Bayesian Analysis (2024) **19**, Number 3, pp. 785–814

# Group Inverse-Gamma Gamma Shrinkage for Sparse Linear Models with Block-Correlated Regressors*

Jonathan Boss[†], Jyotishka Datta[‡], Xin Wang[§], Sung Kyun Park[¶], Jian Kang[‖], and Bhramar Mukherjee[**]

**Abstract.** Heavy-tailed continuous shrinkage priors, such as the horseshoe prior, are widely used for sparse estimation problems. However, there is limited work extending these priors to explicitly incorporate multivariate shrinkage for regressors with grouping structures. Of particular interest in this article, is regression coefficient estimation where pockets of high collinearity in the regressor space are contained within known regressor groupings. To assuage variance inflation due to multicollinearity we propose the group inverse-gamma gamma (GIGG) prior, a heavy-tailed prior that can trade-off between local and group shrinkage in a data adaptive fashion. A special case of the GIGG prior is the group horseshoe prior, whose shrinkage profile is dependent within-group such that the regression coefficients marginally have exact horseshoe regularization. We establish posterior consistency and posterior concentration results for regression coefficients in linear models and mean parameters in sparse normal means models. The full conditional distributions corresponding to GIGG regression can be derived in closed form, leading to straightforward posterior computation. We show that GIGG regression results in low mean-squared error across a wide range of correlation structures and within-group signal densities via simulation. We apply GIGG regression to data from the National Health and Nutrition Examination Survey for associating environmental exposures with liver functionality.

**Keywords:** global-local shrinkage prior, grouped regressors, horseshoe prior, multicollinearity, multipollutant modeling, multivariate shrinkage.

*Dr. Datta acknowledges support from the National Science Foundation (DMS-2015460). Dr. Kang acknowledges support from the National Institutes of Health (R01 DA048993; R01 GM124061; R01 MH105561). Dr. Mukherjee acknowledges support from the National Science Foundation (DMS-1712933) and the National Institutes of Health (R01 HG008773-01).

[†]Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, bossjona@umich.edu

[‡]Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, jyotishka@vt.edu

[§]Department of Epidemiology, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, xwangsph@umich.edu

[¶]Department of Epidemiology, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, sungkyun@umich.edu

[‖]Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, jiankang@umich.edu

[**]Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, bhramar@umich.edu

 https://doi.org/10.1214/23-BA1371

# 1 Introduction

Regression with grouped regressors is a common problem in many biomedical applications. Some examples include metabolomics data, where metabolites are grouped by subpathway membership, neuroimaging data, where adjacent voxels are spatially grouped, and environmental contaminants data, where exposures are grouped by chemical structure, toxicological profile, and pharmacokinetics (see Figure 1). In such cases, leveraging relevant grouping information to construct grouped multivariate shrinkage profiles may help achieve additional variance reduction beyond comparable methods that ignore the grouping structure. The methodological focus of this article will be on grouped multivariate regularization in a continuous shrinkage prior framework.

Ever since the publication of the horseshoe prior (Carvalho et al., 2009, 2010), there has been an explosion of continuous shrinkage priors designed for sparse estimation problems, notably normal-gamma shrinkage (Brown and Griffin, 2010), generalized double

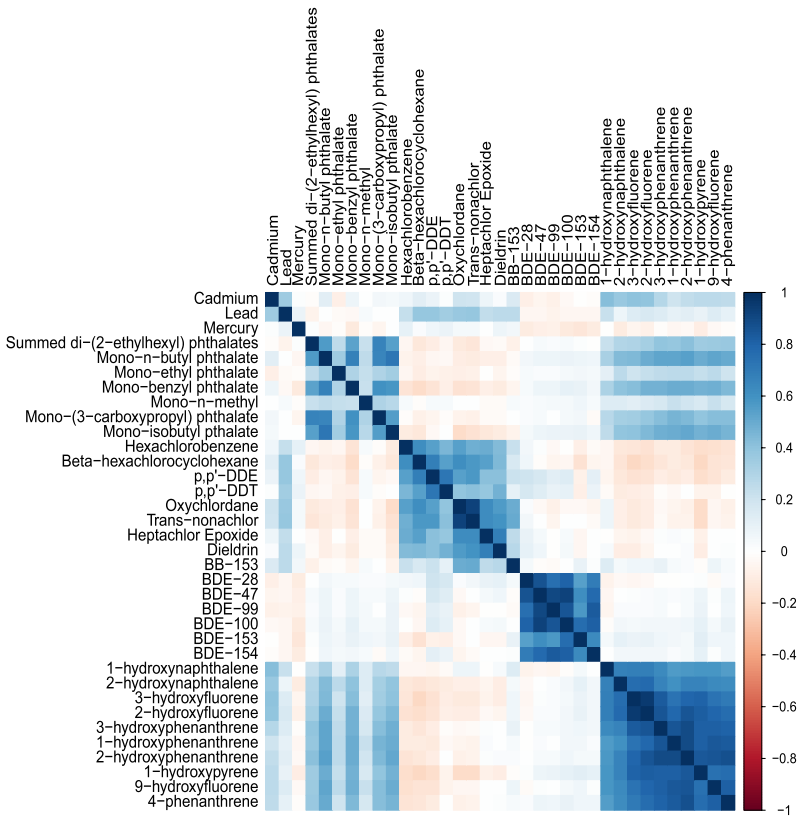


Figure 1: Pairwise Spearman correlation plot between metals, phthalates, organochlorine pesticides, polybrominated diphenyl ethers, and polycyclic aromatic hydrocarbons from the 2003-2004 National Health and Nutrition Examination Survey ($n = 990$).

Pareto shrinkage (Armagan et al., 2013a), Dirichlet–Laplace shrinkage (Bhattacharya et al., 2015), horseshoe+ shrinkage (Bhadra et al., 2017), and normal beta prime (NBP) shrinkage (Bai and Ghosh, 2019; Cadonna et al., 2020), among others. These priors have become increasingly popular for sparse regression problems because of their good theoretical and empirical properties, in addition to their scale mixture representation, which facilitates straightforward and efficient posterior simulation algorithms. The general recipe for constructing a continuous shrinkage prior with good estimation and prediction properties is substantial mass at the origin, to sufficiently shrink null coefficients towards zero, and regularly-varying tails, to avoid overregularizing non-null coefficients (Bhadra et al., 2016). Surveying the continuous shrinkage prior literature on regression with known grouping structure, there are many papers which discuss Bayesian group lasso and its applications (Kyung et al., 2010; Li et al., 2015; Xu and Ghosh, 2015; Hefley et al., 2017; Kang et al., 2019) and several papers which propose extensions to Bayesian sparse group lasso (Xu and Ghosh, 2015), Bayesian group bridge regularization (Mallick and Yi, 2017), and the Normal Exponential Gamma prior with grouping structure (Rockova and Lesaffre, 2014). Xu et al. (2016) introduced the, so called, group horseshoe prior with an emphasis on prediction in Bayesian generalized additive models. However, the group horseshoe prior does not reduce to the horseshoe prior for a group of size one, meaning that the group horseshoe prior, as proposed by Xu et al. (2016), is not a direct generalization of the horseshoe prior. Wei et al. (2020) developed a multivariate Dirichlet-Laplace prior for use in Bayesian additive models with first order interactions. Intuitively, the multivariate Dirichlet-Laplace prior can be thought of as treating the corresponding basis expansion for each regressor as a group. Lastly, although not specifically framed as a grouped multivariate shrinkage prior, Som et al. (2015) proposed a block hyper-g shrinkage prior where the blocks are defined by areas of high collinearity in the regressor space, as in our data example.

Bayesian group lasso-style shrinkage is not generally preferred as a default method for estimation problems, as the Laplacian prior has neither an infinite spike at zero nor regularly-varying tails (Polson and Scott, 2011; Castillo et al., 2015; Bhadra et al., 2016). The multivariate Dirichlet-Laplace prior and the block hyper-g prior, are group/block sparse priors and, therefore, are not designed for problems that require shrinkage at both a group-level and an individual-level. The group horseshoe prior of Xu et al. (2016) has the desired origin and tail behavior marginally, however no hyperparameter in the prior controls how dependent the shrinkage is within a group. Thus, this prior implicitly assumes that the degree of dependence induced by grouped multivariate shrinkage only depends on group size. This assumption is inadequate when we *a priori* believe that, irrespective of group size, some groups have more heterogeneous effect sizes than others. Moreover, this assumption does not avail the opportunity to learn how dependent the shrinkage should be in a data adaptive manner, which is an intrinsic feature in some application areas. For example, in modeling multiple pollutants, this is a relevant consideration as some exposure classes have more homogeneous toxicological profiles than others (Ferguson et al., 2014).

To address these limitations, we propose the group inverse-gamma gamma (GIGG) prior, which extends the horseshoe and normal beta prime (NBP) priors to incorporate grouping structures. The GIGG prior introduces a group level shrinkage parameter, in

addition to the usual global and local shrinkage parameters, such that the induced prior on the product of the group and local shrinkage parameters yields the desired marginal shrinkage profile. This allows the user to control the trade-off between group-level and individual-level shrinkage, leading to relatively low estimation error irrespective of the signal density and the degree of multicollinearity within each group. Additionally, the GIGG prior is constructed such that all parameters have closed-form full conditional distributions, implying that techniques to scale horseshoe regression to large sample sizes and high-dimensional regressor spaces are also applicable to GIGG regression (Bhattacharya et al., 2016; Terenin et al., 2019; Johndrow et al., 2020). Theoretically, we establish posterior consistency and posterior concentration results for regression coefficients with grouping structure in linear regression models and mean parameters with grouping structure in sparse normal means models with respect to several GIGG hyperparameters and correlation structures. To our knowledge, we are the first to apply existing theoretical frameworks for posterior consistency in the sparse linear regression model (Armagan et al., 2013b) and posterior concentration in the sparse normal means model (Datta and Ghosh, 2013) to a non-exchangeable prior, which will be useful for future evaluations of other non-exchangeable priors.

The structure of the paper is as follows. We start with an intuitive explanation of the GIGG prior in Section 2, succeeded by some theoretical results in Section 3. After the methodological and theoretical discussion, we outline computational details, including hyperparameter estimation via marginal maximum likelihood estimation (MMLE) (Section 4). In Section 5, we conduct a simulation study to empirically verify that the intuition and theory developed in Sections 2 and 3 hold for linear regression models with group-correlated regressors. We then apply GIGG regression to data from the 2003-2004 National Health and Nutrition Examination Survey (NHANES) to identify toxicants and metals associated with a biomarker of liver function (Section 6) and conclude with a discussion (Section 7).

# 2 Methods

Throughout the article, $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a multivariate normal distribution with mean parameter $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, $G(a, b)$ denotes a gamma distribution with shape parameter $a$ and rate parameter $b$, and $IG(a, b)$ denotes an inverse-gamma distribution with shape parameter $a$ and scale parameter $b$. Additionally, we will use $\pi(\cdot)$ as general notation for a prior probability measure and $\pi(\cdot \mid \boldsymbol{y})$ as general notation for a posterior probability measure.

## 2.1 Group Inverse-Gamma Gamma (GIGG) Prior

Consider a Bayesian sparse linear regression model

$$\begin{aligned} [\boldsymbol{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2] &\sim N\left(\boldsymbol{C}\boldsymbol{\alpha} + \sum_{g=1}^{G} \boldsymbol{X}_g\boldsymbol{\beta}_g, \sigma^2\boldsymbol{I}_n\right) \\ \pi(\boldsymbol{\alpha}) \propto 1, \;\; [\boldsymbol{\beta} \mid \sigma^2] &\sim \pi(\boldsymbol{\beta} \mid \sigma^2), \;\; \pi(\sigma^2) \propto \sigma^{-2}, \end{aligned} \tag{2.1}$$

where $g = 1, \ldots, G$ indexes the groups, $\boldsymbol{y}$ is an $n \times 1$ vector of centered continuous responses, $\boldsymbol{C}$ is a matrix of adjustment covariates, $\boldsymbol{X}_g$ is an $n \times p_g$ matrix of standardized regressors in the $g$-th group, $\boldsymbol{\beta}_g = (\beta_{g1}, \ldots, \beta_{gp_g})^\top$ is a $p_g \times 1$ vector of regression coefficients corresponding to the $g$-th group, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_G^\top)^\top$ is a $p \times 1$ vector of regression coefficients to employ shrinkage on, and $\boldsymbol{I}_n$ is an $n \times n$ identity matrix. We assume the model is sparse in the sense that many of the entries in $\boldsymbol{\beta}$ are zero. The group inverse-gamma gamma (GIGG) prior is defined as

$$[\beta_{gj}|\tau^2, \gamma_g^2, \lambda_{gj}^2] \sim N(0, \tau^2\gamma_g^2\lambda_{gj}^2)$$

$$[\gamma_g^2|a_g] \sim G(a_g, 1), \quad [\lambda_{gj}^2|b_g] \sim IG(b_g, 1), \quad [\tau^2, \sigma^2] \sim \pi(\tau^2, \sigma^2),$$

where $j = 1, \ldots, p_g$ indexes the regressors within the $g$-th group. In this paper, we will assign $\tau \mid \sigma \sim C^+(0, \sigma)$ and $\pi(\sigma^2) \propto \sigma^{-2}$, where $C^+(0, \sigma)$ is a half-Cauchy distribution (Polson and Scott, 2011). Alternatively, we may also express the prior on $\boldsymbol{\beta}$ as a vector, $[\boldsymbol{\beta}|\tau^2, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}] \sim N(0, \tau^2\boldsymbol{\Gamma}\boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda_{11}^2, \ldots, \lambda_{Gp_G}^2)$ and $\boldsymbol{\Gamma} = \text{diag}(\gamma_1^2, \ldots, \gamma_1^2, \gamma_2^2, \ldots, \gamma_2^2, \ldots, \gamma_G^2, \ldots, \gamma_G^2)$ such that $\gamma_g^2$ is repeated $p_g$ times along the diagonal of $\boldsymbol{\Gamma}$. In the GIGG prior specification, the priors on the group shrinkage parameter, $\gamma_g^2$, and local shrinkage parameter, $\lambda_{gj}^2$, are selected such that the induced prior on the product is a beta prime prior, $\gamma_g^2\lambda_{gj}^2 \sim \beta'(a_g, b_g)$ (see Boss et al. (2023) for distributional definitions). Since the group shrinkage parameter is shared by all $p_g$ observations in the $g$-th group, assigning a beta prime prior on the product ensures normal beta prime shrinkage marginally while the shrinkage is dependent within-group. One point that deserves further clarification is the assignment of the gamma and inverse-gamma priors to the group and local shrinkage parameters, respectively, when either configuration would yield a beta prime prior in the product. The rationale behind this choice is that the inverse-gamma prior is heavier-tailed than the gamma prior, thereby preventing overregularization of large, non-null coefficients due to being grouped with null coefficients.

Setting $a_g = b_g = 1/2$ for all $g$ yields a special case of the GIGG prior which we will call the group horseshoe prior

$$[\beta_{gj}|\tau^2, \gamma_g^2, \lambda_{gj}^2] \sim N(0, \tau^2\gamma_g^2\lambda_{gj}^2)$$

$$\gamma_g^2 \sim G(1/2, 1), \quad \lambda_{gj}^2 \sim IG(1/2, 1), \quad [\tau^2, \sigma^2] \sim \pi(\tau^2, \sigma^2).$$

For a group horseshoe prior with a group of size one, the group shrinkage parameter becomes a local shrinkage parameter. That is, for a group $g'$ of size one,

$$[\beta_{g'1}|\tau^2, \gamma_{g'}^2, \lambda_{g'1}^2] \sim N(0, \tau^2\gamma_{g'}^2\lambda_{g'1}^2)$$

$$\gamma_{g'}^2 \sim G(1/2, 1), \quad \lambda_{g'1}^2 \sim IG(1/2, 1), \quad [\tau^2, \sigma^2] \sim \pi(\tau^2, \sigma^2)$$

can be re-indexed as

$$[\beta_{g'1}|\tau^2, \gamma_{g'1}^2, \lambda_{g'1}^2] \sim N(0, \tau^2\gamma_{g'1}^2\lambda_{g'1}^2)$$

$$\gamma_{g'1}^2 \sim G(1/2, 1), \quad \lambda_{g'1}^2 \sim IG(1/2, 1), \quad [\tau^2, \sigma^2] \sim \pi(\tau^2, \sigma^2),$$

which is equivalent to the horseshoe prior

$$[\beta_{g'1}|\tau^2, \eta^2_{g'1}] \sim N(0, \tau^2\eta^2_{g'1}), \quad \eta_{g'1} \sim C^+(0,1), \quad [\tau^2, \sigma^2] \sim \pi(\tau^2, \sigma^2).$$

It is important to note that this is different from the group horseshoe prior specification described in Xu et al. (2016). The prior in Xu et al. (2016) assigns independent $\beta'(1/2, 1/2)$ priors on both the group and local shrinkage parameters, meaning that the implied prior on the product of the group and local shrinkage parameters is the product of two independent $\beta'(1/2, 1/2)$ random variables. In our construction, the product of the group and local shrinkage parameters is itself $\beta'(1/2, 1/2)$. Consequently, our group horseshoe prior specification has horseshoe regularization marginally, while the group horseshoe prior in Xu et al. (2016) does not. To more clearly distinguish between the two group horseshoe priors, we will refer to the prior in Xu et al. (2016) as the group horseshoe+ prior and the GIGG prior with $a_g = b_g = 1/2$ as the group horseshoe prior for the remainder of the paper.

## 2.2 Marginal Prior Properties

When discussing a proposed shrinkage prior on $\boldsymbol{\beta}$, there are two key features of the marginal prior that need to be investigated. The first is the behavior in a tight neighborhood around zero and the second is the rate at which the prior decays in the extremes. For $\tau^2 = 1$ fixed, Bai and Ghosh (2019) showed that the marginal prior $\pi(\beta_{gj} \mid \tau^2, a_g, b_g)$ has a pole at 0 if and only if $0 < a_g \leq 1/2$, with the pole at zero becoming stronger the closer $a_g$ is to zero. Therefore, one should select $a_g \in (0, 1/2]$ for sparse estimation problems to sufficiently shrink null coefficients towards zero. To clarify the tail behavior we need to introduce the notion of a regularly varying function (Bingham et al., 1989): A positive, measurable function $f$ is said to be regularly varying at $\infty$ with index $\omega \in \mathbb{R}$ if $\lim_{x\to\infty} f(tx)/f(x) = t^\omega$, for all $t > 0$.

**Theorem 2.1.** *Let $\mathcal{B}(a_g, b_g)$ denote the beta function evaluated at $a_g$ and $b_g$ and $\Gamma(b_g + 1/2)$ denote the gamma function evaluated at $b_g + 1/2$. The tails of the marginal prior probability density function of $\beta_{gj}$ decay at the following rate,*

$$\lim_{\beta_{gj}\to\infty} \frac{\pi(\beta_{gj} \mid \tau^2, a_g, b_g)}{r(\beta_{gj}, \tau^2, a_g, b_g)} = 1,$$

$$r(\beta_{gj}, \tau^2, a_g, b_g) = \frac{(2\tau^2)^{b_g}\Gamma(b_g + 1/2)}{\sqrt{\pi}\mathcal{B}(a_g, b_g)}|\beta_{gj}|^{-(1+2b_g)}\left(\frac{\beta^2_{gj}/\tau^2}{1 + \beta^2_{gj}/\tau^2}\right)^{a_g}.$$

*Consequently, the index of regular variation is $\omega = -1 - 2b_g$.*

*Proof.* See the Supplementary Material (Boss et al., 2023).

The concept of regular variation has been extensively discussed in the context of Bayesian robustness and noninformative inference (Dawid, 1973; O'Hagan, 1979; Andrade and O'Hagan, 2006), with the latter being recently elaborated on in the context of global-local shrinkage priors (Bhadra et al., 2016). When the index $\omega < 0$, regular

variation essentially states that the tail of the function decays at a polynomial rate and is therefore considered heavy-tailed. Some examples of priors with regularly varying tails include the student's t prior and the horseshoe prior. Conversely, commonly used priors such as the normal prior and the Laplace prior do not have regularly-varying tails. As a consequence of having exponentially decaying tails, Bayesian linear regression with independent normal priors and Bayesian lasso are prone to overregularizing large signals and are not flexible enough to facilitate conflict resolution between discordant likelihood and prior information (Andrade and O'Hagan, 2006; Polson and Scott, 2011). Theorem 2.1 shows that for any pair of hyperparameters $a_g$ and $b_g$, the marginal GIGG prior has regularly varying tails and, furthermore, that $b_g$ controls the rate at which the tails decay.

## 2.3 Connection to Bayesian LASSO

As pointed out by a reviewer, an interesting connection between the GIGG prior and Bayesian LASSO-type priors can be seen from integrating out the group shrinkage parameter

$$\pi\big(\boldsymbol{\beta}_g \mid \tau^2, \boldsymbol{\lambda}_g^2, a_g\big) = \int_0^\infty \pi\big(\boldsymbol{\beta}_g \mid \tau^2, \gamma_g^2, \boldsymbol{\lambda}_g^2\big)\pi(\gamma_g^2 \mid a_g) d\gamma_g^2$$

$$= \frac{2}{\Gamma(a_g)(2\pi)^{p_g/2}|\tau^2\boldsymbol{\Lambda}_g|^{1/2}} \left(\sqrt{\frac{1}{2\tau^2}\boldsymbol{\beta}_g\boldsymbol{\Lambda}_g^{-1}\boldsymbol{\beta}_g}\right)^{a_g - p_g/2} K_{a_g - p_g/2}\left(\sqrt{\frac{2}{\tau^2}\boldsymbol{\beta}_g\boldsymbol{\Lambda}_g^{-1}\boldsymbol{\beta}_g}\right),$$

where $\boldsymbol{\Lambda}_g = \text{diag}(\boldsymbol{\lambda}_g^2) = \text{diag}(\lambda_{g1}^2, \ldots, \lambda_{gp_g}^2)$ and $K_\zeta(\cdot)$ denotes the modified Bessel function of the second kind with parameter $\zeta$. If $a_g = 1$, then we see that $[\boldsymbol{\beta}_g \mid \tau^2, \boldsymbol{\lambda}_g^2] \sim \mathcal{ML}(\mathbf{0}, \tau^2\boldsymbol{\Lambda}_g)$, has a multivariate-Laplace prior with location parameter $\mathbf{0}$ and scale parameter $\tau^2\boldsymbol{\Lambda}_g$. Recall that for the multivariate Laplace distribution, a diagonal scale does not correspond to independence. Therefore, when $a_g = 1$, we can interpret the GIGG prior as a mixture of multivariate-Laplace priors with the mixing distribution equal to independent inverse-gamma distributions for each $\lambda_{gj}^2$. Moreover, mixing over the local shrinkage parameters implies that the GIGG prior with $a_g = 1$ is a heavy-tailed version of the multivariate-Laplace prior. To connect this result with Bayesian LASSO, we use identity 10.2.17 in Abramowitz and Stegun (1964) and conclude that if $a_g = 1$ and $p_g = 1$, then

$$\pi(\beta_{g1} \mid \tau^2, \lambda_{g1}^2) = \frac{1}{\sqrt{2\tau^2\lambda_{g1}^2}} \exp\left(-\sqrt{\frac{2}{\tau^2\lambda_{g1}^2}}|\beta_{g1}|\right), \quad \beta_{g1} \in (-\infty, \infty).$$

That is, for a group of size one with $a_g = 1$, the GIGG prior can be interpreted as a mixture of Laplace priors, explicitly connecting the GIGG prior with Bayesian LASSO.

## 2.4 Sparse Normal Means

To further elucidate the shrinkage profile of the GIGG prior, we will focus on a special case of the sparse linear regression model called the sparse normal means model ($\boldsymbol{X} =$

$\boldsymbol{I}_n$ and $\boldsymbol{C}$ empty). In the global-local shrinkage prior literature, it is conventional to work with the sparse normal means problem for analytical tractability, even when the ultimate goal is regression (Rockova and Lesaffre, 2014; Bhattacharya et al., 2015), as the posterior mean has a convenient representation, $E[\beta_{gj} \mid y_{gj}, \tau^2, \sigma^2] = (1 - E[\kappa_{gj} \mid y_{gj}, \tau^2, \sigma^2])y_{gj}$. Here, $\kappa_{gj} = \sigma^2/(\sigma^2 + \tau^2\gamma_g^2\lambda_{gj}^2)$ is called a shrinkage factor, because it quantifies how much the posterior mean is shrunk relative to the maximum likelihood estimator $y_{gj}$. Calculating the joint prior distribution for the shrinkage factors in the $g$-th group, $\boldsymbol{\kappa}_g = (\kappa_{g1}, \ldots, \kappa_{gp_g})^\top$, we have

$$\pi\big(\boldsymbol{\kappa}_g \mid \tau^2, \sigma^2, a_g, b_g\big) =$$

$$\frac{\Gamma(a_g + p_g b_g)}{\Gamma(a_g)\big(\Gamma(b_g)\big)^{p_g}} \left(\frac{\tau^2}{\sigma^2}\right)^{p_g b_g} \left(1 + \frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g} \frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g + p_g b_g)} \left(\prod_{j=1}^{p_g} \kappa_{gj}^{b_g - 1}(1-\kappa_{gj})^{-(b_g+1)}\right),$$

where $0 < \kappa_{gj} < 1$ for all $1 \leq j \leq p_g$. Evaluating the prior distribution of $\boldsymbol{\kappa}_g$, we see that the joint density multiplicatively factorizes into "dependent" and "independent" parts where the degree of dependence is governed by the $\sum_{j=1}^{p_g} \kappa_{gj}/(1-\kappa_{gj})$ term. That is, as $a_g + p_g b_g$ goes to zero, the regularization is highly individualistic, whereas if $a_g + p_g b_g$ moves away from zero, then the shrinkage becomes more dependent within the $g$-th group.

Although the dependence between the shrinkage factors in the $g$-th group is controlled by $a_g + p_g b_g$, we can use the marginal prior properties to better understand the primary roles of $a_g$ and $b_g$. From Section 2.2, we know that $a_g \in (0, 1/2]$ should be used for sparse estimation problems, because the pole at the origin of the marginal prior on $\beta_{gj}$ only arises if $a_g \in (0, 1/2]$. Since $a_g$ is heavily restricted in the range of values it can take for sparse estimation problems, then $a_g + p_g b_g$ is primarily determined by the choice of $b_g$. Even setting the restriction on $a_g$ for sparse estimation problems aside, if we interpret $a_g + p_g b_g$ as a weighted sum of hyperparameters, $b_g$ is given more weight than $a_g$ for groups larger than size one, with the weights becoming increasingly disproportionate as group size increases. Therefore, upon simultaneous inspection of the joint prior on the shrinkage factors and the marginal prior properties for the prior on $\beta_{gj}$, $b_g$ offers more control over the dependence of the multivariate shrinkage and $a_g$ offers more control over the strength of the approximate thresholding effect near zero, although these roles are not mutually exclusive. To illustrate this point, Figure 2 visualizes the marginal posterior mean of $\beta_{g1}$ for a group of size two as a function of $a_g$, $b_g$, $y_{g1}$, and $y_{g2}$. When $a_g$ and $b_g$ are close to zero then the thresholding effect on the marginal posterior mean of $\beta_{g1}$ hardly depends on the value of $y_{g2}$, indicating highly individualistic shrinkage. This corroborates our intuition from looking at the joint posterior distribution of the shrinkage weights within the same group. The second major observation is that as $b_g$ moves away from zero, the marginal posterior mean of $\beta_{g1}$ becomes increasingly more dependent on the value of $y_{g2}$. In particular, if we look at the case when $a_g = 0.05$ and $b_g = 2$, we see that when $y_{g2} = 0$ the thresholding effect on $\beta_{g1}$ is much stronger when compared to $y_{g2} = 10$. The last major observation is that as $a_g$ moves towards zero, the thresholding effect becomes stronger, which coincides with a stronger pole at zero in the marginal prior on $\beta_{g1}$.

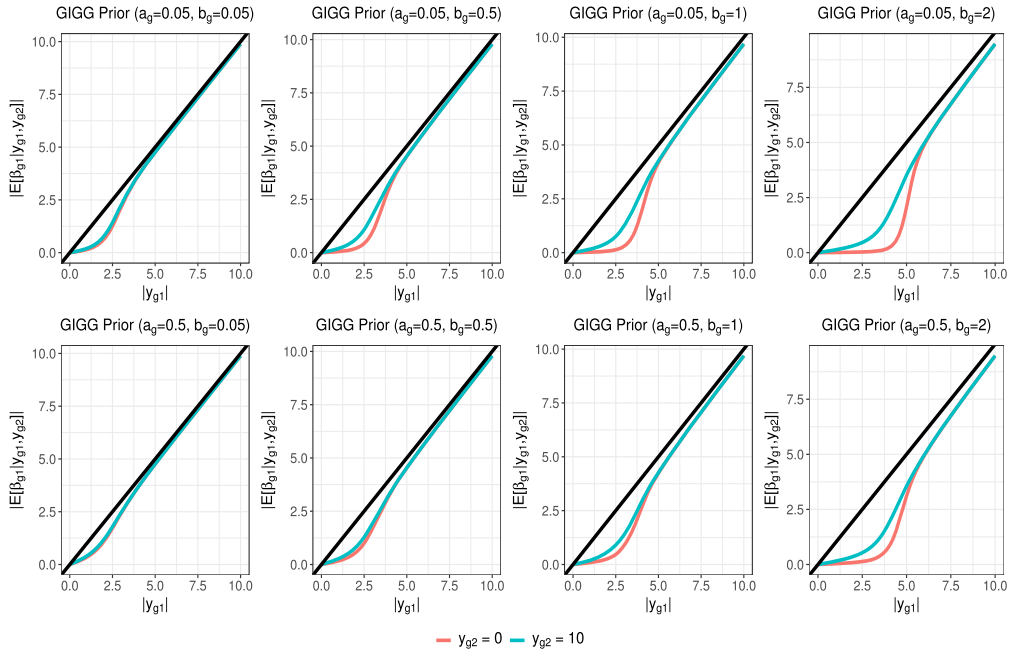


Figure 2: Marginal posterior mean of $\beta_{g1}$ for a group with two observations as $a_g$, $b_g$, $y_{g1}$, and $y_{g2}$ vary. Here, $\tau^2 = 0.2$ and $\sigma^2 = 1$ are fixed.

# 3 Theoretical Properties

In this section, we first prove posterior consistency (Section 3.1) and we then consider posterior concentration properties of GIGG shrinkage across a range of different settings (Section 3.2).

## 3.1 Posterior Consistency

Let $\boldsymbol{X}_n = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{G_n}]$ and $\mathcal{H}_n = \{\boldsymbol{a}, \boldsymbol{b}\}$ denote the collection of hyperparameters where $\boldsymbol{a} = \{a_1, \ldots, a_{G_n}\}$ and $\boldsymbol{b} = \{b_1, \ldots, b_{G_n}\}$. Here, the subscript $n$ in $G_n$ refers to the fact that the number of groups in the regressor space is growing as a function of the sample size. Furthermore, let $\mathcal{A}_n = \{(g, j) : \beta^0_{gj} \neq 0\}$ denote the true active set with cardinality $|\mathcal{A}_n|$. Then, Theorem 3.1 states that the posterior distribution of $\boldsymbol{\beta}_n$ under the GIGG prior is consistent a posteriori for the true $\boldsymbol{\beta}^0_n$. Similarly, we add a subscript $n$ to $\boldsymbol{\beta}^0_n$ and $\boldsymbol{\beta}_n$ to indicate that the total number of regressors, $p_n$, is growing as function of sample size. In the theoretical analysis of our method, letting the number of regressors grow as a function of sample size allows us to consider cases where the number of variables included in the model grows with increasing sample size, in addition to cases where the number of variables does not change as a function of sample size (Ghosal, 1999; Armagan et al., 2013b).

**Theorem 3.1.** *Suppose that* $p_n = o(n)$, $L_n = \sup_{(g,j)} |\beta^0_{gj}| < \infty$, *where* $\beta^0_{gj}$ *indicates the true* $j$*-th regression coefficient in the* $g$*-th group,* $0 < \lim_{n\to\infty} \inf \mathcal{H}_n \leq \lim_{n\to\infty} \sup \mathcal{H}_n < \infty$, *and* $|\mathcal{A}_n| = o(n/\log(n))$. *Further, suppose that the smallest and largest singular values of* $\boldsymbol{X}_n$, *denoted by* $\theta_{n,min}(\boldsymbol{X}_n)$ *and* $\theta_{n,max}(\boldsymbol{X}_n)$, *satisfy* $0 < \liminf_{n\to\infty} \theta_{n,min}(\boldsymbol{X}_n)/\sqrt{n} \leq \limsup_{n\to\infty} \theta_{n,max}(\boldsymbol{X}_n)/\sqrt{n} < \infty$. *Then for any* $\epsilon > 0$,

$$\pi_n(\boldsymbol{\beta}_n : \|\boldsymbol{\beta}_n - \boldsymbol{\beta}^0_n\|_2 < \epsilon \mid \boldsymbol{y}_n, \mathcal{H}_n, \tau^2_n, \sigma^2) \to 1$$

*almost surely as* $n \to \infty$ *provided that* $\tau^2_n = C/(p_n n^\rho \log(n))$ *for some* $\rho, C \in (0, \infty)$.

*Proof.* See the Supplementary Material (Boss et al., 2023).

Of note, the only restriction placed on the values of the hyperparameters in Theorem 3.1 is that they do not converge to the boundary of the hyperparameter space as $n \to \infty$.

**Remark 3.1.** *Theorem 3.1 is a generalization of Theorem 5 in Armagan et al. (2013b) which proved posterior consistency for the NBP prior when* $b_g \in (1, \infty)$. *Restricting* $b_g \in (1, \infty)$ *was done to utilize an argument which required the existence of the second moment of* $\beta_{gj}$, *but does not cover special cases of particular interest such as the horseshoe prior. Therefore, our result extends the existing posterior consistency result from Armagan et al. (2013b) to a more general collection of hyperparameter values with potential grouping structure.*

**Remark 3.2.** *Although Song and Liang (2017) provide an existing theoretical framework for posterior consistency in high-dimensional linear regression when* $\log(p_n) = o(n)$, *this result cannot be directly applied because the GIGG prior is non-exchangeable.*

## 3.2 Concentration Properties of Shrinkage Parameters

In this subsection, we consider posterior concentration properties corresponding to GIGG shrinkage in different settings, which describe the behavior of the posterior distribution for fixed $n$. These concentration properties are important to show for new group global-local shrinkage priors, as Datta and Ghosh (2013) showed that such concentration properties were important for the horseshoe prior. We will consider results for general low-dimensional linear regression models when possible, however, for certain componentwise results, we need to focus on the sparse normal means setting. Separate subsection headers are available to distinguish between the results for linear regression models and the results that are only applicable to sparse normal means models.

### Linear Regression

First, we partially extend the posterior concentration theoretical framework for the sparse normal means model to a low-dimensional linear regression ($p < n$) model with general correlation structure. Going forward, we will drop the subscript $n$ from the notation introduced in the statement of Theorem 3.1 to clarify that the subsequent theoretical results hold for fixed $p$.

**Theorem 3.2.** *Fix $\epsilon \in (0,1)$, $p$, and $n$, such that $p < n$. Further, suppose that the smallest and largest singular values of $\boldsymbol{X}^\top\boldsymbol{X}$, denoted by $\theta_{min}(\boldsymbol{X}^\top\boldsymbol{X})$ and $\theta_{max}(\boldsymbol{X}^\top\boldsymbol{X})$, satisfy $0 < \theta_{min}(\boldsymbol{X}^\top\boldsymbol{X}) \leq \theta_{max}(\boldsymbol{X}^\top\boldsymbol{X}) < \infty$. The full conditional posterior mean corresponding to the GIGG prior is,*

$$E[\boldsymbol{\beta} \mid \boldsymbol{y}, \sigma^2, \tau^2, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}] = \left(\boldsymbol{I}_p + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\frac{\sigma^2}{\tau^2}(\boldsymbol{\Gamma}\boldsymbol{\Lambda})^{-1}\right)^{-1}\hat{\boldsymbol{\beta}}^{OLS}, \quad \hat{\boldsymbol{\beta}}^{OLS} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}.$$

*Then the inequality,*

$$\left\|\hat{\boldsymbol{\beta}}^{OLS} - E[\boldsymbol{\beta} \mid \boldsymbol{y}, \sigma^2, \tau^2, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}]\right\|_2 \geq \left(\frac{1}{1 + \theta_{max}(\boldsymbol{X}^\top\boldsymbol{X})\sigma^{-2}\tau^2 \max_{(g,j)} \gamma_g^2\lambda_{gj}^2}\right)\left\|\hat{\boldsymbol{\beta}}^{OLS}\right\|_2,$$

*holds and we have the following results:*

*a)*

$$\pi\left(\frac{1}{1 + \theta_{max}(\boldsymbol{X}^\top\boldsymbol{X})\sigma^{-2}\tau^2 \max_{(g,j)} \gamma_g^2\lambda_{gj}^2} \geq \epsilon \;\middle|\; \boldsymbol{y}, \mathcal{H}, \tau^2, \sigma^2\right) \to 1 \quad \text{as } \tau^2 \to 0.$$

*b)*

$$\pi\left(\left\|\hat{\boldsymbol{\beta}}^{OLS} - E[\boldsymbol{\beta} \mid \boldsymbol{y}, \tau^2, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \sigma^2]\right\|_2 \geq \epsilon\left\|\hat{\boldsymbol{\beta}}^{OLS}\right\|_2 \;\middle|\; \boldsymbol{y}, \mathcal{H}, \tau^2, \sigma^2\right) \to 1 \quad \text{as } \tau^2 \to 0.$$

*Proof.* See the Supplementary Material (Boss et al., 2023).

Theorem 3.2 states that, irrespective of the correlation structure, $\tau^2 \to 0$ sufficiently shrinks the posterior mean towards zero. The argument used in the proof of Theorem 3.2 can be applied to a litany of other continuous shrinkage priors for which existing posterior concentration results are limited to the sparse normal means model. To supplement these results, we consider the case where we have block diagonal correlation structure, with the blocks defined by the groups, as in Figure 1.

**Corollary 3.1.** *Suppose that the regressors in $\boldsymbol{X}$ satisfy $\boldsymbol{X}_g^\top\boldsymbol{X}_{g'} = \boldsymbol{0}$ for all $g \neq g'$, where $\boldsymbol{0}$ denotes a $p_g \times p_{g'}$ matrix of zeros. If $\tau^2$, $\sigma^2$, and $a_g \in (0,1)$ are fixed, then there exists a constant*

$$\epsilon_g(\tau^2, \sigma^2) = \frac{\sigma^2}{\sigma^2 + \theta_{max}(\boldsymbol{X}_g^\top\boldsymbol{X}_g)\tau^2},$$

*such that for all $\delta \in (0, \epsilon_g(\tau^2, \sigma^2))$*

$$\pi\left(\left\|\hat{\boldsymbol{\beta}}_g^{OLS} - E[\boldsymbol{\beta}_g \mid \boldsymbol{y}, \tau^2, \gamma_g^2, \lambda_{g1}^2, \ldots, \lambda_{gp_g}^2, \sigma^2]\right\|_2 \geq \delta\left\|\hat{\boldsymbol{\beta}}_g^{OLS}\right\|_2 \;\middle|\; \boldsymbol{y}, a_g, b_g, \tau^2, \sigma^2\right) \to 1$$

*as $b_g \to \infty$.*

*Proof.* See the Supplementary Material (Boss et al., 2023).

The conclusion of Corollary 3.1 is that if the hyperparameter $b_g \to \infty$ then there is at least some amount of shrinkage relative to the ordinary least squares estimator in the $g$-th group. If $\tau^2/\sigma^2$ is close to zero, then $\epsilon(\tau^2, \sigma^2) \approx 1$, implying shrinkage of the posterior mean towards zero. Therefore, we can interpret the case when $b_g \to \infty$ and $\tau^2/\sigma^2$ close to zero as shrinkage of the entire $g$-th group towards zero.

### Sparse Normal Means

Although we would ideally consider additional posterior concentration results within the context of a linear regression model, there is not an analytically tractable analog of componentwise shrinkage factors for a general design matrix without any orthogonality. Therefore, we will proceed by considering posterior concentration results within the sparse normal means framework, to make precise statements regarding componentwise shrinkage, as opposed to shrinkage of the entire $L_2$-norm.

One question that arises is whether the dependence induced between the $\beta_{gj}$'s by $\gamma_g^2$ will overly dominate the individual-level shrinkage. As an example, one can conceptualize a case where a group has only one signal, which is overly shrunk by virtue of being grouped with an overwhelming majority of null means. An alternative situation that could occur is a case where few null means are grouped with many non-null means, leading to insufficient shrinkage of the null means toward zero. These two scenarios are described by Som et al. (2016) as the *Conditional Lindley's Paradox* and *Essentially Least Squares Estimation*, respectively. Theorem 3.3a states that if the $gl$-th observation is sufficiently large then there will be minimal shrinkage on $y_{gl}$. This guarantees that group shrinkage will not overly dominate individual shrinkage if the observation is large. Conversely, Theorem 3.3b states that if the global shrinkage parameter converges to zero, then the GIGG prior will sufficiently shrink the $y_{gl}$'s toward zero. Let $\boldsymbol{y}_g = (y_{g1}, \ldots, y_{gp_g})^\top$.

**Theorem 3.3.** *Suppose that* $p_g \in \{2, 3, 4, \ldots\}$.

*a) Fix* $\psi, \delta \in (0, 1)$. *Then there exists a function* $h(p_g, \tau^2, \sigma^2, a_g, b_g, \psi, \delta)$ *such that*

$$\pi(\kappa_{gl} > \psi \mid \boldsymbol{y}_g, \tau^2, \sigma^2, a_g, b_g) \leq \exp\left(-\frac{\psi(1-\delta)}{2\sigma^2} y_{gl}^2 + \frac{\psi\delta}{2\sigma^2} \sum_{j \neq l} y_{gj}^2\right) h(p_g, \tau^2, \sigma^2, a_g, b_g, \psi, \delta).$$

*Consequently, if* $|y_{gl}| \to \infty$, *then* $\pi(\kappa_{gl} \leq \psi \mid \boldsymbol{y}_g, \tau^2, \sigma^2, a_g, b_g) \to 1$.

*b) Fix* $\epsilon \in (0, 1)$. *Then there exists a function* $h(p_g, \sigma^2, \boldsymbol{y}_g, a_g, b_g, \epsilon)$ *such that,*

$$\pi(\kappa_{gl} < \epsilon \mid \boldsymbol{y}_g, \tau^2, \sigma^2, a_g, b_g) \leq \left(\frac{\tau^2}{\sigma^2}\right)^{p_g/2+b_g} \left(\min\left(1, \frac{\tau^2}{\sigma^2}\right)\right)^{-p_g/2} h(p_g, \sigma^2, \boldsymbol{y}_g, a_g, b_g, \epsilon).$$

*Consequently,* $\pi(\kappa_{gl} \geq \epsilon \mid \boldsymbol{y}_g, \tau^2, \sigma^2, a_g, b_g) \to 1$ *as* $\tau^2 \to 0$.

*Proof.* See the Supplementary Material (Boss et al., 2023).

The theoretical statements outlined in Theorem 3.3 were originally discussed for the horseshoe prior (Datta and Ghosh, 2013), but have also been used in the context of several other continuous shrinkage priors (Datta and Dunson, 2016; Bhadra et al., 2017; Bai and Ghosh, 2019), dynamic trend filtering (Kowal et al., 2019), and small area estimation (Tang et al., 2018). We also note that Theorem 3.3 does not restrict the range of values $a_g$ and $b_g$ can take, meaning that Theorem 3.3 applies to a more general class of hyperparameter values than those considered in Bai and Ghosh (2019).

# 4 Computation

## 4.1 Gibbs Sampler

The full conditional updates corresponding to model (2.1), where $\boldsymbol{\beta}$ is endowed with a GIGG prior, are enumerated in the Supplementary Material (Boss et al., 2023). Following Polson and Scott (2011), we assign a half-Cauchy prior scaled by the residual error standard deviation $\tau \mid \sigma \sim C^+(0, \sigma)$ and use a prevalent data augmentation trick,

$$[\tau^2 \mid \nu] \sim IG(1/2, 1/\nu), \quad [\nu \mid \sigma^2] \sim IG(1/2, 1/\sigma^2),$$

to obtain closed form full conditional updates for $\tau^2$ and $\sigma^2$ (Makalic and Schmidt, 2016). There are two major computational bottlenecks for the proposed algorithm. The first is the full conditional update of $\boldsymbol{\beta}$,

$$[\boldsymbol{\beta} \mid \cdot] \sim N\left(\boldsymbol{Q}^{-1}\frac{1}{\sigma^2}\boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\alpha}\right), \boldsymbol{Q}^{-1}\right), \quad \boldsymbol{Q} = \frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{X} + \frac{1}{\tau^2}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda}^{-1}.$$

The second occurs when there are a multitude of group and local parameters that need to be drawn at each iteration of the Gibbs sampler, which is often the case in "large $p$" scenarios. Rather than naïvely sampling from the full conditional distributions there are several strategies to achieve faster posterior computation:

- Draw $\boldsymbol{v} \sim N\big(\sigma^{-2}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\alpha}), \boldsymbol{Q}\big)$, and then solve $\boldsymbol{Q}\boldsymbol{\beta} = \boldsymbol{v}$, rather than explicitly calculating $\boldsymbol{Q}^{-1}$.
- For "small $n$, large $p$" problems, the Woodbury identity can be utilized so that the full conditional update of $\boldsymbol{\beta}$ scales linearly in $p$ (Bhattacharya et al., 2016).
- If $n$ and $p$ are both large, say an order of magnitude of 10,000 each, there are several recently developed approximation approaches, the former of which exploits the ability of the horseshoe prior to shrink $\tau^2\lambda_{gj}^2$ close to zero (Johndrow et al., 2020) while the latter uses a conjugate gradient algorithm to find an approximate solution to $\boldsymbol{Q}\boldsymbol{\beta} = \boldsymbol{v}$ (Nishimura and Suchard, 2022).
- Parallelization can be used within the Gibbs sampler to simultaneously update the shrinkage parameters corresponding to each group (Terenin et al., 2019).

## 4.2 Hyperparameter Selection

If the modeler wants to remain relatively agnostic to the choice of hyperparameters, one can use Marginal Maximum Likelihood Estimation (MMLE) (Casella, 2001), an empirical-Bayes approach executed iteratively within the Gibbs sampler. The $(l+1)$th update is

$$a_g^{(l+1)} = \psi_0^{-1}\Big(E_{a_g^{(l)}}\big[\log(\gamma_g^2) \mid \boldsymbol{y}\big]\Big), \quad b_g^{(l+1)} = \psi_0^{-1}\Big(-\frac{1}{p_g}\sum_{j=1}^{p_g} E_{b_g^{(l)}}\big[\log(\lambda_{gj}^2) \mid \boldsymbol{y}\big]\Big),$$

where $\psi_0(\cdot)$ is the digamma function and the expectation terms can be estimated through standard Monte Carlo methods. The iterative procedure terminates when

$$\sum_{g=1}^{G}\big(a_g^{(l+1)} - a_g^{(l)}\big)^2 + \sum_{g=1}^{G}\big(b_g^{(l+1)} - b_g^{(l)}\big)^2$$

is less than some prespecified error tolerance. However, in our experience it is preferred to fix $a_g = 1/n$ for all $g$ and use MMLE to estimate the $b_g$ hyperparameters. The first reason is that $a_g$ controls the strength of the thresholding effect and choosing $a_g$ close to zero guarantees strong shrinkage of null coefficients towards zero. The second reason is that only estimating one hyperparameter per group is more feasible than estimating two hyperparameters per group, particularly when the number of groups is large. Since $b_g$ primarily controls how dependent the shrinkage is within-group, it is more important to focus estimation on the $b_g$ hyperparameters. We do recognize that setting $a_g = 1/n$ violates a condition in Theorem 3.1 where the infimum of the set of hyperparameters cannot converge to zero as $n \to \infty$. However, for practical purposes, this approach provides an automatic way to set $a_g$ while also yielding similar results to $a_g$ close to zero and fixed as a function of the sample size, such as $a_g = 1/100$.

Although MMLE is useful for problems where the number of groups, $G$, is small relative to the sample size, the estimates for the $a_g$'s and $b_g$'s will become increasingly variable in high-dimensional settings where the number of groups is large. There may also be low-dimensional settings where the user wants to incorporate explicit prior knowledge about the nature of the within-group signal density. In such cases, it may be preferred to fix hyperparameter values in accordance with subject matter expertise. As with the modified MMLE approach, we recommend setting $a_g = 1/n$ for all $g$. To fix $b_g$ we recommend a useful heuristic whereby local, group, and global shrinkage parameters are simulated from the GIGG prior. Using the simulated shrinkage parameters, shrinkage factors can be constructed and the correlation between shrinkage factors within the same group can be empirically calculated. Selecting the hyperparameter $b_g$ is then equivalent to selecting how dependent the shrinkage is within-group, a more easily understandable concept.

Another alternative in high-dimensional cases is to set $a_g = a$ and $b_g = b$ for all $g = 1, \ldots, G$ and $a, b > 0$. While this strategy loses the flexibility of customizing shrinkage for each group, it is at least capable of estimating a global tradeoff between group and

local shrinkage in a manner that is more feasible for a MMLE procedure to reliably estimate. The corresponding MMLE updates for this procedure are

$$a^{(l+1)} = \psi_0^{-1}\left(\frac{1}{G}\sum_{g=1}^{G} E_{a^{(l)}}\left[\log(\gamma_g^2) \mid \boldsymbol{y}\right]\right), \quad b^{(l+1)} = \psi_0^{-1}\left(-\frac{1}{p}\sum_{g=1}^{G}\sum_{j=1}^{p_g} E_{b^{(l)}}\left[\log(\lambda_{gj}^2) \mid \boldsymbol{y}\right]\right).$$

Implementations of GIGG regression with fixed hyperparameters and hyperparameters estimated via MMLE are available on the Comprehensive R Archive Network (CRAN). Out of the strategies for achieving faster computation outlined in Section 4.1, the gigg package implements the approach from Bhattacharya et al. (2016).

# 5 Simulations

## 5.1 Generative Model

The data generative mechanism is linear regression model (2.1), where $\boldsymbol{C}$ includes the intercept term and five adjustment covariates drawn from independent standard normal distributions, $\boldsymbol{\alpha} = (0, 1, 1, 1, 1, 1)^\top$, and $\boldsymbol{X}$ is drawn from a multivariate normal distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Sigma_X}$. $\boldsymbol{\Sigma_X}$ is determined such that the regressors have unit variance and block-diagonal exchangeable correlation structure. Pairwise correlations within each group are $\rho = 0.8$ for the high correlation simulation settings or $\rho = 0.6$ for the medium correlation simulation settings. For all simulation settings, the pairwise correlations across groups are 0.2 and the residual error variance, $\sigma^2$, is fixed such that $\boldsymbol{\beta}^\top \boldsymbol{\Sigma_X} \boldsymbol{\beta} / (\boldsymbol{\beta}^\top \boldsymbol{\Sigma_X} \boldsymbol{\beta} + \sigma^2) = 0.7$.

The first set of simulation settings will be called the fixed coefficient simulation settings, where $n = 500$ and $p = 50$ (see Table 1 for simulation setting details). In the

| Label | Group Sizes | Correlation | Signal Type | Signal Details |
|---|---|---|---|---|
| C10H | 10,10,10,10,10 | 0.8 | Concentrated | Signal concentrated in one of the regressors in all five groups |
| D10H | 10,10,10,10,10 | 0.8 | Distributed | Signal distributed across all regressors within the first group |
| C10M | 10,10,10,10,10 | 0.6 | Concentrated | Signal concentrated in one of the regressors in all five groups |
| D10M | 10,10,10,10,10 | 0.6 | Distributed | Signal distributed across all regressors within the first group |
| C5 | 5,5,5,5,5,5,5,5,5,5 | 0.8 | Concentrated | Signal concentrated in one regressor for five out of ten groups |
| D5 | 5,5,5,5,5,5,5,5,5,5 | 0.8 | Distributed | Signal distributed across all regressors within the first two groups |
| C25 | 25,25 | 0.8 | Concentrated | Signal concentrated in three regressors in the first group and two regressors in the second group |
| D25 | 25,25 | 0.8 | Distributed | Signal distributed across first ten regressors within the first group |
| CL | 30,10,5,3,2 | 0.8 | Concentrated, Large Groups | Signal concentrated in one regressor in the group of size 30 and one regressor in the group of size 10 |
| DL | 30,10,5,3,2 | 0.8 | Distributed, Large Groups | Signal distributed across all regressors within the group of size 30 |
| CS | 30,10,5,3,2 | 0.8 | Concentrated, Small Groups | Signal concentrated in one regressor in the group of size 3 and one regressor in the group of size 2 |
| DS | 30,10,5,3,2 | 0.8 | Distributed, Small Groups | Signal distributed across all regressors within the groups of size 5, 3, and 2 |

Table 1: Fixed coefficient simulation settings where $n = 500$ and $p = 50$. The label column refers to the name of the simulation setting that will be used throughout the rest of the simulation section. The group sizes column shows the sizes of all the groups within each simulation setting. The correlation column lists the pairwise correlations between regressors in the same group. The signal type and signal details columns explain how the signal is distributed among regressors within the active groups.

context of this simulation study, a concentrated signal qualitatively refers to a simulation setting where the signal is contained within few regressors in a group and a distributed signal qualitatively refers to a simulation setting where the signal is shared across many regressors within the same group. The purpose of the fixed coefficient simulation settings with equally sized groups is to ascertain which methods perform well when the within-group signal is sparse or dense, and whether or not the performance depends on group size or strength of the within-group regressor correlations. The purpose of the fixed coefficient simulation settings with groups of different sizes is to determine if the performance depends on whether concentrated or distributed signals are contained within groups of large or small size. Here, the groups of size 30 and 10 are considered the large groups and the groups of size 5, 3, and 2 are considered the small groups.

Beyond the fixed regression coefficient simulation settings, we also consider *random coefficient* simulations in the high correlation setting, where for each simulation iteration a random regression coefficient vector is generated. Here, we have a low-dimensional simulation setting with $n = 500$ and $p = 50$, as well as a high-dimensional simulation setting with $n = 200$ and $p = 500$. All groups in both random coefficient simulation settings contain 10 regressors. To construct a regression coefficient vector, we start by randomly selecting either a concentrated or distributed signal for the first group with even probability to guarantee that each simulation iteration will have at least one true signal. The concentrated and distributed signal magnitudes are selected such that the contribution to $\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{X}} \boldsymbol{\beta}$ is equal, namely the distributed signal is $\beta_{gj} = 0.25$ for $j = 1, \ldots, 10$ and the concentrated signal is $\beta_{g1} = 5.125$ and $\beta_{gj} = 0$ for $j = 2, \ldots, 10$. For the remaining groups, we randomly select a concentrated signal with probability 0.2, a distributed signal with probability 0.2, and no signal with probability 0.6. The goal of the random coefficient simulation settings is to show that, averaged across many combinations of regression coefficient vectors comprised of sparse within-group signals, dense within-group signals, and inactive groups, GIGG regression results in low mean-squared error.

## 5.2 Competing Methods and Evaluation Metrics

Estimation properties will be evaluated based on empirical mean-squared error (MSE), stratified by null and non-null coefficients, across 5000 replicates. That is,

$$\widehat{\mathrm{MSE}} = \frac{1}{5000} \sum_{r=1}^{5000} \left(\hat{\boldsymbol{\beta}}^r - \boldsymbol{\beta}\right)^\top \left(\hat{\boldsymbol{\beta}}^r - \boldsymbol{\beta}\right),$$

where $\hat{\boldsymbol{\beta}}^r$ is the estimate of $\boldsymbol{\beta}$ from simulated dataset $r$. 5000 was selected so that the MSEs listed in the simulation results section are relatively precise. In the random coefficient simulations, calculating the MSE corresponds to an integrated mean-squared error (IMSE) metric averaged across the generative distribution of the regression coefficient vectors. For the fixed coefficient simulations we will consider several special cases of the GIGG prior with fixed hyperparameters, namely all possible combinations of $a_g \in \{1/n, 1/2\}$ and $b_g \in \{1/n, 1/2, 1\}$. That way, we can check whether the intuition gleaned from Figure 2 empirically translates to the regression setting. We will

| $\rho = 0.8$ | Concentrated | | | Distributed | | |
|---|---|---|---|---|---|---|
| **Method** | Null | Non-Null | Overall | Null | Non-Null | Overall |
| Ordinary Least Squares | 3.74 | 0.41 | 4.16 | 8.09 | 2.03 | 10.12 |
| Horseshoe | 0.51 | 0.41 | 0.92 | 0.85 | 2.14 | 2.99 |
| GIGG ($a_g = 1/n, b_g = 1/n$) | 0.11 | 0.30 | **0.40** | 0.04 | 3.60 | 3.63 |
| GIGG ($a_g = 1/2, b_g = 1/n$) | 0.11 | 0.30 | **0.41** | 0.04 | 3.56 | 3.59 |
| GIGG ($a_g = 1/n, b_g = 1/2$) | 0.29 | 0.39 | 0.67 | 0.03 | 1.57 | **1.61** |
| *GIGG ($a_g = 1/2, b_g = 1/2$) | 0.33 | 0.40 | 0.72 | 0.24 | 1.70 | 1.94 |
| GIGG ($a_g = 1/n, b_g = 1$) | 0.53 | 0.49 | 1.03 | 0.03 | 1.43 | **1.46** |
| GIGG ($a_g = 1/2, b_g = 1$) | 0.58 | 0.49 | 1.07 | 0.26 | 1.43 | 1.69 |
| GIGG (MMLE) | 0.23 | 0.36 | **0.59** | 0.04 | 1.36 | **1.40** |
| Group Horseshoe+ | 0.30 | 0.39 | 0.70 | 0.08 | 1.64 | 1.73 |
| Spike-and-Slab Lasso | 0.15 | 0.33 | **0.48** | 0.21 | 4.27 | 4.49 |
| BGL-SS | 2.02 | 0.80 | 2.82 | 0.04 | 1.31 | **1.34** |
| BSGS-SS | 0.23 | 0.42 | 0.65 | 0.04 | 1.84 | 1.88 |

Table 2: Mean-squared errors (MSE) for simulation settings C10H and D10H in Table 1 ($n = 500, p = 50$) with high pairwise correlations ($\rho = 0.8$). Bolded cells indicate the four methods with the lowest overall MSE. Four methods are highlighted to emphasize that GIGG (MMLE) is the best method with respect to MSE for both concentrated and distributed signals aside from methods that only perform well for one of the two settings. *GIGG ($a_g = 1/2$ and $b_g = 1/2$) is equivalent to group horseshoe regression.

also consider the GIGG prior when the hyperparameters $a_g = 1/n$ are fixed and $b_g$ are estimated via MMLE.

The list of competing methods include Ordinary Least Squares (OLS), Horseshoe regression, Group Horseshoe+ regression (Xu et al., 2016), Spike-and-Slab Lasso (Rockova and George, 2018), Bayesian Group Lasso with Spike-and-Slab Priors (BGL-SS) (Xu and Ghosh, 2015), and Bayesian Sparse Group Selection with Spike-and-Slab Priors (BSGS-SS) (Xu and Ghosh, 2015). As a reminder, to avoid confusion with the group horseshoe prior proposed in this paper, we will refer to the group horseshoe prior from Xu et al. (2016) as the group horseshoe+ prior. We will use the posterior mean estimator for all Bayesian methods, with the exception of Spike-and-Slab Lasso, BGL-SS, and BSGS-SS. BGL-SS and BSGS-SS will use the posterior median estimator and Spike-and-Slab Lasso will use the posterior mode estimator. Most methods requiring Markov chain Monte Carlo (MCMC) sampling have 10000 burn-in draws, followed by 10000 posterior draws with no thinning. Some exceptions are BGL-SS and BSGS-SS which have 1000 burn-in draws and 2000 posterior draws with no thinning, due to the relatively slower posterior sampling algorithms. Another exception is group horseshoe+ regression in the high-dimensional random coefficient simulation, which required 100000 burn-in draws to consistently converge.

## 5.3 Simulation Results

Table 2 presents the MSE for simulation settings C10H and D10H and Supplementary Tables 1-3 list the MSEs for the C10M, D10M, C5, D5, C25, and D25 simulation settings (Boss et al., 2023). Because the results for C10H and D10H are similar to C10M, D10M,

| $\rho = 0.8$ | Concentrated | | | Distributed | | |
|---|---|---|---|---|---|---|
| **Method** | Null | Non-Null | Overall | Null | Non-Null | Overall |
| Ordinary Least Squares | 2.02 | 0.08 | 2.11 | 1.58 | 2.85 | 4.43 |
| Horseshoe | 0.19 | 0.06 | 0.25 | 0.16 | 1.04 | 1.20 |
| GIGG ($a_g = 1/n, b_g = 1/n$) | 0.02 | 0.04 | **0.07** | 0.01 | 1.92 | 1.93 |
| GIGG ($a_g = 1/2, b_g = 1/n$) | 0.03 | 0.07 | **0.10** | 0.01 | 1.88 | 1.89 |
| GIGG ($a_g = 1/n, b_g = 1/2$) | 0.06 | 0.05 | 0.10 | 0.01 | 0.99 | **1.00** |
| *GIGG ($a_g = 1/2, b_g = 1/2$) | 0.06 | 0.05 | 0.11 | 0.05 | 0.99 | 1.04 |
| GIGG ($a_g = 1/n, b_g = 1$) | 0.12 | 0.06 | 0.19 | 0.01 | 0.85 | **0.86** |
| GIGG ($a_g = 1/2, b_g = 1$) | 0.13 | 0.06 | 0.19 | 0.05 | 0.83 | **0.88** |
| GIGG (MMLE) | 0.03 | 0.04 | **0.07** | 0.01 | 0.80 | **0.81** |
| Group Horseshoe+ | 0.06 | 0.05 | 0.11 | 0.04 | 1.00 | 1.03 |
| Spike-and-Slab Lasso | 0.04 | 0.03 | **0.07** | 0.08 | 3.29 | 3.36 |
| BGL-SS | 1.26 | 0.22 | 1.48 | 0.02 | 1.36 | 1.38 |
| BSGS-SS | 0.06 | 0.06 | 0.12 | 0.01 | 1.30 | 1.31 |

Table 3: Mean-squared errors (MSE) for simulation settings CL and DL in Table 1 ($n = 500, p = 50$) with high pairwise correlations ($\rho = 0.8$). Bolded cells indicate the four methods with the lowest overall MSE. Four methods are highlighted to emphasize that GIGG (MMLE) is the best method with respect to MSE for both concentrated and distributed signals aside from methods that only perform well for one of the two settings. *GIGG ($a_g = 1/2$ and $b_g = 1/2$) is equivalent to group horseshoe regression.

C5, D5, C25, and D25, we will only focus our discussion around the C10H and D10H simulation settings. The first noteworthy observation is that group horseshoe regression has a uniformly lower MSE than both OLS and horseshoe regression for both null and non-null estimation, although the discrepancy between horseshoe and OLS is much larger than the difference between group horseshoe and horseshoe, particularly for the null coefficients. For GIGG regression with fixed hyperparameters, the top performer is GIGG regression with $b_g = 1/n$ when the signal is concentrated within-group (Null MSE = 0.11, Non-Null MSE = 0.30) and $a_g = 1/n, b_g = 1$ when the signal is distributed within-group (MSE = 1.46), exactly as Figure 2 suggests. However, if the user sets $b_g = 1$ when the signal is concentrated (Null MSE = 0.53, Non-Null MSE = 0.49) or $b_g = 1/n$ when the signal is distributed (Null MSE = 0.04, Non-Null MSE = 3.60), then the "incorrect" prior information results in notably worse MSE compared to the "correct" prior information. That being said, $b_g = 1/2$ appears to be a middle ground where the performance for both concentrated and distributed simulation settings is generally good.

Examining the performance of the competing methods, we note that Spike-and-Slab Lasso does very well for the concentrated signal setting (MSE = 0.48), but struggles when the signal is distributed (MSE = 4.49). Conversely, BGL-SS does poorly when the signal is concentrated (MSE = 2.82), but has good performance when the signal is distributed (MSE = 1.34). Group horseshoe+ regression and BSGS-SS have relatively low MSE across the low-dimensional simulation settings, however, GIGG with MMLE almost always outperforms both methods in the low-dimensional cases with respect to overall MSE. The improved performance for GIGG regression with MMLE over a method like group horseshoe+ regression is precisely because GIGG regression

| $\rho = 0.8$ | Concentrated | | | Distributed | | |
|---|---|---|---|---|---|---|
| **Method** | Null | Non-Null | Overall | Null | Non-Null | Overall |
| Ordinary Least Squares | 2.05 | 0.06 | 2.11 | 2.07 | 0.39 | 2.46 |
| Horseshoe | 0.19 | 0.03 | 0.22 | 0.37 | 0.52 | 0.89 |
| GIGG ($a_g = 1/n, b_g = 1/n$) | 0.02 | 0.03 | **0.04** | 0.02 | 1.06 | 1.08 |
| GIGG ($a_g = 1/2, b_g = 1/n$) | 0.02 | 0.04 | **0.06** | 0.02 | 1.06 | 1.08 |
| GIGG ($a_g = 1/n, b_g = 1/2$) | 0.04 | 0.04 | 0.08 | 0.00 | 0.37 | 0.37 |
| *GIGG ($a_g = 1/2, b_g = 1/2$) | 0.05 | 0.03 | 0.08 | 0.04 | 0.36 | 0.40 |
| GIGG ($a_g = 1/n, b_g = 1$) | 0.06 | 0.05 | 0.11 | 0.00 | 0.32 | **0.33** |
| GIGG ($a_g = 1/2, b_g = 1$) | 0.09 | 0.04 | 0.13 | 0.04 | 0.32 | **0.36** |
| GIGG (MMLE) | 0.02 | 0.03 | **0.05** | 0.00 | 0.32 | **0.32** |
| Group Horseshoe+ | 0.04 | 0.03 | 0.07 | 0.09 | 0.43 | 0.52 |
| Spike-and-Slab Lasso | 0.04 | 0.02 | **0.06** | 0.09 | 1.49 | 1.58 |
| BGL-SS | 0.08 | 0.06 | 0.13 | 0.00 | 0.28 | **0.28** |
| BSGS-SS | 0.02 | 0.03 | 0.06 | 0.00 | 0.45 | 0.45 |

Table 4: Mean-squared errors (MSE) for simulation settings CS and DS in Table 1 ($n = 500, p = 50$) with high pairwise correlations ($\rho = 0.8$). Bolded cells indicate the four methods with the lowest overall MSE. Four methods are highlighted to emphasize that GIGG (MMLE) is the best method with respect to MSE for both concentrated and distributed signals aside from methods that only perform well for one of the two settings. *GIGG ($a_g = 1/2$ and $b_g = 1/2$) is equivalent to group horseshoe regression.

with MMLE is able to data-adaptively control the dependence of the grouped multivariate shrinkage. Group horseshoe+ regression cannot directly control within-group dependence because there are no hyperparameters in the prior specification.

Table 3 shows the MSE results for the CL and DL simulation settings and Table 4 lists the MSE results for the CS and DS simulation settings. As with the other fixed coefficient simulation settings, the same general conclusions hold. Whether or not a concentrated signal is contained in large or small groups, GIGG with MMLE and GIGG with fixed hyperparameters where $b_g = 1/n$ have some of the lowest overall MSEs across all methods. Whether or not a distributed signal is contained in large or small groups, GIGG with MMLE and GIGG with fixed hyperparameters where $b_g = 1$ have some of the lowest overall MSEs. Spike-and-Slab LASSO performed well in the concentrated simulation settings, but BGL-SS only performed well in the distributed setting when the groups containing the true signals were small. Overall, it does not appear that group size and signal distribution within the groups fundamentally change the performance of GIGG with MMLE or GIGG with fixed hyperparameters, within the scope of the data generative parameters that we explored.

Next, Table 5 summarizes the $b_g$ hyperparameter estimates across 5000 simulation iterations for all high correlation simulation settings with $n = 500$, $p = 50$, and $G = 5$. For simulation setting C10H we see that median $b_g$ hyperparameter estimate for groups 1-5 goes from 0.52 in group 1, which contains the smallest signal, to 0.27 and 0.26 for the largest signals. That is, as the concentrated signal becomes stronger, the median $b_g$ estimate starts moving towards zero. Conversely, for all simulation settings with distributed signals, we generally observe that groups with either all signals or all noise regressors tend to result in $b_g$ estimates that are greater than one regardless

| Label | Group Sizes | Group 1 | Group 2 | Group 3 | Group 4 | Group 5 |
|---|---|---|---|---|---|---|
| C10H | 10,10,10,10,10 | 0.52 (0.32-0.84) | 0.35 (0.25-0.71) | 0.29 (0.23-0.55) | 0.27 (0.22-0.43) | 0.26 (0.22-0.44) |
| D10H | 10,10,10,10,10 | 1.84 (1.11-2.59) | 1.19 (0.75-1.80) | 1.19 (0.74-1.81) | 1.19 (0.74-1.81) | 1.19 (0.71-1.81) |
| CL | 30,10,5,3,2 | 0.28 (0.24-0.40) | 0.22 (0.19-0.30) | 1.30 (0.69-2.69) | 1.51 (0.68-3.41) | 1.73 (0.72-4.00*) |
| DL | 30,10,5,3,2 | 1.96 (1.04-2.91) | 1.60 (0.83-2.45) | 2.09 (0.94-3.51) | 2.59 (1.04-4.00*) | 3.13 (1.03-4.00*) |
| CS | 30,10,5,3,2 | 0.66 (0.54-0.88) | 0.79 (0.57-1.22) | 0.92 (0.60-1.54) | 0.16 (0.13-0.23) | 0.14 (0.12-0.21) |
| DS | 30,10,5,3,2 | 0.95 (0.64-1.47) | 1.31 (0.75-2.33) | 2.07 (1.21-2.97) | 1.51 (0.72-2.14) | 1.03 (0.25-1.46) |

Table 5: Median (2.5% Quantile - 97.5% Quantile) $b_g$ estimates for GIGG regression with MMLE in all fixed regression coefficient, high correlation simulations settings with $n = 500$, $p = 50$ and $G = 5$ with 5000 replicates. See Table 1 for the simulation setting details. Here, large groups correspond to groups of size 30 and 10 and small groups correspond to groups of size 5, 3, and 2. $^*b_g$ is capped at four to facilitate numerical stability of the MMLE procedure.

of how large the groups containing the distributed signals are. For the CL simulation setting, we observe that the general trends for concentrated and distributed signals hold, namely that group 1 and group 2, which contain the concentrated signals, have median $b_g$ estimates between 0.2 and 0.3, and groups 3-5, which are null groups, have median $b_g$ estimates greater than one. However, the CS simulation setting is a little more interesting. Group 4 and group 5 in the CS simulation setting are the active groups with concentrated signals and we see that the median $b_g$ estimates are 0.16 and 0.14, respectively. That is, small groups with concentrated signals seem to result in $b_g$ hyperparameter estimates that are even closer to zero compared with larger groups with concentrated signals. Moreover, groups 1-3 in the CS simulation setting are all null groups, and they show a general trend of the median $b_g$ hyperparameter estimates getting smaller, the larger the group is. Specifically, the group of size 30 in the CS simulation setting has a median $b_g$ estimate of 0.66 and the group of size 5 in the CS simulation setting has a median $b_g$ estimate of 0.92. Finally, it is important to mention that $b_g$ is capped at four in our implementation to facilitate numerical stability of the MMLE procedure. For group 5 in the CL simulation setting, $b_g$ was set to four in 131 out of 5000 simulation iterations. For group 4 in the DL simulation setting, $b_g$ was set to four in 350 out of 5000 simulation iterations. For group 5 in the DL simulation setting, $b_g$ was set to four 1195 times out of 5000 simulation iterations. Capping $b_g$ at four was chosen so to facilitate numerical stability of the MMLE procedure, but alternative ceilings on $b_g$ can be considered.

Lastly, we consider the IMSE for the random coefficient simulation settings presented in Table 6. As with the fixed regression coefficient simulations, group horseshoe (Null IMSE = 0.39) and group horseshoe+ regression (Null ISME = 0.36) lead to a substantial improvement in IMSE compared to horseshoe regression in the low-dimensional simulation setting. However, in the low-dimensional simulation setting, we also observe that the additional flexibility of GIGG regression with MMLE to self-adapt to different types of within-group signal distributions results in noticeable improvements in IMSE for the null coefficients (Null IMSE = 0.21). Spike-and-Slab Lasso and BGL-SS struggle in the random coefficient simulation scenario because they tend to only work well when the signal is concentrated or distributed, respectively, leading to unfavorable average performance. The high-dimensional simulation setting shows that GIGG regression

| | Low-Dimensional | | | High-Dimensional | | |
|---|---|---|---|---|---|---|
| **Method** | Null | Non-Null | Overall | Null | Non-Null | Overall |
| Ordinary Least Squares | 8.84 | 3.38 | 12.21 | - | - | - |
| Horseshoe | 0.70 | 1.18 | 1.88 | 86.04 | 215.36 | 301.40 |
| GIGG ($a_g = 1/n, b_g = 1/n$) | 0.09 | 1.79 | 1.88 | 131.32 | 252.04 | 383.36 |
| GIGG ($a_g = 1/2, b_g = 1/n$) | 0.10 | 1.83 | 1.93 | 128.61 | 250.49 | 379.10 |
| GIGG ($a_g = 1/n, b_g = 1/2$) | 0.33 | 1.15 | **1.47** | 80.61 | 213.24 | **293.85** |
| *GIGG ($a_g = 1/2, b_g = 1/2$) | 0.39 | 1.13 | **1.52** | 60.73 | 207.55 | **268.29** |
| GIGG ($a_g = 1/n, b_g = 1$) | 0.69 | 1.11 | 1.79 | 90.12 | 210.61 | 300.74 |
| GIGG ($a_g = 1/2, b_g = 1$) | 0.75 | 1.11 | 1.85 | 53.66 | 203.86 | **257.53** |
| GIGG (MMLE) | 0.21 | 1.06 | **1.27** | 93.11 | 220.90 | 314.01 |
| Group Horseshoe+ | 0.36 | 1.14 | **1.49** | 82.91 | 213.06 | 295.97 |
| Spike-and-Slab Lasso | 0.16 | 3.65 | 3.81 | 159.02 | 344.82 | 503.84 |
| BGL-SS | 2.84 | 2.44 | 5.28 | 1918.84 | 678.36 | 2597.19 |
| BSGS-SS | 0.36 | 1.45 | 1.81 | 2.22 | 254.02 | **256.25** |

Table 6: Integrated mean-squared errors (IMSE) for the random regression coefficient simulation settings with high pairwise correlations ($\rho = 0.8$). The low-dimensional simulation setting has $n = 500$ and $p = 50$ and the high-dimensional simulation setting has $n = 200$ and $p = 500$. Bolded cells indicate the four methods with the lowest overall IMSE. *GIGG ($a_g = 1/2$ and $b_g = 1/2$) is equivalent to group horseshoe regression.

with MMLE does not perform as well as group horseshoe regression, group horseshoe+ regression, and BSGS-SS, likely due to the fact that there is limited sample size to estimate many more group-specific $b_g$ hyperparameters. Note that BSGS-SS has very low Null MSE (Null MSE = 2.22), but very high Non-Null MSE (Non-Null MSE = 254.02) compared with many of the GIGG regression methods. Being a spike-and-slab based method, BSGS-SS has an inherent advantage over continuous shrinkage methods in estimating the null counterpart of sparse parameters because it shrinks coefficients to exact zero. Moreover, because BSGS-SS is based off of sparse group lasso, it shrinks all coefficients much more strongly toward zero than GIGG regression methods. GIGG regression with fixed hyperparameters $a_g = 1/2$ and $b_g = 1$ has the best performance of the continuous shrinkage prior methods, likely because averaging a signal across highly correlated regressors in a high-dimensional setting is preferable to assigning the entire signal to one regressor, with respect to a squared error loss function. The high-dimensional simulations indicate that better strategies to determine hyperparameter values for high-dimensional regression problems could result in improved estimation properties.

# 6 Data Example

The National Health and Nutrition Examination Survey (NHANES) is a collection of studies conducted by the National Center for Health Statistics with the overarching goal of evaluating the health and nutritional status of the United States' populace. Data collection consists of a written survey and physical examination which records demographic, socioeconomic, dietary, and health-related information, including physiological measurements and laboratory tests. We will specifically apply GIGG regression

to a subset of 990 adults from NHANES 2003-2004 with 35 measured contaminants across five exposure classes: metals (3 exposures), phthalates (7 exposures), organochlorine pesticides (8 exposures), polybrominated diphenyl ethers (PBDEs) (7 exposures), and polycyclic aromatic hydrocarbons (PAHs) (10 exposures). Figure 1 illustrates the block diagonal correlation structure of these exposures, where areas of high correlation are mostly contained within exposure class. Gamma glutamyl transferase (GGT), an enzymatic marker of liver functionality, is the outcome of interest. GGT and all environmental exposures were log-transformed to remove right skewness and then subsequently standardized. The final model was adjusted for age, sex, body mass index, poverty-to-income ratio, ethnicity, and urinary creatinine.

Figure 3 presents the estimated percent change in GGT corresponding to a twofold change in each environmental exposure and their associated 95% credible intervals for methods commonly used in multipollutant modeling. Bayesian linear regression with noninformative priors and ridge regression were implemented in R Stan using four chains with no thinning, each with 1000 burn-in draws and 1000 posterior draws. Horseshoe regression and GIGG regression with MMLE used 10000 burn-in samples, followed by 10000 posterior draws with a thinning interval of five. As with the simulation section, GIGG regression with MMLE refers to an implementation of GIGG which fixes $a_g = 1/n$ for all $g$ and then uses MMLE to estimate the $b_g$ hyperparameters. Convergence of the MCMC chains was evaluated using Gelman-Rubin's potential scale reduction factor (PSRF) (Gelman and Rubin, 1992). All methods had a PSRF of $1.00 - 1.01$ for the regression coefficients, indicating that all MCMC chains converged. For GIGG regression with MMLE, the median effective sample size for the $\beta_{gj}$'s was 7309 with an interquartile range (IQR) of 3634-9266 and the effective sample size for $\sigma^2$ was 10000. For the shrinkage parameters, the local shrinkage parameters had a median effective sample size of 9541 with an IQR of 7819-10000, the group shrinkage parameters had a median effective sample size of 1790 with an IQR of 1707-2271, and the global shrinkage parameter had an effective sample size of 983.

Figure 3 compares GIGG regression with Bayesian linear regression (non-informative priors), ridge regression, and horseshoe regression. GIGG is generally more efficient than the other methods, having narrower credible intervals, because GIGG better deals with multicollinearity and homogeneous within-group effect sizes. When there is little multicollinearity and heterogeneous within-group effect sizes, GIGG has similar efficiency to the horseshoe. Further, GIGG allows for different shrinkage on coefficients, unlike ridge regression which overshrinks large coefficients. In detail, the median credible interval length for GIGG regression with MMLE is 21.0% shorter for the PAHs, 63.2% shorter for the PBDEs, and 22.5% shorter for the phthalates compared to horseshoe regression, which are all exposures classes with high pairwise correlations and common estimated effect sizes. However, the metals exposure class, which has weak pairwise correlations and heterogeneous estimated effect sizes, results in a median credible interval length of 0.31 for GIGG regression with MMLE and 0.28 for horseshoe regression. Ridge regression estimates that a twofold change in lead exposure is associated with 1.21% higher GGT (95% CI: 0.09, 2.54), while horseshoe regression estimates 1.76% higher GGT (95% CI: -0.02, 3.68) and GIGG regression with MMLE estimates 2.04% higher GGT (95% CI: 0.01, 3.87). From a computational perspective, GIGG regression with MMLE

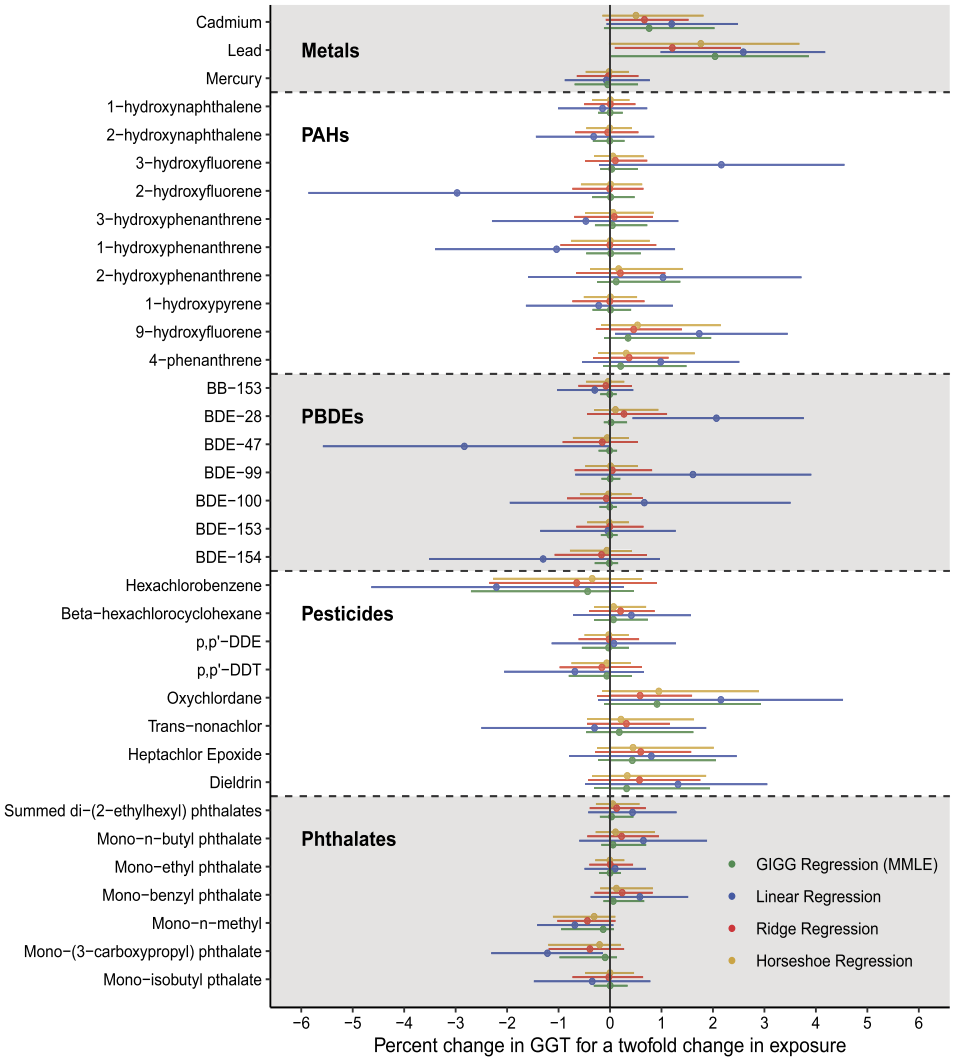


Figure 3: Estimated associations between environmental toxicants (metals, phthalates, pesticides, PBDEs, and PAHs) and gamma glutamyl transferase (GGT) from NHANES 2003-2004 ($n = 990$).

generated a median effective sample size of 559.6 per second for the $\beta_{gj}$'s, compared to a median effective sample size of 791.1 per second for horseshoe regression.

Supplementary Figure 1 provides a focused comparison of the various group shrinkage methods from the simulation study (Boss et al., 2023). GIGG is generally more efficient than the other continuous shrinkage prior methods, having narrower credible intervals than group horseshoe and group horseshoe+. As with the results in Figure 3,

these efficiency gains are attributable to GIGG with MMLE better handling multicollinearity and homogeneous within-group effect sizes. In detail, GIGG regression with MMLE, group horseshoe regression, and group horseshoe+ regression all have very similar performance in terms of point estimation. The 95% credible interval for lead covers zero for group horseshoe+ regression (1.82% higher GGT; 95% CI: -0.02, 3.70), while the 95% credible interval for lead does not cover zero for group horseshoe regression (1.88% higher; 95% CI: 0.01, 3.72) and GIGG regression with MMLE (2.04% higher GGT; 95% CI: 0.01, 3.87). For the PAHs, GIGG regression with MMLE has a 26.1% shorter median credible interval length than group horseshoe regression and a 25.7% shorter median credible interval length than group horseshoe+ regression. For the PBDEs, GIGG regression with MMLE has a 57.7% shorter median credible interval length than group horseshoe regression and a 60.6% shorter median credible interval length than group horseshoe+ regression. Differences in credible interval length for the metals and pesticides among GIGG regression with MMLE, group horseshoe regression, and group horseshoe+ regression were much smaller. The posterior median estimator corresponding to BGL-SS selected both the metals and pesticides groups, despite the fact that no other method identified any pesticides based on 95% credible intervals covering zero or posterior inclusion probabilities being larger than 0.5. BSGS-SS selected lead and cadmium, while the 95% credible intervals for GIGG regression with MMLE and group horseshoe regression only identified lead.

# 7 Discussion

The principal methodological contribution of this paper is to construct a continuous shrinkage prior that improves regression coefficient estimation in the presence of grouped regressors. GIGG regression flexibly controls the relative contributions of individual and group shrinkage to improve regression coefficient estimation, resulting in a relative IMSE reduction of 32.4% compared to horseshoe regression. One of the main limitations of GIGG regression is that regressor groupings must be explicitly specified and regressor groupings may not overlap. Additionally, although the GIGG prior can be imposed on regression coefficients in Bayesian generalized linear models, a theoretical evaluation of the shrinkage properties for non-normal outcome data would be necessary to determine if the GIGG prior is appropriate for such models.

There are several considerations for deciding between a spike-and-slab based bi-level selection method and a grouped multivariate shrinkage prior based method. The first how large the dataset is. Computationally, it is much slower to sample from the posterior corresponding to BSGS-SS than it is to sample from the posterior distribution corresponding to GIGG regression. Therefore, in higher dimensional problems, sampling the posterior corresponding to BSGS-SS may be computationally prohibitive. The second consideration is with regard to the tradeoff between group and local shrinkage. BSGS-SS only has two hyperparameters, so fixing those hyperparameters defines a group-local tradeoff *for all* groups. The GIGG prior is different in that a unique $a_g$ and $b_g$ for each group allows group-local tradeoffs *for each* group. If the expectation is for some groups to have concentrated signals and for others to have distributed signals, then GIGG regression with MMLE is better able to tailor the shrinkage corresponding to each group.

The third is how important variable selection is. There are several techniques to define selection for continuous shrinkage priors, however spike-and-slab based methods define variable selection much more naturally through posterior inclusion probabilities. Therefore, if selection is a primary goal, then a spike-and-slab based method like BSGS-SS might be preferred.

The analysis of multiple pollutant data and chemical mixtures is a key thrust of the National Institute of Environmental Health Sciences, and the GIGG prior provides a useful framework for achieving variance reduction in the presence of group-correlated exposures, characterizing uncertainties in point estimates, and constructing policy relevant metrics, like summary risk scores, in a principled way. However, the generality of the GIGG prior coupled with the relative ease of computation means that, despite its motivation coming from environmental epidemiology, the GIGG prior is applicable to many other areas. For example, in neuroimaging studies, scalar-on-image regression (Kang et al., 2018) has been widely used to study the association between brain activity and clinical outcomes of interest. The whole brain can be partitioned into a set of exclusive regions according to brain functions and anatomical structures. Within the same region, the brain imaging biomarkers tend to be more correlated and have similar effects on the outcome variable. The GIGG prior can be extended for scalar-on-image regression and it has great potential to improve estimating the effects of imaging biomarkers by incorporating brain region information.

In this paper, our focus was sparse estimation, but it is also natural to inquire about uncertainty quantification and variable selection. Based on our simulations, the conclusions of van der Pas et al. (2017) are relevant for the GIGG prior when $0 < a_g \leq 1/2$, but a comprehensive study needs to be carried out. There is no consensus way of defining variable selection for continuous shrinkage priors, however there are several approaches to determine a final active set, including credible intervals covering zero (van der Pas et al., 2017), decoupling shrinkage and selection (DSS) (Hahn and Carvalho, 2015), and penalized credible regions (Zhang and Bondell, 2018). For horseshoe-style shrinkage, variable selection defined through credible intervals covering zero is highly conservative, but works well if one wants to limit the number of false discoveries. The penalized credible region approach searches for the sparsest model that falls within the $100 \times (1-\alpha)\%$ joint elliptical credible region, while DSS constructs an adaptive lasso-style objective function with the goal of sparsifying the posterior mean such that most of the predictive variability is still explained. Since the DSS construction is framed from a prediction perspective, this approach may not be ideal for regression coefficient estimation problems in the presence of correlated regressors. Another crucial point to make is that if one is interested in selection, the posterior mode estimator for the horseshoe prior will result in exact zero estimates, and an approximate algorithm for calculating the joint posterior mode was developed in Bhadra et al. (2019) using the horseshoe-like prior. Therefore, one could conceptualize an extension of the expectation-maximization algorithm developed by Bhadra et al. (2019) using a "GIGG-like" prior. A second option, from a variable selection perspective rather than a model selection perspective, is to ascertain whether or not the marginal posterior modes equal zero, based on the posterior draws (Liu and Ghosh, 2020). Here, a posterior mode equal to zero refers to a regressor that is not selected and a posterior mode not equal to zero refers to a regressor that is selected.

Further work is needed to juxtapose the behavior of all of these different methods for selection and develop novel algorithms for calculating the marginal and joint posterior modes.

## Supplementary Material

Supplementary Material for "Group Inverse-Gamma Gamma Shrinkage for Sparse Linear Models with Block-Correlated Regressors" (DOI: 10.1214/23-BA1371SUPP; .pdf). Distributional definitions, proofs for all theoretical results, full conditional distributions for Gibbs sampling, and group shrinkage methods from the simulation applied to the data example.

## References

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office. MR0167642. 791

Andrade, J. A. A. and O'Hagan, A. (2006). "Bayesian Robustness Modeling Using Regularly Varying Distributions." *Bayesian Analysis*, 1(1): 169–188. MR2227369. doi: https://doi.org/10.1214/06-BA106. 790, 791

Armagan, A., Dunson, D. B., and Lee, J. (2013a). "Generalized Double Pareto Shrinkage." *Statistica Sinica*, 23(1): 119–143. MR3076161. 787

Armagan, A., Dunson, D. B., Lee, J., Bajwa, W. U., and Strawn, N. (2013b). "Posterior consistency in linear models under shrinkage priors." *Biometrika*, 100(4): 1011–1018. MR3142348. doi: https://doi.org/10.1093/biomet/ast028. 788, 793, 794

Bai, R. and Ghosh, M. (2019). "Large-scale multiple hypothesis testing with the normal-beta prime prior." *Statistics*, 53(6): 1210–1233. MR4034859. doi: https://doi.org/10.1080/02331888.2019.1662017. 787, 790, 797

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016). "Default Bayesian analysis with global-local shrinkage priors." *Biometrika*, 103(4): 955–969. MR3620450. doi: https://doi.org/10.1093/biomet/asw041. 787, 790

— (2017). "The Horseshoe+ Estimator of Ultra-Sparse Signals." *Bayesian Analysis*, 12(4): 1105–1131. MR3724980. doi: https://doi.org/10.1214/16-BA1028. 787, 797

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. T. (2019). "The Horseshoe-Like Regularization for Feature Subset Selection." *Sankhya B*. MR4256316. doi: https://doi.org/10.1007/s13571-019-00217-7. 809

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). "Fast sampling with Gaussian scale mixture priors in high-dimensional regression." *Biometrika*, 103(4): 985–991. MR3620452. doi: https://doi.org/10.1093/biomet/asw042. 788, 797, 799

Bhattacharya, A., Pati, D., Pillai, N. S., and B., D. D. (2015). "Dirichlet–Laplace Priors for Optimal Shrinkage." *Journal of the American Statistical Association*, 110(512): 1479–1490. MR3449048. doi: https://doi.org/10.1080/01621459.2014.960967. 787, 792

Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1989). *Regular Variation, vol. 27 of Encyclopedia of Mathematics and its Applications*. Cambridge, UK: Cambridge University Press. MR1015093. 790

Boss, J., Datta, J., Wang, X., Park, S. K., Kang, J., and Mukherjee, B. (2023). *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1371SUPP. 789, 790, 794, 795, 796, 797, 801, 807

Brown, P. J. and Griffin, J. E. (2010). "Inference with normal-gamma prior distributions in regression problems." *Bayesian Analysis*, 5(1): 171–188. MR2596440. doi: https://doi.org/10.1214/10-BA507. 786

Cadonna, A., Frühwirth-Schnatter, S., and Knaus, P. (2020). "Triple the Gamma—A Unifying Shrinkage Prior for Variance and Variable Selection in Sparse State Space and TVP Models." *Econometrics*, 8(2): 20. 787

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). "Handling Sparsity via the Horseshoe." *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, PMLR*, 5: 73–80. 786

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480. MR2650751. doi: https://doi.org/10.1093/biomet/asq017. 786

Casella, G. (2001). "Empirical Bayes Gibbs sampling." *Biostatistics*, 2(4): 485–500. 798

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). "Bayesian linear regression with sparse priors." *Annals of Statistics*, 43(5): 1986–2018. MR3375874. doi: https://doi.org/10.1214/15-AOS1334. 787

Datta, J. and Dunson, D. B. (2016). "Bayesian inference on quasi-sparse count data." *Biometrika*, 103(4): 971–983. MR3620451. doi: https://doi.org/10.1093/biomet/asw053. 797

Datta, J. and Ghosh, J. K. (2013). "Asymptotic Properties of Bayes Risk for the Horseshoe Prior." *Bayesian Analysis*, 8(1): 111–132. MR3036256. doi: https://doi.org/10.1214/13-BA805. 788, 794, 797

Dawid, A. P. (1973). "Posterior Expectations for Large Observations." *Biometrika*, 60(3): 664–667. MR0336889. doi: https://doi.org/10.1093/biomet/60.3.664. 790

Ferguson, K. K., McElrath, T. F., and Meeker, J. D. (2014). "Environmental Phthalate Exposure and Preterm Birth." *JAMA Pediatrics*, 168(1): 61–67. 787

Gelman, A. and Rubin, D. B. (1992). "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science*, 7(4): 457–472. 806

Ghosal, S. (1999). "Asymptotic Normality of Posterior Distributions in High-Dimensional Linear Models." *Bernoulli*, 5(2): 15–331. MR1681701. doi: https://doi.org/10.2307/3318438. 793

Hahn, P. R. and Carvalho, C. M. (2015). "Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective." *Journal of the American Statistical Association*, 110(509): 435–448. MR3338514. doi: https://doi.org/10.1080/01621459.2014.993077. 809

Hefley, T. J., Hooten, M. B., Hanks, E. M., Russell, R. E., and Walsh, D. P. (2017). "The Bayesian Group Lasso for Confounded Spatial Data." *Journal of Agricultural, Biological, and Environmental Statistics*, 22(1): 42–59. MR3607653. doi: https://doi.org/10.1007/s13253-016-0274-1. 787

Johndrow, J. E., Orenstein, P., and Bhattacharya, A. (2020). "Scalable Approximate MCMC Algorithms for the Horseshoe Prior." *Journal of Machine Learning Research*, 21: 1–61. MR4095352. 788, 797

Kang, J., Reich, B. J., and Staicu, A.-M. (2018). "Scalar-on-image regression via the soft-thresholded Gaussian process." *Biometrika*, 105(1): 165–184. MR3768872. doi: https://doi.org/10.1093/biomet/asx075. 809

Kang, K., Song, X., Hu, X. J., and Zhu, H. (2019). "Bayesian adaptive group lasso with semiparametric hidden Markov models." *Statistics in Medicine*, 38(9): 1634–1650. MR3934810. doi: https://doi.org/10.1002/sim.8051. 787

Kowal, D. R., Matteson, D. S., and Ruppert, D. (2019). "Dynamic shrinkage processes." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4): 781–804. MR3997101. doi: https://doi.org/10.1111/rssb.12325. 797

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). "Penalized Regression, Standard Errors, and Bayesian Lassos." *Bayesian Analysis*, 5(2): 369–412. MR2719657. doi: https://doi.org/10.1214/10-BA607. 787

Li, J., Wang, Z., Li, R., and Wu, R. (2015). "Bayesian Group Lasso for Nonparametric Varying-Coefficient Models with Application to Functional Genome-Wide Association Studies." *The Annals of Applied Statistics*, 9(2): 640–664. MR3371329. doi: https://doi.org/10.1214/15-AOAS808. 787

Liu, B. and Ghosh, S. K. (2020). "On empirical estimation of mode based on weakly dependent samples." *Computational Statistics & Data Analysis*, 152: 107046. MR4130895. doi: https://doi.org/10.1016/j.csda.2020.107046. 809

Makalic, E. and Schmidt, D. F. (2016). "A Simple Sampler for the Horseshoe Estimator." *IEEE Signal Processing Letters*, 23(1): 179–182. 797

Mallick, H. and Yi, N. (2017). "Bayesian group bridge for bi-level variable selection." *Computational Statistics & Data Analysis*, 110: 115–133. MR3612612. doi: https://doi.org/10.1016/j.csda.2017.01.002. 787

Nishimura, A. and Suchard, M. A. (2022). "Prior-Preconditioned Conjugate Gradient

Method for Accelerated Gibbs Sampling in "Large n, Large p" Bayesian Sparse Regression." *Journal of the American Statistical Association*. 797

O'Hagan, A. (1979). "On Outlier Rejection Phenomena in Bayes Inference." *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(3): 358–367. MR0557598. 790

Polson, N. G. and Scott, J. G. (2011). "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction." In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics 9*, chapter 17. Oxford, United Kingdom: Oxford University Press. MR3204017. doi: https://doi.org/10.1093/acprof:oso/9780199694587.003.0017. 787, 789, 791, 797

Rockova, V. and George, E. I. (2018). "The Spike-and-Slab LASSO." *Journal of the American Statistical Association*, 113(521): 431–444. MR3803476. doi: https://doi.org/10.1080/01621459.2016.1260469. 801

Rockova, V. and Lesaffre, E. (2014). "Incorporating grouping information in Bayesian variable selection with applications in genomics." *Bayesian Analysis*, 9(1): 221–258. MR3188306. doi: https://doi.org/10.1214/13-BA846. 787, 792

Som, A., Hans, C., and MacEachern, S. N. (2015). "Block Hyper-g Priors in Bayesian Regression." *arXiv*. MR3321977. 787

Som, A., Hans, C. M., and MacEachern, S. M. (2016). "A conditional Lindley paradox in Bayesian linear models." *Biometrika*, 103(4): 993–999. MR3620453. doi: https://doi.org/10.1093/biomet/asw037. 796

Song, Q. and Liang, F. (2017). "Nearly optimal Bayesian Shrinkage for High Dimensional Regression." *arXiv Preprint*. MR4535982. doi: https://doi.org/10.1007/s11425-020-1912-6. 794

Tang, X., Ghosh, M., Ha, N. S., and Sedransk, J. (2018). "Modeling Random Effects Using Global–Local Shrinkage Priors in Small Area Estimation." *Journal of the American Statistical Association*, 113(524): 1476–1489. MR3902223. doi: https://doi.org/10.1080/01621459.2017.1419135. 797

Terenin, A., Dong, S., and Draper, D. (2019). "GPU-accelerated Gibbs sampling: a case study of the Horseshoe Probit model." *Statistics and Computing*, 29(2): 301–310. MR3914622. doi: https://doi.org/10.1007/s11222-018-9809-3. 788, 797

van der Pas, S., Szabó, B., and van der Vaart, A. (2017). "Uncertainty Quantification for the Horseshoe (with Discussion)." *Bayesian Analysis*, 12(4): 1221–1274. MR3724985. doi: https://doi.org/10.1214/17-BA1065. 809

Wei, R., Reich, B. J., Hoppin, J. A., and Ghosal, S. (2020). "Sparse Bayesian Additive Nonparametric Regression with Application to Health Effects of Pesticides Mixtures." *Statistica Sinica*, 30: 55–79. MR4285485. doi: https://doi.org/10.5705/ss.202017.0315. 787

Xu, X. and Ghosh, M. (2015). "Bayesian Variable Selection and Estimation for Group

Lasso." *Bayesian Analysis*, 10(4): 909–936. MR3432244. doi: https://doi.org/10.1214/14-BA929. 787, 801

Xu, Z., Schmidt, D. F., Makalic, E., Qian, G., and Hopper, J. L. (2016). "Bayesian Grouped Horseshoe Regression with Application to Additive Models." In on Artificial Intelligence 2016, A. J. C. (ed.), *AI 2016: Advances in Artificial Intelligence*, chapter 3. Hobart, Australia: Springer. MR3595648. doi: https://doi.org/10.1007/978-3-319-50127-7_19. 787, 790, 801

Zhang, Y. and Bondell, H. D. (2018). "Variable Selection via Penalized Credible Regions with Dirichlet–Laplace Global-Local Shrinkage Priors." *Bayesian Analysis*, 13(3): 823–844. MR3807868. doi: https://doi.org/10.1214/17-BA1076. 809

**Acknowledgments**

We would like to thank Mike Kleinsasser for assisting with the development of the gigg R package.