Computational Statistics and Data Analysis 177 (2023) 107568

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

Shared Bayesian variable shrinkage in multinomial logistic regression

Md Nazir Uddin [a], Jeremy T. Gaskins [b,*]

[a] Medpace Inc, 5375 Medpace Way, Cincinnati, OH 45227, USA
[b] Department of Bioinformatics and Biostatistics, University of Louisville, 485 E. Grey Street, Louisville, KY 40202, USA

ARTICLE INFO

Article history:
Received 12 July 2021
Received in revised form 19 April 2022
Accepted 10 July 2022
Available online 20 July 2022

Keywords:
Variable selection
Shared shrinkage
Bayesian analysis
Baseline categorical regression

ABSTRACT

Multiple Bayesian approaches have been explored for variable selection in the multinomial regression framework. While there are a number of studies considering variable selection in the regression paradigm with a numerical response, the research is limited for a categorical response variable. The proposed approach develops a method for leveraging the features of the global-local shrinkage framework to improve variable selection in baseline categorical logistic regression by introducing new shrinkage priors that encourage similar predictors to be selected across the models for different response levels. To that end, the proposed shrinkage priors share information across response models through the local parameters that favor similar levels of shrinkage for all coefficients (log-odds ratios) of a predictor. Different shrinkage approaches are explored using the horseshoe and normal gamma priors within this setting and compared to a spike-and-slab setup and other shrinkage priors that fail to share information across models. The performance of the approach is investigated in both simulations and a real data application.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

Variable selection in statistical modeling is a common challenge. This problem arises when multiple variables can be involved in building the statistical model that describes the relationship between a response variable of interest and a set of predictor variables. The main challenge is to identify a meaningful subset of predictors. This work explores different Bayesian variable shrinkage approaches in an unordered categorical response model.

The two most widely used probabilistic models for multicategory response variable are the multinomial logistic (MNL) regression and the multinomial probit (MNP) model. The natural extension of binary logistic regression to the multicategory setting is MNL regression. Suppose the random variable C_i is an unordered categorical response with K categories, $C_i \in \{1, 2, \ldots, K\}$ $(i = 1, 2, \ldots, n)$, and there are P predictors for each observation $X_i = (X_{i0}, X_{i1}, \ldots, X_{iP})$. Here $X_{i0} = 1$ to provide the intercept term for i^{th} subject. The MNL regression consists of $(K - 1)$ logistic regression models of the form

$$log\left[\frac{P(C_i = j)}{P(C_i = K)}\right] = X_i \kappa_j,$$

* Corresponding author.
E-mail address: jeremy.gaskins@lousville.edu (J.T. Gaskins).

https://doi.org/10.1016/j.csda.2022.107568
0167-9473/© 2022 Elsevier B.V. All rights reserved.

for $j = 1, \ldots, K-1$. In this model, $\kappa_j = (\kappa_{j0}, \kappa_{j1}, \ldots, \kappa_{jP})$ is the vector of log-odds ratios of class j versus class K. Without loss of generality, we let class K serve as the baseline class, and this model is known as the baseline category logit model. By combining all $(K-1)$ models, the probability that observation i will be in the jth class $(j = 1, \ldots, K)$ comes from

$$P(C_i = j \mid X_i) = \frac{e^{X_i \kappa_j}}{\sum_{l=1}^{K} e^{X_i \kappa_l}} = \pi_{ij}. \tag{1}$$

Note that to make the equation (1) identifiable, we define $\kappa_K = 0$ since K is the reference class. Further information about the MNL model and other categorical response models can be found in Agresti (2015).

In contrast to the MNL, the multinomial probit model uses $\Phi(\cdot)$, the cumulative distribution function of a standard normal, as the link function so that $\Phi\{P(C_i = j)/P(C_i = K)\} = X_i \kappa_j$ for each j. While this model can fit using Metropolis-Hastings sampling (as in Li et al., 2021), it is often difficult to appropriately tune the sampler and to ensure convergence. More commonly, the MNL model can be reframed by considering C_i in terms of normally distributed latent variables H_i. The response C_i is determined by through the $(K-1)$-dimensional H_i through

$$C_i = \begin{cases} K, & \text{if } max(H_i) < 0 \\ j, & \text{if } max(H_i) = H_{ij} > 0. \end{cases}$$

The latent variable H_i is modeled as

$$H_i = X_i \kappa + \varepsilon_i, \qquad \varepsilon_i \sim MVN(\mathbf{0}, \Psi), \tag{2}$$

where κ is $(K-1) \times P$ coefficient matrix and Ψ is a $(K-1) \times (K-1)$ positive definite matrix subject to identifiability constraints. The MNP has been applied by many authors, including McCulloch et al. (2000) and Imai and Van Dyk (2005). Although it is relatively easy to conceptualize and fit a (Bayesian) hierarchical model framework connecting the predictors to latent variables to the response, interpretation is not straightforward due to the fact that the coefficients act on the latent H_i rather than on the response C_i directly. Conversely, it is easy to interpret the MNL model as the coefficients are standard log-odds ratios, but Bayesian parameter estimation is not simple due to a lack of conjugacy in κ. Dow and Endersby (2004) and Kropko (2008) have directly compared the MNP and MNL models and argue that there is no general rule in favor of one model over the other. Our perspective is to prefer MNL over MNP due to its more straight-forward interpretation and because the Polya-Gamma (PG) data augmentation strategy developed by Polson et al. (2013) facilitates conjugate sampling. However, many aspects of our proposed methodology could be implemented in an MNP approach.

In the case of a large number of predictors P, it is typically believed that there are many irrelevant predictors that do not have any effect in determining the class membership of a subject. The main objective is to get rid of these irrelevant variables thorough Bayesian variable selection. In the Bayesian paradigm, there are two popular prior structures that are commonly used for sparse estimation in the standard linear regression model: (i) discrete choice selection priors and (ii) continuous shrinkage priors.

Discrete choice priors, the most common of which is the spike-and-slab (S&S), explicitly characterize each predictor as active or inactive through a discrete binary model that determines whether κ_p, the coefficient of the p^{th} predictor, will be non-zero or zero. For the κs that are active (non-zero), a conditional prior specifies a continuous distribution for these coefficients. This model originates from Mitchell and Beauchamp (1988), and numerous authors have explored various adjustments to this general approach (e.g., George and McCulloch, 1993; Liang et al., 2008; Ročková and George, 2014; Chen and Walker, 2019; Posch et al., 2020, and many others) by adjusting some combination of the structure of the variable inclusion probabilities, the continuous prior distribution for the non-zero coefficients, and the sampling/estimation strategy.

In contrast to the discrete selection models, the global-local (GL) shrinkage framework (Polson and Scott, 2010) encompasses the most widely-used continuous shrinkage priors. In a usual linear regression setting, the hierarchical structure of the prior distribution of a regression coefficient vector $\kappa = (\kappa_1, \ldots, \kappa_P)$ is expressed as

$$\kappa_p \mid \delta_p^2, \phi^2 \sim N\left(0, \delta_p^2 \phi^2\right),$$

with $\delta_p^2 \sim \pi_L(\delta_p^2)$ and $\phi^2 \sim \pi_G(\phi^2)$. Here, δ_p^2 is the local variance component which is predictor-specific. This allows deviation in the degree of shrinkage across the predictors, and ϕ^2 is the global variance component which determines the overall level of shrinkage towards zero. Both distributions π_L, for local shrinkage parameter, and π_G, for the global parameter, should have large mass close to zero to allow aggressive shrinkage, but the local distribution should also be heavy-tailed to leave large signals unshrunk. Some of the most common variable shrinkage methods within the GL framework include the following: Horseshoe (HS; Carvalho et al., 2010), Normal Gamma (NG; Griffin and Brown, 2010), Dirichlet-Laplace (Bhattacharya et al., 2015), and Horseshoe+ (Bhadra et al., 2017). The interested reader is encouraged to consider the survey paper on different Bayesian shrinkage methods by Bhadra et al. (2019). While shrinkage methods do not provide an explicit characterization of each variable as active or not, they are typically computationally faster than the selection methods

and their Markov chains tend to mix better, while providing comparable or more accurate parameter estimation. Arguably, coefficient shrinkage has become the preferred approach for high-dimensional Bayesian linear regression modeling.

While the literature is rich for variable selection and shrinkage for linear regression models, there has been limited work in the context of categorical regression. Zens (2019) proposed a variable shrinkage method in MNL regression model with a Normal Gamma prior structure. Recently, Bhattacharyya et al. (2021) explored Bayesian variable shrinkage with a horseshoe prior on the coefficients. Additionally, Polymeropoulos (2020) considered a selection-based approach using a mixture of g-priors combined with discrete predictor choice. Similarly, Tüchler (2008) used the spike-and-slab model for variable selection in a mixed effects MNL framework. Despite the importance of the unordered categorical regression problem, there is has been little work investigating the performance of variable selection in this context.

In the case of correlated continuous multi-outcome responses, Kundu et al. (2021) consider GL variable shrinkage and propose sharing the local parameter across the regression models for each response so that similar levels of shrinkage are applied across the models. Consequently, if a particular covariate is active in the regression model for one outcome (i.e., it is relatively unshrunk due to a large local parameter), it is more likely to play a substantial role in the other outcome models. Instead of multiple continuous responses, we have a multinomial response variable with K categories, but the MNL yields $K-1$ models that must each be fit. The main goal of this research work is leveraging the features in the GL framework to improve variable selection by developing new shrinkage priors that encouraging similar predictors to be selected across models for different response levels. Similarly to the strategy in Kundu et al. (2021), this may be achieved by sharing the local parameters across category models.

The remainder of this manuscript is divided into the following five sections. In section 2, we introduce the proposed shared shrinkage framework. Parameter estimation through Markov chain Monte Carlo (MCMC) is described in section 3 along with the data augmentation strategy required to gain conjugate sampling. In section 4, we perform simulation studies to evaluate the performance of our model. In particular, we consider the coverage rates of the coefficient credible intervals as an additional criteria for model performance evaluation. In section 5, we implement our model on a real data set consisting of patients at risk of Alzheimer's Disease. Section 6 provides some concluding remarks and discussion. This manuscript also has a corresponding supplementary Appendix file with additional details, tables, and figures.

2. Shared shrinkage (SS) prior in MNL

2.1. Shared shrinkage framework in MNL

Inspired by the shared shrinkage methods introduced in Kundu et al. (2021), we seek to explore variable selection in the MNL regression problem (1) that explicitly shares information across the different outcome levels. Again, the MNL model has $(K-1)$ logistic regression equations, and it may be believed that the κ_j $(j=1,2,\ldots,K-1)$ vectors are sparse in the sense that many of the entries κ_{jp} $(p=1,2,\ldots,P)$ are zero or nearly zero.

We assume that $\kappa_{jp} \sim N\left(0, \delta_p^2\phi^2\right)$, where δ_p is the shrinkage parameter of the p^{th} predictor. This δ_p is the predictor-specific local parameter which plays the key role in determining the most relevant predictors, and it will obtain large values for the most relevant predictors leading to little-to-no shrinkage of the associated regression coefficients. Conversely, a small value of δ_p will lead κ_{jp} to be near zero across all $j=1,\ldots,K-1$ models, indication this covariate is an irrelevant predictor. As recommended by Polson and Scott (2010), the distribution of the local parameter δ_p should have large variance to induce a thick tail in the κ_{jp} distribution and also substantial mass near zero to accommodate aggressive shrinkage for irrelevant predictors. By sharing the local parameter across the $K-1$ classes, this approach encourages similar predictors to be selected across these models for different response levels.

The global parameter ϕ contributes to determining the overall level of shrinkage to all coefficients. In particular, it shared across the $(K-1)$ regression models as we would expect similar magnitudes of these log-odds ratios across the classes. This is in contrast to the choice of Kundu et al. (2021) where each (normal) regression model had a response-specific parameter ϕ_j to allow for different scaling of the different responses. While our model can be easily extended to allow unique ϕ for each model j, we do not believe this is necessary as these log-odds model are implicitly on the same scale across j.

The above shared shrinkage framework is applied to the variances of the regression coefficients, but we do not wish to perform any shrinkage on the intercept terms. To that end, a normal distribution with mean 0 and variance 100 is the prior for κ_{j0} $(j=1,\ldots,K-1)$ to provide a relatively disperse, yet proper, prior.

Obviously, this shared shrinkage choice imposes sparsity in the log-odds ratios between class j and the reference class K. However, an additional useful property of our shared shrinkage framework is that we obtain a level of sparsity in the log-odds ratio between any two classes j and j'. The parameter κ_{jp} represents the log-odds of class j relative to the reference class K due to the predictor p, and as we have constrained the $\kappa_{Kp}=0$ for the reference class K, this log-odds is equivalent to $\kappa_{jp}-\kappa_{Kp}$. For any two classes j and j' $(j\neq j'\neq K)$, $\kappa_{jp}-\kappa_{j'p}$ represents the log-odds of class j relative to class j' due to the predictor p, but note that the implied distribution for this effect is $N(0, 2\delta_p^2\phi^2)$ under the shared shrinkage models. Consequently, a small δ_p does not only imply that the p^{th} predictor is not associated with the difference between class j and the baseline class K, but that it is unassociated with differences between any two response class.

Again, the crucial feature of our method is that the local parameter δ_p facilitates sharing information about the importance of the p^{th} predictor variables across all K categories of the outcome. This occurs by shrinking the original log-odds

ratio κ_{jp} for category j relative to the (fixed) baseline K and by shrinking all other pairwise log-odds comparisons. It is important to recognize that the previous shrinkage methods in the literature (Zens, 2019; Bhattacharyya et al., 2021) do not have this feature. In their work, observing a large coefficient in κ_{jp} will have no impact on the magnitude of $\kappa_{j'p}$ $(j \neq j')$. Similarly, the previous selection methods (Tüchler, 2008; Polymeropoulos, 2020) also have unique indicator vectors for each of the $K-1$ classes, and the priors do not favor similarity across these vectors. To our knowledge, our approach is the only Bayesian methodology proposed that encourages similarity in variable importance without restricting exact agreement in variable selection.

2.2. Horseshoe shared shrinkage (HS SS)

The Horseshoe prior is one of the most widely applied methods in the GL shrinkage framework because of its theoretical properties and its practical performance in shrinking a large number of predictors effectively. Therefore, we choose to implement the horseshoe prior structure proposed by Carvalho et al. (2010) in the MNL shared shrinkage framework. The proposed hierarchical prior setup is as follows

$$\begin{aligned} \kappa_{jp} &\sim N\left(0, \delta_p^2\phi^2\right) \\ \delta_p, \phi &\sim C^+(0,1). \end{aligned}$$

Here, $C^+(0,1)$ is the half Cauchy distribution with density $f(x) \propto 1/\left(1+x^2\right)$, $x>0$. As noted previously, this δ_p plays the key role in determining the most relevant predictors, and it is expected to have a larger value (less shrinkage) for meaningful predictors and a smaller value (more shrinkage) for unnecessary predictors that do not have any effect on response. This shrinkage method is referred to as the Horseshoe Shared Shrinkage (HS SS) model.

One difficulty is that HS prior as proposed by Carvalho et al. (2010) is not conjugate, even to a normal response variable, which makes the corresponding Gibbs sampler difficult and time consuming. To achieve conjugacy of the hyperparameters, Makalic and Schmidt (2015) proposed a data augmentation strategy by introducing auxiliary variables. If $\aleph^2 \mid \tau \sim IG\left(\frac{1}{2}, \frac{1}{\tau}\right)$ and $\tau \sim \left(\frac{1}{2}, 1\right)$, then the marginal distribution is $\aleph \sim C^+(0,1)$. With this equivalence result, the prior structure of HS SS can be equivalently stated as

$$\begin{aligned} \kappa_{jp} \mid \delta_p^2, \phi^2 &\sim N\left(0, \delta_p^2\phi^2\right) \\ \delta_p^2 \mid \eta_p &\sim IG\left(\frac{1}{2}, \frac{1}{\eta_p}\right) \\ \phi^2 \mid \xi &\sim IG\left(\frac{1}{2}, \frac{1}{\xi}\right) \\ \eta_1, \eta_2, \cdots, \eta_P, \xi &\sim IG\left(\frac{1}{2}, 1\right). \end{aligned}$$

Here, IG stands for the inverse gamma distribution with probability density function $f(x \mid a, b) \propto x^{-a-1} exp\,(-b/x)$, $x>0$.

2.3. Normal Gamma shared shrinkage (NG SS)

The NG variable shrinkage model proposed by Griffin and Brown (2010) assume a gamma distribution on the local variance parameter to achieve a thick-tailed distribution on the normally distributed regression coefficients. We adopt a modified parameterization of the NG prior structure similar to that applied by Zens (2019). Our normal gamma shared shrinkage model (NG SS) considers the following hierarchy:

$$\begin{aligned} \kappa_{jp} \mid \delta_p^2, \phi &\sim N\left(0, \frac{2}{\phi}\delta_p^2\right) \\ \delta_p^2 &\sim Gamma\,(\theta, \theta) \\ \phi &\sim Gamma\,(c_0, c_1). \end{aligned}$$

Here, the hyperparameter θ plays the key role in the distribution of the regression coefficients κ_{jp}. The mean and variance of δ_p^2 are 1 and $1/\theta$, respectively. Consequently, a small value of θ will lead to more shrinkage due to the large variability in local parameters δ. The θ can be sampled using Metropolis Hasting (MH) or other non-conjugate sampling steps, but Zens (2019) argued that fixing $\theta = 0.05$ yielded effective variable selection with more efficient computation. The choice of fixed θ is further discussed in detail by Bitto and Frühwirth-Schnatter (2019). We follow the recommendation of Zens (2019) throughout and fix $\theta = 0.05$. The values of hyperparameters c_0 and c_1 are both chosen to be 0.01.

While the MNL shared shrinkage framework is general, we have focused on its implementation with the two most common global-local shrinkage models. However, this model can be implemented with other prior structures. For instance,

the Dirichlet-Laplace model (Bhattacharya et al., 2015), which places a Dirichlet prior on the vector of local priors, is one such choice. However, as shown in Kundu et al. (2021), this model loses conjugacy when implemented in a shared shrinkage setting, so we do not further explore its performance. Similarly, the Bayesian Lasso (BL) model (Park and Casella, 2008) is a common global-local shrinkage model that can be extended to our framework by sharing the exponentially distributed local parameter across models. Importantly, the Bayesian Lasso is equivalent to the NG model with θ fixed at 1. As this choice induces lighter tails for κ_{jp} than when $\theta = 0.05$, we would expect such a model to have worse performance than NG SS as it may over-shrinking large signals. We do include this BL choice in our simulation study, and we do generally see this to be the case.

3. Markov Chain Monte Carlo (MCMC) computations

3.1. Polya-Gamma data augmentation

Inference in these Bayesian models is performed by drawing posterior samples using the MCMC methods. In this section, we present the posterior conditionals of each model with the necessary sampling steps.

An important challenge here is that the data likelihood (1) as usually framed is not conjugate to the multivariate normal prior on the κ_j vectors. With the categorical response variable, the easiest and most efficient way is to introduce the data augmentation strategy developed by Polson et al. (2013). For $j = 1, \ldots, K-1$, we let $R_{ij} = I(C_i = j)$ be an indicator variable such that $P(R_{ij} = 1) = \pi_{ij}$ with probabilities π_{ij} from (1). Polya-Gamma (PG) latent variables ω_{ij} can be introduced for each R_{ij} to obtain conjugacy in the κ_j vectors. That is, $\omega_{ij} \sim PG(1, B_{ij})$ where PG is the Polya-Gamma distribution due to Polson et al. (2013) depending on parameter $B_{ij} = x_i^T \kappa_j - D_{ij}$, $D_{ij} = log \sum_{h \neq j} e^{x_i^T \kappa_h}$. The key result for PG distribution is

$$\frac{\left(e^{B_{ij}}\right)^{R_{ij}}}{\left(1+e^{B_{ij}}\right)} = 2^{-1} e^{(R_{ij}-1/2)B_{ij}} \int_0^\infty e^{-w_{ij}B_{ij}^2/2}\, p\left(w_{ij} \mid 1, 0\right) d\omega_{ij}. \tag{3}$$

Here $p\left(\omega_{ij} \mid 1, 0\right)$ is the $PG(1, 0)$ density. While the left side of the equation (3) is not in a conjugate form, the right side becomes conjugate to the multivariate normal (MVN) density. By sampling with respect to the data augmented likelihood, we can sample the regression coefficient vectors from conjugate MVN steps.

For a given $j = 1, \ldots, K-1$, the likelihood function for $\kappa_j \mid \kappa_{-j}$ becomes

$$\prod_{i=1}^{n} e^{(R_{ij}-\frac{1}{2})(x_i'\kappa_j - D_{ij})} e^{\frac{1}{2}(x_i'\kappa_j - D_{ij})^2 \omega_{ij}}\, PG(\omega_{ij} \mid 1, 0),$$

which again is clearly conjugate to any MVN prior for κ_j. We let κ_{-j} denote the collection of κ vectors excluding the j^{th} class.

3.2. Sampling steps for HS SS model

The MCMC samples of the parameters in the HS SS model are obtained by cycling through the following conditional sampling distributions.

(i) For each class $j = 1, 2, \ldots, K-1$,
  (a) Sample the PG random variable $\omega_{ij} \mid \cdot \sim PG\left(1, B_{ij}\right)$ where $B_{ij} = x_i^T \kappa_j - log \sum_{h \neq j} e^{x_i^T \kappa_h}$ for $i = 1, \ldots, n$.
  (b) Sample the κ_j vector for class j conditionally on the current value of the other regression vectors. The sampling distribution is $\kappa_j \mid \kappa_{-j}, \cdot \sim MVN\left(V_{\kappa_j} X' \left(z_j^\star + \Omega_j D_j\right), V_{\kappa_j}\right)$, where $V_{\kappa_j} = \left(\Delta_j + X'\Omega_j X\right)^{-1}$ and Δ_j is the $(P+1) \times (P+1)$ diagonal matrix with $\left[\left(\delta_0^2\right)^{-1}, \left(\delta_1^2\phi^2\right)^{-1}, \left(\delta_2^2\phi^2\right)^{-1}, \ldots, \left(\delta_P^2\phi^2\right)^{-1}\right]$ entries. As noted previously, $\delta_0^2 = 100$ is fixed to account for the prior variance of the intercept. Ω_j is the $n \times n$ diagonal matrix with entries $\left(\omega_{1j}, \omega_{2j}, \ldots, \omega_{nj}\right)$, $z_j^\star$ is the n-vector with entries $R_{ij} - 1/2$, and $D_j = \left(D_{1j}, D_{2j}, \ldots, D_{mj}\right)$ is the n-vector with elements $D_{ij} = log \sum_{h \neq j} e^{x_i^T \kappa_h}$.

(ii) For $p = 1, 2, \ldots, P$, sample the local parameters $\delta_p^2 \mid \cdot \sim IG\left(\frac{K}{2}, \frac{1}{\eta_p} + \frac{1}{2\phi^2} \sum_{j=1}^{K-1} \kappa_{jp}^2\right)$.

(iii) Sample the global parameter, $\phi^2 \mid \cdot \sim IG\left(\frac{(K-1)P+1}{2}, \frac{1}{\xi} + \frac{1}{2} \sum_{j=1}^{K-1} \sum_{p=1}^{P} \frac{\kappa_{jp}^2}{\delta_p^2}\right)$.

(iv) For $p = 1, 2, \ldots, P$, sample the data augmentation variables for the local parameters $\eta_p \mid \cdot \sim IG\left(1, 1 + \frac{1}{\delta_p^2}\right)$.

(v) Sample the data augmentation variable for the global shrinkage parameter $\xi \mid \cdot \sim IG\left(1, 1 + \frac{1}{\phi^2}\right)$.

Table 1
True values for the MNL regression coefficients in the simulation setup.

Study I							
Class	Intercept	X_1	X_2	X_3	X_4	$\cdots$	X_{100}
Class 1	-0.50	2.00	2.50	0.00	0.00	$\cdots$	0.00
Class 2	0.25	0.75	-1.50	0.00	0.00	$\cdots$	0.00
Study II							
Class 1	-0.50	2.00	2.50	0.00	0.00	$\cdots$	0.00
Class 2	0.25	0.75	-1.50	1.00	0.00	$\cdots$	0.00
Study III							
Class 1	-0.50	2.00	2.50	0.00	0.00	$\cdots$	0.00
Class 2	0.25	0.00	0.00	0.75	-1.50	$\cdots$	0.00

3.3. Sampling steps for NG SS model

The MCMC sampling steps of NG SS are similar to the HS SS model except for the necessary modifications to the distribution of the global and local parameters.

(i) The steps for updating the ω_{ij} and κ_j are same as in HS SS model.

(ii) For $p = 1, 2, \ldots, P$, sample the local parameters $\delta_p^2 \mid \cdot \sim GIG\left(\theta - \frac{1}{2},\ \frac{1}{2}\phi \sum_{j=1}^{K-1} \kappa_{jp}^2,\ 2\theta\right)$. Here, $GIG(\varpi, \Upsilon, \Psi)$ is the generalized inverse Gaussian distribution with density $f(x) \propto x^{\varpi-1} e^{-\frac{1}{2}\left(\frac{\Upsilon}{x} + \Psi x\right)}$, $x > 0$.

(iii) Sample the global parameter $\phi \mid \cdot \sim Gamma\left(c_0 + \frac{(K-1)P}{2},\ c_1 + \sum_{j=1}^{K-1} \sum_{p=1}^{P} \frac{\kappa_{jp}^2}{4\delta_{jp}^2}\right)$.

R functions to implement these MCMC algorithms are available at https://github.com/jeremygaskins/BCLshrink. We also note that the Bayesian Lasso shared shrinkage (BL SS) model has the same sampler as NG SS with the choice of $\theta = 1$.

4. Simulation studies

4.1. Simulation set up

We have conducted three simulation studies to demonstrate our methodology and to compare it against other approaches. Here, we consider scenarios with $K = 3$ categories for the outcome variable, and in each case, we perform analysis on 100 independently replicated data sets. Each data set contains 100 predictors of which half are binary, generated from Bernoulli distribution with success probability of 0.20 or 0.50, and half are continuous from a mean zero, unit variance normal distribution. As predictors are often correlated, we use an auto-regression correlation structure with $\rho = 0.85$ for the correlated predictors. The binary predictors are placed in the odd positions and continuous predictors in even positions. For improved stability, the binary predictors are shifted to take values of 0.5 for success and -0.5 for failure.

In simulation study I, we allow the first two predictors (one binary and one continuous) to determine the response categories. Study II is designed to check the sensitivity of the shrinkage methods where the set of active predictors differ slightly between classes. In this case, the third predictor is active in the regression model for class 2 (versus $K = 3$), but not in the logistic model for class 1 (versus class 3). In simulation study III, we consider even more dissimilarity in the predictors for the response categories by allowing different pairs of predictors (one binary and one continuous for each) to be active in class 1 (vs $K = 3$) and class 2 (vs $K = 3$). Table 1 shows the true regression coefficients values used in the two simulation studies. In all studies, we consider $n = 500$ observations and either (A) $P = 10$ predictors or (B) $P = 100$ predictors as a higher-dimension setting.

4.2. Competitor models

To compare the performance of our proposed models, we consider unique shrinkage models based on both horseshoe and normal gamma structures as competitor models. The idea of unique shrinkage is analogous to placing unique versions of the GL shrinkage prior on each of the $K - 1$ binary logistic models that define the MNL model. As noted previously, this is the usual practice when applying Bayesian variable shrinkage in the MNL model. Under this choice, each coefficient κ_{jp} has its own local parameter δ_{jp}, and there is only one global parameter ϕ for all $j = 1, \ldots, K - 1$ class. No information about variable importance is shared between the κ_1 and κ_2 vectors.

- **Horseshoe Unique Shrinkage (HS US) model:** Rather than sharing shrinkage information across $K - 1$ regression models, each coefficient has a unique δ_{jp} shrinkage parameter from the usual HS local distribution. The prior hierarchy for HS

US can be written as $\kappa_{jp} \sim N\left(0, \delta_{jp}^2 \phi^2\right)$ and $\delta_{jp}, \phi \sim C^+(0,1)$. The relevant Gibbs sampling steps are straight-forward adaptations of those listed in Section 3.2, and we include the details in Appendix A.1. We note that this model is same as that used by Bhattacharyya et al. (2021).

- **Normal Gamma Unique Shrinkage (NG US) model:** Making similar adaptations to the NG SS as above gives us the model hierarchy as $\kappa_{jp} \sim N\left(0, 2\delta_{jp}^2/\phi\right)$, $\delta_{jp}^2 \sim G(\theta, \theta)$, and $\phi \sim G(c_0, c_1)$. Zens (2019) applied this hierarchical structure in the MNL framework. Again, sampling from the NG US model involves the obvious adjustments to the NG SS sampler; details are provided in Appendix A.
- **Bayesian Lasso Shared Shrinkage (BL SS) and Bayesian Lasso Unique Shrinkage (BL US) models:** Changing $\theta = 0.05$ to $\theta = 1$ in the NG SS and NG US models produce equivalent versions based on the Bayesian Lasso, rather than the Normal Gamma shrinkage model. All steps are the same as in the NG SS/US choices with the new value of θ.
- **Spike-and-slab (S&S) model:** As noted in the introduction, discrete choice variable selection models that include a mixture of a zero point mass and continuous distribution are a common variable selection choice in Bayesian regression modeling. To implement a version of this S&S strategy in the MNL context, we use a prior distribution for κ_{jp} of $(1-\pi_j)I(\kappa_{jp}=0) + \pi_j N(0, \tau^2)$ where π_j represent the probability of a non-zero coefficient in class j.
  We note that similar S&S variable selection models have used in the context of MNL by Tüchler (2008) and Polymeropoulos (2020). Similar to our shrinkage models, we implement this approach by using the PG data augmentation to faciliate sampling. We assume a Beta$(1,1)$ prior for the sparsity parameters π_j, and we fix $\tau^2 = 100$ to provide a moderately disperse distribution for the non-zero coefficients. We provide the details for sampling this model in Appendix A.
- **No Shrinkage (NS) model:** In the NS competitor, we assume the prior mean 0 and variance 100 for all regression coefficients, $\kappa_{jp} \sim N(0, 100)$. This provides a reference method that does not use any sophisticated shrinkage methodology and will instead approximate maximum likelihood estimation by using a moderately disperse prior. Sampling from this model involves using steps (i) and (ii) from the HS SS model with Δ_j having $1/100$ as its diagonal elements.

4.3. Model performance evaluation

For the simulation studies, estimation accuracy is evaluated by considering the bias and mean square error (MSE) of the regression coefficients. We let κ_{jp}^0 be the true value of the population parameter for p^{th} predictor at j^{th} model, and for each of the $M = 100$ replicated data sets, $\hat{\kappa}_{jp}^m$ is the estimate (posterior mean) from the m^{th} data set. The bias and MSE of a particular coefficient are averaged across the data sets as bias $= \frac{1}{M}\sum_{m=1}^{M}\left(\hat{\kappa}_{jp}^m - \kappa_{jp}^0\right)$ and MSE$= \frac{1}{M}\sum_{m=1}^{M}\left(\hat{\kappa}_{jp}^m - \kappa_{jp}^0\right)^2$.

In addition to investigating the bias and MSE, we also evaluate the empirical coverage rate (CvR) of the credible intervals for the regression coefficients. The coverage rate as a function of the level of significance $1-\alpha$ is defined as CvR(α) and estimated by

$$\frac{1}{M}\sum_{m=1}^{M} I\left[L_{jp}^m(\alpha) \le \kappa_{jp}^0 \le U_{jp}^m(\alpha)\right].$$

Here, $L_{jp}^m(\alpha)$ to $U_{jp}^m(\alpha)$ is the $100(1-\alpha)\%$ credible interval (CI) for κ_{jp}^0 from the m^{th} data set. Throughout, we use equal-tailed intervals estimated by selecting the $\alpha/2$ and $1-\alpha/2$ quantiles from the posterior samples. For the active (non-zero) predictors, a good performing method will have CvR(α) close to the nominal rate of $100(1-\alpha)\%$ for all choices of α. In contrast, for predictors whose coefficient is zero, which will refer to as "fake predictors," we would expect the shrinkage methods to yield posterior distributions that are almost entirely within a small neighborhood of the true coefficient value zero. That is, we would want and expect over-coverage of these null effects. Recall that the shrinkage methods do not provide interpretable variable selection parameters since the posterior is a continuous distribution. As is commonly done, we will investigate the 95% CI and consider the predictor to be active in model j if its interval excludes zero. For the S&S method, variables are selected if they have a greater than 50% posterior probability of being non-zero (see Appendix A for additional details on inference under the S&S model).

Throughout we will report the bias, MSE and CvR for the coefficients of the first four predictors from each class; as shown in Table 1, this set will contain the active predictors for each setting. An average of absolute bias, MSE and CvR across the remaining predictors X_5 to X_P, all of which are fake predictors, is obtained by averaging the accuracy measures across these remaining estimands. To make the values in the results tables more readable, we scale the bias and MSE by a factor of 100 in all cases to highlight the differences between the methodologies.

While our interest is mainly in parametric inference and variable selection, we also investigate the differences across the methods in terms of predictive accuracy. To that end, we generate an additional data set for each setting and predict the class outcomes in this out-of-sample test data set using the estimates from the $M = 100$ (training) data sets. For each observation $i = 1, \ldots, n_{test} = 500$, we compute the predicted class probabilities $\hat{\pi}_{ij}$ from equation (1) using the estimated $\hat{\kappa}_j$ from each method. For each observation in the test data, we predict the class based on $\text{argmax}_{j=1,\ldots,K}\{\hat{\pi}_{ij}\}$ and measure this "Classification Accuracy" as the percent agreement with the true class C_i across all test observations i and all M simulated data sets.

Table 2
MSE of all real and some fake coefficients for simulation study I. All values are scaled by 100.

Study I, Case A ($P = 10$)									
Class	Predictors	HS SS	HS US	NG SS	NG US	BL SS	BL US	NS	S&S
Class 1	X_1	21.15	22.27	19.22	21.35	23.59	23.22	21.09	37.40
	X_2	11.34	12.90	10.04	11.22	14.25	15.99	25.42	9.42
	X_3	1.26	1.82	0.65	1.09	3.32	3.95	11.31	0.30
	X_4	0.81	1.53	0.18	0.47	4.06	5.45	18.25	0.00
Class 2	X_1	10.87	17.09	10.81	20.43	9.91	11.31	14.53	44.53
	X_2	4.68	5.22	4.42	4.70	4.86	5.32	8.94	5.25
	X_3	1.86	2.59	0.96	1.38	2.09	5.81	13.03	0.00
	X_4	1.18	2.31	0.32	0.85	4.24	5.54	13.69	0.00
All fake		0.93	1.50	0.31	0.58	3.40	4.28	12.76	0.02
Study I, Case B ($P = 100$)									
Class 1	X_1	32.35	42.19	27.46	34.17	101.51	99.73	7995.09	147.27
	X_2	9.85	11.26	9.86	10.01	19.05	22.82	19592.01	9.57
	X_3	0.05	0.09	0.26	0.35	0.55	0.66	547.09	0.00
	X_4	0.01	0.02	0.15	0.24	1.62	2.35	886.71	0.00
Class 2	X_1	11.94	37.57	10.42	30.06	21.59	32.28	315.78	55.40
	X_2	5.33	5.79	5.97	5.82	5.72	5.29	1233.42	5.50
	X_3	0.09	0.12	0.47	0.56	1.15	1.36	158.28	0.00
	X_4	0.02	0.08	0.17	0.51	1.37	2.49	137.22	0.00
All fake		0.06	0.08	0.20	0.24	0.66	2.49	431.06	0.00

We also consider two additional predictive accuracy measures that use the full predicted class probabilities, instead of only whichever is largest. To that end we utilize the predictive log-score (Gaskins, 2019; Zhou et al., 2015; Gneiting and Raftery, 2007) which is defined as

$$L^{test} = E^{post}\left\{\sum_{i=1}^{n_{test}}\sum_{j=1}^{K} r_{ij}\, log\left(\frac{e^{X_i\kappa_j}}{\sum_{l=1}^{K} e^{X_i\kappa_l}}\right)\right\}. \tag{4}$$

The right hand side of (4) represents a posterior expectation with respect to the MCMC sample from the m^{th} training data. Here, X_i and $r_{ij} = I(C_i^{test} = j)$ correspond to patient i in the test data. As this is basically a log-likelihood for the test data evaluated at the training data parameter estimates, model with higher log-score indicates better model fit. Additionally, we measure the similarity between the estimated class probability vector $(\hat{\pi}_{i1}, \ldots, \hat{\pi}_{iK})$ and the true probability vector $(\pi_{i1}, \ldots, \pi_{iK})$ based on the true parameter values from Table 1 using the Kullback-Leibler (KL) divergence, summed over all test data observations:

$$KL^{test} = \sum_{i=1}^{n_{test}}\sum_{j=1}^{K} \pi_{ij}\, log\left(\frac{\pi_{ij}}{\hat{\pi}_{ij}}\right).$$

Smaller values represent better agreement between the estimated and true probability vectors. Measuring the accuracy using the log-score and KL divergence will favor model with both high accuracy and that are well-calibrated.

For each data set, we run the MCMC algorithm of Section 3 for a total of 15,000 iterations, excluding the first 5000 as burn-in, for all models except the S&S approach. Since the S&S model is much more computationally time-consuming, we ran it for only 2500 iterations with a burn-in of 500 iterations. In the cases when $P = 10$, the average computation time for the shrinkage models is approximately 0.88 minutes per 1000 iterations, whereas S&S requires approximately 2.28 minutes per 1000 iterations. When the number of predictors is increased to $P = 100$, the average time is 3.22 minutes per 1000 iterations for the shrinkage models, compared to 85.22 minutes for S&S. Consequently, we only run the S&S algorithms for a fraction of the iterations in the other choices, although the overall computing times require are comparable.

4.4. Simulation results: Study I

In this section, the results are provided for simulation study I. The MSE of the first four predictors are shown in Table 2, and a corresponding table of the biases is reported in Appendix B, Table B.1. In case A ($P = 10$), NG SS has the smallest MSE among the shrinkage methods. We see the shared shrinkage methods are better (lower MSE) than their corresponding unique shrinkage method with the same GL prior structure. Coefficients for the fake predictors are consistently estimated to

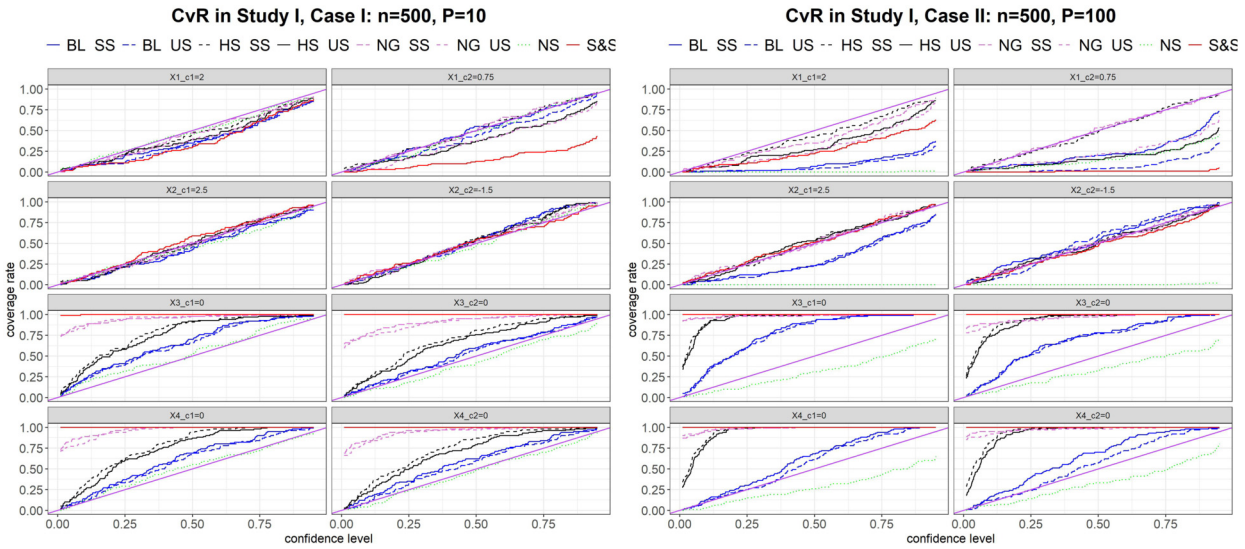

Fig. 1. Posterior credible interval coverage rates for study I. (For interpretation of the colors in the figures, the reader is referred to the web version of this article.)

be near zero with the NG and SS choices performing better than HS and US across all combinations. Due its thinner tails the BL methods are substantially worse than the corresponding NG version as expected. Even with the small P, failing to utilize a shrinkage prior yields substantially poorer estimation in the NS method for the active predictor X_2 and especially for the fake predictors. While the S&S has the strongest estimation for the zero coefficients, it has substantially larger estimation error for the real predictors, particular for X_1. Investigation of the bias (Table B.1) shows that on average the shrinkage methods display a slight bias due to pulling the point estimates closer to zero; this is the expected behavior in a shrinkage framework. For the real predictors X_1 and X_2, the US methods tend to have slightly higher bias than the NG choices.

In case B ($P = 100$) where the number of predictors is increased, much of the conclusions are the same. Sharing information about predictor performance leads to lower MSE and smaller bias than the unique shrinkage methods. For larger P the horseshoe prior produces much better estimation for the zero coefficients relative to NG. S&S accurately estimates the zero coefficients but displays larger estimation error for the non-zero coefficients. Failing to shrinkage the coefficients in the NS choices yields completely erroneous and unstable inference as this model massively overfit the data.

In addition to the bias and MSE of the predictors, we are also interested in the coverage rates of the corresponding credible intervals. Table B.3 in Appendix B shows the coverage rates for the 20%, 50%, 80% and 95% intervals for cases A and B in each method for the first four predictors. As we noted earlier, we want the intervals for the real predictors X_1 and X_2 to achieve their nominal rates, whereas we want to over-cover the null effects X_3 and X_4. Fig. 1 shows the full trend of CvR(α) across all values of nominal interval probability. Briefly, we note that all shrinkage methods obtain the nominal coverage rate for the real predictors, except for the smaller coefficient of X_1 in class 2 ($\kappa_{21} = 0.75$). For both small and large P, the unique shrinkage methods yield intervals that undercover the true value as they overshrink this smaller effect size. By sharing information across models, the SS choices learn that X_1 is a relevant predictor, and this protects from overshrinking and undercovering. However, when the true predictor is zero, the US methods show much more aggressive overcoverage than the shared shrinkage approaches as they yield posteriors that are more concentrated near zero.

When a binary variable selection decision is required, one typically uses the 95% CI to select variables whose intervals exclude zero. (For the S&S model, we consider a variable selected if its posterior probability of being non-zero is greater than 50%.) These results are shown in Table 3. For small P (case A), all methods correctly select X_2 as relevant predictor for both outcome models, and all methods recognize X_1 is important for distinguishing category 1 (versus $K = 3$). However, it is clear that conclusions about the role of X_1 for category 2 will be greatly impacted by failing to share information across models, as the US and S&S methods have substantially lower power for this method. The results are similar in case B ($P = 100$) except that the overall power across all methods is reduced. Additionally, NS produces a large number of false positives (incorrectly selected coefficients), and relative to other shrinkage approaches, the BL versions also have higher false positive rates.

In Table 4, we compare the out-of-sample predictive accuracy across all methods. Despite the difference in parameter estimation, we see relative little difference in the classification accuracy across methods. For small P, all methods have approximately 70–72% accuracy, and the performance is similar for large P, except for the much poorer performing NS. However, when we consider the estimated class probability through the log-score (instead of the predicted class indicator), we see better performance for the SS methods. Additionally, when we measure the similarity between the estimated classification probabilities to the true model classification probabilities through the Kullback-Leibler divergence, we see more

Table 3
Percentage of replicated data sets in which predictor is chosen as significant in Study I. The Average False Positives row indicates the average number of coefficients from predictors X_5 to X_P that are incorrectly chosen to be non-zero. Recall the true predictors are X_1 and X_2 for both classes.

Study I, Case A ($P = 10$)									
		HS SS	HS US	NG SS	NG US	BL SS	BL US	NS	S&S
Class 1	X_1(%)	99	98	97	97	100	100	100	95
	X_2(%)	100	100	100	100	100	100	100	100
	X_3(%)	1	1	0	0	1	2	5	0
	X_4(%)	0	0	0	0	1	1	7	0
Class 2	X_1(%)	62	28	66	20	44	44	77	15
	X_2(%)	100	100	100	100	100	100	100	100
	X_3(%)	0	1	0	0	3	3	9	0
	X_4(%)	0	1	0	0	1	3	7	0
Average False Positives		0.01	0.02	0.00	0.00	0.08	0.14	0.67	0.00
Study I, Case B ($P = 100$)									
Class 1	X_1(%)	81	79	91	93	87	91	100	58
	X_2(%)	100	100	100	100	100	100	100	100
	X_3(%)	0	0	0	0	0	0	29	0
	X_4(%)	0	0	0	0	0	0	35	0
Class 2	X_1(%)	58	2	50	3	21	6	82	1
	X_2(%)	100	100	100	100	100	100	100	100
	X_3(%)	0	0	0	0	0	0	30	0
	X_4(%)	0	0	0	0	0	1	20	0
Average False Positives		0.00	0.00	0.03	0.01	0.04	0.08	56.71	0.00

Table 4
Classification accuracy (%), predictive log-score, and Kullback-Leibler divergence for study I.

	Study I, Case A ($P = 10$)							
	HS SS	HS US	NG SS	NG US	BL SS	BL US	NS	S&S
Classification accuracy	71.60	71.42	71.62	71.49	71.36	71.26	70.73	71.26
(SE)	(0.07)	(0.07)	(0.06)	(0.07)	(0.07)	(0.08)	(0.09)	(0.08)
Predictive log-score	-312.6	-315.2	-311.4	-314.3	-315.7	-317.0	-319.7	-317.8
(SE)	(0.5)	(0.5)	(0.4)	(0.5)	(0.6)	(0.6)	(0.8)	(0.6)
Kullback-Leibler	5.12	6.32	4.14	5.33	7.65	8.52	12.93	6.92
(SE)	(0.27)	(0.00)	(0.24)	(0.27)	(0.28)	(0.29)	(0.41)	(0.46)
	Study I, Case B ($P = 100$)							
Classification accuracy	71.55	71.28	71.30	70.93	70.06	69.73	61.63	70.61
(SE)	(0.08)	(0.10)	(0.08)	(0.09)	(0.11)	(0.11)	(0.28)	(0.13)
Predictive log-score (SE)	-313.2	-319.4	-316.3	-321.4	-333.4	-336.8	-1413.3	-326.0
(SE)	(0.6)	(0.6)	(0.6)	(0.6)	(0.8)	(0.8)	(53.0)	(0.9)
Kullback-Leibler (SE)	6.15	9.11	8.76	11.52	23.98	26.92	257.93	13.64
(SE)	(0.44)	(0.49)	(0.40)	(0.39)	(0.63)	(0.60)	(4.05)	(0.92)

accurate performance in the shared shrinkage approach. As with the parameter estimation, NG SS is best in case A and HS SS in case B.

An important consideration in all of these settings is MCMC convergence. In Figures B.1–B.6 in Appendix B.1, we show the traceplots for the same set of four predictors in a representative data set for the case B setting. These plots show that the posterior samples are well-mixed and that there is no evidence of convergence failure for the shrinkage methods. Additionally, we compute the effective sample size (ESS) for these coefficients (averaged over the 100 simulated data sets) and report them in Table 4. As a general rule, we would like to have an effective sample size of at least 1000 for each parameter of interest, as this indicates that our MCMC samples contain information equivalent to 1000 independently posterior draws. Our choice of 15,000 iterations typically achieves this criteria for the shrinkage methods. However, in the non-shrinkage (default independent prior) choice, the higher autocorrelation for the large P case significantly reduces the ESS.

As is often the case, the spike-and-slab approach is computationally slower by many orders of magnitude. As noted earlier, we have chosen to limit the number of iterations for S&S based on the computation time needed for convergence of the corresponding shrinkage models. This implies that the S&S MCMC chains may not have been run long enough to ensure full burn-in and/or posterior samples that are well-mixed across the chain. The traceplots (Figure B.6) and ESS (Table B.2) indicate that this may be the case. The average ESS for S&S are less than 500 for all non-zero coefficients, implying that much longer MCMC chains are required to gain well-mixed posterior samples in these data. However, as this slow mixing is a critical feature of the S&S methods and as users must obtain inference in finite time, we believe using comparable computation time provides the most fair comparison across models.

Table 5
MSE of all real and some fake coefficients for simulation study II. All values are scaled by 100.

Study II, Case A ($P = 10$)									
		HS SS	HS US	NG SS	NG US	BL SS	BL US	NS	S&S
Class 1	X_1	20.58	20.97	19.09	20.57	22.10	21.77	21.05	35.26
	X_2	8.96	10.28	7.74	8.61	10.91	11.59	22.70	7.71
	X_3	5.49	2.13	5.72	1.00	5.91	4.63	12.79	0.03
	X_4	1.21	1.83	0.30	0.43	4.02	5.12	16.69	0.00
Class 2	X_1	11.34	15.25	11.46	19.88	10.06	10.92	15.69	40.46
	X_2	4.96	5.76	4.62	5.15	5.35	3.01	9.63	5.34
	X_3	18.61	15.81	22.62	18.42	12.90	11.84	14.43	41.41
	X_4	1.69	2.71	0.53	1.08	4.60	6.07	14.24	0.01
All fake		1.35	1.92	0.44	0.68	3.88	4.71	12.99	0.04
Study II, Case B ($P = 100$)									
Class 1	X_1	29.59	38.33	26.80	31.15	89.68	89.39	7042.34	138.66
	X_2	7.90	9.20	8.04	8.22	14.41	17.27	17338.65	7.89
	X_3	3.36	0.05	3.89	0.18	2.11	0.81	635.28	0.00
	X_4	0.03	0.03	0.18	0.19	1.71	2.26	833.35	0.00
Class 2	X_1	11.81	35.58	10.52	28.30	19.34	29.90	363.44	55.08
	X_2	5.19	5.74	5.62	5.88	5.33	5.58	1105.20	4.92
	X_3	39.18	34.17	27.87	24.98	28.76	26.31	667.92	76.74
	X_4	0.05	0.13	0.31	0.61	1.64	3.00	159.18	0.00
All fake		0.06	0.09	0.22	0.26	0.74	0.90	384.57	0.00

Table 6
Percentage of replicated data sets in which predictor is chosen as significant in Study II. The Average False Positives row indicates the average number of coefficients from predictors X_5 to X_P that are incorrectly chosen to be non-zero. Recall the true predictors are X_1 and X_2 for both classes and X_3 in class 2 only.

Study II, Case A ($P = 10$)									
		HS SS	HS US	NG SS	NG US	BL SS	BL US	NS	S&S
Class 1	X_1(%)	99	100	98	98	100	100	100	95
	X_2(%)	100	100	100	100	100	100	100	100
	X_3(%)	3	0	1	0	3	2	4	0
	X_4(%)	0	0	0	0	0	0	6	0
Class 2	X_1(%)	66	31	69	20	64	53	78	20
	X_2(%)	100	100	100	100	100	100	100	100
	X_3(%)	68	74	50	56	81	82	92	55
	X_4(%)	0	0	0	0	2	5	10	0
Average False Positives		0.03	0.02	0.00	0.00	0.14	0.15	0.84	0.00
Study II, Case B ($P = 100$)									
Class 1	X_1(%)	87	84	89	90	90	95	100	60
	X_2(%)	100	100	100	100	100	100	100	100
	X_3(%)	0	0	0	0	0	0	32	0
	X_4(%)	0	0	0	0	0	0	34	0
Class 2	X_1(%)	59	4	55	7	22	8	83	1
	X_2(%)	100	100	100	100	100	100	100	100
	X_3(%)	30	38	39	44	55	57	94	19
	X_4(%)	0	0	0	0	0	1	24	0
Average False Positives		0.00	0.00	0.00	0.00	0.03	0.05	56.70	0.00

4.5. Simulation results: Study II

As shown in Table 1, the simulation study II specifies the third predictor X_3 to have non-zero coefficient in the logistic model for class 2 ($\kappa_{23} = 1$) but is zero for the class 1 model ($\kappa_{13} = 0$). We wish to investigate how the models perform when we consider a predictor that is active in only one of the models. The MSEs are shown in Table 5 and the variable selection results in Table 6. In the Appendix, we also include tables summarizing the bias (Table B.4), computational performance (B.5), the predictive accuracy (Table B.6), and the coverage rates (Table B.7). We briefly describe some key results, particularly those related to X_3.

In this setting, most of the key results remain the same as before. Shared shrinkage produces better estimates for the coefficients of X_1 and X_2 and for the zero coefficients than US; again, HS is better than NG for large P but slightly worse

Table 7
MSE of real and some fake predictors for simulation study III. All values are scaled by 100.

Study III, Case A ($P = 10$)									
	Predictors	HS SS	HS US	NG SS	NG US	BL SS	BL US	NS	S&S
Class 1	X_1	18.59	15.27	17.91	13.98	19.66	18.48	21.25	13.49
	X_2	12.80	9.98	12.30	9.10	15.47	13.28	17.99	8.57
	X_3	3.96	1.69	3.39	0.73	4.87	3.91	11.69	0.06
	X_4	8.23	2.32	6.59	0.97	10.40	6.75	17.24	0.01
Class 2	X_1	10.50	3.14	10.9	1.85	9.49	6.09	14.36	0.20
	X_2	7.50	1.75	7.19	0.77	7.53	4.56	7.51	0.00
	X_3	14.11	14.36	19.08	17.45	10.74	10.46	14.16	34.67
	X_4	11.70	7.24	9.80	5.41	12.04	10.00	15.79	3.51
All fake		1.94	1.93	0.54	0.89	4.30	4.77	13.30	0.04
Study III, Case B ($P = 100$)									
Class 1	X_1	21.41	16.02	21.58	16.59	69.26	59.68	8352.55	31.92
	X_2	12.86	9.49	14.55	8.99	40.16	24.20	12605.73	7.72
	X_3	1.62	0.06	2.23	0.17	1.48	0.71	547.44	0.00
	X_4	6.80	0.03	8.57	0.31	17.43	4.36	1093.11	0.00
Class 2	X_1	10.52	0.47	9.78	0.98	7.57	2.32	150.27	0.00
	X_2	9.09	0.11	9.69	0.60	21.22	6.25	87.20	0.07
	X_3	29.00	27.9	22.42	21.41	18.72	18.03	532.24	51.25
	X_4	11.50	4.64	10.71	5.64	22.20	12.23	1236.07	5.05
All fake		0.08	0.08	0.22	0.25	0.79	0.88	389.03	0.00

for small P; BL is worse than both HS and NG. With respect of X_3, it is unsurprising to see the US methods doing better than SS for these coefficients as they can use a small δ_{31}^2 to aggressively shrink κ_{31} and a larger δ_{32}^2 to avoid shrinking κ_{32}. With a single shrinkage parameter δ_3^2, SS must split the difference in choosing a moderate value that tends to undershrink κ_{31} and overshrink κ_{32}. Consequently, there is somewhat poorer performance in the estimation of the X_3 effects. However, this estimation loss should be weighed against the improvement in MSE seen in the SS methods over US across the totality of predictors. In particular, the increase in the MSE associated with X_3 in the SS vs US models is similar or smaller than the decrease associated with the smaller κ_{12} coefficient. In particular, this conclusion is supported by the variable selection results in Table 6. While the US methods have a slightly higher selection rates κ_{23} relative to SS, their selection rates for κ_{21} are substantially lower than SS. Hence, the SS choices will select variables closer to the true set than US will.

Considering the coverage rates in Figure B.7, all methods correctly overshrink the null κ_{31} coefficient with the SS choices displaying coverage closer to the nominal rates than their corresponding US versions. For the non-zero κ_{32} all methods slightly undershrink in case A ($P = 10$) and show more substantial undercoverage in case B ($P = 100$). In terms of out-of-sample predictions (Table B.6), all methods continue to have equivalent classification accuracy, but SS yields probability estimates that better capture the true class (lower log-score) and better recover the true model probabilities (lower KL). Overall, these findings from study II indicates that the shared shrinkage approaches still perform superior to the unique shrinkage approach when there are slight deviations to our general assumption that the same set of predictors are involved in the models across all classes.

4.6. Simulation results: Study III

In Study III we now consider a setting where different pairs of predictors are included in the models for the two non-baseline classes. This represents a worst case scenario for our model. Predictors X_1 and X_2 are active for class 1, and X_3 and X_4 are active for class 2, relative to baseline class 3. Similar to studies I and II, we reported the MSE (Table 7) and variable selection (Table 8). The Appendix contains tables for bias, ESS, prediction accuracy, and coverage rates. (Tables B.8–B.11).

Generally speaking, the unique shrinkage methods produce better estimation for the coefficients of the first four predictors. They show smaller MSE than the SS methods for the active predictors (X_1 and X_2 for class 1; X_3 and X_4 for class 2), as well for the zero coefficients within these four. This is not surprising as the SS will chose moderate local parameters to try to split the difference between overshrinking for the zero coefficient in one class and undershrinking the non-zero coefficient in the other. However, despite the slightly poorer parameter estimation, we see almost no difference between the methods in terms of variable selection (Table 8). Out-of-sample predictions slightly favor US (Table B.10) although the differences between US and SS are fairly small.

We have also considered an additional simulation study where $P > n$. This comes from the same situation as in Study I using $n = 200$ and $P = 500$. The details of this simulation are contained in Appendix C. With this higher dimension setting, all methods perform less favorably than before, but HS SS performs the best in terms of variable selection and our predictive measures.

Table 8
Percentage of replicated data sets in which predictor is chosen as significant in Study III. The Average False Positives row indicates the average number of coefficients from predictors X_5 to X_P that are incorrectly chosen to be non-zero. Recall the true predictors are X_1 and X_2 in class 1 and predictors X_3 and X_4 for class 2.

		HS SS	HS US	NG SS	NG US	BL SS	BL US	NS	S&S
Study III, Case A ($P = 10$)									
Class 1	X_1(%)	100	100	99	99	100	100	100	100
	X_2(%)	100	100	100	100	100	100	100	100
	X_3(%)	0	0	0	0	2	0	0	0
	X_4(%)	2	0	1	0	7	3	5	0
Class 2	X_1(%)	4	1	5	0	4	2	6	0
	X_2(%)	5	0	5	0	8	2	5	0
	X_3(%)	50	48	25	31	67	67	77	30
	X_4(%)	97	100	96	98	100	100	100	100
Average False Positives		0.03	0.03	0.00	0.00	0.21	0.24	0.78	0.00
Study III, Case B ($P = 100$)									
Class 1	X_1(%)	98	97	98	97	98	98	100	94
	X_2(%)	100	100	100	100	100	100	100	100
	X_3(%)	0	0	0	0	0	0	33	0
	X_4(%)	1	0	3	0	18	1	38	0
Class 2	X_1(%)	3	0	4	0	5	1	24	0
	X_2(%)	7	0	9	0	32	7	22	0
	X_3(%)	13	12	16	15	30	33	81	3
	X_4(%)	95	99	96	98	38	100	99	99
Average False Positives		0.00	0.00	0.00	0.00	0.05	0.07	56.78	0.00

To conclude we make a few final comments about the main conclusions from this series of simulation studies. First, compared to the standard methods (Zens, 2019; Bhattacharyya et al., 2021) that use unique shrinkage parameters for each log-odds coefficient, the strategy of sharing the local shrinkage parameter across logistic regression model yields improved point estimation, variable selection, and out-of-sample estimation of the class probabilities when the set of predictors is the same or mostly the same across models (settings I and II). In the worst case setting (III) with different predictors in each outcome model, parameter estimation and out-of-sample class probability estimation are slightly degraded, but there is little impact on variable selection. In all settings with $P = 100$, the no shrinkage choice yields highly inaccurate and inefficient estimation, and the S&S estimates are less accurate than the shrinkage method. As noted earlier, the algorithm for S&S is substantially slower and poorer mixing, so some of this poorer performance may be due to insufficient posterior samples. However, S&S is run for an equivalent (or slightly longer) computational time, our results remain valid in the sense that we learn about the accuracy of S&S within the time required to achieve inference under Bayesian shrinkage. Between the three shrinkage options, the NG choice performs favorably for smaller dimensions with HS superior when we consider a moderate or large number of predictors. The BL choice was consistently dominated by these two models, so we do not advocate its use.

5. Alzheimer's disease classification example

We demonstrate the performance of our methodology in a real data application by using the publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) data set. The data can be accessed from http://adni.loni.usc.edu/ website. Among its many components, this data include demographics, neuropsychological testing scores from a variety of diagnostic surveys, magnetic resonance imaging (MRI) and positron emission tomography (PET) summaries, and cerebro-spinal fluid measures from patients at baseline enrollment, as well as at 6- and 12-month followup visits. As with most real observational data set, the ADNI data contains many missing values. We exclude patients who do not have information at one or more of three time points: baseline, month 06, and month 12. We consider only variables with fewer than 20% missing observations. A table of the $P = 61$ predictor variables considered in this analysis is given in the Appendix D, Table D.1. To accommodate these missing values in the predictor space, we incorporate a basic data augmentation step where we sample the missing predictors conditional on the observed predictors and the class C_i. Details about the missing data imputation can be found in Appendix E. All continuous predictor variables are standardized for computational efficiency. All binary predictors are shifted to 0.50 for 1 and -0.50 for 0.

The resulting data set used for our analysis considers $n = 562$ subjects and $P = 61$ predictor variables. The categorical response variable is diagnosis status (DX) at the 12-month visit, which has three levels: Mild Cognitive Impairment (MCI), Dementia (DM), and Cognitively Normal (CN). Of the 562 patients, 288 (51%) are classified as having MCI at the 12 month visit, 96 (17%) as having DM, and the remaining 184 (32%) are CN which we treat as the reference class. A more detailed discussion of the ADNI data can be found in Weiner et al. (2013).

Table 9
Test data prediction accuracy for the ADNI data in 10-fold cross validation.

	HS SS	HS US	NG SS	NG US	S&S	RF
Classification Accuracy	93.20	93.10	93.80	93.89	91.50	88.45
(SE)	(0.82)	(0.81)	(0.89)	(0.81)	(0.62)	(1.30)
Predictive log-score	-34.02	-37.95	-34.16	-37.20	-33.01	
(SE)	(1.29)	(1.92)	(1.17)	(1.27)	(1.40)	

Table 10
Posterior mean and credible interval for the meaningful predictors using the full data set. Bold face designates meaningful predictors (95% CI excluding zero for shrinkage methods; posterior inclusion probability > 0.5 for S&S). The remaining 57 predictors were not meaningful in either class model, so their estimates are excluded from the table.

Logit equation	Meaningful predictors	HS SS	NG SS	NG US	S&S
Class 1 (MCI) vs Class 3 (CN)	PTEDUCAT	**2.17 (1.06, 3.49)**	**2.17 (1.12, 3.46)**	**1.80 (0.90, 2.84)**	0.60 (0.00, 3.31)
	MMSE.bl	-1.19 (-2.56, 0.04)	-1.18 (-2.67, 0.34)	-0.37 (-1.83, 0.12)	0.00 (0.00, 0.00)
	LDELTOTAL.bl	**-10.32 (-14.96, -6.80)**	**-9.71(-13.98, -6.40)**	**-8.51 (-11.95, -5.81)**	**-8.89 (-15.82, -5.47)**
	FAQ.m12	**6.24 (0.16,14.62)**	4.52 (-0.26, 12.00)	0.91 (-0.13, 3.95)	**7.61 (0.00, 16.00)**
Class 2 (DM) vs Class 3 (CN)	PTEDUCAT	**2.16 (0.94, 3.56)**	**2.16 (1.01, 3.53)**	**1.99 (0.74, 2.93)**	0.61 (0.00, 3.39)
	MMSE.bl	**-3.97 (-5.65, -2.37)**	**-3.95 (-5.80, -1.82)**	**-3.31 (-5.13, -1.54)**	**-2.76 (-3.69, -2.00)**
	LDELTOTAL.bl	**-10.84 (-15.65, -7.07)**	**-10.14 (-14.60, -6.52)**	**-8.93 (-12.69, -5.83)**	**-9.27 (-16.19, -5.60)**
	FAQ.m12	**7.96 (1.67,16.35)**	**6.26 (1.12, 13.85)**	**2.75 (1.04, 5.91)**	**9.27 (1.32, 17.68)**
Class 1 (MCI) vs Class 2 (CN)	PTEDUCAT	0.00 (-0.40, 0.42)	0.00 (-0.30, 0.52)	0.01 (-0.43, 0.47)	0.00 (-0.25, 0.25)
	MMSE.bl	**2.78 (1.90, 3.75)**	**2.76 (1.64, 3.84)**	**2.93 (1.41, 4.33)**	**2.74 (2.00, 3.59)**
	LDELTOTAL.bl	0.52 (-0.65, 1.73)	0.43 (-0.86, 1.71)	0.41 (-1.09, 1.87)	**0.38 (-0.65, 1.44)**
	FAQ.m12	**-1.72 (-2.41, -1.04)**	**-1.74 (-2.48, -0.97)**	**-1.83 (-2.67, -0.97)**	**-1.65 (-2.18, -1.16)**

To verify our model's performance, we perform 10 fold cross validation (CV) with a training and test data split ratio of 80:20 at each fold. This leads to a training data with $n_{train} = 455$ out of 568 patients. Predictive accuracy is measured using the log-score as defined in equation (4) and the percent correct classification. We utilize the same models as in the simulation study, except that we exclude the NS and BL versions due to their consistently uncompetitive performance relative to other models. For each of the remaining 4 shrinkage models, we run 5 independent chains using the training data for 70,000 iterations excluding the first 10,000 samples as burn-in. As the S&S is computationally more expensive, we only run it for 35,000 iterations (dropping 5000) in each of the 5 chains. Despite running for fewer iterations, we note that the S&S chains take approximately twice as long as the shrinkage methods. We also obtain the classification accuracy using a Random Forest (RF) algorithm as an additional competitor, although we do not consider a log-score measure as there is no posterior sample to define the expectation over.

Table 9 shows prediction results. Clearly, all of the shrinkage models are highly accurate with 93–94% accuracy; S&S is slightly lower at 91.5% and the random forest shows the lowest performance at 88.5%. In terms of the predictive log-score, S&S curiously has the highest/best value, although this result may not be trustworthy due to convergences issues. The two shared shrinkage models clearly due better than the two unique shrinkage models, with the horseshoe shared shrinkage version as the best.

To better understand the difference between the models, we refit the top 4 models (HS SS, NG SS, NG US, S&S) to the full data to investigate the resulting coefficient estimates. For simplicity, we only show the coefficients corresponding to predictors that have a 95% CI that excludes zero in any of these four models. We refer to these as the meaningful predictors. Table 10 shows the posterior mean and the 95% credible interval for the 4 predictors that meet this criteria, and bold typeface used to indicate these coefficients are significantly different from zero. The meaningful predictors are total number of years of education (PTEDUCAT), Mini Mental State Ratio score on the Mini Mental State Ratio diagnostic test (MMSE) at baseline, Activities of Daily Living (FAQ) at month 12, and Logical Memory - Delayed Recall (LDELTOTAL) at baseline. In addition to the effects for class 1 (MCI) and class 2 (DM) relative to baseline (CN), we also report the estimated log-odds ratio for MCI relative to DM by consider the posterior behavior of $\kappa_{1p} - \kappa_{2p}$.

We begin by interpreting the effects in the best fitting HS SS model. We see that low baseline scores on the Logical Memory - Delayed Recall (LDELTOTAL) instrument are strongly associated with patients being in one of the two impaired categories, relative to the control CN patients. Similarly, large current scores in the Activities of Daily Living instrument (FAQ.m12) are also meaningfully associated with distinguishing impairment from cognitively normal. Surprisingly, patients with higher levels of education were also found to be associated with higher levels of impairment, conditionally on the other survey scores. Comparing the two impaired classes (MCI vs DM), we see the higher baseline scores in the Mini Mental State Ratio examination (MMSE.bl) and low current FAQ scores are associated with mild impairment as opposed to Alzheimer's.

Comparing the estimates in Table 10 across the four models, we see almost identical inference under the two shared shrinkage models. The unique shrinkage NG model also produces estimates that are mostly similar to the SS results except in a few components. NG US shows more aggressive shrinkage for baseline MMSE in class 1 and the FAQ effect for both

models. S&S does not find an education effect in the data unlike the shrinkage model, as it places a high probability (greater than 50%) on these coefficients being zero. Additionally, S&S estimates a larger role for FAQ than the other methods.

We briefly comment here about the MCMC convergence behavior for these four models, and further details and discussion is available in Appendix F. We include a table with effective sample sizes of the data likelihood and the key parameters (Table F.1), as well as trace plots for these 8 coefficients (Table F.1–F.4). In particular, we see reasonably effective mixing in the shrinkage models, although there are some concerns regarding autocorrelation for the coefficients associated with the FAQ instrument. The S&S results show major mixing problems with very different results across the five chains. Consistent with our discussion from the simulation section, the S&S model fails to produce trustworthy inference when given the same computational resources as the shrinking methods.

6. Conclusion

In this work, we have implemented Bayesian shrinkage methods in a categorical regression model by sharing local shrinkage information about each predictor's importance across classes. To the best of our knowledge, our proposal is the first to flexibly share variable importance information across all levels of the categorical outcome. Given that the $K-1$ models correspond to different levels of the same categorical outcome, it is especially critical that we share information about predictor importance, and other methodologies will have inefficient inference by failing to leverage this property. We believe that this is an important and novel consideration that has been missing in the scarce literature for categorical regression modeling, and the empirical results from the simulations and the ADNI data analysis further support this approach by showing more accurate parameter estimation and category prediction from our shared shrinkage method.

We considered the proposed method by utilizing the horseshoe, normal gamma, and Bayesian Lasso prior structures through the GL framework, although other choices can easily be incorporated as discussed in Section 2. The coverage rate shows that the unique shrinkage methods more aggressively shrink the posterior distributions than the shared shrinkage methods. In some cases, the unique shrinkage achieve 90% coverage rate even with 20% credible interval for a null effect. However, the shared shrinkage strategy was still found to overcover true zeros while typically maintaining nominal coverage of real effects, even those real effects that were only moderately large. When using the 95% credible intervals to determine significant variables in the model, we find higher selection rates in the SS approaches relative to unique shrinkage when the predictor is active in multiple outcome models, and when predictors have differential effects across classes, variable selection with SS is similar to the US results. In terms of parameter estimation accuracy (MSE), the shared shrinkage approaches performed better than the unique shrinkage approaches.

While we have considered the use of the shared shrinkage framework in the context of the baseline categorical logit model for categorical data, our approach could similarly be used for a multinomial probit case. As noted in equation (2), each observation's $(K-1)$ latent variables are associated with the predictors through a $(K-1)\times P$ regression coefficient matrix, and a similar shared sparsity framework that shares the same local parameter across the $(K-1)$ coefficients associated with predictor p. As noted previously, the MNP model can be difficult to sample due to the identifiability constraints in the latent variable covariance matrix and difficult to interpret since the coefficients act of the latent space. Consequently, we have not pursued the MNP model in this project.

Acknowledgements

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see http://www.adni-info.org.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2022.107568.

References

Agresti, A., 2015. Foundations of Linear and Generalized Linear Models. John Wiley & Sons.

Bhadra, A., Datta, J., Polson, N.G., Willard, B., 2017. The horseshoe+ estimator of ultra-sparse signals. Bayesian Anal. 12, 1105–1131.

Bhadra, A., Datta, J., Polson, N.G., Willard, B., 2019. Lasso meets horseshoe: a survey. Stat. Sci. 34, 405–427.

Bhattacharya, A., Pati, D., Pillai, N.S., Dunson, D.B., 2015. Dirichlet–Laplace priors for optimal shrinkage. J. Am. Stat. Assoc. 110, 1479–1490.

Bhattacharyya, A., Pal, S., Mitra, R., Rai, S., 2021. Applications of Bayesian shrinkage prior models in clinical research with categorical responses. https://www.researchsquare.com/article/rs-137866/v1.

Bitto, A., Frühwirth-Schnatter, S., 2019. Achieving shrinkage in a time-varying parameter model framework. J. Econom. 210, 75–97.

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

Chen, S., Walker, S.G., 2019. Fast Bayesian variable selection for high dimensional linear models: marginal solo spike and slab priors. Electron. J. Stat. 13, 284–309. https://doi.org/10.1214/18-EJS1529.

Dow, J.K., Endersby, J.W., 2004. Multinomial probit and multinomial logit: a comparison of choice models for voting research. Elect. Stud. 23, 107–122.

Gaskins, J., 2019. Hyper Markov laws for correlation matrices. Stat. Sin. 29, 165–184.

George, E.I., McCulloch, R.E., 1993. Variable selection via gibbs sampling. J. Am. Stat. Assoc. 88, 881–889. http://www.jstor.org/stable/2290777.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102, 359–378.

Griffin, J.E., Brown, P.J., 2010. Inference with normal-gamma prior distributions in regression problems. Bayesian Anal. 5, 171–188.

Imai, K., Van Dyk, D.A., 2005. A Bayesian analysis of the multinomial probit model using marginal data augmentation. J. Econom. 124, 311–334.

Kropko, J., 2008. Choosing Between Multinomial Logit and Multinomial Probit Models for Analysis of Unordered Choice Data. Ph.D. thesis. University of North Carolina, Chapel Hill.

Kundu, D., Mitra, R., Gaskins, J.T., 2021. Bayesian variable selection for multioutcome models through shared shrinkage. Scand. J. Stat. 48, 295–320.

Li, X., Zhang, S., Wu, Y., Wang, Y., Wang, W., 2021. Exploring influencing factors of intercity mode choice from view of entire travel chain. J. Adv. Transp. 2021.

Liang, F., Paulo, R., Molina, G., Clyde, M.A., Berger, J.O., 2008. Mixtures of g priors for bayesian variable selection. J. Am. Stat. Assoc. 103, 410–423. https://doi.org/10.1198/016214507000001337.

Makalic, E., Schmidt, D.F., 2015. A simple sampler for the horseshoe estimator. IEEE Signal Process. Lett. 23, 179–182.

McCulloch, R.E., Polson, N.G., Rossi, P.E., 2000. A Bayesian analysis of the multinomial probit model with fully identified parameters. J. Econom. 99, 173–193.

Mitchell, T.J., Beauchamp, J.J., 1988. Bayesian variable selection in linear regression. J. Am. Stat. Assoc. 83, 1023–1032.

Park, T., Casella, G., 2008. The bayesian lasso. J. Am. Stat. Assoc. 103, 681–686. https://doi.org/10.1198/016214508000000337.

Polson, N.G., Scott, J.G., 2010. Shrink globally, act locally: sparse Bayesian regularization and prediction. Bayesian Stat. 9, 501–538.

Polson, N.G., Scott, J.G., Windle, J., 2013. Bayesian inference for logistic models using Pólya–gamma latent variables. J. Am. Stat. Assoc. 108, 1339–1349.

Polymeropoulos, A., 2020. Bayesian Variable Selection in Multinomial Logistic Regression: a Conditional Latent Approach. Ph.D. thesis. University of Milano-Bicocca.

Posch, K., Arbeiter, M., Pilz, J., 2020. A novel bayesian approach for variable selection in linear regression models. Comput. Stat. Data Anal. 144, 106881. https://doi.org/10.1016/j.csda.2019.106881. https://www.sciencedirect.com/science/article/pii/S0167947319302361.

Ročková, V., George, E.I., 2014. Emvs: the em approach to bayesian variable selection. J. Am. Stat. Assoc. 109, 828–846. https://doi.org/10.1080/01621459.2013.869223.

Tüchler, R., 2008. Bayesian variable selection for logistic models using auxiliary mixture sampling. J. Comput. Graph. Stat. 17, 76–94.

Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Liu, E., et al., 2013. The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. Alzheimer's Dement. 9, e111–e194.

Zens, G., 2019. Bayesian shrinkage in mixture-of-experts models: identifying robust determinants of class membership. Adv. Data Anal. Classif. 13, 1019–1051.

Zhou, Z., Matteson, D.S., Woodard, D.B., Henderson, S.G., Micheas, A.C., 2015. A spatio-temporal point process model for ambulance demand. J. Am. Stat. Assoc. 110, 6–15.