The Grouped Horseshoe Distribution and Its Statistical Properties

VIRGINIA X. HE AND MATT P. WAND

University of Technology Sydney

12th July, 2024

Abstract

The Grouped Horseshoe distribution arises from hierarchical structures in the recent Bayesian methodological literature aimed at selection of groups of regression coefficients. We isolate this distribution and study its properties concerning Bayesian statistical inference. Most, but not all, of the properties of the univariate Horseshoe distribution are seen to transfer to the grouped case.

Keywords: Additive models; Bayesian statistical inference; variable selection.

1 Introduction

Since around 2010, numerous continuous distributions have been proposed for use as prior distributions of coefficients in Bayesian regression-type models. Table 1 of Bai & Ghosh (2018) provides seven such examples, all of which correspond to scale mixtures of Normal density functions with various polynomial-tailed density functions. In this article we focus on one of these examples known as the *horseshoe* prior. The underlying Horseshoe distribution (Carvalho *et al.*, 2010) corresponds to the mixing distribution being $F_{1,1}$ for variance parameter scale mixing or Half-Cauchy for standard deviation scale mixing.

Most of this literature is concerned with variable selection for individual coefficients. The *grouped* extension is concerned with simultaneous selection of a group of variables. For example, in additive model selection (e.g. Schiepl *et al.*, 2012; He & Wand, 2024) a group of variables corresponds, typically, to a set of spline basis functions of a continuous predictor. Grouped variable selection is an attractive mechanism for deciding between the continuous predictor having a linear or non-linear effect. Our focus in this article is the grouped extension of the horseshoe prior as proposed by Xu *et al.* (2016).

Our first goal is determination of the underlying multivariate density function corresponding to grouped horseshoe variable selection. This involves integrating out the scale mixing density function and leads to a family of distributions, indexed by dimension, that we label the *Grouped Horseshoe* distribution. We derive an expression for the Grouped Horseshoe density function in terms of the generalized exponential integral functions. As for the ordinary Horseshoe distribution, the Grouped Horseshoe density function is shown to have a pole at the origin.

We then investigate the grouped extensions of the various Bayesian statistical inference properties of horseshoe priors studied by Carvalho *et al.* (2010). The score function behaviour and robustness to large signals property of horseshoe priors, studied in Section 2 of Carvalho *et al.* (2010), is shown to extend to the grouped situation. However, the super-efficiency property based on risk rates of convergence, studied in Section 3.3 of Carvalho *et al.* (2010), does not extend to the grouped situation.

Our main results are laid out in Sections 2–3. The topic of Section 4 is *thresholding*, which is concerned with practical data-based rules for deciding whether or not a coefficient parameter in a Bayesian model is set to zero. In this section we also investigate use of the Grouped Horseshoe distribution for Bayesian generalized additive model selection as considered by the authors in He & Wand (2024). Our conclusions are summarized in Section 5. An online supplement provides full derivations of all results.

1.1 Notation

For a logical proposition $\mathcal{P}$ we let $I(\mathcal{P}) = 1$ if $\mathcal{P}$ is true and $I(\mathcal{P}) = 0$ if $\mathcal{P}$ is false. The Euclidean norm of column vector $\boldsymbol{a}$ is denoted by $\|\boldsymbol{a}\| \equiv \sqrt{\boldsymbol{a}^T \boldsymbol{a}}$. If $\boldsymbol{v}$ is a random vector then $\mathfrak{p}(\boldsymbol{v})$ denotes the density function of $\boldsymbol{v}$. If f is a smooth function that maps $\mathbb{R}^d$ to $\mathbb{R}$ then $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ denotes the $d \times 1$ vector of partial derivatives of $f(\boldsymbol{x})$ with respect to the entries of $\boldsymbol{x}$.

2 Density Function Explicit Form

Section 2.2 of Xu *et al.* (2016) introduced the grouped horseshoe model. The underlying distribution, which we call the Grouped Horseshoe distribution, corresponds to setting $\sigma = \tau = G = 1$ and $s_1 = d$ in equation (6) of Xu *et al.* (2016). This leads to the $d \times 1$ random vector $\boldsymbol{x}$ having a (standard) Grouped Horseshoe distribution if and only if

$$\boldsymbol{x}|\lambda \sim N(\boldsymbol{0}, \lambda^2 \boldsymbol{I}_d) \quad \text{where} \quad \mathfrak{p}(\lambda) = \frac{2I(\lambda > 0)}{\pi(1 + \lambda^2)}. \tag{1}$$

Let E_ν denote the *generalized exponential integral* function, given by

$$E_\nu(x) \equiv \int_1^\infty \exp(-xt)/t^\nu \, dt, \quad x, \nu \in \mathbb{R}$$

(e.g. 8.19.3 of Olver *et al.* 2023).

Result 1. *Let $\boldsymbol{x}$ be a $d \times 1$ random vector having a Grouped Horseshoe distribution as defined according to (1). Then the density function of $\boldsymbol{x}$, denoted by $\mathfrak{p}_{HS,d}(\boldsymbol{x})$, is*

$$\mathfrak{p}_{HS,d}(\boldsymbol{x}) = \frac{\Gamma\left(\frac{1}{2}(d+1)\right)}{\sqrt{2\pi^{d+2}}} \, \exp\left(\|\boldsymbol{x}\|^2/2\right) E_{(d+1)/2}\left(\|\boldsymbol{x}\|^2/2\right) \Big/ \|\boldsymbol{x}\|^{d-1}, \quad \boldsymbol{x} \in \mathbb{R}^d.$$

A derivation of Result 1 is given in Section S.1 of the supplement.

A simple consequence of Result 1 is

$$\mathfrak{p}_{HS,1}(x) = \frac{1}{\sqrt{2\pi^3}} \, \exp\left(x^2/2\right) E_1\left(x^2/2\right), \quad x \in \mathbb{R}.$$

which matches an expression given in the appendix of Carvalho *et al.* (2010) for the ordinary Horseshoe distribution. For $d = 2$ we have

$$\mathfrak{p}_{HS,2}(x_1, x_2) = \frac{1}{2\sqrt{2\pi^3}} \, \exp\left((x_1^2 + x_2^2)/2\right) E_{3/2}\left((x_1^2 + x_2^2)/2\right) \Big/ \sqrt{x_1^2 + x_2^2}, \quad (x_1, x_2) \in \mathbb{R}^2.$$

which is displayed in Figure 1. It is apparent from Figure 1 that $\mathfrak{p}_{HS,2}$ has a pole at the origin. We formalise this behaviour for general $d \in \mathbb{N}$ in Section 3.1.

3 Statistical Properties

We now investigate various statistical properties of the Grouped Horseshoe distribution. A particular focus is Bayesian statistical inference where a parameter vector has a Grouped Horseshoe prior.

3.1 Pole at the Origin Existence

In Carvalho *et al.* (2010), $\mathfrak{p}_{HS,1}$ is shown to have a pole at the origin. This is shown to provide some inferential advantages when $\mathfrak{p}_{HS,1}$ is used as a prior density function. Result 2, which is derived in Section S.2 of the online supplement, shows that the Grouped Horseshoe density function has a pole at the origin for any dimension.

Result 2. *For each $d \in \mathbb{N}$, $\lim_{\boldsymbol{x} \to \boldsymbol{0}} \mathfrak{p}_{HS,d}(\boldsymbol{x}) = \infty$.*
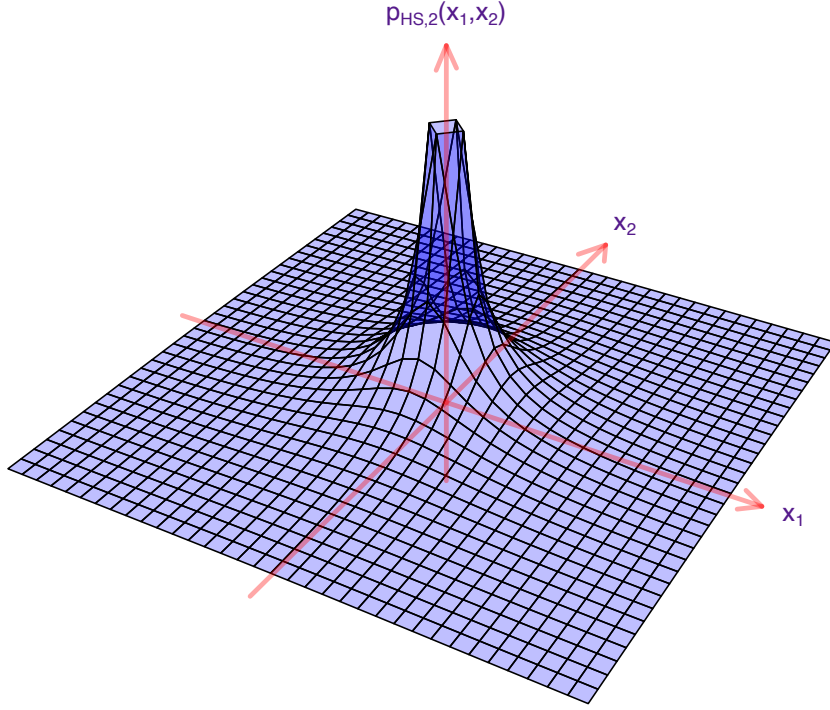
Figure 1: *Perspective plot of the bivariate Grouped Horseshoe density function:* $\mathfrak{p}_{HS,2}$.

3.2 Score Function and Tail Robustness

Consider the model

$$\boldsymbol{y}|\boldsymbol{\theta} \sim N(\boldsymbol{\theta}, \boldsymbol{I}_d) \quad \text{with prior} \quad \mathfrak{p}(\boldsymbol{\theta}) = \mathfrak{p}_{HS,d}(\boldsymbol{\theta}/\tau)/\tau^d \tag{2}$$

for some $\tau > 0$ that is fixed and known. In their Section 2, Carvalho *et al.* (2010) consider the $d = 1$ version of (2) and prove that the score function

$$\frac{d\log\{\mathfrak{p}(y)\}}{dy} \quad \text{converges to 0 as } |y| \to \infty. \tag{3}$$

As argued there, (3) implies a type of robustness to large signals which Carvalho *et al.* (2010) refer to as *tail robustness*. Result 3, which is proven in Section S.3, shows that grouped horseshoe priors also possesses this property.

Result 3. *For model (2), the tail behaviour of the score function is given by*

$$\nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\} \sim -\frac{(d+1)\boldsymbol{y}}{\|\boldsymbol{y}\|^2} \quad \text{for} \quad \|\boldsymbol{y}\| \gg 1.$$

Consequently

$$\lim_{\|\boldsymbol{y}\|\to\infty} \nabla_{\boldsymbol{y}} \log\{\mathfrak{p}(\boldsymbol{y})\} = \boldsymbol{0} \quad \text{and} \quad E\big\|\boldsymbol{y} - E(\boldsymbol{\theta}|\boldsymbol{y})\big\| \leq b_\tau$$

for some $b_\tau < \infty$ that depends on τ.

3.3 Risk Convergence Rates

Consider the model

$$\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n|\boldsymbol{\theta} \quad \text{independently distributed as } N(\boldsymbol{\theta}, \sigma^2\boldsymbol{I}_d) \text{ with prior } \mathfrak{p}(\boldsymbol{\theta}) = \mathfrak{p}_{HS,d}(\boldsymbol{\theta}). \tag{4}$$

Suppose that the true sampling distribution of the $\boldsymbol{y}_i$ is $N(\boldsymbol{\theta}^0, \sigma^2\boldsymbol{I}_d)$. In the $d = 1$ case, Theorem 4 of of Carvalho *et al.* (2010) states rates of convergence results for the so-called Cesàro-average

risk of the Bayes estimator of θ, which they denote by R_n. The rates differ depending on whether $\theta^0 = 0$ or $\theta^0 \neq 0$ where θ^0 is the value of θ according to the sampling distribution of the y_i. The horseshoe prior is shown to lead to a super-efficient risk rate when $\theta^0 = 0$.

We now provide a Grouped Horseshoe distribution extension of Theorem 4 of Carvalho *et al.* (2010). The risk quantity R_n has a definition analogous to that given in Section 3.3 of Carvalho *et al.* (2010) for the d-variate extension of the set-up treated there.

Result 4. *Consider model (4) and suppose that the $\boldsymbol{y}_i$ have sampling distribution $N(\boldsymbol{\theta}^0, \sigma^2 \boldsymbol{I}_d)$. Let R_n be Cesàro-average risk of the Bayes estimator of $\boldsymbol{\theta}$. When $\boldsymbol{\theta}^0 = \boldsymbol{0}$ we have*

$$R_n \leq \begin{cases} \dfrac{\log(n)}{2n} - \dfrac{\log\{\log(n)\}}{n} + O\left(\dfrac{1}{n}\right) & \text{if } d = 1, \\[3mm] \dfrac{\log(n)}{2n} + O\left(\dfrac{1}{n}\right) & \text{if } d \geq 2. \end{cases}$$

When $\boldsymbol{\theta}^0 \neq \boldsymbol{0}$ we have $R_n \leq d\log(n)/(2n) + O(1/n)$ for all $d \in \mathbb{N}$.

Section S.4 provides the full derivational details of Result 4. For $d = 1$ and $\boldsymbol{\theta}^0 = \boldsymbol{0}$, super-efficiency corresponds to presence of the $-\log\{\log(n)\}/n$ term in the upper bound on R_n. Result 4 shows that this term only arises in the $d = 1$ case. The Bayes estimator is not super-efficient for $d \geq 2$.

4 Thresholding

Consider a Bayesian model that contains specifications of the form

$$\boldsymbol{\theta}|\sigma_\theta \quad \text{has density function} \quad \mathfrak{p}_{\text{HS},d}(\boldsymbol{\theta}/\sigma_\theta)/\sigma_\theta^d \quad \text{where } \boldsymbol{\theta} \text{ is } d \times 1. \tag{5}$$

From results in Section 2, specification (5) is equivalent to

$$\boldsymbol{\theta}|\sigma_\theta, \lambda \sim N(\boldsymbol{0}, \sigma_\theta^2 \lambda^2 \boldsymbol{I}_d), \quad \mathfrak{p}(\lambda) = \frac{2I(\lambda > 0)}{\pi(1 + \lambda^2)} \tag{6}$$

The introduction of the auxiliary variable λ is important for the upcoming approach to thresholding. In the scalar case, Carvalho *et al.* (2010) develop a thresholding approach for deciding between

$$\theta = 0 \quad \text{and} \quad \theta \neq 0, \quad \theta \in \mathbb{R}.$$

In this section we describe and evaluate the extension of their approach to deciding between

$$\boldsymbol{\theta} = \boldsymbol{0} \quad \text{and} \quad \boldsymbol{\theta} \neq \boldsymbol{0}, \quad \boldsymbol{\theta} \in \mathbb{R}^d.$$

The Carvalho *et al.* (2010) approach involves the following result concerning a simple "side" model:

Result 5. *For the Bayesian model*

$$\boldsymbol{y}|\boldsymbol{\psi} \sim N(\boldsymbol{\psi}, \tau_1^2 \boldsymbol{I}_d), \quad \boldsymbol{\psi}|\lambda \sim N(\boldsymbol{0}, \lambda^2 \tau_2^2 \boldsymbol{I}_d), \quad \mathfrak{p}(\lambda) = \frac{2I(\lambda > 0)}{\pi(1 + \lambda^2)}, \quad \tau_1, \tau_2 > 0 \text{ fixed} \tag{7}$$

the posterior mean of $\boldsymbol{\psi}$ is

$$E(\boldsymbol{\psi}|\boldsymbol{y}) = E\left(\frac{\lambda^2 \tau_2^2}{\tau_1^2 + \lambda^2 \tau_2^2}\bigg|\boldsymbol{y}\right) \boldsymbol{y}.$$

A derivation of Result 5 is given in Section S.5.

For general Bayesian models containing (5) or, equivalently, (6) forms Result 5 suggests the following rule:

$$\text{decide that } \boldsymbol{\theta} = \mathbf{0} \quad \text{if and only if} \quad E(\gamma_{\text{GHS}}|\boldsymbol{y}) < \tfrac{1}{2} \quad \text{where} \quad \gamma_{\text{GHS}} \equiv \frac{\lambda^2 \sigma_\theta^2}{\sigma_\varepsilon^2 + \lambda^2 \sigma_\theta^2}. \qquad (8)$$

To better understand the efficacy of (8), we ran a simulation study similar to that in Section 4 of our recent article, He & Wand (2024), on generalized additive model selection. The study involved the Bayesian generalized additive model given by equation (9) in He & Wand (2024) where $d_\circ$ is the number of candidate predictors that may have a zero or linear effect and $d_\bullet$ is the number of candidate predictors that may have a zero, linear or non-linear effect. The study involved both the set-up in He & Wand (2024), with the Laplace-Zero and Grouped Lasso-Zero priors that are used in that article, and an alternative version with the likelihood taking the form

$$\boldsymbol{y}|\beta_0, \boldsymbol{\beta}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_\bullet}, \sigma_\varepsilon^2 \sim N\left(\mathbf{1}_n \beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{d_\bullet} \boldsymbol{Z}_j \boldsymbol{u}_j, \sigma_\varepsilon^2 \boldsymbol{I}_n\right)$$

where $\boldsymbol{\beta}$ is a $(d_\circ + d_\bullet) \times 1$ vector of linear effects coefficients and, for each $1 \le j \le d_\bullet$, $\boldsymbol{u}_j$ is a $K_j \times 1$ vector of spline coefficients for the jth non-linear effect. Section 2 of He & Wand (2024) contains fuller details, including the definition of the spline basis $\boldsymbol{Z}_j$ matrices.

Let β_j denote the jth entry of $\boldsymbol{\beta}$. Rather than imposing Laplace-Zero distributions on the β_j, as conveyed by equation (6) of He & Wand (2024), we instead consider the independent scalar Horseshoe specifications

$$\mathfrak{p}(\beta_j|\sigma_\beta) = \mathfrak{p}_{\text{HS},1}(\beta_j/\sigma_\beta)/\sigma_\beta, \quad 1 \le j \le d_\circ + d_\bullet. \qquad (9)$$

Similarly, rather than imposing a Grouped Lasso-Zero distribution on $\boldsymbol{u}_j$, as conveyed by equation (7) of He & Wand (2024), we instead consider the independent Grouped Horseshoe prior specifications

$$\mathfrak{p}(\boldsymbol{u}_j|\sigma_{u_j}) = \mathfrak{p}_{\text{HS},K_j}(\boldsymbol{u}_j/\sigma_{u_j}) \big/ \sigma_{u_j}^{K_j}, \quad 1 \le j \le d_\bullet. \qquad (10)$$

Note that (9) has the auxiliary variable representation

$$\beta_j|\sigma_\beta, \lambda_{\beta j} \overset{\text{ind.}}{\sim} N(0, \sigma_\beta^2 \lambda_{\beta j}), \quad \mathfrak{p}(\lambda_{\beta j}) = \frac{2I(\lambda_{\beta j} > 0)}{\pi(1 + \lambda_{\beta j}^2)}, \quad 1 \le j \le d_\circ + d_\bullet,$$

where $\overset{\text{ind.}}{\sim}$ denotes "independently distributed as". Similarly, (10) has the auxiliary variable representation

$$\boldsymbol{u}_j|\sigma_{uj}, \lambda_{uj} \overset{\text{ind.}}{\sim} N\big(\mathbf{0}, \sigma_{uj}^2 \lambda_{uj} \boldsymbol{I}_{K_j}\big), \quad \mathfrak{p}(\lambda_{uj}) = \frac{2I(\lambda_{uj} > 0)}{\pi(1 + \lambda_{uj}^2)}, \quad 1 \le j \le d_\bullet.$$

From (8), the (Grouped) Horseshoe analogues of He & Wand (2024)'s $\gamma_{\beta j}$ and γ_{uj} are

$$\gamma_{\beta_j, \text{HS}} \equiv \frac{\lambda_{\beta_j}^2 \sigma_\beta^2}{\sigma_\varepsilon^2 + \lambda_{\beta_j}^2 \sigma_\beta^2} \quad \text{and} \quad \gamma_{u_j, \text{GHS}} \equiv \frac{\lambda_{uj}^2 \sigma_{uj}^2}{\sigma_\varepsilon^2 + \lambda_{uj}^2 \sigma_{uj}^2}.$$

Therefore, for the $d_\bullet$ predictors that can have a zero, linear or non-linear effect the classification rule that arises from (8) is

$$\text{the effect is zero if } \max\big\{E(\gamma_{\beta_j, \text{HS}}|\boldsymbol{y}), E(\gamma_{u_j, \text{GHS}}|\boldsymbol{y})\big\} \le \tfrac{1}{2},$$

$$\text{the effect is linear if } E(\gamma_{\beta_j, \text{HS}}|\boldsymbol{y}) > \tfrac{1}{2} \text{ and } E(\gamma_{u_j, \text{GHS}}|\boldsymbol{y}) \le \tfrac{1}{2}, \qquad (11)$$

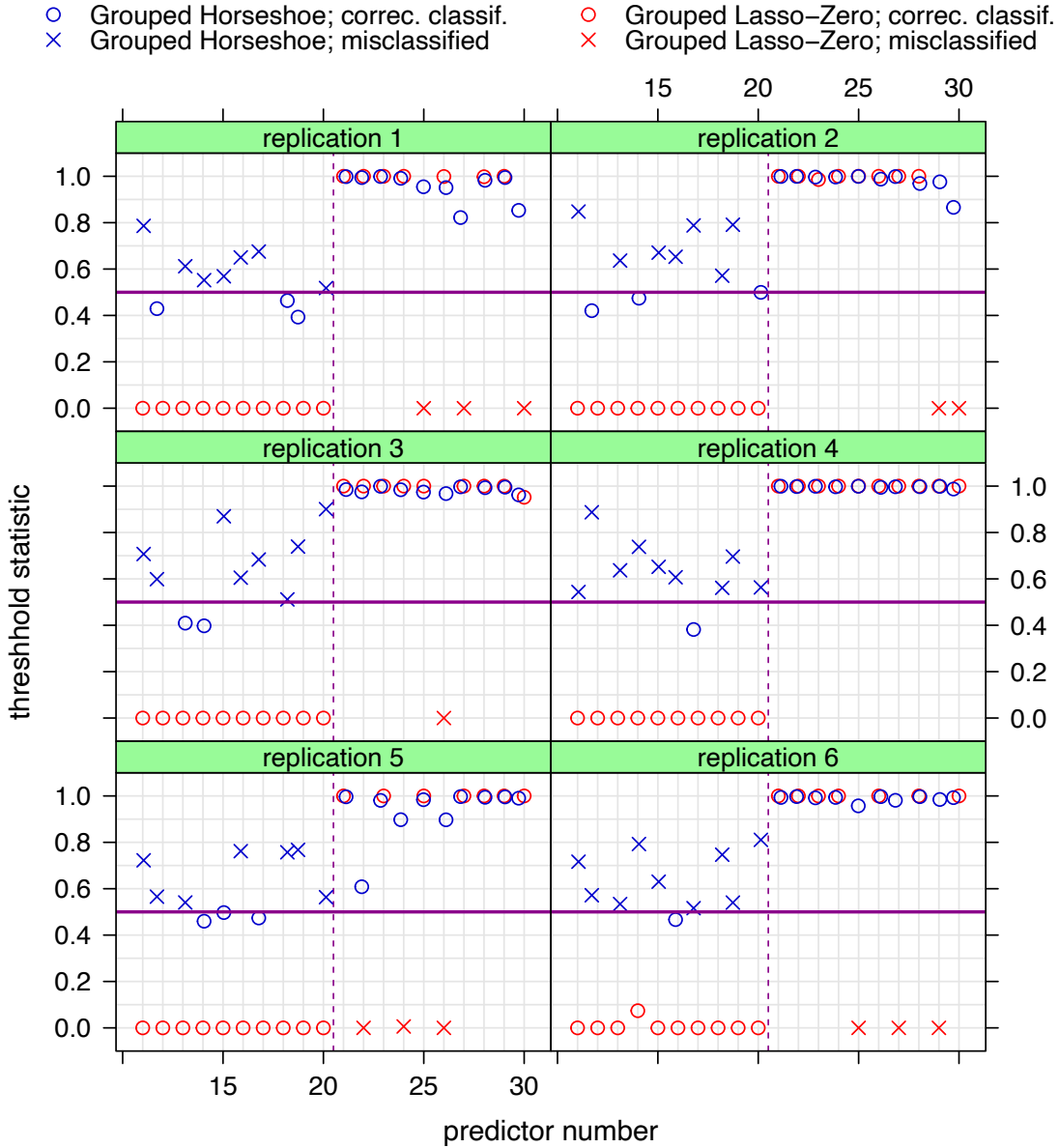$$\text{otherwise the effect is non-linear.}$$

Figure 2: *The results of linear effect versus non-linear effect classification for six replications of the simulation study described in Section 4 of He & Wand (2024) with $n = 500$ and $\sigma_\varepsilon = 2$. The blue symbols correspond to the Markov chain Monte Carlo-approximate $E(\gamma_{u_j,\mathrm{GHS}}|\boldsymbol{y})$ values for $21 \leq j \leq 30$, which are the predictors that are simulated to have linear effects (left of the vertical dashed line) or non-linear effects (right of the vertical dashed line). A blue circle indicates correct classification using (11), whilst a blue cross indicates misclassification. The red symbols are similar, but for the $E(\gamma_{u_j}|\boldsymbol{y})$ statistics corresponding to strategy described in Section 3.5.2 of He & Wand (2024). The classification border of $\frac{1}{2}$ is shown by the horizontal purple line.*

Note that this rule is analogous to the $\tau = \frac{1}{2}$ rule given in Section 3.5.2 of He & Wand (2024) for Laplace-Zero and Grouped Lasso-Zero priors, with $\gamma_{\beta_j,\mathrm{HS}}$ and $\gamma_{u_j,\mathrm{GHS}}$ instead of their $\gamma_{\beta j}$ and γ_{uj}.

Figure 2 shows the approximate, based on Markov chain Monte Carlo sampling, values of $E(\gamma_{u_j,\mathrm{GHS}}|\boldsymbol{y})$ and $E(\gamma_{uj}|\boldsymbol{y})$ for six replications of the simulation study described in Section 4 of He & Wand (2024) with $n = 500$ and $\sigma_\varepsilon = 2$. The data are simulated so that the effect of the jth

predictor is

$$
\begin{array}{llll}
\text{zero} & \text{for} & j \in \{1, 2, \ldots, 10\}, & \\
\text{linear} & \text{for} & j \in \{11, 12, \ldots, 20\} \text{ and} & \qquad (12) \\
\text{non-linear} & \text{for} & j \in \{21, 22, \ldots, 30\}. &
\end{array}
$$

The horizontal axis corresponds to $j = 11, 12, \ldots, 30$, which is concerned with linear versus non-linear classification. Correct classifications are shown as circles and incorrect classifications are shown as crosses. For these replications, use of the Grouped Lasso-Zero prior results in a misclassification rate of $12/120 = 10\%$. For the Grouped Horseshoe prior the misclassification rate is $47/120 = 39.2\%$, which is about four times worse. We see from Figure 2 that most of the Grouped Lasso-Zero threshold statistics are close to 1 when the true effect is non-linear and close to 0 when the true effect is linear. In contrast, most of the Grouped Horseshoe threshold statistics are close to 1 when the true effect is non-linear, but scattered between 0.4 and 0.8 when the true effect is linear. This last-mentioned behavior means that many predictors that have a linear effect are misclassified as having a non-linear effect when the Grouped Horseshoe prior is used.

Figure 3 differs from Figure 2 in that the sample size is quadrupled to $n = 2,000$ and the error standard deviation is decreased to $\sigma_\varepsilon = 0.25$. This should make linear versus non-linear classification much easier and use of the Grouped Lasso-Zero prior leads to perfect performance for these six replications. However, for the Grouped Horseshoe prior the more favorable conditions do not seem to help and the threshold statistics are still scattered between 0.4 and 0.8 when the true effect is linear, leading to a $49/120 = 40.8\%$ misclassification rate.

We experimented with a possible remedy to the poor performance of thresholding the $E(\gamma_{u_j,\text{GHS}}|\boldsymbol{y})$ statistics at $\frac{1}{2}$. This involved applying k-means clustering (e.g. MacQueen, 1967) to the $E(\gamma_{u_j,\text{GHS}}|\boldsymbol{y})$ observations, with the number of clusters fixed at 2. The function `kmeans()` within the R computing environment (R Core Team, 2024) was used to obtain the two clusters and corresponding classification rule. As an example, for the analysis correponding to replication 1 of Figure 3, the k-means threshold is 0.8031. The results from use of this k-means alternative to the Figure 3 analyses are shown in Figure 4. The misclassification rate drops to $13/120 = 10.83\%$.

This experimental k-means approach to thresholding for generalized additive model selection has some promise, but relies on situations where there are many candidate predictors of various effect types. If there are only 3–6 candidate predictors, say, such that most of them have strongly non-linear effects then k-means threshold choice may not be viable.

Figures 2–4 are based on only six replications. They also omit the zero versus linear/non-linear classifications based on the $E(\gamma_{\beta_j,\text{HS}}|\boldsymbol{y})$ statistics and their Laplace-Zero counterparts, which has similar results regarding Horseshoe versus Laplace-Zero priors. To get a more complete picture, we ran an adaptation of the simulation study described in Section 4 of He & Wand (2024). The data were generated in exactly the same manner as there, with 30 candidate predictors having "true" effects as described by (12), the sample size ranging over $n \in \{500, 1000, 2000\}$ and the error standard deviation ranging over $\sigma_\varepsilon \in \{0.25, 0.5, 1, 2\}$.

The side-by-side boxplots in Figure 5 facilitate comparison of

   A. the Laplace-Zero/Grouped Lasso-Zero prior approach of He & Wand (2024) with the classification border set to $\frac{1}{2}$,

   B. use of the rule (11) involving (Grouped) Horseshoe priors, the $E(\gamma_{\beta_j,\text{HS}}|\boldsymbol{y})$ and $E(\gamma_{u_j,\text{GHS}}|\boldsymbol{y})$ threshold statistics and also with the classification border set to $\frac{1}{2}$,

   C. the same as B., but with the classification border based on k-means clustering.

We see that A. is clearly superior to B. Also, C. offers a big improvement on B., but does not perform as well as A. In the He & Wand (2024) Bayesian generalized additive model selection setting use of (Grouped) Horseshoe priors does not compete very well with use of Laplace-Zero/Grouped Lasso-Zero priors.
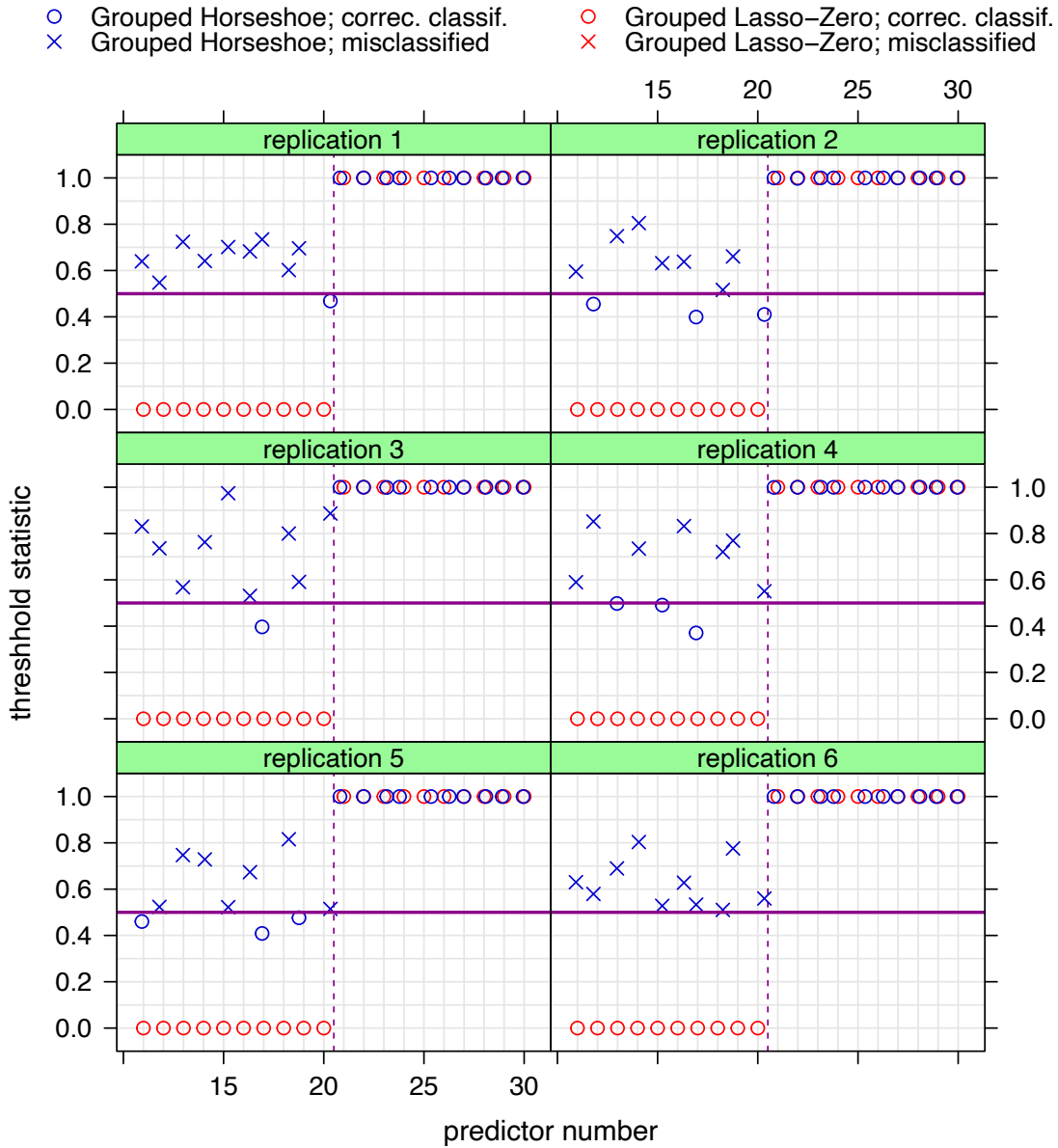
Figure 3: *Similar to Figure 2 but with a lower sample size, $n = 2,000$, and a higher error standard deviation, $\sigma_\varepsilon = 0.25$.*

5 Conclusions

In this article we have conducted a thorough investigation into the statistical properties of the Grouped Horseshoe distribution. We have shown that most of the properties possessed by the univariate Horseshoe distribution extend to the grouped situation. Our investigation was motivated by our interest in Bayesian generalized additive model selection, as described in our recent He & Wand (2024) article. The numerical studies in Section 4, concerned with using Result 5 to carry out generalized additive model selection with (Grouped) Horseshoe priors, reveal some performance concerns compared with a spike-and-slab benchmark. Perhaps this research can lead to the development of better selection rules based on (Grouped) Horseshoe priors.
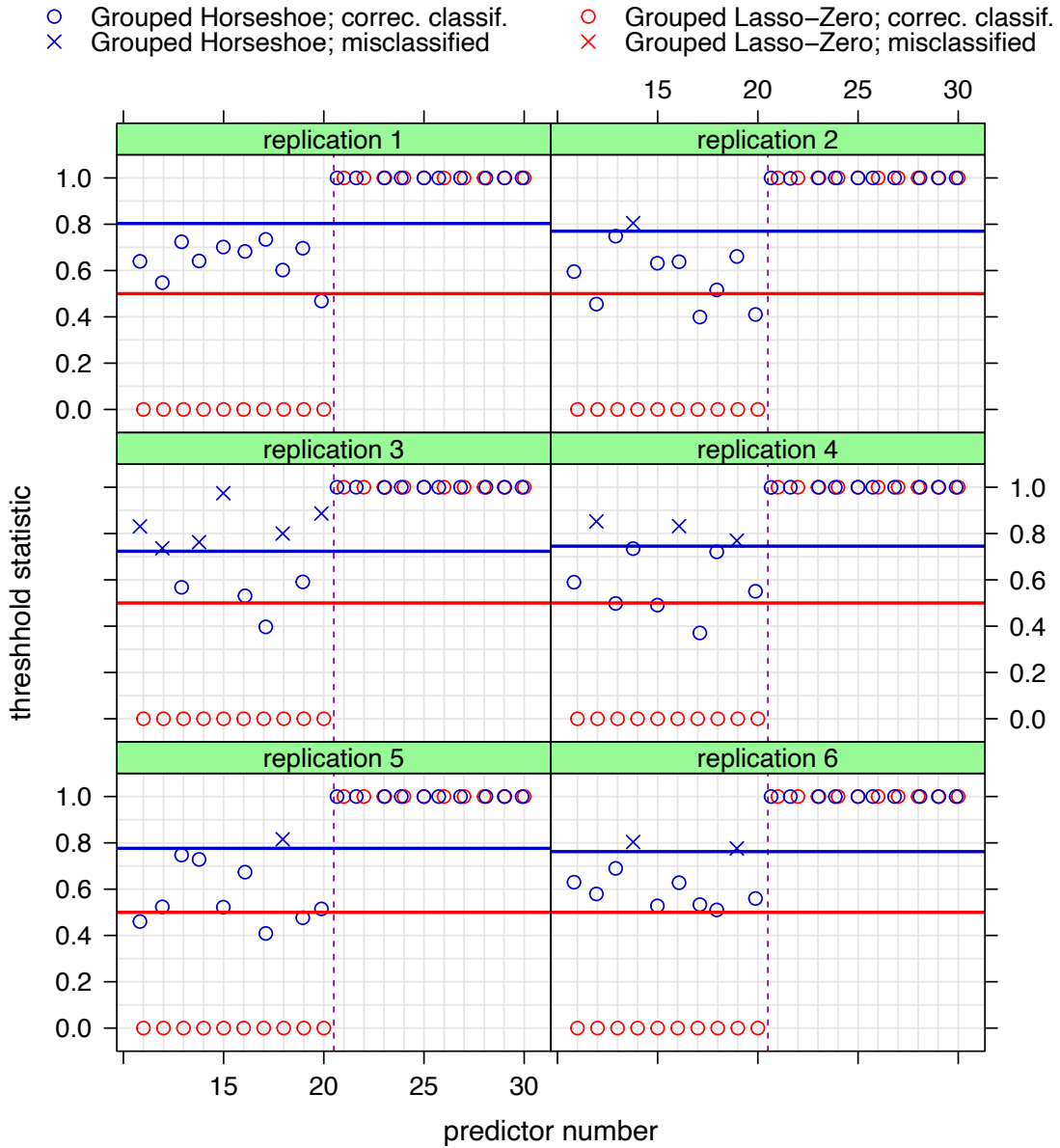
Figure 4: *Similar to Figure 3 but with the thresholding of the Grouped Horseshoe $E(\gamma_{u_j,\text{GHS}}|\boldsymbol{y})$ statistics based on k-means clustering. For each replication, the horizontal blue line shows the classification border arising from k-means clustering. The horizontal red line at $\frac{1}{2}$ is the threshold for the Grouped Lasso-Zero statistics.*
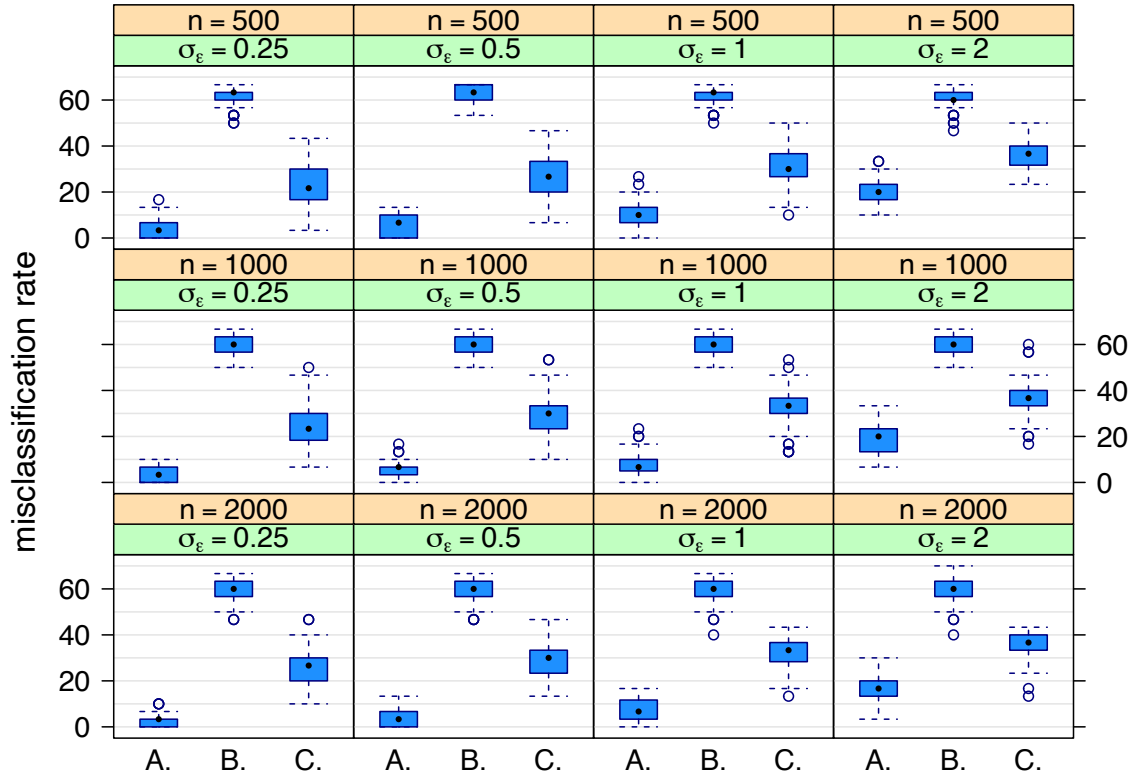
Acknowledgements

Figure 5: *Side-by-side boxplots of the misclassification rates for the comparative performance simulation study described in the text in the case of the response variable being Gaussian. Each panel corresponds to a different combination of sample size and error standard deviation. Within each panel, the side-by-side boxplots compare the zero, linear, non-linear effect misclassification rate across each of three methods: A. Laplace-Zero/Grouped Lasso-Zero priors with classification border at $\frac{1}{2}$, B. (Grouped) Horseshoe priors with classification border at $\frac{1}{2}$, C. (Grouped) Horseshoe priors with classification border determined via k-means clustering.*

References

Bai, R. & Ghosh, M. (2018). High dimensional multivariate posterior consistency under global-local shrinkage priors. *Journal of Multivariate Analysis*, **167**, 157–170.

Carvalho, C.M., Polson, N.G. & Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, **97**, 465–480.

He, V.X. and Wand, M.P. (2024). Bayesian generalized additive model selection including a fast variational option. *Advances in Statistical Analysis*, in press.

MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, pp. 281–297, Berkeley, California: *University of California Press*.

R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`

Scheipl, F., Fahrmeir, L. & Kneib, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, **107**, 1518–1532.

Xu, Z., Schmidt, D.F., Enes, M., Qian, G. & Hopper, J.L. (2016). Bayesian grouped horseshoe regression with application to additive models. In B.H. Kang & Q. Bai (eds.), *AI 2016: Advances in Artificial Intelligence*, pp. 229–240, Cham, Switzerland: Springer.