Structured Shrinkage Priors

Maryclare Griffin¹ and Peter D. Hoff⁵

¹Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA; ⁵Department of Statistical Science, Duke University, Durham, NC

ABSTRACT

In many regression settings the unknown coefficients may have some known structure, for instance they may be ordered in space or correspond to a vectorized matrix or tensor. At the same time, the unknown coefficients may be sparse, with many nearly or exactly equal to zero. However, many commonly used priors and corresponding penalties for coefficients do not encourage simultaneously structured and sparse estimates. In this article we develop structured shrinkage priors that generalize multivariate normal, Laplace, exponential power and normal-gamma priors. These priors allow the regression coefficients to be correlated in a priori without sacrificing elementwise sparsity or shrinkage. The primary challenges in working with these structured shrinkage priors are computational, as the corresponding penalties are intractable integrals and the full conditional distributions that are needed to approximate the posterior mode or simulate from the posterior distribution may be nonstandard. We overcome these issues using a flexible elliptical slice sampling procedure, and demonstrate that these priors can be used to introduce structure while preserving sparsity. Supplementary materials for this article are available online.

1. Introduction

Shrinkage prior-based penalized estimates of regression coefficients are ubiquitous and useful. When we observe an $n \times 1$ vector of responses $y$ and an $n \times p$ matrix of regressors $X$ and the data are high dimensional, that is $p$ is large relative to $n$, traditional regression methods can fail. They may produce estimates of the $p \times 1$ vector of regression coefficients $\beta$ which have prohibitively large variance or are not unique because the data provide relatively little information about the unknown regression coefficients.

Using a prior for $\beta$ that reflects our a priori knowledge, we can obtain better estimates of $\beta$. When our a priori knowledge involves similarities and differences among elements of $\beta$, we might assume a structured mean-zero multivariate normal prior with covariance matrix $\Sigma$. Alternatively, when our a priori knowledge involves magnitudes of elements of $\beta$, we might assume a sparsity inducing mean-zero independent Laplace prior with variance $\sigma^2$ in order to encode information about the origin and tail behavior of $\beta$. This is useful when $\beta$ is expected to be sparse, as the posterior mode of $\beta$ under this prior corresponds to the $\ell_1$ or lasso penalized estimate of $\beta$ (Tibshirani 1996; Park and Casella 2008).

When our a priori knowledge involves both similarities and differences among and magnitudes of elements of $\beta$, it would be desirable to assume a structured sparsity inducing shrinkage prior. As an example, consider the analysis of brain-computer interface data. Brain-computer interfaces (BCIs) are used to detect changes in subjects’ cognitive state from contemporaneous electroencephalography (EEG) measurements at different channels corresponding to physical locations on the subject’s skull, which can be collected noninvasively at high temporal resolution (Makeig et al. 2012; Wolpaw and Wolpaw 2012). We consider the P300 speller, a specific BCI device which is designed to detect when a subject is viewing a specified target letter (Forney et al. 2013). For an individual subject, we consider 20 indicators $y$ for whether or not the subject was viewing a specified target letter during trial $i$ and 20 contemporaneous EEG measurements $x_{jk}$ from time point $j$ and channel $k$, for $j = 1, \ldots, 208$ time points and $k = 1, \ldots, 8$ channels. A total of 240 indicators and 240 contemporaneous EEG measurements are available, but we consider a subset of 20 indicators and contemporaneous EEG measurements because performance with such little data is of special interest and especially challenging. Scientifically, we expect to observe a P300 wave in contemporaneous EEG measurements during trials when the target letter is present, which is characterized by a sharp rise and then dip before returning to equilibrium. We expect that the wave will begin shortly after the target letter is shown, and will be observed earlier and more clearly on some channels than others.

Estimated regression coefficients that describe the relationship between EEG measurements from time point $j$ and channel $k$ and whether or not the subject was viewing the target letter can be obtained by regressing the indicators $y$ on an intercept and measurements for each time point and channel $x_{jk}$ using a logistic regression model, which assumes $y_{ij} \sim \text{Binomial}(1 + \exp(-y_{jk} - \beta_{jk}x_{ijk}))^{-1}$. Because EEG data are noisy, making the P300 wave difficult to observe, regression coefficients $\beta_{jk}$ may be
poorly estimated. It can be desirable to assume a model for the regression coefficients that incorporates the scientific context suggesting that $\beta$ should be sparse and structured, as only $\beta_{jk}$ that correspond to time points where the P300 wave occurs are expected to be nonzero and pairs of coefficients that correspond to time points where the P300 wave occurs suggesting that regression coefficients that incorporates the scientific context poorly estimated. It can be desirable to assume a model for the

Figure 1. Estimated logistic regression coefficients $\hat{\beta}_{logit,jk}$ using the first 20 trials and their sample autocorrelations over time and sample correlations across channels.

This suggests a need for structured shrinkage priors that (i) can incorporate a priori knowledge of structure by allowing elements of $\beta$ to covary with nondiagonal prior covariance matrix $\Sigma = \nabla[\beta]$, (ii) can incorporate a priori knowledge of sparsity by allowing elementwise shrinkage of $\beta$, and (iii) span the range of common structured and sparsity inducing priors by generalizing multivariate normal and independent Laplace priors. However, existing approaches fail to satisfy all three of these criteria. We can see this by considering the normal scale-mixture representations of many common priors, which are used in Bayesian approaches to sparse regression (Polson and Scott 2010). Letting “$\omega$” be the Hadamard elementwise product, $z \sim \text{normal}(0, \Omega)$ and $s$ be a vector of stochastic scales that are independent of $z$, a prior distribution for $\beta$ has a normal scale-mixture representation if there exists a density $p(s|\theta)$ such that $\beta = s \circ z$. These priors are interpretable from a data generating perspective and have easy-to-compute moments. Specifically the prior covariance matrix of $\beta$ that encodes the a priori knowledge of structure is $\Sigma = \mathbb{E}[ss'] \circ \Omega$.

Much literature focuses on the unstructured case where $s$ is a vector of independent, identically distributed elements and $\Omega \propto I_p$. This includes the Laplace prior, bridge/exponential power priors, normal-gamma priors, Dirichlet-Laplace priors, and horseshoe priors (Park and Casella 2008; Griffin and Brown 2010; Carvalho et al. 2010; Bhattacharya et al. 2015). These priors can incorporate a priori knowledge of sparsity by modeling a separate stochastic scale for every element of $\beta$ and choosing distributions for $s$ that yield possibly sparse posterior modes. However, they do not allow elements of $\beta$ to covary. Because the marginal covariance matrix $\Sigma$ is an elementwise product of $\mathbb{E}[ss']$ and $\Omega$, elements of $\beta$ are uncorrelated a priori under these priors. The same limitation afflicts structured shrinkage priors which model elements of $s$ as correlated but continue to assume that $\Omega \propto I_p$ (van Gerven et al. 2010; Kall and Griffin 2014; Wu et al. 2014; Zhao et al. 2016; Kowal et al. 2017).

A different strand of literature includes all elliptically contoured prior distributions. It sets $s = s I_p$ and models elements of $z$ as potentially correlated with covariance matrix $\Omega$. When $s^2$ is assumed to be exponentially distributed and $\Omega \propto I_p$, the prior corresponding to the group lasso introduced in Yuan and Lin (2006) is obtained. The more general multivariate Laplace prior introduced by van Gerven et al. (2009) is obtained by assuming exponentially distributed $s^2$ and allowing arbitrary $\Omega$. These priors can incorporate a priori knowledge of structure in two ways: by treating elements of $\beta$ as grouped through their shared stochastic scale $s$ and, when $\Omega$ is allowed to take on arbitrary values, by encoding a priori knowledge of structure through specification of $\Omega$ that does not satisfy $\Omega \propto I_p$, as $\Sigma \propto \Omega$. However, as these priors only have a single shrinkage factor $s$, they shrink all elements of $\beta$ jointly and posterior modes based on these priors will only produce sparse estimates of $\beta$ for which $\beta = 0$ (Simon et al. 2013). Additionally, these priors do not generalize their independent counterparts. For instance, the multivariate Laplace prior with $\Omega = \omega^2 I_p$ does not correspond to the independent Laplace prior, and the corresponding penalty is the group lasso as opposed to the lasso penalty. This is not appropriate when only some elements of $\beta$ are expected to be sparse.

At least one set of priors can incorporate a priori knowledge of structure and sparsity by also modeling elements of the
inverse covariance matrix of $\Omega$. This includes the prior distribution that yields the fused lasso penalty $\lambda_1||\beta||_1 + \lambda_2 \sum_{j=1} ||\beta_j||_1$ and priors that correspond to more general structured penalties of the form $\beta^T Q^{-1} \beta + \lambda_1 ||\beta||_1$, where $Q^{-1}$ is positive semidefinite (Kyung et al. 2010; Ng and Abugharbieh 2011; de Brecht and Yamagishi 2012). The penalties corresponding to these priors are very popular, as they yield computationally feasible posterior mode optimization problems. Their main limitation is that relating the prior parameters $\lambda_1$ and $\lambda_2$ or $Q$ and $\lambda$ to the prior moments of $\beta$ is prohibitively challenging. This makes understanding exactly how flexible these priors are and specifying values or priors for these parameters difficult, as it is unclear how to relate the kind of a priori knowledge we might have to the prior parameter values.

One last class of relevant priors is the class of structured normal-gamma priors introduced in Griffin and Brown (2012a, 2012b), which is obtained by assuming $\beta \sim C \circ z$, where $C$ is a $p \times q$ rectangular matrix with $p < q$ and $s_i^2$ and $z_j$ are independent gamma and normal random variables, respectively. Both elementwise and structured shrinkage can be simultaneously encouraged by setting $C = [I_p, D]$, where $D$ is a $p \times (q - p)$ matrix with columns that encode groups of elements of $\beta$ that should be penalized jointly. However, the theory that justifies the use of these structured priors is specific to generalizing independent normal-gamma priors.

We construct a class of "structured Hadamard product" (SHP) priors that can incorporate a priori knowledge of structure and sparsity by allowing nondiagonal $\Omega$ and, accordingly, $\Sigma$. The main challenge in using such priors is computational; their development has been limited as a result. The marginal priors for $\beta$ are intractable integrals and correspond to penalties $-\log(\int p(\beta|\Omega, s)p(s|\theta)\, ds)$ with no simple closed form. Their use requires computationally demanding Markov chain Monte Carlo (MCMC) algorithms. Two exceptions are Finegold and Drton (2011), which develops a multivariate $t$-distribution by assuming independent inverse-gamma squared scales $s_i^2$, and Roy et al. (2021), which develops the structured product normal (SPN) prior by assuming normal scales $s \sim$ normal $(0, \Psi)$.

This article proceeds as follows. In Section 2, we describe the SPN prior and introduce the novel structured normal-gamma (SNG) and structured power/bridge (SPB) priors which generalize the independent normal-gamma priors in Caron and Doucet (2008) and Griffin and Brown (2010) and power/bridge priors in Frank and Friedman (1993) and Polson et al. (2014), respectively. We do not consider a structured generalization of the horseshoe prior because the corresponding prior covariance matrix $\Sigma = E[ss'] \circ \Omega$ is not finite, which makes it difficult to understand how a priori knowledge of structure is incorporated. We discuss properties of these priors in Section 3. Several properties and how they differ across the SPN, SNG, and SPB priors are reviewed in Table 1. In Section 4, we describe how the elliptical slice sampling method of Murray et al. (2010) can be used to overcome computational issues, regardless of the distribution assumed for elements of $s$, and discuss estimation of hyperparameters. We focus on problems where the log-likelihood of the data given $\beta$ and any additional nuisance parameters $\phi$, denoted by $-h(y|X, \beta, \phi)$, can be written as conditionally quadratic in $\beta$, that is exp$(-h(y|X, \beta, \phi)) \propto \exp\left[-\frac{1}{2}(\beta'A\beta - 2\beta'c)\right]$ for some positive definite matrix $A$ and real valued vector $c$. This includes linear regression models and certain latent variable representations of logistic and negative binomial regression models (Polson et al. 2013). In Section 5, we use SHP priors to analyze the data depicted in Figure 1. A discussion follows in Section 6.

### Table 1. Properties of specific SHP priors.

| Property | SNG | SPB | SPN |
|----------|-----|-----|-----|
| Generalizes an independent shrinkage prior | ✓ | ✓ | ✓ |
| Generalizes a Laplace prior | $c = 1$ | $q = 1$ | ✓ |
| Generalizes a normal prior | $c \rightarrow \infty$ | $q \rightarrow 2$ | ✓ |
| Infinite spike or pole at zero | $c \leq 1/2, c \neq 0$ | ✓ |
| Quadratic scale log full conditional | ✓ | ✓ | ✓ |
| Arbitrary covariance structure achievable | ✓ | ✓ | ✓ |

### 2. Structured Shrinkage Priors

We define several SHP prior distributions for $\beta$ of the form $\beta = s \circ z$, where "$\circ$" is the elementwise Hadamard product and $z \sim$ normal$(0, \Omega)$ and $s$ is a vector of stochastic scales that are independent of $z$. Because the scales of elements of $s$ are not separately identifiable from diagonal elements of $\Omega$, we parametrize $s$ such that $E[s_j^2] = 1$ for $j = 1, \ldots, p$. These priors are mean $\theta$ and have prior variance $\Sigma = E[ss'] \circ \Omega$.

**Structured Product Normal (SPN) Prior.** When $s \sim$ normal$(0, \Psi)$ and $\Psi$ is a positive definite matrix with diagonal elements equal to 1, the SPN prior is obtained. The parameters of the prior distribution for $s$, denoted by $\theta$, correspond to the off-diagonal elements of $\Psi$. Elements of the hyperparameters $\Omega$ and $\Psi$ can be related to prior moments of $\beta$. Diagonal elements of $\Omega$ correspond to prior variances of $\beta$, and off diagonal elements of $\Omega$ and $\Psi$ determine covariances and fourth-order prior cross moments of $\beta$, as shown in the appendix. Originally discussed as a sparsity inducing penalty in Hoff (2016) and later implemented as a prior distribution in Roy et al. (2021), the SPN prior is uniquely computationally simple to work with as the full conditional distributions of $z$ and $s$ are both multivariate normal distributions when the log-likelihood $-h(y|X, \beta, \phi)$ is quadratic or conditionally quadratic in $\beta$. This is described in greater detail in Section 4. The SPN prior is also appealing as it is the only prior we define that has correlations among elements of $s$. To reduce the number of freely varying parameters and to facilitate use of the SPN prior in settings where it is challenging to estimate or formulate prior opinions about fourth-order prior cross moments of $\beta$, we also define a special case of the SPN prior which we call the symmetric SPN (sSPN) prior. The sSPN prior requires that all elements of $\Psi$ be positive and that the correlation matrix corresponding to $\Omega$ has elements that are equal in magnitude to the elements of $\Psi$, that is $|\omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}}| = \psi_{ij}$. Under this constraint, $\Psi$ is a deterministic function of $\Omega$ and $\theta$ is an empty vector.

**Structured Normal-Gamma (SNG) Prior.** When the squared stochastic scales $s_i^2$ are independent gamma random variables $s_i^2 \sim$ gamma$(c, c)$ with $E[s_i^2] = 1$ for fixed $c \in (0, \infty)$
and the stochastic scales \( s_j = \sqrt{s_j^2} \) are strictly positive, the structured normal-gamma (SNG) prior is obtained. The SNG prior generalizes the normal-gamma prior of Griffin and Brown (2010), which is obtained by setting \( \Omega \propto I_p \). The fixed shape parameter \( c \) parameterizes the prior class and \( \theta \) is an empty vector. The value chosen for \( c \) determines the prior fourth order moments of \( \beta \); SNG priors with smaller values of \( c \) have lighter tails.

### Structured Power/Bridge (SPB) Prior

When the squared stochastic scales \( s_j^2 \) are independently distributed according to a polynomially tilted positive \( \alpha \)-stable distribution with index of stability \( \alpha = q/2 \) and \( \mathbb{E}[s_j^2] = 1 \) for fixed \( q \in (0, 2) \) and the stochastic scales \( s_j = \sqrt{s_j^2} \) are strictly positive, the structured power/bridge (SPB) prior is obtained. The SPB prior generalizes the bridge or exponential power prior discussed in Polson et al. (2014). The fixed shape parameter \( q \) parameterizes the prior class and \( \theta \) is an empty vector. Working with this prior is especially computationally challenging because the polynomially tilted positive \( \alpha \)-stable random variable can be represented as a rate mixture of generalized gamma random variables (Devroye 2009). A more detailed description of this representation and how it enables computation under the SPB prior is provided in the appendix.

As with the SNG prior, the value chosen for \( q \) for the SPB prior determines the prior fourth order moments of \( \beta \); SPB priors with larger values of \( q \) have lighter tails.

### Relationships Between Priors

Figure 2 shows relationships between the SPN, SNG, and SPB priors, the independent Laplace prior, and the multivariate normal prior. When \( c = q = 1 \), the SNG and SPB priors are equivalent and generalize the independent Laplace prior. In the limit as \( c \to \infty \) or \( q \to 2 \), the SNG and SPB priors generalize the multivariate normal prior. The SPN prior does not generalize the Laplace prior or the multivariate normal prior, but is equivalent to the SNG prior with \( c = 1/2 \) when \( \Psi \) and \( \Omega \) are diagonal.

### 3. Properties

#### 3.1. Univariate Prior Properties

Introducing structure does not alter the marginal prior distributions of elements \( \beta_j \). For instance, if we assume a structured prior that generalizes the independent Laplace prior for \( \beta \), for example an SNG prior with \( c = 1 \) or an SPB prior with \( q = 1 \), the marginal prior distribution of any element \( \beta_j \) is Laplace. This follows from the stochastic representation of elements \( \beta \) under these priors, \( \beta = s \circ z \). Recall that \( s \) is a vector of stochastic scales, \( z \sim \text{normal}(0, \Omega) \), \( s \) and \( z \) are independent of each other, and that all three SHP priors are obtained by assuming different distributions for \( s \). If we consider an individual element \( \beta_j = sjzj \) under the SPN or SPB priors, introducing structure does not affect \( sj \) at all and does not affect the marginal distribution of \( z_j \), as the marginal distribution of an element of a correlated normal vector is still normal. Similarly, if we consider an individual element \( \beta_j = sjzj \) under the SNG prior, introducing structure does not affect the marginal distributions of \( sj \) and \( z_j \), again because the marginal distribution of an element of a correlated normal vector is still normal.

Because introducing structure does not alter the marginal prior distributions of elements \( \beta_j \), the marginal prior distributions of elements \( \beta_j \) retain the same sparsity inducing properties of the corresponding independent priors. For instance, independent normal-gamma priors with \( c \leq 1/2 \) are known to have an infinite spike or pole at \( b_j = 0 \), which has been viewed in the literature as a sufficient condition for the recovery of sparse signals (Carvalho et al. 2010; Griffin and Brown 2010; Bhattacharya et al. 2015). Because introducing structure does not alter the marginal prior distributions of elements \( \beta_j \), the marginal prior distributions of elements \( \beta_j \) under SNG priors with \( c \leq 1/2 \) and under the SPN prior will also have an infinite spike or pole at \( b_j = 0 \). Proofs are provided in the Appendix.

#### 3.2. Joint Prior Properties

##### 3.2.1. Range of Achievable \( \Sigma \)

Introducing structure while retaining elementwise shrinkage can come at a cost. Specifically, under the SNG, SPB and SPN priors, preserving elementwise shrinkage limits how correlated elements of \( \beta \) can be. Recall that under all three SHP priors, the prior covariance matrix of \( \beta \) is \( \Sigma = \mathbb{E}[ss'] \circ \Omega \). Under the SNG and SPB priors, \( \mathbb{E}[ss'] \) is constant for fixed values of \( c \) or \( q \). Diagonal elements of \( \mathbb{E}[ss'] \) are equal to 1, whereas off-diagonal elements are less than one in absolute value. Thus, \( \mathbb{E}[ss'] \) shrinks off-diagonal elements of \( \Omega \), reducing dependence. When \( \beta \in \mathbb{R}^2 \), we can explicitly calculate the maximum marginal prior correlation \( \rho \) under the SNG and SPB priors as a function of \( c \) or \( q \). Under the SNG prior, the maximum correlation is equal to \( c^{-1}(\Gamma(c+1/2)/\Gamma(c))^2 \) and under the SPB prior, the maximum
correlation is equal to \((\pi/2)(\Gamma(2/q)/\sqrt{\Gamma(1/q)\Gamma(3/q)})^2\). When \(q = c = 1\) and both priors are equivalent, the maximum correlation is equal to \(\pi/4 \approx 0.79\).

We plot the maximum correlation as a function of kurtosis, a measure of tail behavior, under both priors in Figure 3. We observe greater reductions in the maximum correlation when the kurtosis is higher, and under the SNG prior relative to a SPB prior with equal kurtosis. The restricted range of \(\Sigma\) under the SNG and SPB priors is similar to the restricted range of the variance-covariance matrix of the alternative multivariate t-distribution introduced in Finegold and Drton (2011). Intuitively, the restricted range of \(\Sigma\) relates to the conflict that arises between the properties of the marginal joint density \(p(\beta|\Omega,\Theta)\) needed to preserve elementwise shrinkage, specifically concentration of the density along the axes, and the properties of the marginal joint density needed to encourage structure, for example concentration of the density along the 45 degree line line when \(p = 2\) and elements of \(\beta\) are expected to be similar to each other.

In contrast, the unrestricted SPN prior can accommodate an arbitrary prior covariance \(\Sigma\). Given a positive semidefinite prior covariance \(\Sigma\), it is always possible to find at least one pair of positive semidefinite covariance matrices \(\Omega\) and \(\Psi\) that satisfy \(\Sigma = \Omega \circ \Psi\) (Styan 1973). The sSPN prior is less flexible. It is easy to simulate a positive semidefinite covariance matrix \(\Sigma\) for which the corresponding values of \(\Omega\) or \(\Psi\) satisfying \(|\omega_{ij}| = |\psi_{ij}|\) are not positive semidefinite, but challenging to explicitly characterize the class of covariance matrices \(\Sigma\) that correspond to nonpositive semidefinite values of \(\Omega\) or \(\Psi\).

### 3.2.2. Copulas

Even when all three SHP priors share the same prior covariance matrix \(\Sigma\), the induced dependence structures vary widely. We compare the induced dependence structures by considering \(\beta \in \mathbb{R}^2\) with unit marginal variances and correlation \(\rho = 0.5\), and examining corresponding copula densities. Let \(F_{j}^{\text{SNG,1}}(\beta_j)\) refer to the marginal prior CDF of \(\beta_j\) corresponding to one of the SHP prior distributions. We can always write \(\beta_j \overset{d}{=} F_j^{-1}(u_j)\), where \(u_j\) is a random variable with uniform margins. The joint distribution of the \(p \times 1\) vector \(\mathbf{u}\) is called the copula of \(\beta\), and it characterizes the induced dependence structure. Even if the inverse CDFs \(F_j^{-1}(u_j)\) are not known, the copula density can be approximated by simulating values of \(\beta\), transforming simulated values of \(\beta_1\) and \(\beta_2\) into percentiles \(u_1\) and \(u_2\), and computing a kernel bivariate density estimate of the percentiles. Figure 4 shows numerical approximations to copula densities for several SHP priors with unit marginal prior variances and marginal prior correlation \(\rho = 0.5\).

The first four panels in the top and bottom rows show copula densities under SNG and SPB priors with increasingly heavy tails, and the final panels on the top and bottom show copula densities under SPN priors with \(\omega_{12} = \rho^{0.5}\) and \(\omega_{12} = \rho^{0.1}\). The parameters of the SPB priors have been chosen to ensure that the kurtosis, a measure of tail behavior, of each SPB prior is equal to the kurtosis of the SNG prior above. For example, the SNG prior with \(c = 0.3\) has the same kurtosis as the SPB prior with \(q = 0.65\).

The dependence structures induced by the SNG and SPB priors are similar. As \(c\) or \(q \to 0\) and the SNG and SPB priors become less normal and heavier tailed, the copulas display increasingly strong orthant dependence. This means that as \(c\) or \(q \to 0\), the priors concentrate more strongly around values of \(\beta\) that have the same sign. The SNG prior displays especially strong orthant dependence and appears to converge to a uniform distribution over the positive and negative orthants as \(c \to 0\). Additionally, as \(c\) or \(q \to 0\) the copula densities become more concentrated around the axes where at least one element of \(\beta\) is nearly or exactly equal to zero, suggesting that these priors are still encouraging elementwise shrinkage.

Like the SNG and SPB priors, the SPN priors concentrate in the upper-right and lower-left corners, where \(\beta_1\) and \(\beta_2\) have the same sign and are large in magnitude. However, in comparison to the SNG and SPB priors, both SPN priors also concentrate strongly at the origin, where \(\beta_1 = \beta_2 = 0\), and less strongly along the axes, where \(\beta_1 = 0\) or \(\beta_2 = 0\). Both SPN priors also concentrate more about values of \(\beta\) that are equal in magnitude but opposite in sign. Intuitively, this is due to the fact that \(\beta\) is made up of one strongly correlated component and one very weakly correlated component, both of which can take any value on \(\mathbb{R}\) when \(\omega_{12} = \rho^{0.1}\) and \(\psi_{12} = \rho^{0.9}\).

#### 3.2.3. Conditional Prior Distributions

We can explore how introducing structure can inform shrinkage of elements of \(\beta\) by examining the induced conditional prior distributions. Figure 5 shows conditional prior distributions \(p(\beta_1|\beta_2,\Omega,\Theta)\), for \(\beta \in \mathbb{R}^2\) with unit marginal prior variance, marginal prior correlations \(\rho = \{0,0.5\}\), and \(\beta_2 = \{0,2\}\). Again, the first four panels in the top and bottom rows show conditional prior distributions under SNG and SPB priors with increasingly heavy tails, and the final panels on the top and bottom show conditional prior distributions under SPN priors with \(\omega_{12} = \rho^{0.5}\) and \(\omega_{12} = \rho^{0.1}\). The parameters of the SPB priors have been chosen to ensure that the kurtosis, a measure of tail behavior, of each SPB prior is equal to the kurtosis of the SNG prior above it, for example, the SNG prior with \(c = 0.3\) has the same kurtosis as the SPB prior with \(q = 0.65\).

Introducing structure via a positive prior correlation between \(\beta_1\) and \(\beta_2\) can allow us to use knowledge of \(\beta_2\) to better estimate \(\beta_1\). We see that sparsity inducing SHP priors concentrate more strongly about zero than their independent counterparts when \(\beta_2 = 0\), and shift their mass toward positive nonzero values of \(\beta_1\) when \(\beta_2 = 2\). The conditional priors reflect differing amounts of
Figure 4. Approximate copula densities for $\beta \in \mathbb{R}^2$ with unit marginal prior variances and marginal prior correlation $\rho = 0.5$. Copula density approximations were made by simulating 1,000,000 values from the corresponding prior, transforming simulated values of $\beta_1$ and $\beta_2$ into percentiles $u_1$ and $u_2$, and computing a kernel bivariate density estimate of the percentiles.

Figure 5. Conditional prior distributions $p(\beta_1 | \beta_2, \rho, \theta)$ for $\beta \in \mathbb{R}^2$ with unit marginal prior variances, marginal prior correlation $\rho \in \{0, 0.5\}$ and $\beta_2 \in \{0, 2\}$ obtained via simulation.

origin versus tail dependence. The nearly normal SNG and SPB priors on the far left display relatively more tail dependence, as the structured conditional priors for $\beta_1$ differ more from their independent counterparts when $\beta_2 = 2$ than when $\beta_2 = 0$. At the other end of the spectrum, the heavier-than-Laplace tailed SNG and SPB priors and the sSPN prior display more origin dependence than tail dependence, as the structured conditional priors for $\beta_1$ concentrate much more strongly about $\beta_1 = 0$ when $\beta_2 = 0$ than their independent counterparts, and still have a mode near $\beta_1 = 0$ even when $\beta_2 = 2$. This is especially striking under the SNG priors with $c \leq 1/2$ and the sSPN prior, which retain a very sharp peak at $\beta_1 = 0$ even when $\beta_2 = 2$. The asymmetric SPN prior with $\omega_{12} = \rho^{0.1}$ is in between; the full conditional distribution of $\beta_1$ concentrates much more strongly about $\beta_1 = 0$ when $\beta_2 = 0$ and shifts markedly to toward $\beta_1 = 2$ when $\beta_2 = 2$.

3.2.4. Joint Marginal Prior Contours

The joint marginal prior contours, which can be interpreted as contours of a penalty function, provide a powerful tool for studying the properties of different priors. Figure 6 shows Monte Carlo approximations of the contours of the joint marginal priors $\int p(\beta | s, \Omega) p(s | \theta) ds$, for $\beta \in \mathbb{R}^2$ with marginal prior variance-covariance matrix $\Sigma = (1 - \rho) I_2 + \rho 1_2 1_2'$ and marginal prior correlation $\rho = 0.5$.

Again, the first four panels in the top and bottom rows show contours of SNG and SPB priors with increasingly heavy tails. The parameters of the SPB priors have been chosen to ensure that the kurtosis, a measure of tail behavior, of each SPB prior is equal to the kurtosis of the SNG prior above it, for example the SNG prior with $c = 0.3$ has the same kurtosis as the SPB prior with $q = 0.65$. The rightmost panels show contours of SPN priors. The top panel is a symmetric sSPN prior with $\omega_{12} = \rho^{0.5}$.
and $\psi_{12} = \rho^{0.5}$, whereas the bottom panel is an asymmetric SPN prior with $\omega_{12} = \rho^{0.1}$ and $\psi_{12} = \rho^{0.9}$.

All three SHP priors encourage similar estimates of $\beta_1$ and $\beta_2$ by pushing contours away from the origin when $\beta_1$ and $\beta_2$ have opposite signs. The SPN, SNG with $c \leq 1$ and SPB with $q \leq 1$ priors encourage sparse estimates of $\beta$ by retaining discontinuities of the log marginal prior on the axes. The contours do not necessarily keep the same shape as the value of prior density changes. Contours closer to the origin are more similar to their independent counterparts, with relatively more encouragement of sparsity than structure, whereas contours farther from the origin tend to encourage relatively more structure and less sparsity. This is especially evident under the SPN prior with $\rho_0 = \rho^{0.1}$. It is clear from the contours that these priors are not log-concave when $\rho \neq 0$ and correspond to nonconvex penalties for SNG priors with $c \leq 1$, SPB priors with $q \leq 1$, and SPN and sSPN priors. Under the SNG priors with $c \leq 1/2$ and the SPN priors, the contours do not cross the axes. The joint marginal distribution of $\beta$ under the SNG prior with $c \leq 1/2$ has an infinite spike or pole along the axes, where at least one element is equal to 0.

**Proposition 3.1.** For the SNG prior with $c \leq 1/2$ and any $b_j = 0$, then $p(b|c, \Omega) = +\infty$.

This is a multivariate analogue of the univariate marginal prior’s spike or infinite pole at $b = 0$ under the SNG prior with $c \leq 1/2$. The joint marginal SPN prior behaves similarly.

**Proposition 3.2.** For the SPN prior, if any $b_j = 0$, then $p(b|\Psi, \Omega) = +\infty$.

Proofs are given in the Appendix. This suggests that the SPN and SNG priors with $c \leq 1/2$ may recover sparse signals well (Carvalho et al. 2010; Griffin and Brown 2010; Bhattacharya et al. 2015). However, the presence of an infinite spike or pole at $b_j = 0$ also makes the already challenging problem of computing the posterior mode intractable, as the log marginal prior is not only nonconvex but also has infinitely many modes.

### 3.3. Posterior Properties

One last perspective on the properties of all three SHP priors can be gained by examining how the posterior mode of $\hat{\beta}$ relates to the unpenalized OLS estimate $\hat{\beta}_{OLS}$. This can offer an intuitive understanding of the properties of estimators obtained under the three different priors. In the sparsity inducing penalty literature this is often characterized by the thresholding function, which is defined in the context of linear model for $y$ with a single covariate $x$ that satisfies $x'x = 1$ as

$$
\hat{\beta} = \arg\min_{\beta} \frac{1}{2\phi^2} (\hat{\beta}_{OLS} - \beta)^2 - \log \left( \int p(\beta|s, \omega) p(s|\theta) \, ds \right).
$$

(1)

where $\phi^2$ refers to assumed variance of deviations of $\hat{\beta}_{OLS} - \beta$. For example, (1) gives the soft-thresholding function when $\beta$ has a mean-zero Laplace distribution.

Because the advantage of working with any of the three SHP priors is the ability to introduce a priori dependence across elements of vector valued $\beta$, examining the corresponding univariate thresholding functions for scalar $\beta$ by given by (1) does not fully explore how the posterior mode of $\beta$ relates to the unpenalized OLS estimate $\hat{\beta}$. Accordingly, we define and examine a bivariate thresholding function. We continue to assume a linear model for $y$, but assume an orthogonal design matrix comprised of two covariates $X$ that satisfies $X'X = \text{diag}(1_2)$ instead of a single covariate $x$. The bivariate thresholding function relates the posterior mode of both $\beta_1$ and $\beta_2$ to the noise variance $\phi^2$, prior parameters $\Omega$ and $\theta$, and OLS estimates $\hat{\beta}_{OLS,1}$ and $\hat{\beta}_{OLS,2}$, and is given by

$$
\hat{\beta} = \arg\min_{\beta} \frac{1}{2\phi^2} \left\| \beta_{OLS} - \beta \right\|_2^2 - \log \left( \int p(\beta|s, \omega) p(s|\theta) \, ds \right).
$$

(2)

The bivariate thresholding function can be approximated using a Gibbs-within-EM algorithm described in the Section 4. Figure 7 shows approximate bivariate thresholding functions for $\beta \in \mathbb{R}^2$ with unit marginal prior variances, marginal prior correlation $\rho = 0.5$, noise variance $\phi^2 = 0.1$, $\hat{\beta}_{OLS,2} \in \{-0.5, 0, 0.5, 1\}$ and $\hat{\beta}_{OLS,1} \in [0, 1]$, computed from 1,001,000 iterations.
Gibbs sampler iterations, with the first 1000 samples discarded as burn-in.

Examination of the bivariate thresholding functions confirms that the SPN prior, SNG prior with \( c \leq 1 \) and SPB prior with \( c \leq 1 \) can yield possibly sparse posterior mode estimates of \( \beta \), and that the introduction of structure encourages or discourages sparse estimates \( \beta_1 \) depending not only on the observed value of \( \beta_{OLS,1} \) but also the observed value of \( \beta_{OLS,2} \). There are a few interesting trends that relate to previously identified properties of the SPN, SNG and SPB priors. Both SPN priors shrink \( \beta_1 \) toward 0 less aggressively when \( \beta_{OLS,2} = -0.5 \) than when \( \beta_{OLS,2} = 0 \), which reflects the tendency of the SPN priors to encourage estimates of \( \beta \) with similar absolute magnitudes. Additionally, the value of \( \beta_{OLS,2} \) appears to affect the estimate \( \beta_1 \) more when \( \beta_{OLS,1} \) is smaller under the SPN prior, SNG prior with \( c \leq 1 \) and SPB prior with \( c \leq 1 \). This is consistent with what we observed when examining the copulas and conditional prior distributions; as the tails become heavier, the Laplace and heavier-than-Laplace prior distributions display more dependence at the origin than the tails. Last, the SNG prior with \( c = 0.3 \) not only induces especially strong shrinkage of \( \beta_1 \) when \( \beta_{OLS,2} \) is small, but also induces aggressive inflation of \( \beta_1 \) when \( \beta_{OLS,1} \) is small and \( \beta_{OLS,2} \) is very large. This makes sense in the context of the prior conditional distributions under the SNG and SPB priors with \( c = 0.3 \) and \( q = 0.65 \). Although these two priors have equally heavy tails, the SNG prior is much more concentrated at the origin and has heavier tails.

4. Computation

4.1. Posterior Approximation

The posterior mode of \( \beta \) maximizes the integral
\[
\int p(y | X, \beta, \phi)p(\beta | \Omega, s)p(s | \theta)ds \quad \text{over} \quad \beta
\]
and can be thought of as a penalized estimate of \( \beta \), where the penalty is given by
\[
-\log(\int p(\beta | \Omega, s)p(s | \theta)ds).
\]
This integral is generally intractable when \( \Omega \) is not diagonal, but can be maximized using an MCMC within Expectation Maximization (EM) algorithm (Dempster et al. 1977). Given an initial value \( \beta^{(0)} \), this algorithm proceeds by iterating the following until \( ||\beta^{(i+1)} - \beta^{(i)}||_2 \) converges:

- using MCMC to simulate \( M \) draws \( s^{(1)}, \ldots, s^{(M)} \) from the full conditional distribution of \( s \) given \( \beta^{(i)} \) and \( \theta \), set
  \[
  \hat{E}(\frac{1}{N}) \{ s^{(i)} | \beta^{(i)}, \theta \} = \frac{1}{M} \sum_{j=1}^{M} (\frac{1}{N}) \{ s^{(j)} | \beta^{(i)}, \theta \};
  \]
- set
  \[
  \beta^{(i+1)} = \arg\min_{\beta} h(y | X, \beta, \phi) + \frac{1}{2} \beta'(\Omega^{-1} \circ \hat{E}(\frac{1}{N}) \{ s^{(i)} | \beta^{(i)}, \theta \}) \beta.
  \]

Alternative posterior summaries can be obtained by simulating \( M \) values \( \beta^{(1)}, \ldots, \beta^{(M)} \) from the joint posterior distribution of \( (\beta, s) \) given \( y, X, \phi \) and \( \theta \), starting with an initial value \( \beta^{(0)} \) and iteratively simulating \( s^{(i)} \) from the full conditional distribution of \( s \) given \( \beta^{(i)} \) and \( \theta \) and simulating \( \beta^{(i+1)} \) from the full conditional distribution of \( \beta \) given \( y, X, s^{(i)}, \phi \). When the log-likelihood is quadratic or conditionally quadratic in \( \beta \), the full conditional distribution \( p(\beta | s, y, X, \phi) \) is a multivariate normal distribution.

The full conditional distribution \( p(s | \beta, \Omega, \theta) \) can be written as proportional to
\[
p(s | \beta, \Omega, \theta) \propto \left( \prod_{j=1}^{p} |\Sigma_j|^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} (1/s_j) (\Omega^{-1} \circ (\beta \beta') (1/s)) p(s | \theta) \right] \right)
\]
where “/” is applied elementwise. The choices \( p(s | \theta) \) that yield the SPN, SNG, and SPB models do not yield standard distributions when \( \Omega \) is not a diagonal matrix.

We simulate from the full conditional distribution (3) using generalized elliptical slice sampling (Nishihara et al. 2014). First, we unify all three models by recognizing that simulating \( s \) according to (3) is equivalent to simulating unconstrained \( s' \) according to
\[
p(s' | \beta, \Omega, \theta) \propto \left( \prod_{j=1}^{p} |\Sigma_j|^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} (1/f (s')) (\Omega^{-1} \circ (\beta \beta')) (1/f (s')) p(f (s') | \theta) \right] \right)
\]
and setting \( s = f(s') \), where \( f(s') = s' \) when the SPN prior is used and \( f(s') = |s'| \) when the SNG or SPB priors are used. Such use of the absolute value transformation is common in the generalized elliptical slice sampling literature (Nishihara et al. 2014). Then we perform a change of variables from \( s' \) to \( s'' = V^{-1/2} (s' - m) \), where \( m \) and \( V \) may depend on the current values of \( \beta, \Omega, \) and \( \theta \) but not \( s \) and \( V^{-1/2} \) refers to the symmetric square root of the pseudoinverse \( V' \) of \( V \). Applying a change of variables when simulating from conditional distributions is a standard practice, as seen in (Murray et al. 2010).

Simulating \( s \) according to (3) is equivalent to simulating \( s'' \) according to

\[
p(s''|\beta, \Omega, \theta) \propto |\text{diag} \left( |V^{1/2}s'' + m| \right)|^{-1} \exp \left\{ -\frac{1}{2} \left( 1/f \left( V^{1/2}s'' + m \right) \right)' \left( \Omega^{-1} \circ (\beta\beta') \right) \left( 1/f \left( V^{1/2}s'' + m \right) \right) \right\} \times p(f (V^{1/2}s'' + m) | \theta)
\]

and setting \( s = f (V^{1/2}s'' + m) \), where “/” and “|” are applied elementwise. In turn, simulating \( s'' \) according to (5) is equivalent to simulating four \( p \times 1 \) vectors \( r, u, \) and \( \pi \) according to

\[
p(s'' = u \sin(\pi) + r \cos(\pi) | \beta, \Omega, \theta) \quad \prod_{j=1}^{p} \exp \left\{ -\frac{1}{2} \left( r_j^2 + u_j^2 \right) \right\} / \exp \left\{ -s_j^2/2 \right\}.
\]

For fixed \( m \) and \( V \), iteration \( i \) of a Gibbs sampler simulating from (6) is as follows:

1. Simulate \( r_{ij}^{(i)} \) \( \sim \) normal\((0, 1) \) for \( j = 1, \ldots, p \).
2. Set \( u_{ij}^{(i)} = (s'^{(i-1)} - t_i \cos(\pi^{(i-1)})) \).
3. Set \( r_{ij}^{(i)} = (s'^{(i-1)} - t_i \sin(\pi^{(i-1)})) \).
4. Simulate \( \pi_j^{(i)} \) according to (6) for \( j = 1, \ldots, p \) using univariate slice sampling.
5. Set \( s''_{ij}^{(i)} = u_{ij}^{(i)} \sin(\pi_{ij}^{(i)}) + r_{ij}^{(i)} \cos(\pi_{ij}^{(i)}) \).
6. Set \( s_{ij}^{(i)} = f (V^{1/2}s''_{ij}^{(i)} + m) \).

Univariate slice sampling is described in the appendix. We can think of steps 1–4 as generating a proposal for \( s''^{(i)} \) and step 5 as generating an adjustment for the proposal that reflects the relationship between the proposal distribution and the target full conditional distribution. Validity of the proposed algorithm follows from Murray et al. (2010). The vector \( m \) and matrix \( V \) can be chosen to improve mixing and decrease computational burden. Choices of \( m \) and \( V \) that provide better approximations to the full conditional distribution \( p(s|\beta, \Omega, \theta) \) are likely to result in better mixing, whereas choices of \( m \) and \( V \) that are easier to compute will improve the speed of the sampling process. Here, \( m \) and \( V \) are chosen to approximate the mode of \( p(s'|\beta, \Omega, \theta) \) and corresponding Hessian. We compute \( m \), an approximate mode of \( p(s'|\beta, \Omega, \theta) \), by performing coordinate descent with a large convergence threshold and a small number of maximum iterations. We describe how \( m \) can be obtained via coordinate descent under the SHP priors in the appendix.

### 4.1.1. Simulating from the Posterior under the SPN Prior

As noted in Section 2, simulation from the joint posterior distribution under the SPN prior is simple. In the linear regression setting with known unit variance where \( -h(y|X, \beta) \propto \frac{1}{2} (\beta'X'X\beta - 2\beta'X'y) \), then

\[
z|X, y, s, \Omega \sim \text{normal} \left( \left( (X'X) \circ (ss') + \Omega^{-1} \right)^{-1} (s \circ (X'y)), \ \left( (X'X) \circ (ss') + \Omega^{-1} \right)^{-1} \right),
\]

\[
s|X, y, z, \Psi \sim \text{normal} \left( \left( (X'X) \circ (zz') + \Psi^{-1} \right)^{-1} (z \circ (X'y)), \ \left( (X'X) \circ (zz') + \Psi^{-1} \right)^{-1} \right).
\]

Both full conditional distributions are multivariate normal distributions. As a result, a straightforward Gibbs sampler can be used to simulate from the joint posterior distribution.

### 4.2. Hyperparameter Estimation

In the previous section, we presented a general approach for simulating from the posterior distribution of \( \beta \) under the SPN, SNG and SPB priors given hyperparameters \( \Omega \) and, in the case of the SPN prior, \( \Psi \). When the hyperparameters are unknown, one approach is to assume prior distributions for \( \phi, \Omega, \) and/or \( \Psi \). For all three priors, a conjugate inverse-Wishart prior for \( \Omega \) is a natural choice. For the SPN prior, a conjugate inverse-Wishart prior for \( \Psi \) is likewise natural. There are situations where conjugate inverse-Wishart priors may have undesirable properties (Schuurman et al. 2016), in which case alternative prior distributions for \( \Omega \) and, when the SPN prior is used, \( \Psi \), could be used.

Alternatively, hyperparameter estimates can be obtained via maximum marginal likelihood estimation (MMLE) or the method of moments. We can compute the MMLE of the unknown variance components \( \Omega \) and, in the case of the SPN prior, \( \Psi \), using a Gibbs-within-EM algorithm as described in the Appendix. However, maximum marginal likelihood estimation of hyperparameters can converge prohibitively slowly in practice (Roy and Chakraborty 2016). Furthermore, the Gibbs step can be prohibitively computationally demanding when the data are high dimensional. Fortunately, method of moments type estimates of the unknown variance components can be obtained under the SNG and SPB priors for fixed \( c \) and \( q \), respectively, and under the symmetric sSPN prior, so long as \( y \) is linearly related to \( X\beta \). As noted in the Introduction, the prior moments are easy to compute under all three priors

\[
\mathbb{E} [\beta'] = \mathbb{E} [s'] \circ \mathbb{E} [z] \quad \text{and} \quad \Sigma = \mathbb{E} [ss'] \circ \mathbb{E} [zz'].
\]

Furthermore, under the sSPN, SNG, and SPB priors, the hyperparameters are correspond to second order moments of \( \beta \). When \( y \) is linearly related to \( X\beta \), a positive semidefinite estimate of \( \Sigma \) can be obtained using methods from Perry (2017). Under the sSPN prior, estimates of \( \Omega \) and \( \Psi \) can be obtained from an estimate \( \hat{\Sigma} \) by projecting \( \sqrt{\hat{\Sigma}} \) and \( \text{sign}(\hat{\Sigma})\sqrt{\hat{\Sigma}} \) onto the positive semidefinite cone, where \( \sqrt{\cdot}, | \cdot | \) and \( \text{sign}(\cdot) \) are applied elementwise. Under the SNG and SPB priors, an
estimate of $\Omega$ can be obtained from an estimate $\hat{\Sigma}$ by projecting $\hat{\Sigma}/(1 - \mathbb{E}[\Sigma_1]) I_p + \mathbb{E}[\Sigma_1]I_p I_p^\top$ onto the positive semidefinite cone, where “/” is applied elementwise, $\mathbb{E}[\Sigma_1]^2 = c^{-1}(\Gamma(c + 1/2)/\Gamma(c))^2$ under the SNG prior and $\mathbb{E}[\Sigma_1]^2 = (\pi/2)(\Gamma(2/q)/\sqrt{\Gamma(1/q)\Gamma(3/q)})^2$ under the SPB prior.

4.3. Prior Selection

Our focus is on understanding the properties of each novel prior and providing a method for using them in practice. However, we would be remiss to omit guidance on when SNG, SPB, and SPN priors should be used as opposed to more common existing alternatives, how to choose between the SNG, SPB, and SPN priors when their use is warranted, and, in the case of the SNG and SPB priors, how to choose $c$ or $q$.

Ideally, our understanding of the scientific problem would inform our choice of prior. Alternatively, we recommend first choosing a criterion for selecting a prior that is appropriate to the specific application and goals of analysis, for example, Deviance Information Criterion (DIC), the widely applicable information criterion (WAIC), log marginal likelihood, mean log conditional predictive ordinate (CPO)/leave-one-out cross-validation error (LOO) (Roos and Held 2011; Piironen and Vehtari 2017; Vehtari et al. 2017). We then recommend performing an initial analysis that compares the chosen criterion under an independent normal, multivariate normal, and independent sparsity inducing prior of interest, for example, a Laplace prior. If both the multivariate normal and independent sparsity inducing prior provide improved fit compared to an independent normal prior, we recommend considering the SPN, SNG, and/or SPB priors for a range of values of $c$ and/or $q$ and selecting the prior that is best according to the chosen criterion.

5. Application

In this section we return to the subset of BCI data discussed in Section 1. We assume the following model for the indicators of whether or not the subject was viewing the target letter $y$ and EEG measurements at time point $k$ and channel $j$. $x_{jk}$ during the first 20 trials,

$$y_i \sim \text{Binomial}(1 + \exp(-y_{jk} - \beta_{jk}x_{jk}))^{-1},$$

where each $y_{jk}$ is a time point and channel specific intercept and each $\beta_{jk}$ is a regression coefficient that describes the relationship between EEG measurements from time point $j$ and channel $k$ and whether or not the subject was viewing the target letter, for $j = 1, \ldots, p_1$ and $k = 1, \ldots, p_2$. Because simulating from the posterior distribution of $\beta$ becomes increasingly computationally demanding as the number of unknown regression coefficients $\beta_{jk}$ increases and because we want to explore the results of using several different SNG and SPB priors, we consider a reduced set of time points, specifically every eighth time point. This reduces $p_1$ from 208 to 26. We continue to consider all channels, $p_2 = 8$.

Letting $\beta = \text{vec}(B)$ refer to a vectorized $p_1 \times p_2$ matrix $B$ with elements $b_{jk} = \beta_{jk}$, we assume that $\mathbb{E}[\beta] = 0$ and $\mathbb{V}[\beta] = \Sigma$, where $\Sigma$ is separable, that is, $\Sigma = \Sigma_2 \otimes \Sigma_1$. The covariance matrices $\Sigma_1$ and $\Sigma_2$ characterize relationship of regression coefficients $B$ over time and across channels, respectively. Because the scales of $\Sigma_1$ and $\Sigma_2$ are not separately identifiable, we assume that the diagonal elements of $\Sigma_1$ are exactly equal to 1. We also assume that $\Omega_1$ and $\Psi_1$ have autoregressive structures of order one, with $\omega_{1ij} = [i-j]$ and $\psi_{1ij} = [i-j]$. We note that the corresponding marginal variance $\Sigma_2$ is autoregressive of order one under the SPN prior but not the SNG and SPB priors. If $\Omega_1$ and $\Psi_1$ have autoregressive structure of order one with parameters $\rho_{12}$ and $\rho_{1q}$, then $\Sigma_1$ has autoregressive structure of order one with parameter $\rho = \rho_{1q}$. In contrast, the matrix $\Sigma_1 = \Omega_1 \circ \mathbb{E}[ss^\top]$ does not have autoregressive structure of order one.

We assume $\rho_{12} \sim \text{beta}((p_1 + 1)/2, (p_1 + 1)/2)$ and $\Omega_2^{-1} \sim \text{Wishart}(p_2 + 2, \kappa^{-1} I_{p_2})$, where $\kappa = \frac{1}{p_2 p_2^\top} \sum_{j=1}^{p_1} \sum_{k=1}^{p_2} (\hat{\beta}_{\logit,jk} - \bar{\hat{\beta}}_{\logit})^2$, $\bar{\hat{\beta}}_{\logit} = \frac{1}{p_2 p_2^\top} \sum_{j=1}^{p_1} \sum_{k=1}^{p_2} \hat{\beta}_{\logit,jk}$, and $\hat{\beta}_{\logit,jk}$ are the logistic regression estimates of $\beta_{jk}$ depicted in Figure 1. When using the SPN prior, we also assume $\Psi_2^{-1} \sim \text{Wishart}(p_2 + 2, I_{p_2})$ and $\rho_{1q} \sim \text{beta}((p_1 + 1)/2, (p_1 + 1)/2)$. We assume conjugate inverse-Wishart priors for convenience, as our primary goal is to compare estimates under the SNG, SPB, and SPN priors for $B$. We use the latent variable representation of the logistic regression model introduced in Polson et al. (2013), simulate from the full conditional distribution of $\beta$ jointly, use the elliptical slice sampling procedure described in Section 4 to simulate from the full conditional distribution of $s$, and use univariate slice sampling to simulate from the full conditional distributions of $\rho_{12}$ and $\rho_{1q}$. We simulate 20 chains of 13,500 samples from the posterior distribution, discard the initial 1000 from each chain as burn-in, and thin by a factor of 5. The remaining 50,000 samples have minimum effective sample sizes shown in the appendix.

Figure 8 shows posterior mean estimates of elements of $B$ under a multivariate normal prior (SNO) and nine SHP priors. Because shrinkage priors are often used to select nonzero elements of $B$, approximate 90% intervals that do not include zero are also indicated. In cases where a sparse point estimate of $B$ is desired, the methods of Hahn and Carvalho (2015), Li and Pati (2017), Piironen et al. (2020), or Woody et al. (2021) can alternatively be used. The estimated posterior means and approximate 90% intervals that do not include zero display similar structure across channels and over time. They also show increasing shrinkage of individual elements of $B$ as $c$ or $q \to 0$. Even the SNG and SPB priors with $c = 0.3$ and $q = 0.65$, which encourage sparsity very aggressively, produce estimates are structured, that is, estimates of elements of $B$ corresponding to the same channel or time point tend to be similar regardless of whether they are nearly equal to zero or large in magnitude. In comparison to the normal prior, the more aggressive sparsity inducing priors facilitate identification of especially strong signals; for instance, they suggest that whether or not the subject is viewing the target letter is strongly correlated with EEG measurements at the 5th and 12th time points on nearly all of the channels.

Figure 9 shows posterior mean estimates of channel-by-channel correlation matrices corresponding to $\Sigma_2$ and indications of approximate 90% intervals that do not include zero.
as well as posterior mean estimates of correlations across consecutive time points \(\sigma_{1,i(i+1)}\) and approximate 90% intervals. There is evidence of dependence, especially across channels. The amount of dependence across channels and over time under the SNG and SPB priors is decreasing as \(c\) or \(q\) \(\to\) 0. Surprisingly, the SPN prior estimates weak correlations despite being able to accommodate arbitrary \(\Sigma\). This may be due to especially strong shrinkage of elements of \(\Sigma\) under the assumed prior distributions for the hyperparameters.

The available prior information, which suggests that \(B\) is sparse with strong signals observed across several channels at a subset of time points that roughly correspond to the expected timing of the P300 wave, can be used to assess the relative merits of the different priors. The SNG and SPB priors with \(c = 10\) and \(q = 1.75\), respectively, produce estimates that are not sparse and so similar to the estimates produced under a normal prior that there is no added benefit relative to a multivariate normal prior. Meanwhile, the SNG prior with \(c = 0.3\) and the SPN prior...

Figure 8. Approximate posterior means of \(B\). Boxes enclosing cells indicate approximate 90% intervals that do not include zero.

Figure 9. Approximate posterior mean channel-by-channel correlation matrices \(\Sigma_2\), with boxes enclosing cells with approximate 90% intervals that do not include zero, and correlations of consecutive time points \(\sigma_{1,12}\) with approximate 90% intervals.
produce correlations that are not distinguishable from zero and thus provide no added benefit relative to an independent sparsity inducing prior and are implausibly sparse. This leaves the SNG and SPB priors with \( c = 1 \) and \( q = 1 \), which are equivalent and generalize the independent Laplace prior, the SNG prior with \( c = 0.5 \) and the SPB priors with \( q = 0.78 \), and \( q = 0.65 \). The SNG prior with \( c = 0.5 \) misses signals that occur later in each trial that are likely to correspond to the known timing of the P300 wave. The SPB priors with \( q = 0.78 \) and \( q = 0.65 \) provide estimates that are qualitatively very similar to the estimates obtained under the SPB prior with \( q = 1 \). Accordingly, we recommend the multivariate Laplace prior, which corresponds to the SNG prior with \( c = 1 \) and the SPB prior with \( q = 1 \); it is the simplest SHP prior that produces estimates that align with the available scientific context.

For the purposes of demonstration, we also consider three criteria for prior selection: WAIC, DIC, and LOO. Approximations to these criteria are shown in Figure 10. Regardless of the criteria considered, the sparsity inducing independent product normal prior and the structured multivariate normal prior both outperform the independent normal prior and therefore suggest that SHP priors are worth considering. All structured priors outperform their independent counterparts. A detailed comparison of the estimates of \( B \) obtained under the structured priors we consider and their independent counterparts is included in the Appendix. Among structured priors, less sparsity inducing priors perform better and the multivariate normal prior performs best in this specific setting according to these criteria.

To conclude, we compare approximate 90% intervals that do not include zero to “ground truth” 90% confidence intervals that do not include zero based on simple logistic regression estimates computed from all 240 trials. Table 2 shows that normal and nearly normal priors provide the best true positive rates, while the multivariate Laplace prior and SPB priors with \( q < 1 \) provide the best true positive rates relative to false positive rates.

6. Discussion

We introduce the SHP class of novel structured shrinkage priors for regression coefficients and show that they can encourage both sparsity and structure, which can be difficult to simultaneously model using existing prior distributions. We provide a parsimonious and general approach to simulation from the posterior distributions under SHP priors based on elliptical slice sampling and demonstrate how SHP priors can improve interpretability of estimated regression coefficients relative to multivariate normal or independent priors.

We have focused on the development of structured shrinkage prior distributions for regression coefficients, however, the same distributions could be used to model errors as in Finegold and Drton (2011) or as alternatives to Gaussian processes. We could also construct shrinkage priors for covariance matrices by extending Daniels and Pourahmadi (2002) or building on the matrix-\( F \) prior distribution for covariance matrices introduced in Mulder and Pericchi (2018), which assumes that \( \Sigma \sim \mathcal{W} \mathcal{S} \), where \( \mathcal{W} \) has an inverse-Wishart distribution with scale \( I_p \) and degrees of freedom \( \delta + k - 1 \) and \( \mathcal{S} \) has a Wishart distribution with scale \( \Xi \) and degrees of freedom \( \nu \). We could define distributions for covariance matrices according to \( \Sigma \sim \mathcal{W} \mathcal{S} \mathcal{O} \), for \( s \sim \text{normal}(0, \Psi) \), \( s \sim \text{Gamma}(c, c) \), or \( s \sim \text{I} \text{ind} \text{Gamma}(c, c) \), independently distributed according to a polynomially tilted positive \( \alpha \)-stable distribution with index of stability \( \alpha = q/2 \) and \( \mathbb{E}[s^2] = 1 \). Also, the elliptical slice sampling procedure we use to construct a tractable Gibbs sampler for simulating from \( p(s|\beta, \theta) \) under the SPN, SNG, and SPB priors can be used.

### Table 2

| SNO | SNG | SPN | SPB |
|-----|-----|-----|-----|
| \( c = 0.5 \) | \( c = 0.3 \) | \( q = 0.65 \) | \( q = 0.78 \) | \( q = 1 \) |
| TP% | 18.18 | 0.00 | 0.00 | 18.18 | 13.64 |
| FP% | 17.68 | 10.98 | 3.05 | 15.85 | 10.98 |

Figure 10. Approximate criteria for prior selection.
to perform posterior inference under other novel structured generalizations of other shrinkage priors that can be represented using Hadamard products involving a normal random vector, including the horseshoe and Dirichlet-Laplace priors (Polson and Scott 2010; Bhattacharya et al. 2015). Popularity of the horseshoe prior suggests that a structured generalization may be especially valuable. Additional extensions might be to assess whether or not the conditions for the convergence of elliptical slice sampling algorithms provided in Natarovskii et al. (2021) hold for the proposed elliptical slice sampling procedure and to improve scalability of the proposed elliptical slice sampling procedure with respect to $p$, the total number of penalized covariates. As shown in the appendix, seconds per posterior draw increases and minimum effective sample size decreases under these structured priors as more penalized covariates are used. The time costs of using structured priors as opposed to their independent counterparts increases with the number of penalized covariates and is greater for priors that encourage sparsity more aggressively.

Another extension to this work would treat $c$ and $q$ as unknown under the SNG and SPB priors, respectively, and perform maximum marginal likelihood estimation of $c$ or $q$ or assume prior distributions for $c$ and $q$.

Last, an extension might consider modifications of the proposed priors that continue to allow for structure in the presence of extreme elementwise sparsity, for example, extensions of SNG and SPB priors that allow for correlated scales and extensions of the SPN prior that can encourage shrinkage more aggressively by considering elementwise products of more than two normal vectors.

**Supplementary Materials**

Supplementary material available online includes proofs of all the propositions and additional numerical results. A stand-alone package for implementing the methods described in this article can be downloaded from https://github.com/maryclare/sspcomp.

**Disclosure Statement**

No potential conflict of interest was reported by the author(s).

**Funding**

This work was partially supported by NSF grants DGE-1256082 and DMS-1505136.

**References**

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015), “Dirichlet–Laplace Priors for Optimal Shrinkage,” *Journal of the American Statistical Association*, 110, 1479–1490. [2,4,7,13]

Caron, F., and Doucet, A. (2008), “Sparse Bayesian Nonparametric Regression,” in *International Conference on Machine Learning*, pp. 88–95. [3]

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), “The Horseshoe Estimator for Sparse Signals,” *Biometrika*, 97, 465–480. [2,4,7]

Daniels, M. J., and Pourahmadi, M. (2002), “Bayesian Analysis of Covariance Matrices and Dynamic Models for Longitudinal Data,” *Biometrika*, 89, 553–566. [12]

de Brecht, M., and Yamagishi, N. (2012), “Combining Sparseness and Smoothness Improves Classification Accuracy and Interpretability,” *NeuroImage*, 60, 1550–1561. [3]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society, Series B*, 39, 1–38. [8]

Devroye, L. (2009), “Random Variate Generation for Exponentially and Polynomially Tiled Stable Distributions,” *ACM Transactions on Modeling and Computer Simulation*, 19, 1–20. [4]

Finegold, M., and Drton, M. (2011), “Robust Graphical Modeling of Gene Networks Using Classical and Alternative t-distributions,” *Annals of Applied Statistics*, 5, 1057–1080. [3,5,12]

Forney, E., Anderson, C., Davies, P., Gavin, W., Taylor, B., and Roll, M. (2013), “A Comparison of EEG Systems for Use in p300 Spellers by Users with Motor Impairments in Real-World Environments,” in *Proceedings of the Fifth International Brain-Computer Interface Meeting*. [1]

Frank, I. E., and Friedman, J. H. (1995), “A Statistical View of Some Chemometrics Regression Tools,” *Technometrics*, 35, 109–135. [3]

Griffin, J. E., and Brown, P. J. (2010), “Inference with Normal-Gamma Prior Distributions in Regression Problems,” *Bayesian Analysis*, 5, 171–188. [2,3,4,7]

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, 435–448. [10]

Hoff, P. D. (2016), “Equi-variant and Scale-Free Tucker Decomposition Models,” *Bayesian Analysis*, 11, 627–648. [3]

Kalli, M., and Griffin, J. E. (2014), “Time-Varying Sparsity in Dynamic Regression Models,” *Journal of Econometrics*, 178, 779–793. [2]

Kowal, D. R., Mattheson, D. S., and Ruppert, D. (2017), “Dynamic Shrinkage Processes,” ArXiv preprint. arXiv:1707.00763, pp. 1–45. [2]

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010), “Penalized Regression, Standard Errors, and Bayesian Lassos,” *Bayesian Analysis*, 5, 309–412. [3]

Li, H., and Pati, D. (2017), “Variable Selection using Shrinkage Priors,” *Computational Statistics and Data Analysis*, 107, 107–119. [10]

Makeig, S., Kothe, C., Mullen, T., Bigdely-Shamlo, N., Zhang, Z., and Kreutz-Delgado, K. (2012), “Evolving Signal Processing for Brain – Computer Interfaces,” *Proceedings of the IEEE*, 100, 1567–1584. [1]

Mulder, J., and Pericchi, L. R. (2018), “The Matrix $t$-Prior for Estimating and Testing Covariance Matrices,” *Bayesian Analysis*, 13, 1189–1210. [12]

Murray, I., Adams, R. P., and Mackay, D. C. (2010), “Elliptical Slice Sampling,” *Journal of Machine Learning Research: Wc&P*, 9, 541–548. [3,9]

Natarovskii, V., Rudolf, D., and Sprungk, B. (2021), “Geometric Convergence of Elliptical Slice Sampling,” *Proceedings of the 38th International Conference on Machine Learning*, 139, 7969–7978. [13]

Ng, B., and Abujharbi, R. (2011), “Modeling Spatiotemporal Structure in fMRI Brain Decoding Using Generalized Sparse Classifiers,” *IEEE International Workshop on Pattern Recognition in NeuroImaging*, pp. 65–68. [3]

Nishihara, R., Murray, I., and Adams, R. P. (2014), “Parallel MCMC with Generalized Elliptical Slice Sampling,” *Journal of Machine Learning Research*, 15, 2087–2112. [8,9]

Park, T., and Casella, G. (2008), “The Bayesian Lasso,” *Journal of the American Statistical Association*, 103, 681–686. [1,2]

Perry, P. O. (2017), “Fast Moment-based Estimation for Hierarchical Models,” *Journal of the Royal Statistical Society, Series B*, 79, 267–291. [9]

Piironen, J., Paasiniemi, M., and Vehtari, A. (2020), “Projective Inference in High-Dimensional Problems: Prediction and Feature Selection,” *Electronic Journal of Statistics*, 14, 2155–2197. [10]

Piironen, J., and Vehtari, A. (2017), “Comparison of Bayesian Predictive Methods for Model Selection,” *Statistics and Computing*, 27, 711–735. [10]

Polson, N. G., and Scott, J. G. (2010), “Shrink Globally, Act Locally: Bayesian Sparsity and Regularization,” in *Bayesian Statistics 9*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, pp. 501–538, Oxford: Oxford University Press. [2,13]
Polson, N. G., Scott, J. G., and Windle, J. (2013), "Bayesian Inference for Logistic Models using polya-gamma Latent Variables," Journal of the American Statistical Association, 108, 1339–1349. [3,10]
——— (2014), "The Bayesian Bridge," Journal of the Royal Statistical Society, Series B, 76, 713–733. [3,4]
Roos, M., and Held, L. (2011), "Sensitivity Analysis in Bayesian Generalized Linear Mixed Models for Binary Data," Bayesian Analysis, 6, 259–278. [10]
Roy, A., Reich, B. J., Guinness, J., Shinohara, R. T., and Staicu, A.-M. (2021), "Spatial Shrinkage via the Product Independent Gaussian Process Prior," Journal of Computational and Graphical Statistics, 30, 1068–1080. [3]
Roy, V., and Chakraborty, S. (2016), "Selection of Tuning Parameters, Solution Paths and Standard Errors for Bayesian Lassos," Bayesian Analysis, 12, 753–778. [9]
Schuurman, N. K., Grasman, R. P., and Hamaker, E. L. (2016), "A Comparison of Inverse-Wishart Prior Specifications for Covariance Matrices in Multilevel Autoregressive Models," Multivariate Behavioral Research, 51, 185–206. [9]
Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013), "A Sparse-Group Lasso," Journal of Computational and Graphical Statistics, 22, 231–245. [12]
Styan, P. H. (1973), "Hadamard Products and Multivariate Statistical Analysis," Linear Algebra and Its Applications, 6, 217–240. [5]
Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society, Series B, 58, 267–288. [1]
van Gerven, M., Cseke, B., Oostenveld, R., and Heskes, T. (2009), "Bayesian Source Localization with the Multivariate Laplace Prior," Advances in Neural Information Processing Systems, 22, 1–9. [2]
van Gerven, M. A. J., Cseke, B., de Lange, F. P., and Heskes, T. (2010), "Efficient Bayesian Multivariate fMRI Analysis Using a Sparsifying Spatio-Temporal Prior," Neuroimage 50, 150–161. [2]
Vehtari, A., Gelman, A., and Gabry, J. (2017), "Practical Bayesian Model Evaluation using Leave-One-Out Cross-Validation and waic," Statistics and Computing, 27, 1413–1432. [10]
Wolpaw, J., and Wolpaw, E. W. (2012), Brain-Computer Interfaces: Principles and Practice, Oxford: Oxford University Press. [1]
Woody, S., Carvalho, C. M., and Murray, J. S. (2021), "Model Interpretation through Lower-Dimensional Posterior Summarization," Journal of Computational and Graphical Statistics, 30, 144–161. [10]
Wu, A., Park, M., Koyejo, O., and Pillow, J. W. (2014), "Sparse Bayesian Structure Learning with Dependent Relevance Determination Prior," in Advances in Neural Information Processing Systems, pp. 1628–1636. [2]
Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables," Journal of the Royal Statistical Society, Series B, 68, 49–67. [2]
Zhao, S., Gao, C., Mukherjee, S., and Engelhardt, B. E. (2016), "Bayesian Group Factor Analysis with Structured Sparsity," Journal of Machine Learning Research, 17, 1–47. [2]