE data set. All variables are continuous variables.

To perform the real data analysis, we used hold-out validation methods by dividing data into two subsets, where the training data contains 75% of the samples and the testing data contains 25% of the samples.

All predictors were expanded using Legendre polynomials with degrees $K = \{2, 4, 6, 8, 10\}$ for each of the following methods: BHS, BHSG, HBHSG, lasso-BIC. We also tested the BHS-NE with original non-expanded predictors as the benchmark.

Table 2. Mean and standard deviation (in parentheses) of squared prediction error for BHS, HBHSG, BHS-NE and lasso-BIC of Function 2, when sample size $n = \{100, 200\}$, signal-to-noise-ratio SNR $= \{1, 5, 10\}$, and the highest degree of Legendre polynomial expansions $K = \{3, 6, 9, 12\}$.

n	SNR	Degree	BHS	HBHSG	BHS-NE	lasso-BIC
100	1	3	0.5008(0.1204)	0.5122(0.1357)	0.9182(0.0756)	0.6512(0.2203)
		6	0.4968(0.1518)	0.5279(0.1721)	0.9208(0.0771)	0.6850(0.2236)
		9	0.5524(0.1757)	0.5816(0.1877)	0.9192(0.0757)	0.9774(1.8780)
		12	0.6591(0.2845)	0.6874(0.3629)	0.9199(0.0760)	18.150(20.220)
	5	3	0.3055(0.0382)	0.2947(0.0360)	0.8579(0.0488)	0.3523(0.0682)
		6	0.1728(0.0354)	0.1578(0.0321)	0.8593(0.0505)	0.2345(0.0656)
		9	0.1357(0.0495)	0.1246(0.0669)	0.8582(0.0497)	0.5216(0.9430)
		12	0.1882(0.1893)	0.1819(0.2279)	0.8582(0.0492)	3.4360(4.0510)
	10	3	0.2823(0.0293)	0.2739(0.0271)	0.8490(0.0415)	0.3189(0.0561)
		6	0.1363(0.0234)	0.1245(0.0206)	0.8501(0.0431)	0.1781(0.0430)
		9	0.0859(0.0381)	0.0792(0.0475)	0.8491(0.0423)	0.2571(0.4638)
		12	0.1287(0.1521)	0.1353(0.2316)	0.8488(0.0414)	1.6830(2.0180)
200	1	3	0.3306(0.0491)	0.3271(0.0485)	0.8452(0.0379)	0.3985(0.0765)
		6	0.2330(0.0560)	0.2220(0.0574)	0.8450(0.0372)	0.3417(0.1006)
		9	0.2125(0.0646)	0.1973(0.0661)	0.8453(0.0372)	0.3517(0.1147)
		12	0.2329(0.0783)	0.2185(0.0740)	0.8449(0.0372)	0.3820(0.1175)
	5	3	0.2545(0.0188)	0.2500(0.0179)	0.8170(0.0023)	0.2795(0.0275)
		6	0.1098(0.0154)	0.1051(0.0150)	0.8168(0.0231)	0.1464(0.0269)
		9	0.0602(0.0128)	0.0551(0.0133)	0.8172(0.0233)	0.1034(0.0289)
		12	0.0639(0.0178)	0.0580(0.0152)	0.8168(0.0231)	0.1168(0.0313)
	10	3	0.2442(0.0135)	0.2404(0.0128)	0.8136(0.0216)	0.2623(0.0187)
		6	0.0937(0.0109)	0.0906(0.0104)	0.8134(0.0215)	0.1157(0.0167)
		9	0.0406(0.0076)	0.0374(0.0085)	0.8138(0.0217)	0.0667(0.0150)
		12	0.0413(0.0104)	0.0373(0.0093)	0.8134(0.0215)	0.0721(0.0182)

The mean and standard deviation (in parentheses) of squared prediction errors are shown in Table 4. We see that there clearly exists a nonlinear pattern in the data because the mean squared prediction error decreases as the degree of expansions increases. The HBHSG on average has the lowest MSPE, especially when the degree of polynomials grows. The BHSG has the smallest MSPE when each group has few variables ($K = 2$).

In conclusion, we have proposed the grouped Bayesian horseshoe method and the hierarchical Bayesian horseshoe method for performing both group-wise and within group selection. We have shown the good performance of the hierarchical Bayesian grouped horseshoe model in terms of the mean squared prediction error on both simulated data and real data. The proposed methods outperform the regular Bayesian horseshoe method when it is applied to nonlinear functions and additive models. Even when there is no underlying group structure, it is

Table 3. Mean and standard deviation (in parentheses) of squared prediction error for BHS, HBHSG, BHS-NE and lasso-BIC of Function 3, when sample size $n = \{100, 200\}$, signal-to-noise-ratio SNR $= \{1, 5, 10\}$, and the highest degree of Legendre polynomial expansions $K = \{3, 6, 9, 12\}$.

n	SNR	Degree	BHS	HBHSG	BHS-NE	lasso-BIC
100	1	3	0.1564(0.0705)	0.1336(0.0699)	0.5514(0.0678)	0.2398(0.1146)
		6	0.1809(0.0789)	0.1645(0.0824)	0.5521(0.0677)	0.3128(0.1434)
		9	0.2142(0.0989)	0.2019(0.1079)	0.5514(0.0669)	0.6967(2.5440)
		12	0.2876(0.3553)	0.2808(0.4541)	0.5520(0.0670)	18.090(20.070)
	5	3	0.0329(0.0158)	0.0262(0.0141)	0.4951(0.0260)	0.0555(0.0247)
		6	0.0366(0.0172)	0.0318(0.0163)	0.4953(0.0261)	0.0732(0.0308)
		9	0.0421(0.0197)	0.0394(0.0194)	0.4951(0.0260)	0.1852(0.5761)
		12	0.0647(0.0969)	0.0618(0.1343)	0.4954(0.0256)	3.3320(4.3170)
	10	3	0.0179(0.0084)	0.0139(0.0073)	0.4896(0.0211)	0.0287(0.0127)
		6	0.0201(0.0086)	0.0173(0.0081)	0.4899(0.0212)	0.0386(0.0143)
		9	0.0230(0.0105)	0.0215(0.0105)	0.4896(0.0211)	0.0977(0.2842)
		12	0.0357(0.0521)	0.0339(0.0674)	0.4899(0.0209)	1.4920(1.9740)
200	1	3	0.0756(0.0318)	0.0629(0.0297)	0.4995(0.0259)	0.1235(0.0495)
		6	0.0905(0.0374)	0.0809(0.0355)	0.4994(0.0258)	0.1709(0.0702)
		9	0.0993(0.0413)	0.0894(0.0374)	0.4996(0.0262)	0.1918(0.0735)
		12	0.1056(0.0432)	0.0955(0.0396)	0.4993(0.0258)	0.2089(0.0763)
	5	3	0.0166(0.0069)	0.0133(0.0064)	0.4786(0.0110)	0.0289(0.0113)
		6	0.0192(0.0082)	0.0173(0.0080)	0.4787(0.0110)	0.0382(0.0125)
		9	0.0206(0.0087)	0.0189(0.0080)	0.4787(0.0110)	0.0456(0.0165)
		12	0.0218(0.0094)	0.0201(0.0082)	0.4785(0.0110)	0.0502(0.0178)
	10	3	0.0091(0.0038)	0.0071(0.0034)	0.4763(0.0093)	0.0150(0.0060)
		6	0.0105(0.0048)	0.0095(0.0045)	0.4763(0.0094)	0.0209(0.0073)
		9	0.0112(0.0050)	0.0104(0.0044)	0.4763(0.0093)	0.0242(0.0081)
		12	0.0118(0.0056)	0.0111(0.0047)	0.4762(0.0093)	0.0266(0.0084)

competitive with the regular Bayesian horseshoe method. The results of real data analysis also support their promising performance.

References

1. Alcalá, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F.: Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. Journal of Multiple-Valued Logic and Soft Computing 17(2-3), 255–287 (2010)
2. Breheny, P., Huang, J.: Penalized methods for bi-level variable selection. Statistics and its interface 2(3), 369–380 (July 2009)
3. Breiman, L.: Better subset regression using the nonnegative garrote. Technometrics 37(4), 373–384 (1995)

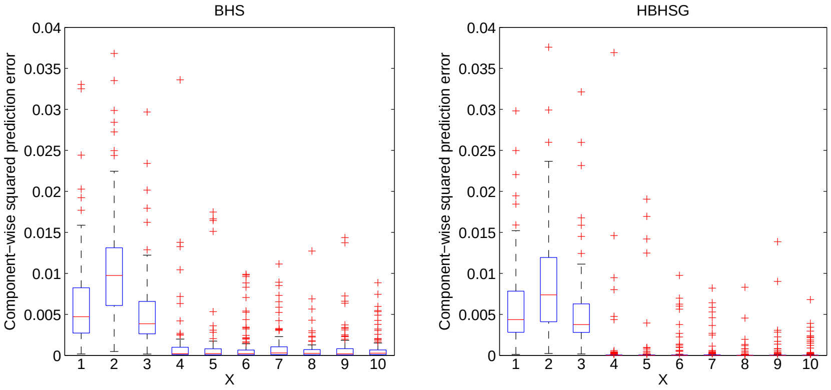

Fig. 1. Boxplots of component-wise prediction error for BHS and HBHSG when there are $p = 10$ predictors, $n = 100$ samples, SNR $= 5$, $K = 3$ degree of Legendre polynomial expansions. The first three components are associated with the response variable.

Table 4. Mean and standard deviation of squared prediction errors for Electrical-Maintenance data set.

Degree	BHS	BHSG	HBHSG	BHS-NE	lasso-BIC
2	26866.9(2636)	26855.0(2637)	26879.8(2637)	27309.7(2489)	26858.6(2632)
4	13405.1(1285)	13437.1(1285)	13394.0(1281)	27314.3(2493)	13488.9(1291)
6	12939.2(1257)	13061.4(1255)	12939.9(1254)	27312.0(2498)	13038.6(1252)
8	11019.9(1141)	11054.0(1097)	10978.4(1136)	27307.9(2492)	11067.9(1100)
10	9970.9(1096)	10057.1(1075)	9958.4(1097)	27316.6(2492)	10022.4(1079)

4. Bühlmann, P., Geer, S.V.D.: Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media (2011)
5. Carvalho, C.M., Polson, N.G., Scott, J.G.: Handling sparsity via the horseshoe. In: JMLR. vol. 5, pp. 73–80 (2009)
6. Carvalho, C.M., Polson, N.G., Scott, J.G.: The horseshoe estimator for sparse signals. Biometrika 97(2), 465–480 (2010)
7. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association 96(456), 1348–1360 (2001)
8. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(1), 55–67 (Feburary 1970)
9. Huang, J., Ma, S., Xie, H., Zhang, C.H.: A group bridge approach for variable selection. Biometrika 96(2), 339–355 (2009)
10. Makalic, E., Schmidt, D.F.: A simple sampler for the horseshoe estimator. IEEE Signal Processing Letters 23(1), 179–182 (2016)
11. Park, T., Casella, G.: The Bayesian lasso. Journal of the American Statistical Associationt 103(482), 681–686 (June 2008)

12. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68(1), 49–67 (2006)
13. Zhao, P., Rocha, G., Yu, B.: The composite absolute penalties family for grouped and hierarchical variable selection. The Annals of Statistics 37(6A), 3468–3497 (December 2009)
14. Zou, H.: The adaptive lasso and its oracle properties. Journal of the American statistical association 101(476), 1418–1429 (2006)
15. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67(2), 301–320 (2005)

Appendix: Full conditional distributions

The hierarchical specification of the complete model of hierarchical Bayesian horseshoe model (HBHSG) is given in Equation (7). By using the decomposition [10] , the hierarchical representation becomes:

$$\begin{aligned}
\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2 &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbb{1}_n),\\
\boldsymbol{\beta}|\sigma^2,\tau^2,\lambda_1^2,\cdots,\lambda_G^2,\delta_1,\cdots,\delta_p &\sim \mathcal{N}(\mathbf{0},\sigma^2\tau^2\mathbf{D}_{\boldsymbol{\lambda}}\mathbf{D}_{\boldsymbol{\delta}}),\\
\text{where } \mathbf{D}_{\boldsymbol{\lambda}} = \operatorname{diag}(\lambda_1^2\mathbf{I}_{s_1},\cdots,\lambda_G^2\mathbf{I}_{s_G}),\ &\mathbf{D}_{\boldsymbol{\delta}} = \operatorname{diag}(\delta_1^2,\cdots,\delta_p^2),\\
\lambda_g^2|t_g &\sim \mathcal{IG}\left(\frac{1}{2},\frac{1}{t_g}\right),\ t_g \sim \mathcal{IG}\left(\frac{1}{2},1\right),\ g=1,\cdots,G,\\
\delta_j^2|c_j &\sim \mathcal{IG}\left(\frac{1}{2},\frac{1}{c_j}\right),\ c_j \sim \mathcal{IG}\left(\frac{1}{2},1\right),\ j=1,\cdots,p,\\
\tau^2|v &\sim \mathcal{IG}\left(\frac{1}{2},\frac{1}{v}\right),\ v \sim \mathcal{IG}\left(\frac{1}{2},1\right),\\
\sigma^2 &\sim \frac{1}{\sigma^2}d\sigma^2.
\end{aligned}$$

The full conditional distributions of $\boldsymbol{\beta}$, σ^2, $\lambda_1^2,\cdots,\lambda_G^2$, $\delta_1^2,\cdots,\delta_p^2$, τ are:

$$\boldsymbol{\beta}|\sigma^2,\tau^2,\lambda_1^2,\cdots,\lambda_G^2,\delta_1^2,\cdots,\delta_p^2 \sim \mathcal{N}\left(\mathbf{A}^{-1}\mathbf{X}^T\mathbf{y},\sigma^2\mathbf{A}^{-1}\right), \text{where } \mathbf{A} = \mathbf{X}^T\mathbf{X} + (\tau^2\mathbf{D}_{\boldsymbol{\lambda}}\mathbf{D}_{\boldsymbol{\delta}})^{-1}$$

$$\sigma^2|\boldsymbol{\beta},\tau^2,\lambda_1^2,\cdots,\lambda_G^2,\delta_1^2,\cdots,\delta_p^2 \sim \mathcal{IG}\left(\frac{n-1+p}{2},\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\boldsymbol{\beta}^T(\tau^2\mathbf{D}_{\boldsymbol{\lambda}}\mathbf{D}_{\boldsymbol{\delta}})^{-1}\boldsymbol{\beta}}{2}\right),$$

$$\lambda_g^2|\boldsymbol{\beta},\sigma^2,\tau^2,t_g,\delta_1^2,\cdots,\delta_p^2 \sim \mathcal{IG}\left(\frac{s_g+1}{2},\frac{\boldsymbol{\beta}_g^T(\mathbf{D}_{\boldsymbol{\delta}_g})^{-1}\boldsymbol{\beta}_g}{2\sigma^2\tau^2}+\frac{1}{t_g}\right),\ t_g|\lambda_g^2 \sim \mathcal{IG}\left(1,\frac{1}{\lambda_g^2}+1\right),$$

$$\delta_j^2|\boldsymbol{\beta},\sigma^2,\tau^2,\lambda_1^2,\cdots,\lambda_G^2,c_j \sim \mathcal{IG}\left(1,\frac{\beta_j^2}{2\sigma^2\tau^2\lambda_{gj}^2}+\frac{1}{c_j}\right),\ c_j|\delta_j^2 \sim \mathcal{IG}\left(1,\frac{1}{\delta_j^2}+1\right),$$

$$\tau^2|\boldsymbol{\beta},\sigma^2,\tau^2,\lambda_1^2,\cdots,\lambda_G^2,\delta_1^2,\cdots,\delta_p^2,v \sim \mathcal{IG}\left(\frac{p+1}{2},\frac{\boldsymbol{\beta}^T(\mathbf{D}_{\boldsymbol{\lambda}}\mathbf{D}_{\boldsymbol{\delta}})^{-1}\boldsymbol{\beta}}{2\sigma^2}+\frac{1}{v}\right),$$

$$v|\tau^2 \sim \mathcal{IG}\left(1,\frac{1}{\tau^2}+1\right).$$