A Horseshoe Pit mixture model for Bayesian screening with an application to light sheet fluorescence microscopy in brain imaging

Francesco Denti*[1], Ricardo Azevedo[2,3], Chelsie Lo[2,3], Damian Wheeler[3], Sunil P. Gandhi[2,3,4], Michele Guindani[1] and Babak Shahbaba[1]

[1]Department of Statistics, University of California, Irvine
[2]Department of Neurobiology and Behavior, University of California, Irvine
[3]Translucence Biosystems, Inc, Irvine
[4]Center for the Neurobiology of Learning and Memory, University of California, Irvine

Finding parsimonious models through variable selection is a fundamental problem in many areas of statistical inference. Here, we focus on Bayesian regression models, where variable selection can be implemented through a regularizing prior imposed on the distribution of the regression coefficients. In the Bayesian literature, there are two main types of priors used to accomplish this goal: the spike-and-slab and the continuous scale mixtures of Gaussians. The former is a discrete mixture of two distributions characterized by low and high variance. In the latter, a continuous prior is elicited on the scale of a zero-mean Gaussian distribution. In contrast to these existing methods, we propose a new class of priors based on discrete mixture of continuous scale mixtures providing a more general framework for Bayesian variable selection. To this end, we substitute the observation-specific local shrinkage parameters (typical of continuous mixtures) with mixture component shrinkage parameters. Our approach drastically reduces the number of parameters needed and allows sharing information across the coefficients, improving the shrinkage ef-

*fdenti@uci.edu

fect. By using half-Cauchy priors, this approach leads to a cluster-shrinkage version of the Horseshoe prior. We present the properties of our model and showcase its estimation and prediction performance in a simulation study. We then recast the model in a multiple hypothesis testing framework and apply it to a neurological dataset obtained using a novel whole-brain imaging technique.

Keywords: Bayesian inference; Variable selection; Mixture models; Neuroscience.

1 Introduction

Variable selection plays a central role in many statistical inference problems, where n measurements of a dependent random variable $\boldsymbol{Y}$ are modeled as a function of p covariates, $\boldsymbol{X}$. Broadly speaking, variable selection aims to provide a parsimonious and generalizable representation of the functional relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$ by designating a subset of variables in $\boldsymbol{X}$ as interesting and relevant, while discarding the remaining ones as unrelated to the outcome (i.e., noise). Variable selection is of pivotal importance when dealing with modern large-scale inference analyses, especially the so-called "small n large p" problems. To tackle such problems, a large class of methodologies have been proposed based on *regularization* (or *shrinkage*) of the parameters. This procedure aims at identifying a meaningful subsets of variables in $\boldsymbol{X}$ by shrinking to zero the effect of the variables deemed as irrelevant, usually assumed to be the vast majority (see, for example, Tibshirani, 2013; George and McCulloch, 1997, and the references therein).

Within the Bayesian framework, the regularization process involves the specification of shrinkage priors for the regression coefficients $\boldsymbol{\beta} = \{\beta_j\}_{j=1}^{p}$. To achieve this goal, two main approaches have been proposed in the literature: the spike-and-slab (or two-group) models (Mitchell and Beauchamp, 1988; McCulloch and George, 1993) and the continuous shrinkage scale mixture models (Polson et al., 2012). The first approach models the parameter of interest as a discrete mixture between a point mass at 0 (or a distribution centered at zero with low variance) and a "flat" distribution with large variance. This way, the resulting model-based clustering can discriminate between relevant and irrelevant coefficients. Despite being a well-established topic for more than two decades, this is still an active area of research. For example, Rockova and George (2016) have recently proposed the spike-and-slab Lasso, where the two competing distributions are assumed to be double Exponentials, in the spirit of the Bayesian Lasso of Park and Casella (2008). The second approach relies on the specification of continuous scale mixtures, e.g. by placing hierarchical priors on the variance parameter of Gaussian distributions (refer to Bhadra et al., 2019a, for a recent review). The scale parameter is often decoupled into two (or more) terms, commonly referred to as the global (i.e., shared by all the regression coefficients) and the local shrinkage parameters. Several shrinkage models can be seen as a special case of this global-local shrinkage paradigm, including the Bayesian Lasso

(Park and Casella, 2008), the Normal-Gamma (Griffin and Brown, 2010), the Horseshoe (Carvalho et al., 2010), or the Horseshoe+ (Bhadra et al., 2017). The selection between relevant and irrelevant variables needs to be done ex-post, usually by thresholding a proxy of the posterior probability of relevance: $\mathbb{P}[\beta_j \neq 0|\text{data}]$.

In this paper, we aim at bridging the gap between these two alternatives by proposing a discrete mixture of continuous scale mixtures to perform different types of Bayesian screening. The proposed approach allows to combine the regularization effect typical of continuous shrinkage priors while inducing a segmentation of the coefficients as in the spike-and-slab case. Indeed, the model will automatically detect groups of coefficients driven by different levels of sparsity, imposing an adaptive regularization within each group. Moreover, the discrete mixture greatly reduces the complexity of the model, avoiding the usual specification of a local shrinkage parameter for each variable. For example, we will show how a unique, shared mixture component shrinkage parameter is often sufficient to model all the null coefficients, enabling at the same time sharing of information across the parameters.

Our proposed approach is related to several other methods employing mixture models to improve the efficacy of the variable selection and shrinkage processes. For example, Shahbaba and Johnson (2013) proposed a scale mixture of Gaussian distributions capable of ranking the covariates (e.g., a treatment indicator) by modeling the sampling variances via the Dirichlet Process (DP, Ferguson, 1973). This Bayesian nonparametric specification groups the variance parameters according to their magnitude, therefore inducing a ranking in terms of relevance among the groups of corresponding coefficients. Our model extends this idea with the possibility of regularizing the estimates by adopting shrinkage priors to model the sampling variances. Therefore, the ranking property inherited by our discrete mixture specification is appealing since it complements the lack of an explicit screening solution typical of the continuous shrinkage methods. MacLehose and Dunson (2010) proposed to model regression coefficients as binary mixtures between a traditional double Exponential centered in zero (Bayesian Lasso) and a double Exponential with non-zero location parameter. They employ DPs for both the location and scale parameters. This model was extended by Yang et al. (2011), who adopted mixtures of Bayesian elastic-nets (Li and Lin, 2010). In contrast, our approach focuses only on the modeling of variances, fixing the center of the mixture kernel distributions to 0. This choice allows to exploit the properties of the continuous scale mixture family. Recently, Ding and Karabatsos (2021) explored the effects of the combination between shrinkage priors and a covariate dependent DP mixtures. The authors jointly model the conditional response distributions and the covariates with a nonparametric mixture, inducing a partition over the observations. We instead consider fixed covariates and employ the mixture to cluster the regression coefficients into shrinkage profiles. Our proposal is also related to the Dirichlet-t distribution introduced in Finegold and Drton (2011, 2014) within the Bayesian robust graphical modeling framework, where a DP with Inverse Gamma base measure models the scale parameters of a multivariate Normal. Following the discussion in Shahbaba (2014), we adopt different base measures and recast their modeling framework for screening purposes. See Appendix A for further discussion.

The article proceeds as follows. In Section 2, we present the model and investigate its theoretical properties. In Section 3, we discuss how to perform posterior inference. In Section 4, we conduct several simulation studies. In Section 5 we showcase how our model performs on an innovative whole-brain imaging dataset obtained using light sheet fluorescence microscopy to detect classes of activation in the brain. Finally, in Section 6 we summarize advantages and shortcomings of our proposed method and discuss future directions.

2 A discrete mixture of continuous scale mixtures

We introduce our Bayesian variable selection framework by considering a linear regression model. More specifically, we assume that an outcome vector $\boldsymbol{Y}$ of length n is modeled as

$$\boldsymbol{Y} = \beta_0 \mathbb{1}_n + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{1}$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is the matrix of covariates, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of regression coefficients, $\beta_0 \in \mathbb{R}$ is the intercept, and $\boldsymbol{\varepsilon}$ is the noise term. We assume homoschedastic and uncorrelated errors, i.e. $\boldsymbol{\Sigma} = \sigma^2 \mathbb{I}_n$. Here, $\mathcal{N}_k(\boldsymbol{a}, \boldsymbol{A})$ indicates a multivariate Normal distribution of dimension k with mean vector $\boldsymbol{a}$, covariance matrix $\boldsymbol{A}$, and density function $\phi_k(\boldsymbol{a}, \boldsymbol{A})$. In the univariate case, we let $\mathcal{N}_1 \equiv \mathcal{N}$ and $\phi_1 \equiv \phi$. Without loss of generality, we assume that the outcome variable is centered, and therefore $\beta_0 = 0$.

To adopt a fully Bayesian setting, we need to specify prior distributions for the variance parameter σ^2 and the coefficients $\boldsymbol{\beta}$. A common choice for the prior distribution of σ^2 is the Jeffreys prior $\pi(\sigma^2) \propto 1/\sigma^2$. More care is needed to specify the prior distribution for $\boldsymbol{\beta}$. This choice is crucial, especially if we want to enforce any regularizing effect on the parameters. In the usual global-local shrinkage parameter models (Polson et al., 2012), the regression coefficients are assumed to be distributed as a continuous scale mixture of Gaussian distributions, i.e. $\beta_j | \tau, \boldsymbol{\lambda}_p, \sigma^2 \sim \mathcal{N}(0, \sigma^2 \cdot \tau^2 \cdot \lambda_j^2)\ \forall j = 1, \dots, p$ and with λ_j assumed stochastic. Here, the parameter $\tau \in \mathbb{R}^+$ is called the *global* shrinkage parameter, while the vector $\boldsymbol{\lambda}_p = \{\lambda_j\}_{j=1}^p$, $\lambda_j \in \mathbb{R}^+$ contains all the *local* shrinkage parameters. Conditioning on the variance of the data, σ^2, guarantees a unimodal posterior (Park and Casella, 2008).

We modify this model by proposing a discrete mixture of continuous scale mixtures of Gaussians. As a result, the large number of local shrinkage parameters is substituted by a more parsimonious set of L *mixture component* shrinkage parameters. More specifically, we assume

$$\beta_j | \tau, \boldsymbol{\lambda}_L, \boldsymbol{\pi}, \sigma^2 \sim \sum_{l=1}^{L} \pi_l \, \mathcal{N}(0, \sigma^2 \cdot \tau^2 \cdot \lambda_l^2), \qquad j = 1, \dots, p, \tag{2}$$

where $\boldsymbol{\pi}$ is the vector of mixture weights and the elements of the vector $\boldsymbol{\lambda}_L = \{\lambda_l\}_{l=1}^L$ assume the role of *mixture component* shrinkage parameters. The specification in (2) is very general and encompasses many known models. In particular, when $L = 2$ and $\lambda_1 \approx 0$, we recover the spike-and-slab framework, while when $L = p$ and $\pi_l = \delta_p(l)\ \forall l, p$

(i.e., inducing p different singleton clusters) we recover the continuous shrinkage framework.

From another perspective, we can consider the Normal mean estimation problem $Y_j \sim \mathcal{N}(\beta_j, \sigma^2)$, for $j = 1, \ldots, n$, by simply adopting $\boldsymbol{X} = \mathbb{I}_n$ and $\beta_0 = 0$ in model (1). This scenario is often considered for hypothesis testing, where the task is to detect the test statistics that depart from the standard Gaussian distribution specified under the null hypothesis ($H_{0,j} : \beta_j = 0$). Suppose we adopt the classical global-local shrinkage prior for $\boldsymbol{\beta}$ to induce sparsity and set $\sigma^2 = 1$. One can easily show that $Y_j|\lambda_j, \tau \sim \mathcal{N}(0, 1 + \tau^2\lambda_j^2)$. In our discrete mixture of continuous scale mixture model, the induced sampling distribution is itself a mixture:

$$Y_j|\tau, \boldsymbol{\lambda}_L, \boldsymbol{\pi} \sim \sum_{l=1}^{L} \pi_l \, \mathcal{N}_1(0, 1 + \tau^2\lambda_l^2). \tag{3}$$

In the multiple hypothesis testing setting, we can see $\boldsymbol{Y}$ as a vector of n properly standardized test statistics corresponding to n different null hypotheses. Thus, model (3) can be interpreted as an extension of the classical two-group model (Efron, 2007). Without loss of generality, let us assume that the first mixture component is characterized by the smallest scale parameter $\lambda_{(1)} = \min_l \lambda_l$. One can impose this constraint *a priori* or recover the mixture component with smallest scale parameter after the model estimation. In the case where the product $\tau\lambda_{(1)} \approx 0$, the corresponding mixture component represents the null distribution, resembling a theoretical standard Gaussian. The product $\tau\lambda_{(1)}$ is allowed to be different from zero to reflect a departure from the theoretical null. The remaining mixture components define the alternative distribution, which can be decomposed into degrees of relevance according to the magnitude of the parameters remaining $\boldsymbol{\lambda} \setminus \lambda_{(1)}$.

Whether we are adopting our model to perform variable selection or hypothesis testing, we need to elicit prior distributions for both $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}_L$ to complete the Bayesian specification. In addition, we can also specify a distribution for the global shrinkage parameter τ. The prior distribution for the weights changes if we assume a countable or uncountable number of mixture components. If we assume L to be finite, we can simply set $\boldsymbol{\pi} \sim \mathit{Dirichlet}(a_1, \ldots, a_L)$. Notice that even $L > p$ is a viable option since one has to distinguish between mixture components and active components, i.e. the actual clusters found in the dataset. For example, this argument has been recently discussed in Malsiner-Walli et al. (2016), where the authors advocate for the use of sparse mixture models. Setting the hyperparameters $a_l = \epsilon \ \forall l$ with ϵ small (≤ 0.05) allows the model to parsimoniously select the number of active components needed to describe the data. Another possibility is to specify a nonparametric model via a DP mixture model:

$$\beta_j|\tau, \boldsymbol{\lambda}_\infty, \sigma^2 \sim \mathcal{N}(0, \tau^2\sigma^2\lambda_j^2), \quad \lambda_j|G \sim G, \quad G \sim DP(\alpha, H), \tag{4}$$

where $DP(\alpha, H)$ indicates a Dirichlet Process with concentration parameter α and base measure H. Adopting the Stick Breaking (SB) representation of Sethuraman (1994),

model (4) becomes

$$\beta_j|\tau, \boldsymbol{\lambda}_\infty, \sigma^2, \boldsymbol{\pi} \sim \sum_{l=1}^{+\infty} \pi_l \phi(\beta_j; 0, \sigma^2 \cdot \tau^2 \cdot \lambda_l^2), \quad \lambda_l \sim H, \quad \boldsymbol{\pi} \sim SB(\alpha), \tag{5}$$

where the weights $\boldsymbol{\pi}$ are defined as $\pi_1 = u_1$, $\pi_l = u_l \prod_{q<l}(1-u_q)$ for $l > 1$ and $u_l \sim Beta(1, \alpha)$ for $l \geq 1$. Different nonparametric priors, such as the Pitman-Yor process, can be adopted as well.

The introduction of mixture component shrinkage parameters is beneficial mainly for two reasons. First, this specification can improve the effectiveness of the regularization with respect to common global-local scale mixtures models. A discrete mixture allows the model to borrow information across all the parameters and self-adapt to the different degrees of sparsity characterizing subsets of the coefficients. Second, the model-based clustering nature of our approach enables the ranking of groups of coefficients into several *shrinkage profiles*, as in Shahbaba and Johnson (2013), improving on the typical binary solutions (e.g., relevant vs. irrelevant) and providing more complete information. In Section 5, we will present a successful application of this concept.

In what follows, we will adopt a Half-Cauchy prior for the mixture component shrinkage parameters: $\lambda_l \sim \mathcal{C}^+(0,1)$, $\forall l$ following the suggestion Shahbaba (2014) made in the context of robust modeling. We call the model following from this specification Horseshoe Pit (HSP), in the spirit of the Horseshoe (HS) prior introduced by Carvalho et al. (2010).

2.1 The mixture component and the cluster shrinkage factors

Consider again the Normal mean estimation framework, and define $\kappa_j = 1/(1+\tau^2\lambda_j^2) \in (0,1)$. It follows that $\mathbb{E}[\beta_j|Y_j] = (1 - \mathbb{E}[\kappa_j|Y_j]) \cdot Y_j$, and $\mathbb{E}[\beta_j|\lambda_j, \tau, Y_j] = \frac{\tau^2\lambda_j^2}{1+\tau^2\lambda_j^2} \cdot Y_j$, where κ_j is known as the *shrinkage factor* for observation j, which can be interpreted as a proxy of the complement of the posterior probability of relevance in the two-group model (Carvalho et al., 2010). It is interesting to see how this key quantity changes under our model specification. For the conditional model, the posterior expected values of the coefficients become

$$\begin{aligned} \mathbb{E}[\beta_j|\boldsymbol{\pi}, \boldsymbol{Y}] &= \sum_{l=1}^{L} \mathbb{E}[r_l(Y_j)(1-\kappa_l^*)|\boldsymbol{Y}] \cdot Y_j, \\ \mathbb{E}[\beta_j|\tau, \boldsymbol{\lambda}_L, \boldsymbol{\pi}, \boldsymbol{Y}] &= \left(\sum_{l=1}^{L} r_l(Y_j)(1-\kappa_l^*) \right) \cdot Y_j = (1-\tilde{\kappa}_j) \cdot Y_j, \end{aligned} \tag{6}$$

where $r_l(Y_j) = \frac{\pi_l \phi(Y_j; 0, 1+\tau^2\lambda_l^2)}{\sum_{l=1}^{L} \pi_l \phi(Y_j; 0, 1+\tau^2\lambda_l^2)}$. See Appendix B for the derivation of (6). Here, we distinguish between the *mixture component shrinkage factors* (MCSF - one for every mixture component) defined as $\kappa_l^* = 1/(1+\tau^2\lambda_l^2)$ and the *cluster shrinkage factors* (CSF

- one for every coefficient) $\tilde{\kappa}_j = \sum_{l=1}^{L} r_l(Y_j)\kappa_l^*$. Each CSF is a function of a convex combination of the L MCSFs and directly controls the amount of shrinkage that affects each parameter β_j. Simultaneously, the weights of the convex combination depend on the components of the marginal sampling distribution $\phi(Y_j; 0, 1+\tau^2\lambda_l^2)$. It becomes clear how the model structure takes advantage of the the sharing of statistical strength across parameters. Indeed, the posterior mean for β_j is the result of two effects. Given its mixture nature, the shrinkage is affected by all the other mixture components parameters through information sharing. However, since the mixture is driven by weights that directly depend on each data point's contribution to the marginal likelihood, we retain an observation-specific effect in the shrinkage process. These simultaneous effects help the estimating procedure to place more emphasis on shrinkage profiles that better describe the data points in $\boldsymbol{Y}$.

As a simple example, consider a sample of 1,000 observations generated from a linear regression model with a true vector of coefficients $\boldsymbol{\beta}$ composed of 100 zeros and 200 realizations generated in equal proportions from two Normal distributions centered around zero with variances 100 and 1, respectively. The error noise is set to $\sigma = 0.5$. and we fix ex-ante $\tau^2 = 0.05$. Given this dataset, we compare the HS and the HSP model in terms of estimated variances (and, therefore, shrinkage effects). The four panels of Figure 1 show the posterior means (left column) and medians (right column) for the quantity $\hat{\lambda}_j\tau\sigma$ plotted against the true coefficients, transformed as $\tilde{\beta}_j = sign(\beta_j)\sqrt{\beta_j}$ to ease the visual comparison. The points in the bottom panels are colored according to the resulting best partition (more detail on this in the next section). By comparing the two panels, the cluster-specific shrinkage becomes evident. The HSP can capture the different magnitudes of the coefficients and rank them in three different clusters. In particular, one profile includes all the noisy coefficients, providing a clear solution to the variable selection problem. It is also interesting to compare the posterior mean with the median in the HSP case. While the three distinct shrinkage profiles are clear in the latter case, the behavior of the posterior means reveals the presence of both a cluster effect and an observation-specific effect.

3 Posterior Inference

3.1 The latent membership labels and posterior conditional distribution

To conduct posterior inference, we need to rely on MCMC techniques since the posterior distribution is not directly available in closed form. To simplify posterior simulation, we augment model (2) with the latent membership labels $\boldsymbol{z} = \{z_j\}_{j=1}^{p}$, where $z_j \in \{1, \ldots, L\}$, linking each coefficient with a cluster. In other words, $z_j = l$ if the j-th coefficients has been assigned to the l-th cluster. We obtain

$$\beta_j|\tau, \boldsymbol{\lambda}_L, z_j, \sigma^2 \sim \mathcal{N}(0, \sigma^2 \cdot \tau^2 \cdot \lambda_{z_j}^2), \qquad z_j|\boldsymbol{\pi} \sim \sum_{l=1}^{L} \pi_l \delta_l(\cdot). \tag{7}$$

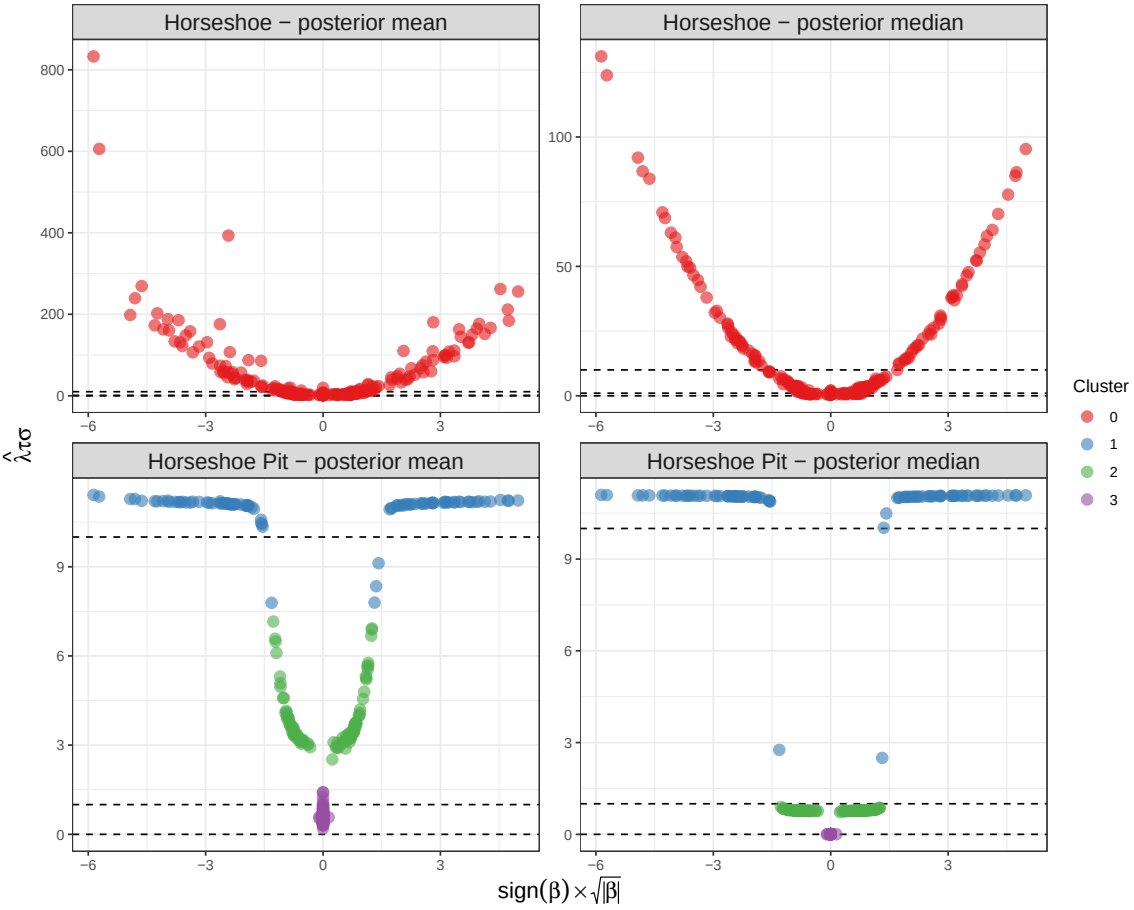

Figure 1: A comparison between the estimated individual mean and median standard deviations $\hat{\lambda}_j\tau\sigma$ against the true coefficients for the HS (top panels) and the HSP (bottom panels) models. The colors in the bottom panels describe the partitions into different magnitudes.

Once the auxiliary membership labels are introduced in the model, it is straightforward to derive the full conditional for the corresponding Gibbs sampler. Both the global and the mixture component shrinkage parameters can be efficiently sampled following a parameter augmentation strategy (Makalic and Schmidt, 2016) or via slice sampler (as in the Supplementary Material of Polson et al., 2014). The details of the Gibbs sampler are deferred to Appendix C.

This data augmentation not only is usefult to conduct feasible posterior inference, but also provides more insights regarding the behavior of the MCSFs. Indeed, we can derive the conditional posterior distribution for the l-th MCSF κ_l^* under the Horseshoe Pit

prior, obtaining

$$p(\kappa_l^*|\boldsymbol{z}, \boldsymbol{Y}, \tau^2, \sigma^2) \propto \frac{(\kappa_l^*)^{-\frac{1}{2}}(1-\kappa_l^*)^{-\frac{1}{2}}}{\tau^2\kappa_l^* + 1 - \kappa_l^*}(\kappa_l^*)^{\frac{n_l}{2}} \cdot \exp\left[-\frac{\kappa_l^*}{2\sigma^2}\sum_{j:z_j=l} Y_j^2\right]. \tag{8}$$

It is crucial to note how all the observations that are grouped in the l-th cluster explicitly contribute to the conditional posterior distribution of κ_l^*. Without loss of generality, we set $\sigma^2 = \tau^2 = 1$. Moreover, define $S_l = \sum_{j:z_j=l} Y_j^2$. Then, the distribution in (8) simplifies into $p(\kappa_l^*|\boldsymbol{z}, \boldsymbol{Y}) \propto (\kappa_l^*)^{\frac{n_l-1}{2}}(1-\kappa_l^*)^{-\frac{1}{2}} \exp\left[-\frac{\kappa_l^*}{2}S_l\right]$. Whenever $L = p$, $n_l = 1$, and $S_l = Y_j^2$, we recover the case of the Horseshoe model.

We exemplify the behavior of this posterior in the two panels of Figure 2, where we report different shapes of the posterior density function for different combinations of (n_l, S_l). We expect our model to group parameters characterized by similar magnitude, so the top panel shows what happens when $S_l = 0$ and n_l grows. Ideally, the more observations with 0 magnitude are assigned to the same cluster, the stronger is the shrinkage effect of the MCSF κ_l^*, concentrating all the mass around 1. Clearly, high values of S_l and low values of n_l will result in small κ_l^*, as we can see in the bottom panel. In other words, the MCSF is lower when the relative cluster is comprised of few observations of great magnitude. In Appendix E, we report a diagram that depicts this behavior, which explains how the sharing of information is exploited by our model to tune the amount of shrinkage according to the observed data.

3.2 Postprocessing of the results

Once the posterior sample has been collected, we can estimate the cluster-shrinkage factors by means of the membership labels. We map each coefficient β_j to the assigned local shrinkage parameter via z_j constructing the vector $(\lambda_{z_1}, \ldots, \lambda_{z_J})$. It is then straightforward to compute $\hat{\bar{\kappa}}_j = 1/(1+\tau^2\lambda_{z_j}^2)$. One of the main advantages of our model is that, once the MCMC sample of size T is collected, it allows the estimation of the best partition that groups the different coefficients into classes of similar magnitude. Let $\boldsymbol{z}^{(t)} = \{z_1^{(t)}, \ldots, z_n^{(t)}\}$ be the realization of the membership labels at iteration $t = 1, \ldots, T$. With this information, we can estimate the Posterior Probability Coclustering (PPC) matrix, whose entries are defined as $\widehat{PPC}_{j,j'} = \sum_{t=1}^{T} \mathbb{1}_{\left(z_j^{(t)} = z_{j'}^{(t)}\right)}/T$, for $j, j' = 1, \ldots, p$. In other words, $\widehat{PPC}_{j,j'}$ estimates the proportion of times that coefficients j and j' have been assigned to the same cluster along the MCMC iterations. Hierarchical clustering can be applied directly to the $\widehat{PPC}$ matrix for fast solutions as in Medvedovic et al. (2004).

The resulting partition is easy to interpret. The HSP prior allows for a model-based clustering driven by the cluster-shrinkage parameter vector $\boldsymbol{\lambda}_L$. Therefore, the clusters in the solution specified by the optimal partition $\hat{\boldsymbol{z}}$ can be described as classes of different magnitude. Therefore, whenever a regularization problem is addressed, we can explicitly

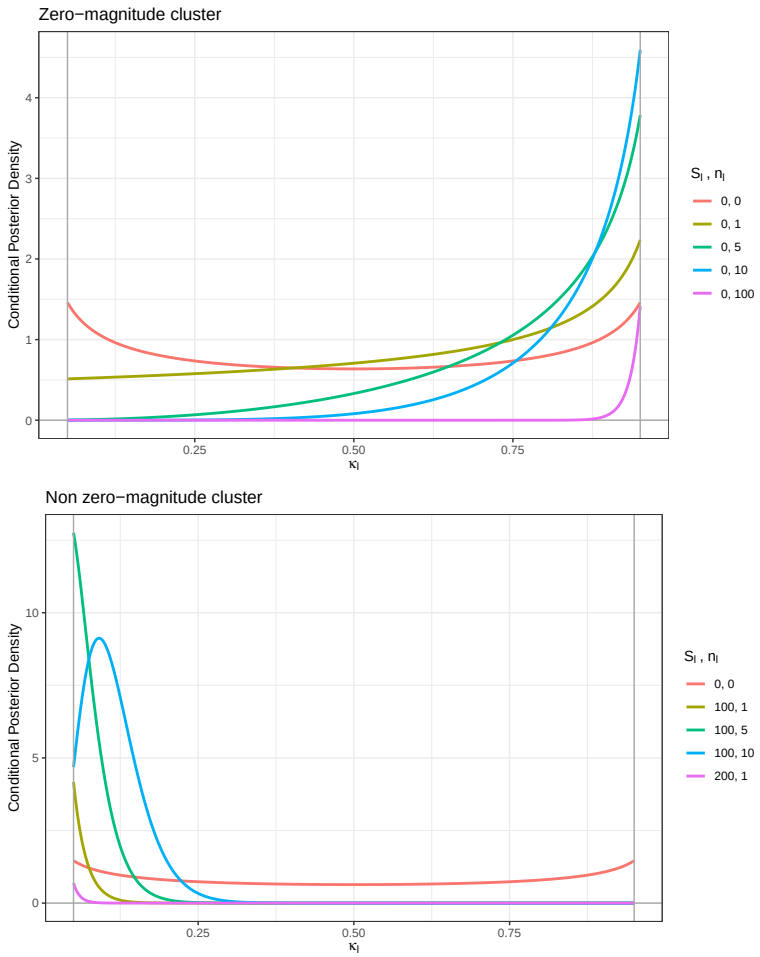

Figure 2: Posterior densities of the MCSF κ_l^*, changing according to the number of coefficients assigned to the cluster and to their magnitudes.

identify the subgroup of coefficients characterized by the smallest magnitude that can be deemed as irrelevant, similarly to the null component in of the two-group model. In a linear regression framework, this means that we are able to identify the set of indexes that indicate the least relevant covariates, say $\mathcal{B}_0 = \{j \in \{1, \ldots, p\} : \beta_j = 0\}$, inducing a variable selection solution. Moreover, the model also allows the classification of the remaining parameters into subsets of different magnitudes, yielding an interpretable ranking.

4 Simulation study

4.1 Estimating and Predictive Performance

We compare the estimating and predictive performance of the HSP model and the HS in a linear regression framework. To estimate the model under the Horseshoe prior specification, we employ the `R` package `bayesreg` (Makalic and Schmidt, 2016).

Our experiment consists of five scenarios, characterized by different values of the ratio n/p, describing the proportion between the sample size and number of variables. Specifically, we consider the following five ratios: $n/p \in \{(1000, 500) = 2, (200, 150) = 1.33, (200, 200) = 1, (500, 1000) = 0.5, (200, 500) = 0.4\}$. Under each scenario, we generate $K = 100$ training and test datasets as follows. We first sample n independent observations from a multivariate Gaussian as $X_{i,k} \sim \mathcal{N}_p(0, \mathbb{I}_p)$, $i = 1, \dots, n$ creating the design matrix $\boldsymbol{X}_k$, $k = 1, \dots, K$. A matrix of the same size is generated to be used as test set. Then, we sample the regression coefficients $\boldsymbol{\beta}_k$ organized in three different blocks: $\beta^{(1)}_{j_1,k} \sim \mathcal{N}(0, 400)$ for $j_1 = 1, \dots, 50$, $\beta^{(2)}_{j_2,k} \sim \mathcal{N}(0, 9)$ for $j_2 = 1, \dots, 50$, and $\beta^{(3)}_{j_3,k} \sim \delta_0$ for $j_3 = 1, \dots, p - 100$. That is, for a fixed number of covariates $p > 100$, we generate 50 coefficients of high magnitude ($\sigma^{(1)} = 20$), 50 coefficients of low magnitude ($\sigma^{(2)} = 3$), and $p - 100$ coefficients identically equal to zero. Finally, we set $Y_k = X_k\boldsymbol{\beta}_k + \boldsymbol{\epsilon}_k$, with $\boldsymbol{\epsilon}_k \sim \mathcal{N}_n(0, \mathbb{I}_n)$.
For the mixture weights, we adopt a sparse mixture specification using $L = 50$ mixture components and $a = 0.05$. To quantify the performance of the two models, for each dataset we compute the mean squared error between the posterior mean $\hat{\boldsymbol{\beta}}_k$ and the ground truth, defined as $\text{MSE}(\boldsymbol{\beta}_k, \hat{\boldsymbol{\beta}}_k) = \sum_{j=1}^{p}(\boldsymbol{\beta}_{j,k} - \hat{\boldsymbol{\beta}}_{j,k})^2/p$. The same measure is adopted to compare the prediction errors obtained in test set: $\text{MSE}(\boldsymbol{Y}_k, \hat{\boldsymbol{Y}}_k)$. Moreover, we focus on the estimating performance within each block of coefficients, to assess the shrinking properties of the HSP compared to the classical HS. We compute the ratio between the MSE obtained on each block of parameters, namely $\text{MSER}(\boldsymbol{\beta}^{(l)}) = \text{MSE}^{HSP}\left(\boldsymbol{\beta}^{(l)}_k, \hat{\boldsymbol{\beta}}^{(l)}_k\right)/\text{MSE}^{HS}\left(\boldsymbol{\beta}^{(l)}_k, \hat{\boldsymbol{\beta}}^{(l)}_k\right)$ for $l = 1, 2, 3$. Therefore, the MSER provides a relative measure of comparison between the two models shrinkage effects across different levels of magnitude.
Figure 3 shows the boxplots of the resulting error measures obtained over the 100 datasets for each scenario. Each row corresponds to a different quantity: MSE over the coefficients, MSER over the parameter blocks, and MSE over the outcome variable. Within every row, each panel correspond to one of the scenarios. The HSP performs consistently better than the simple HS model, especially when both n/p and n are small. Almost all the MSER values are below 1, indicating a better estimating performance of the HSP model. However, when the sample size is larger than the number of variables (Scenarios 1 and 2), the estimating performance of the parameters belonging to the first two blocks are comparable, with MSER values mostly between 0.75 and 1. The most important gain given by the HSP model is in precision of the estimation in the third group of coefficients. In other words, the HSP effectively detects and shrinks to zero the true null coefficients. The MSER boxplots shows that the errors made on the third block of parameters by the

HSP is almost negligible when compared to the classic HS. In Scenario 5 ($n/p = 0.4$), the HS fails to accurately estimate the regression coefficients, resulting in a large average prediction MSE.

We report in Appendix D another simulation study, where we investigate the effect of the cluster shrinkage as a function of the number of null parameters.

4.2 Multivariate HSP: application to a simulated f-MRI dataset

To show how the HSP can be seamlessly adapted to a multivariate regression setting, we now employ our model to an artificially designed functional magnetic resonance imaging (f-MRI) dataset. In a f-MRI experiment, the level of the blood-oxygen-level dependent (Bold) signal is measured for the entire brain. The blood flow captured by the Bold level is used as a proxy to study neuronal activation. The main goal of the f-MRI analysis is to investigate the activation of the brain regions in association with a certain task (stimulus) performed by the subject under study. In particular, a 3D image of the brain is partitioned into voxels according to a three-dimensional grid. The Bold level is then measured for each voxel over time. Therefore, the typical f-MRI dataset is a dynamic 3D image that can be expressed as a four-dimensional array of coordinates (x, y, z, t). Through the `R` package `neuRosim` (Welvaert et al., 2011), we simulate a f-MRI dataset that closely resembles realistic situations. To this end, we first create an event-specific stimulus function over $T = 100$ time stamps of 2 seconds each. This function is then convoluted with a double-Gamma hemodynamic response function (HRF), to mimic the hemodynamic delay between the neuron activation and the corresponding difference in the metabolic status. We use the same event-related onset as suggested in Welvaert et al. (2011). The resulting explanatory variable $X(t)$, $t = 1, \dots, T$ is shown in the Appendix E: the red bars represent the stimuli, and the black line the resulting response function.

According to our simulation scenario, the response function $X(t)$ affects three different, manually specified, brain regions (Top Right - TR, Top Left - TL and Bottom -BT). The Bold reactions in the three subsets have different magnitudes, being $\beta_{BT} = 5$, $\beta_{TL} = 10$ and $\beta_{TR} = 10$. On top of the signal, we specify a rician white noise along the temporal dimension and a Gaussian markov random field for modeling the spatial correlation structure. For the parameters of the error specification, we resort to the default values of the functions in `neurosim`. We also adopt the same signal-to-noise ratio, specified as SNR = 3.67. We focus on a particular brain slice ($z = 13$) of dimension 64×64. The ground truth is highlighted in the left panel of Figure 4.

To apply the HSP model in this context, we need to rewrite our model in a multivariate fashion. Let $Y_\nu(t)$ be the Bold level at time $t = 1$ measured in voxel ν. We define, for each voxel $\nu = 1, \dots, V$, the model $Y_\nu(t) = \beta_0 + X(t)\beta_\nu + \epsilon_\nu$, and then assume model (7) for the regression coefficients β_ν. We employ a sparse mixture model, adopting a Dirichlet prior for the mixture weights with parameters $a_l = 0.05 \, \forall l$ and run the Gibbs sampler for 10'000 iterations after discarding the burn-in period. To compare our method to other

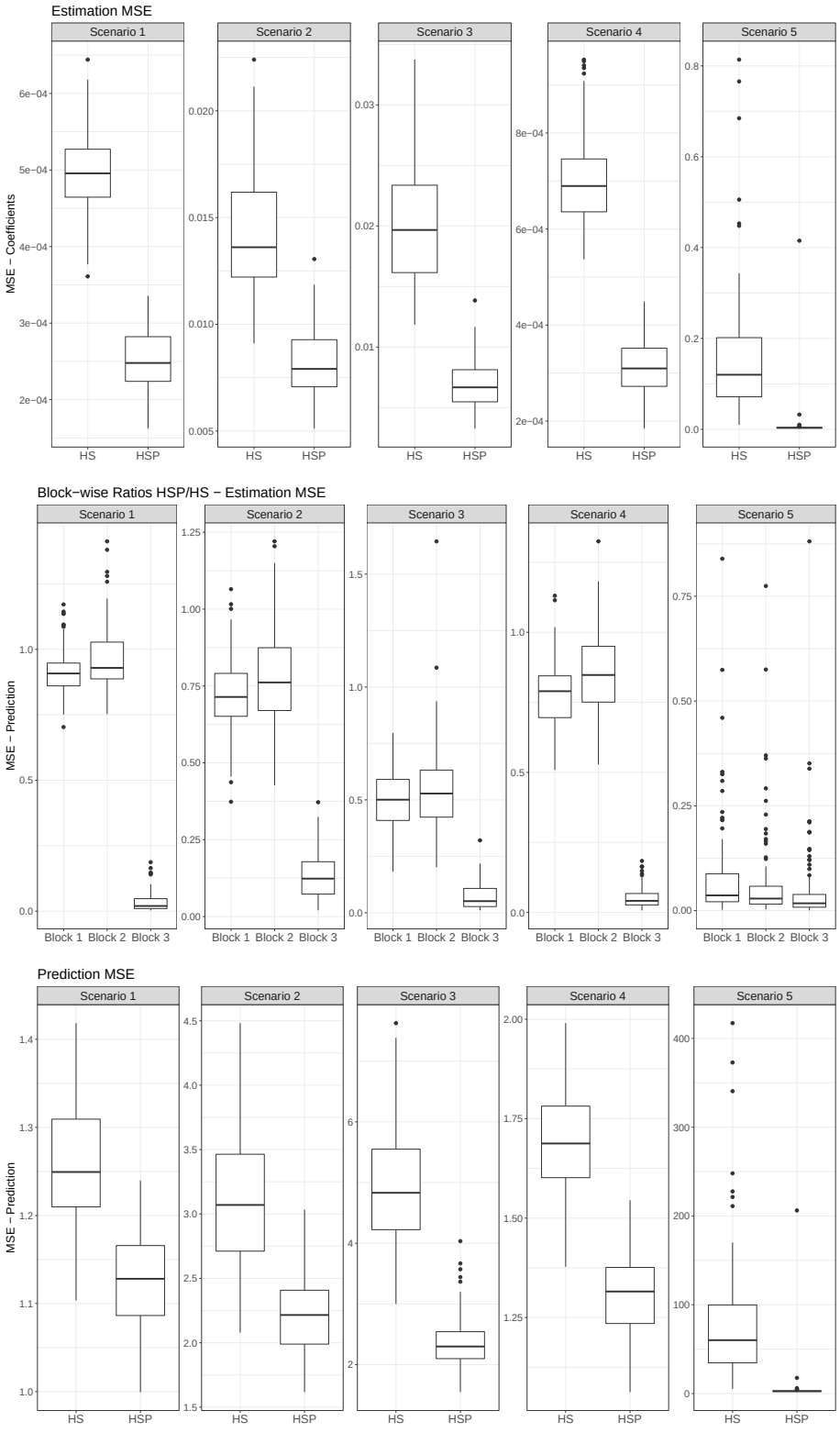

Figure 3: Boxplots of estimating and predictive performance of the HSP model compared to the HS model. The first row of plots shows the MSE computed on the coefficients, the second row shows the relative MSE in three different blocks of coefficients, the third row shows the predictive MSE computed on the outcome variable.

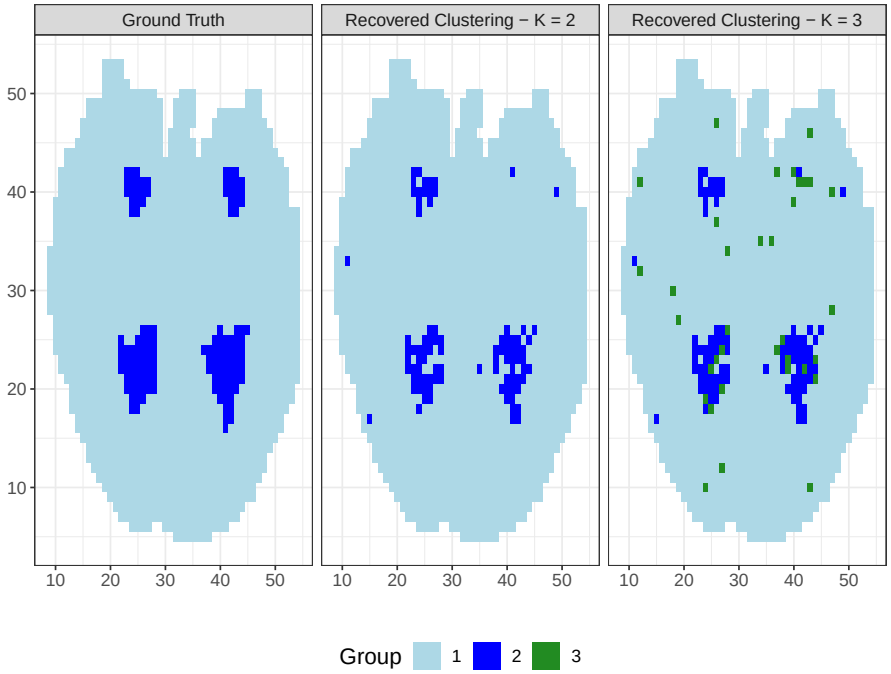

Figure 4: Clustering results of the multivariate HSP model. The left panel reports the ground truth. The central and right panels displays the detected voxels when two and three clusters are chosen in the post-processing procedure, respectively.

models, we run OLS and Bayesian HS voxel-specific models. We showcase the heatmaps with the MLE estimates (for OLS) and the posterior means (HS and HSP) in Figure 5. Compared with the other methods, the estimates of the inactive voxels under the HSP (right panel) are better detected and regularized, as we can see from the uniformity of the brain-image background, while the true signal is mostly recovered. To understand how strong is the shrinkage of the coefficients induced by the HSP model, we provide a comparison of the coefficients' magnitude in the Appendix E. The HSP aggressively shrinks to zero all the small coefficients, especially when compared to non-regularizing methods such as the OSL. The departure from the black line suggests that the HSP perform a stronger regularization than the HS model as well.

Finally, we can provide a clear distinction between active and inactive voxels exploiting the clustering nature of the HSP mixture specification. We threshold the resulting posterior coclustering matrix looking for two and three clusters. The results are shown in the central and right panels of Figure 4. Cluster 1 contains all the voxels characterized by smallest *cluster shrinkage* parameter. Thus, these voxels can be deemed as inactive. In the right plot, we can also distinguish between two additional clusters characterized by low and high magnitude groups of coefficients, therefore providing an informative ranking between the types of voxel activations as lowly active (Cluster 3) and highly active (Cluster 2), respectively.

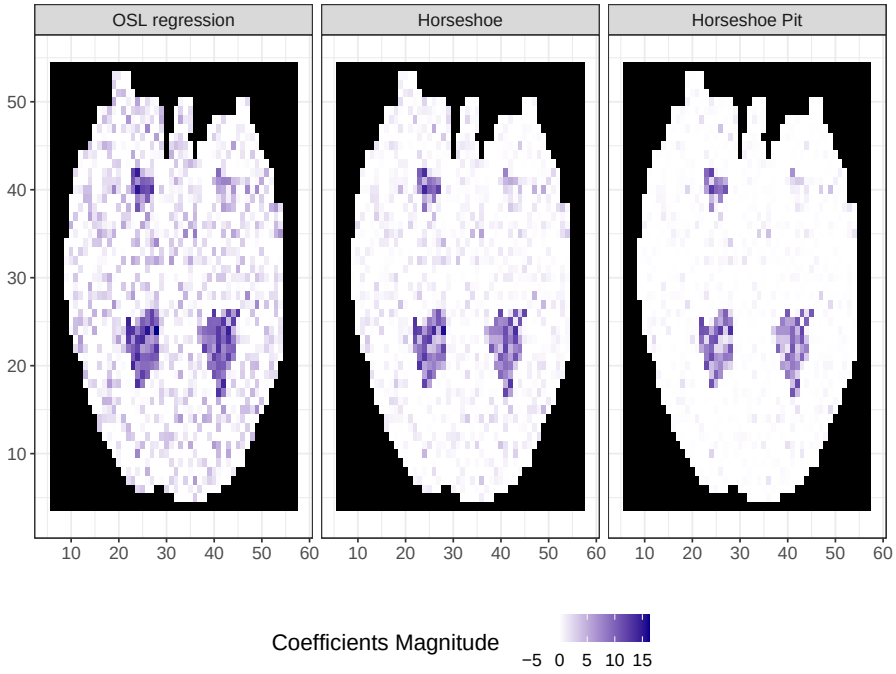

Figure 5: Brain heatmaps colored according to the magnitude of the coefficient estimates of the three different models: simple OLS, Horhseshoe (HS) and Horseshoe Pit (HSP).

5 Application to Brain Wide Anatomics in Mouse

We apply our method to a real problem involving cellular-resolution mapping of differences in neuronal activity across brain regions. Advanced methods in optical tissue clearing and light sheet fluorescence microscopy (LSFM) provide sub-micron resolution three-dimensional snapshots of fluorescence labeled postmortem mouse brains (Richardson and Lichtman, 2015; Renier et al., 2016). We leverage this novel imaging technology to measure differences in neuronal activity by labeling immediate-early gene (IEG) protein products with fluorescent antibodies. IEGs are rapidly transcribed following neuronal activity and their protein products can be measured within minutes to assess a neuron's recent activity.

Antibody labeling of IEGs is used extensively in thin-section microscopy, and more recently in intact, optically cleared tissues (Lin et al., 2008; Renier et al., 2016). This novel methodology differs from the classical f-MRI experiment: the data are not dynamic, but are collected at a very high resolution. Thus, we employ the HSP model in a multiple testing framework to detect differentially activated regions. As we will show, our methods is able to overcome the classical binary classification (relevant vs. non-relevant) and provide more insights ranking of the differential signals according to their importance. Specifically, to test the efficacy of our method we looked at the effect of light exposure

on neuronal activity across brain regions.

The brain activity is assessed by measuring up-regulated IEG Npas4. The experiment was devised as follows: 14 mice were stationed in the dark for 24h and then exposed to ambient light. The brains of 6 (baseline group) mice were examined 0-15min after light exposure. The brains of the other 8 mice (light-exposed group) were examined 30-120min after light exposure, within the window of Npas4 protein up-regulation (Ramamoorthi et al., 2011). This experimental design aimed to detect brain regions differentially activated between the baseline and light-exposed groups. A more detailed description of the experiment and an example of the resulting brain imaging is reported in Appendix F.

To identify activated neurons, we used brainQuant3D augmented with a custom machine-learning enabled classifier to segment only activated neurons (Schneider et al., 2019). Prior to segmentation, intensity data are standardized to remove variation in background fluorescence across samples as $\texttt{intensity} = \frac{i-\tilde{\mu}}{\tilde{\sigma}}$ where i is the voxel intensity and $\tilde{\mu}$ and $\tilde{\sigma}$ represent the image total intensity mean and standard deviation, respectively. We aligned the imaged brain volumes to the annotated Allen reference brain atlas (Allen Institute, ARAv2). Brain regions in the ARAv2 are hierarchically organized and a terminal region (`id`) is assigned to each neuron. We refer to the `parent` region of a neuron c (i.e., its closest ancestor) as $\text{PID}(c)$. In total, we analyzed the distribution of over 300,000 neurons. In summary, this workflow allows us to obtain the location of active neurons in a common three-dimensional reference space and extract their `intensity` and `volume` with remarkable precision. Because Npas4 protein induction is not binary, the degree of induction within a neuron is used to weigh each count and avoid arbitrary intensity thresholds. The intensity per unit of volume ($\texttt{iov}=\frac{\texttt{intensity}}{\texttt{volume}}$) is the main variable of interest.

We expect that light exposure induces widespread visually evoked activity. To exemplify, Figure 6 displays the brains of two representative mice sampled from the two different experimental conditions (baseline and light-exposed, respectively). Every dot represents a neuron c, colored according to its parent regions $\text{PID}(c)$. The size of each dot corresponds to the neuron's volume. In Appendix F, we report the main summary statistics of the frequencies of neurons recorded in the various `parents` across all mice, stratified by exposure level along with additional descriptive violin plots. We can appreciate that the activated neuron count is higher in the light-exposed group of mice.

First, we seek to remove the potential distortion in the intensity given by specific mouse effects and the potential influence of ancestor areas. As a simple solution, we regress the variable `iov` on all possible interactions between the mouse identifiers and the ancestor identifiers. We denote the resulting residual for each neuron as r_c. To take into account the frequency of the neurons, we multiply r_c by the density of neurons per unit of `parent` volume. This way, we obtain a new variable of interest: $\tilde{r}_c = r_c * \frac{n_{\text{PID}(c)}}{\text{Vol}_{\text{PID}(c)}}$. Finally, we filter out all the brain regions that contain less than 15 neurons. A total of $N = 281$ regions remain in our analysis after this step.

We start our comparison by testing the differential activation of each brain regions using

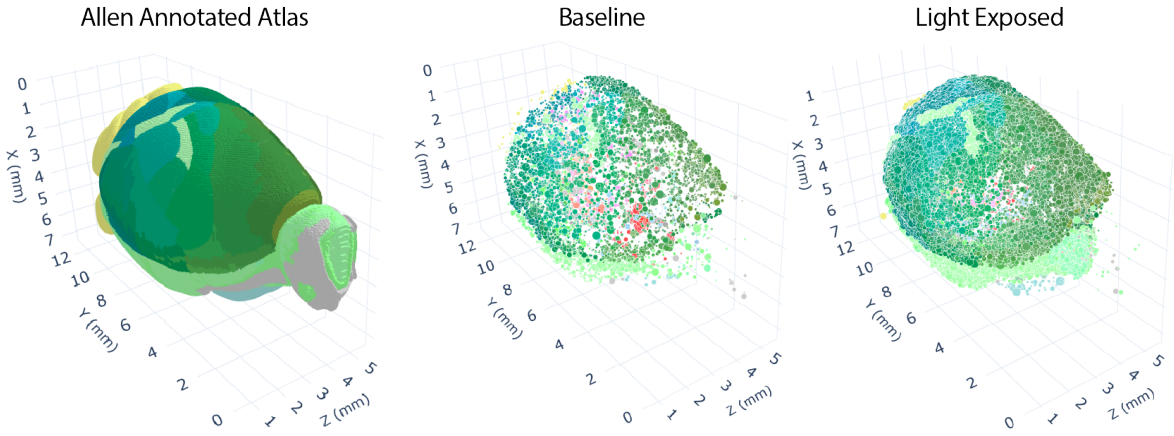

Figure 6: Comparison between detected Npas4 expressing neurons in brains of two representative mice exposed to different experimental conditions (Allen annotated atlas- left, baseline- middle, light-exposed- right). Each dot represents a neuron, the size and color of which represent the neuron's volume and `parent` partition, respectively.

a two-sided Welch's t-test, comparing the average $\tilde{r}$ baseline vs. light-exposed expression values in each region. We obtain the t-statistics $\boldsymbol{t} = \{t_i\}_{i=1}^N$, the values of degrees of freedom estimated by the Welch–Satterthwaite equation $\boldsymbol{d} = \{d_i\}_{i=1}^N$, and the corresponding p-values $\boldsymbol{p} = \{p_i\}_{i=1}^N$. The p-values are post-processed following Benajmini and Hochberg (`BH`, 1995) and thresholded at 5% to detect the activated regions. The results provide a benchmark for later comparisons. To obtain a second benchmark, we also employ the Efron's empirical Bayes two-group model (`locfdr`, Efron, 2007). To do so, we first transform the t-statistics to z-scores: $z_i = \Phi^{-1}\left(F_{T_{d_i}}(t_i)\right)\ \forall i$, where Φ and F_{T_d} denotes the c.d.f. of the standard normal distribution and a Student-t distribution with d degrees of freedom, respectively. Then, we threshold the resulting local false discovery rate at 0.20, as suggested in the literature.

Finally, we apply the HSP model directly to the z-scores:

$$z_i|\beta_i, \sigma^2 \sim \mathcal{N}(\beta_i, \sigma^2), \qquad \beta_i|\lambda, \tau, \sigma \sim \sum_{l\geq 1} \pi_l\, \phi(0, \lambda_l^2\, \tau^2\, \sigma^2), \quad i = 1, \ldots, N. \tag{9}$$

As previously mentioned, the expression in (9) can be seen as a multi-group model in a multiple hypothesis testing framework. Within this setting, we will interpret the component characterized by the lowest variance as representative of the null distribution. In contrast, the other components (once ranked in increasing order) represent different degrees of relevance (Shahbaba and Johnson, 2013). To fit model (9), we adopt a Bayesian nonparametric approach employing a DP stick–breaking representation over the mixture weights. We fix $\tau^2 = 0.001$ and run 150,000 iterations as burn-in period and collect 100,000 as posterior sample.

We post-process the posterior coclustering matrix with the Medvedovic approach partitioning the z-scores into four tiers of relevance ranging from no activation (`Tier 4`) to clear activation (`Tier 1`). Figure 7 reports both the posterior mean of the coefficients and the estimated posterior probability of relevance $1 - \bar{\kappa}_i$. The elements in both panels are colored according to the tier to which they are assigned. As expected, in the left panel we see that the different levels are associated with the increasing magnitude of z-scores. We report both the posterior means (dots) and median (crosses) in the right panel. This plot helps interpret the tiers of relevance: we notice the shift from `Tier 2` to `Tier 3` in the posterior means occurring around 0.5. Therefore, our method can be seen as an extension of the two-group model, automatically detecting the null group. Moreover, after having filtered out the irrelevant observations, we can partition the remaining ones into different sets of increasing importance, capturing more information from the z-scores.

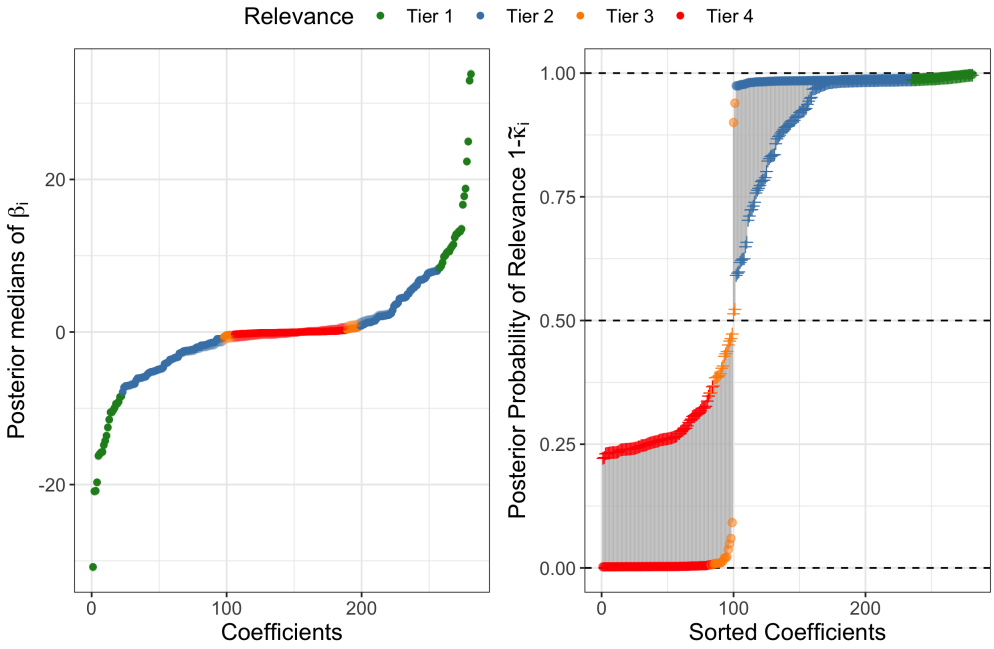

Figure 7: The left panel displays the estimated posterior mean superimposed onto the original data (transparent dots). The right panel shows the mean and median posterior probability of relevance, approximated as the complement to one of the cluster shrinkage factors $\tilde{\kappa}_i$, linked with a gray vertical line to highlight the difference.

It is interesting to compare the results obtained by `BH`, `locfdr`, and HSP. In Table 1 we report the confusion matrices comparing the allocations in classes of relevance of the different methods. The `locfdr` method is the most conservative, detecting only 38 regions that are part of HSP's `Tier 1`. HSP and `BH` are more concordant. Overall, HSP detects 38 additional regions assigned to `Tier 2`, while we also observe a separation

HSP Tier	1	2	3	4
`BH - Relevant`	47	95	0	0
`BH - Irrelevant`	0	38	18	83
`locfdr - Relevant`	38	0	0	0
`locfdr - Irrelevant`	9	133	18	83

Table 1: Comparison of the results obtained with the `BH` and `locfdr` procedures vs. the HSP allocation in tiers of relevance computed on all the brain regions.

between the `BH` relevant regions into the top two tiers. Given this result, we sought to identify whether HSP was better at selecting known visually responsive regions.

Visual cortex comprises a collection of posterior cortex regions that provide low level visual feature extraction and are strongly modulated by visual activity (Hübener, 2003). Like other cortical regions, visual cortex regions can be divided into layers (Hübener, 2003). Compared to `BH`, HSP identifies 3 more regions from the visual regions laminae. Moreover, HSP identifies 4 additional regions of hippocampus, which is known to be involved in visual memory formation. For this type of data, HSP appears to model the underlying activity more faithfully. However, further investigations are needed to reveal more regional patterns of statistically significant activation.

6 Discussion

This paper has introduced a novel shrinkage prior for variable selection and multiple hypothesis testing. Our approach consists in adopting a discrete mixture model, reminiscent of the two-group models, where each mixture component is itself a continuous scale mixture distribution. In this way, we can retain the strong shrinkage properties of the continuous mixtures while performing model-based clustering typical of the discrete mixtures. The clustering enables the detection of the irrelevant coefficients and potentially segments the relevant ones into classes of relevance, according to the magnitude of each coefficient. The combination of the two approaches shows promising results, especially in targeting and regularizing the null coefficients via the clustering-induced shrinkage structure. Half-Cauchy priors are adopted for the shrinkage parameters, mimicking the Horseshoe model. We have derived theoretical results regarding the shrinkage properties and showcased the potential of our approach in simulated scenarios. Our proposal paves the way for many future research questions. First, different continuous scale mixture types can be considered to improve the HSP model: either taking into consideration a double Exponential prior generalizing the model in Rockova and George (2016), or more refined HS distributions such as the HS-like distribution (Bhadra et al., 2019b) or the Regularized HS (Piironen and Vehtari, 2017). Second, when massive datasets are analyzed, the devised Gibbs sampler could be too expensive to employ. Efficient MCMC alternatives can be explored, such as the two algorithms for horseshoe estimation recently

proposed in Johndrow et al. (2020). Alternatively, an approximate inference method such as mean-field variational Bayes (Neville et al., 2014) can be adopted.

References

Yoav Benajmini and Yosef Hochberg. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *J R Statist Soc B*, 57(1):289–300, 1995. ISSN 00359246. doi: 10.2307/2346101. URL http://www.stat.purdue.edu/$\sim$doerge/BIOINFORM.D/FALL06/BenjaminiandYFDR.pdf%5Cnhttp://engr.case.edu/ray_soumya/mlrg/controlling_fdr_benjamini95.pdf.

Anindya Bhadra, Jyotishka Datta, Nicholas G. Polson, and Brandon Willard. The horseshoe+estimator of ultra-sparse signals. *Bayesian Analysis*, 12(4):1105–1131, 2017. ISSN 19316690. doi: 10.1214/16-BA1028.

Anindya Bhadra, Jyotishka Datta, Nicholas G. Polson, and Brandon Willard. Lasso meets horseshoe: A survey. *Statistical Science*, 34(3):405–427, 2019a. ISSN 21688745. doi: 10.1214/19-STS700.

Anindya Bhadra, Jyotishka Datta, Nicholas G. Polson, and Brandon T. Willard. The Horseshoe-Like Regularization for Feature Subset Selection. *Sankhya B*, 2019b. ISSN 09768394. doi: 10.1007/s13571-019-00217-7.

Anirban Bhattacharya, Debdeep Pati, Natesh S. Pillai, and David B. Dunson. Dirichlet–Laplace Priors for Optimal Shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490, 2015. ISSN 1537274X. doi: 10.1080/01621459.2014.960967.

Anirban Bhattacharya, Antik Chakraborty, and Bani K. Mallick. Miscellanea fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, 103(4):985–991, 2016. ISSN 14643510. doi: 10.1093/biomet/asw042.

Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010. ISSN 00063444. doi: 10.1093/biomet/asq017.

Dawei Ding and George Karabatsos. Dirichlet process mixture models with shrinkage prior. *Stat*, 10(1), 2021. ISSN 2049-1573. doi: 10.1002/sta4.371. URL http://arxiv.org/abs/2010.11385.

Bradley Efron. Size, power and false discovery rates. *Annals of Statistics*, 35(4):1351–1377, 2007. ISSN 00905364. doi: 10.1214/009053606000001460.

Thomas S. Ferguson. A Beyesian Analysis of Some non-parametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

Michael Finegold and Mathias Drton. Robust graphical modeling of gene networks using classical and alternative t-distributions. *Annals of Applied Statistics*, 5(2 A):1057–1080, 2011. ISSN 19326157. doi: 10.1214/10-AOAS410.

Michael Finegold and Mathias Drton. Robust Bayesian graphical modeling using dirichlet t-distributions. *Bayesian Analysis*, 9(3):521–550, 2014. ISSN 19316690. doi: 10.1214/13-BA856.

Edward I. George and Robert E. McCulloch. Approaches for bayesian variable selection. *Statistica Sinica*, 7(2):339–373, 1997. ISSN 10170405.

Jim E. Griffin and Philip J. Brown. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010. ISSN 19360975. doi: 10.1214/10-BA507.

Mark Hübener. Mouse visual cortex, aug 2003. ISSN 09594388.

Hemant Ishwaran and Lancelot F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001. ISSN 1537274X. doi: 10.1198/016214501750332758. URL http://www.tandfonline.com/doi/abs/10.1198/016214501750332758.

James Johndrow, Paulo Orenstein, and Anirban Bhattacharya. Scalable approximate MCMC algorithms for the horseshoe prior. *Journal of Machine Learning Research*, 21, 2020. ISSN 15337928.

Qing Li and Nan Lin. The Bayesian elastic net. *Bayesian Analysis*, 5(1):151–170, 2010. ISSN 19360975. doi: 10.1214/10-BA506.

Yingxi Lin, Brenda L. Bloodgood, Jessica L. Hauser, Ariya D. Lapan, Alex C. Koon, Tae Kyung Kim, Linda S. Hu, Athar N. Malik, and Michael E. Greenberg. Activity-dependent regulation of inhibitory synapse development by Npas4. *Nature*, 455(7217):1198–1204, oct 2008. ISSN 14764687. doi: 10.1038/nature07319. URL https://www.nature.com/articles/nature07319.

Richard F. MacLehose and David B. Dunson. Bayesian semiparametric multiple shrinkage. *Biometrics*, 66(2):455–462, 2010. ISSN 0006341X. doi: 10.1111/j.1541-0420.2009.01275.x.

Enes Makalic and Daniel F. Schmidt. High-Dimensional Bayesian Regularised Regression with the BayesReg Package. *ArXiv Preprint*, 2016. URL http://arxiv.org/abs/1611.06649.

Gertraud Malsiner-Walli, Sylvia Frühwirth-Schnatter, and Bettina Grün. Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*, 26(1-2):303–324, 2016. ISSN 15731375. doi: 10.1007/s11222-014-9500-2.

Robert E. McCulloch and Edward I. George. Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993. ISSN 0162-1459. URL http://apps.isiknowledge.com.www.lib.ncsu.edu:2048/full{_}record.do?product=WOS{&}search{_}mode=GeneralSearch{&}qid=66{&}SID=3AGpa7HOGgHh8kJ75fp{&}page=1{&}doc=6{%}5Cnhttp://www.jstor.org.www.lib.ncsu.edu:2048/stable/pdfplus/2290777.pdf.

Mario Medvedovic, K. Y. Yeung, and R. E. Bumgarner. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20(8):1222–1232, 2004. ISSN 13674803. doi: 10.1093/bioinformatics/bth068.

T. J. Mitchell and J. J. Beauchamp. Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 83(404):1023, 1988. ISSN 01621459. doi: 10.2307/2290129.

Sarah E. Neville, John T. Ormerod, and M. P. Wand. Mean field variational bayes for continuous sparse signal shrinkage: Pitfalls and remedies. *Electronic Journal of Statistics*, 8:1113–1151, 2014. ISSN 19357524. doi: 10.1214/14-EJS910.

Trevor Park and George Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008. ISSN 01621459. doi: 10.1198/016214508000000337.

Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051, 2017. ISSN 19357524. doi: 10.1214/17-EJS1337SI.

Nicholas G. Polson, James G. Scott, Bertrand Clarke, and C. Severinski. Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction. *Bayesian Statistics*, 9, 2012. doi: 10.1093/acprof:oso/9780199694587.003.0017.

Nicholas G. Polson, James G. Scott, and Jesse Windle. The Bayesian bridge. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 76(4):713–733, 2014. ISSN 14679868. doi: 10.1111/rssb.12042.

Kartik Ramamoorthi, Robin Fropf, Gabriel M. Belfort, Helen L. Fitzmaurice, Ross M. McKinney, Rachael L. Neve, Tim Otto, and Yingxi Lin. Npas4 regulates a transcriptional program in CA3 required for contextual memory formation. *Science*, 334 (6063):1669–1675, dec 2011. ISSN 10959203. doi: 10.1126/science.1208049. URL http://science.sciencemag.org/.

Nicolas Renier, Eliza L. Adams, Christoph Kirst, Zhuhao Wu, Ricardo Azevedo, Johannes Kohl, Anita E. Autry, Lolahon Kadiri, Kannan Umadevi Venkataraju, Yu Zhou, Victoria X. Wang, Cheuk Y. Tang, Olav Olsen, Catherine Dulac, Pavel Osten, and Marc Tessier-Lavigne. Mapping of Brain Activity by Automated Volume Analysis of Immediate Early Genes. *Cell*, 165(7):1789–1802, jun 2016. ISSN 00928674. doi: 10.1016/j.cell.2016.05.007. URL http://www.cell.com/cell/fulltext/S0092-8674(16)30555-4http://linkinghub.elsevier.com/retrieve/pii/S0092867416305554.

Douglas S. Richardson and Jeff W. Lichtman. Clarifying Tissue Clearing. *Cell*, 162 (2):246–257, 2015. ISSN 00928674. doi: 10.1016/j.cell.2015.06.067. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867415008375.

Veronika Rockova and Edward I George. The Spike-and-Slab LASSO. *Journal of the American Statistical*, 2016.

Håvard Rue. Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 63(2):325–338, 2001. ISSN 13697412. doi: 10.1111/1467-9868.00288.

Christine A. Schneider, Dario X. Figueroa Velez, Ricardo Azevedo, Evelyn M. Hoover, Cuong J. Tran, Chelsie Lo, Omid Vadpey, Sunil P. Gandhi, and Melissa B. Lodoen. Imaging the dynamic recruitment of monocytes to the blood–brain barrier and specific brain regions during toxoplasma gondii infection. *Proceedings of the National Academy of Sciences*, 116(49):24796–24807, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1915778116. URL https://www.pnas.org/content/116/49/24796.

J. Sethuraman. A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4: 639–650, 1994. URL http://www3.stat.sinica.edu.tw/statistica/j4n2/j4n27/..{%}5Cj4n216{%}5Cj4n216.htm.

Babak Shahbaba. Comment on Robust Bayesian Graphical Modeling Using Dirichlet t-Distributions. *Bayesian Analysis*, 9(3):557–560, 2014.

Babak Shahbaba and Wesley O. Johnson. Bayesian nonparametric variable selection as an exploratory tool for discovering differentially expressed genes. *Statistics in Medicine*, 32(12):2114–2126, 2013. ISSN 02776715. doi: 10.1002/sim.5680.

Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 2013.

Marijke Welvaert, Joke Durnez, Beatrijs Moerkerke, Geert Verdoolaege, and Yves Rosseel. neuRosim: An R package for generating fmri data. *Journal of Statistical Software*, 44(10):1–18, 2011. URL http://www.jstatsoft.org/v44/i10/.

Hongxia Yang, David Dunson, and DL Banks. The Multiple Bayesian Elastic Net. *Stat.Duke.Edu*, pages 1–39, 2011. URL http://stat.duke.edu/{~}hy35/MBEN.pdf.

Supplementary Material for "A Horseshoe Pit mixture model for Bayesian screening with an application to light sheet fluorescence microscopy in brain imaging"

Francesco Denti*[1], Ricardo Azevedo[2,3], Chelsie Lo[2,3], Damian Wheeler[3], Sunil P. Gandhi[2,3,4], Michele Guindani[1] and Babak Shahbaba[1]

[1]Department of Statistics, University of California, Irvine
[2]Department of Neurobiology and Behavior, University of California, Irvine
[3]Translucence Biosystems, Inc, Irvine
[4]Center for the Neurobiology of Learning and Memory, University of California, Irvine

*fdenti@uci.edu

Appendix A: the Dirichlet-HS distribution and robust modeling

To provide an additional perspective, we can relate our regularization prior to the "classical" vs. "alternative" paradigm discussed by Finegold and Drton (2011) in terms of robust modeling. Suppose a generic random variable $\boldsymbol{Y}$ is distributed according to a continuous scale mixture of Normals. In that case, it can be equivalently represented as $\boldsymbol{Y} = \boldsymbol{X}\rho$, where $\boldsymbol{X}$ has the multivariate standard Gaussian distribution and ρ is a random scale parameter. It is well-known that, if $\rho \sim Gamma(\nu/2, \nu/2)$, then $\boldsymbol{Y}$ is distributed as a multivariate Student-t. In the "classical" case, ρ is univariate and shared across all the coordinates of $\boldsymbol{X} = \{X_j\}_{j=1}^p$. This case is equivalent to a regularization model with a unique, random global shrinkage parameter $\rho = \tau$. The opposite situation, where each entry X_j is paired with a unique scale ρ_j, is referred to as the "alternative" multivariate Student-t. This case corresponds to the classical global-local shrinkage models with deterministic $\bar{\tau}$, where $\rho_j = \bar{\tau}\lambda_j$. In Finegold and Drton (2014), the authors propose a third distribution, Dirichlet-t, assuming $\tau_j \sim G$ and $G \sim DP(\alpha, Gamma(\nu/2, \nu/2))$. This last option is directly linked with our proposal.

Following their definition, once we assume a fixed global shrinkage parameter $\tau = \bar{\tau}$ and $\lambda_l \sim \mathcal{C}^+$, both models (2) and (4) in the main paper can be seen as the parametric and nonparametric versions of a novel Dirichlet-HS distribution, respectively. Shahbaba (2014) suggested to use this structure in the context of robust modeling, given the appealing properties of the Horseshoe distribution. This argument can be generalized to Dirichlet-ρ distributions, defined by considering different specifications for the scale parameter ρ. Notice that, despite similar names, these distributions are essentially different from the one proposed by Bhattacharya et al. (2015). Even in this context, the gain is twofold. First, we obtain a distribution that is more flexible than the "classic" one. Second, we allow sharing of statistical strength across the elements of the random vector even without assuming a distribution for τ, for which special care is needed to carry out posterior simulation (Piironen and Vehtari, 2017). Moreover, we gain in terms of computational properties by greatly reducing the number of parameters. Finegold and Drton (2014) highlight that the structure of a distribution such as the generic Dirichlet-ρ interpolates between the two extreme model specifications ("classical" vs. "alternative"). We add that, even in the simple case of a parametric mixture, model (2) in the main paper gives us direct control of where the resulting distribution takes place between the two extremes by tweaking the prior over the mixture weights. Figure 8 provides an example with the Dirichlet-HS distribution with $L = 10$. The overfitted mixture case ($a \leq 0.05$) is the closest one to the "classical" model, which can be recovered for $a \to 0$. As a increases, if L is large enough, the likelihood of sampling distinct values of ρ_j for all j increases, and therefore we get closer to the "alternative" model.

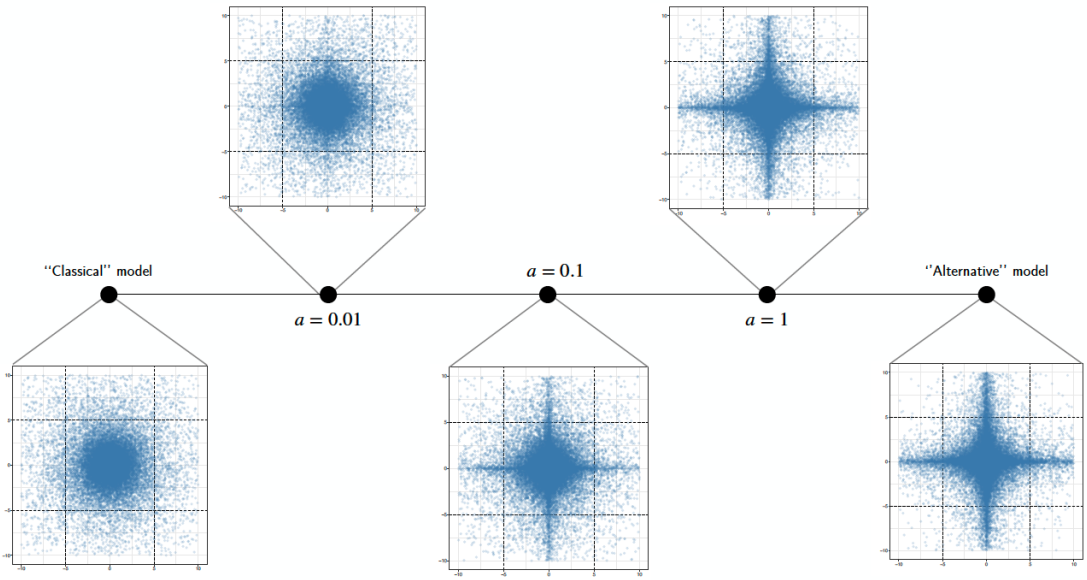

Figure 8: Parametric Dirichlet-HS realizations induced by different specification of the Dirichlet distribution adopted for $L = 10$ mixture weights.

Appendix B: Distributional derivations

Marginal Likelihood

Let $\sigma^2 = 1$ without loss of generality. Then, the sampling distribution for Y_j, $j = 1, \ldots, n$ after marginalizing out the parameter vector $\boldsymbol{\beta}$ is obtained as:

$$\begin{aligned}
\pi(Y_j|\tau, \boldsymbol{\lambda}, \boldsymbol{\pi}) &= \int \pi(Y_j|\boldsymbol{\beta})\pi(\boldsymbol{\beta}|\tau, \boldsymbol{\lambda}, \boldsymbol{\pi})d\boldsymbol{\beta} \\
&= \int \pi(Y_j|\beta_j)\pi(\beta_j|\tau, \boldsymbol{\lambda}, \boldsymbol{\pi})d\beta_j \\
&= \int \frac{1}{\sqrt{2\pi}} e^{-\frac{(Y_j-\beta_j)^2}{2}} \left(\sum_{l=1}^{L} \pi_l \frac{1}{\sqrt{2\pi\tau^2\lambda_l^2}} e^{-\frac{\beta_j^2}{2\tau^2\lambda_l^2}} \right) d\beta_j \\
&= \sum_{l=1}^{L} \pi_l \frac{1}{2\pi\sqrt{\tau^2\lambda_l^2}} \int e^{-\frac{(Y_j-\beta_j)^2}{2} - \frac{\beta_j^2}{2\tau^2\lambda_l^2}} d\beta_j \\
&= \sum_{l=1}^{L} \pi_l \frac{e^{-\frac{Y_j^2}{2}}}{2\pi\sqrt{\tau^2\lambda_l^2}} \int e^{-0.5\left(\frac{1+\tau^2\lambda_l^2}{\tau^2\lambda_l^2}\beta_j^2 - 2\beta_j Y_j\right)} d\beta_j \\
&= \sum_{l=1}^{L} \pi_l \frac{1}{2\pi\sqrt{\tau^2\lambda_l^2}} \sqrt{\frac{\tau^2\lambda_l^2}{1+\tau^2\lambda_l^2}} e^{-\frac{Y_j^2}{2}\left(1 - \frac{\tau^2\lambda_l^2}{1+\tau^2\lambda_l^2}\right)} \\
&= \sum_{l=1}^{L} \frac{\pi_l}{\sqrt{2\pi(1+\tau^2\lambda_l^2)}} e^{-\frac{Y_j^2}{2(1+\tau^2\lambda_l^2)}}.
\end{aligned}$$

Therefore, $Y_j|\tau, \boldsymbol{\lambda}, \boldsymbol{\pi} \sim \sum_{l=1}^{L} \pi_l \mathcal{N}(0, 1+\tau^2\lambda_l^2)$. If σ^2 is supposed to be stochastic, we would obtain $Y_j|\tau, \boldsymbol{\lambda}, \boldsymbol{\pi}, \sigma^2 \sim \sum_{l=1}^{L} \pi_l \mathcal{N}(0, \sigma^2(1+\tau^2\lambda_l^2))$.

Posterior mean

Again, let us suppose $\sigma^2 = 1$. We can derive the posterior mean of the parameter β_j as follows.

$$\begin{aligned}
\mathbb{E}\left[\beta_j|\tau,\boldsymbol{\lambda},\boldsymbol{\pi},Y_j\right] &= \int \beta_j \frac{\pi(Y_j,\beta_j,\tau,\boldsymbol{\lambda},\boldsymbol{\pi})}{\pi(Y_j,\tau,\boldsymbol{\lambda},\boldsymbol{\pi})}d\beta_j = \frac{\pi(\tau,\boldsymbol{\lambda},\boldsymbol{\pi})}{\pi(Y_j,\tau,\boldsymbol{\lambda},\boldsymbol{\pi})}\int \beta_j \pi(Y_j|\beta_j)\pi(\beta_j|\tau,\boldsymbol{\lambda})d\beta_j \\
&= \frac{1}{\pi(Y_j|\tau,\boldsymbol{\lambda},\boldsymbol{\pi})}\int \beta_j \pi(Y_j|\beta_j)\pi(\beta_j|\tau,\boldsymbol{\lambda})d\beta_j \\
&= \sum_{l=1}^{L}\pi_l \frac{e^{-\frac{Y_j^2}{2}\left(1-\frac{\tau^2\lambda_l^2}{1+\tau^2\lambda_l^2}\right)}}{\pi(Y_j|\tau,\boldsymbol{\lambda},\boldsymbol{\pi})\sqrt{2\pi(1+\tau^2\lambda^2)}}\int \beta_j \phi\left(\beta_j; Y_j\frac{\tau^2\lambda_l^2}{1+\tau^2\lambda_l^2}, \frac{\tau^2\lambda_l^2}{1+\tau^2\lambda_l^2}\right)d\beta_j \\
&= \sum_{l=1}^{L}\pi_l \frac{e^{-\frac{Y_j^2}{2(1+\tau^2\lambda_l^2)}}}{\pi(Y_j|\tau,\boldsymbol{\lambda},\boldsymbol{\pi})\sqrt{2\pi(1+\tau^2\lambda_l^2)}} Y_j \frac{\tau^2\lambda_l^2}{1+\tau^2\lambda_l^2} \\
&= \sum_{l=1}^{L}\pi_l \frac{\phi(Y_j;0,1+\tau^2\lambda_l^2)}{\pi(Y_j|\tau,\boldsymbol{\lambda},\boldsymbol{\pi})} Y_j(1-\kappa_l^*) \\
&= \frac{\sum_{l=1}^{L}\pi_l\phi(Y_j;0,1+\tau^2\lambda_l^2)(1-\kappa_l^*)}{\pi(Y_j|\tau,\boldsymbol{\lambda},\boldsymbol{\pi})} Y_j \\
&= \frac{\sum_{l=1}^{L} r_l(Y_j)(1-\kappa_l^*)}{\sum_{l=1}^{L} r_l(Y_j)} Y_j.
\end{aligned}$$

where $1-\kappa_k^* = \frac{\tau^2\lambda_l^2}{1+\tau^2\lambda_l^2}$ and $r_l(Y_j) = \pi_l\phi(Y_j;0,1+\tau^2\lambda_l^2)$.

Appendix C: Gibbs Sampler for the HSP model

1. Let $\mathbf{\Lambda}_Z = \tau^2 \cdot diag(\lambda^2_{z_1}, \ldots, \lambda^2_{z_p})$ Sample $\beta \sim \mathcal{N}_p\left(A^{-1}\boldsymbol{X}'\boldsymbol{Y}, \sigma^2 A^{-1}\right)$, where $A = (\boldsymbol{X}'\boldsymbol{X} + \mathbf{\Lambda}_Z)$. To efficiently sample from this distribution we follow Makalic and Schmidt (2016) and employ the algorithm of Rue (2001) when $p/n \leq 2$, while we use Bhattacharya et al. (2016) otherwise.

2. Sample z_j according to $\pi(z_j = l) \propto \pi_l \cdot \phi(\beta_j; 0; \sigma^2 \cdot \tau^2 \cdot \lambda^2_l)$.
Then, compute $n_l = \#\{j : z_j = l\}$ for $l = 1, \ldots, L$.

3. Introduce the auxiliary variable u_{λ_l}. Let $t_l = 1/\lambda^2_l$. Then, sample $u_{\lambda_l} \sim \mathcal{U}\left(0, 1/(1+t_l)\right)$ and
$$t_l \sim G\left(\frac{(n_l+1)}{2}, \frac{\sum_{j:z_j=l}\beta^2_l}{2\tau^2\sigma^2}\right)\mathbb{1}_{t_l \in [0, 1/u_{\lambda_l}-1]}.$$

4. Introduce the auxiliary variable u_τ. Let $t^* = 1/\tau^2$. Then, sample $u_\tau \sim \mathcal{U}\left(0, 1/(1+t^*)\right)$ and
$$t^* \sim G\left(\frac{(p+1)}{2}, \frac{\sum_j \beta^2_j/\lambda^2_{z_j}}{2\sigma^2}\right)\mathbb{1}_{t^* \in [0, 1/u_\tau - 1]}.$$

5. Sample the error variance
$$\sigma^2 \sim IG\left(\frac{n+p}{2}, \frac{\sum_n (y_n - X_n\beta)^2 + \sum_j \beta^2_j/\tau\lambda_{z_j}}{2}\right).$$

6. Sample the mixture weights $\boldsymbol{\pi}$ from $Dir(a_1+n_1, \ldots, a_J+n_L)$ for finite mixture models; for nonparametric mixtures, use the corresponding step for the stick-breaking construction from blocked Gibbs sampler of Ishwaran and James (2001).

Appendix D: The HSP "gravitational pull" toward zero

To show how the shrinkage effect of the HSP can exploit the similarities in the data, we consider five different datasets that we index with $s = 1, \ldots, 5$ and specify the HSP model for estimating the means. Each dataset is characterized by different sample size $n_s \in \{350, 400, 500, 600, 800\}$. Under each scenario, we generate 300 non-zero means from Normal distributions with standard deviations 10 and 3, in equal proportions. Then, we generate additional "null" observations β_i, with $i = 101, \ldots, n_s$ from a $\mathcal{N}(0, \sqrt{0.001})$ distribution. Finally, the different datasets are generated according to $\boldsymbol{Y}^s \sim \mathcal{N}_{n_s}(\boldsymbol{\beta}^s \mathbb{I}_{n_s}, \mathbb{I}_{n_s})$, where $\boldsymbol{Y}^s$ is the target variable of length n_s under scenario s. Here, we want to assess how increasing the presence of small/negligible mean values affects the shrinkage profiles and, in turn, how they affect the posterior mean and median estimates. To isolate the clustering effect, we fix $\tau^2 = 0.001$ and set $\sigma^2 = 1$ to reflect null distribution specification typical of the two-group model. We employ the HSP model using a Dirichlet process, reflecting the definition of Dirichlet-HS model in the spirit of Finegold and Drton (2014).

The results, focusing on the coefficients values ranging between -2.5 and 2.5, are reported in Figure 9. As the number of null observations increases, the precision of the "shrinkage focus" of the model improves, imposing more robust regularization on the observations. To support this statement, we discuss two effects that are evident from the two panels. First, as more null data points are added to the dataset, the gravitational pull towards zero that the mode sets on the estimates becomes stronger, imposing wider regularization. Second, more null observations help to reduce the unnecessary shrinkage imposed on the non-null ones. This effect is highlighted by the different smoothing lines in the right panel: the median estimates leave the shrinkage-affected area and reach the black symmetric line faster as the number of null observations increases.

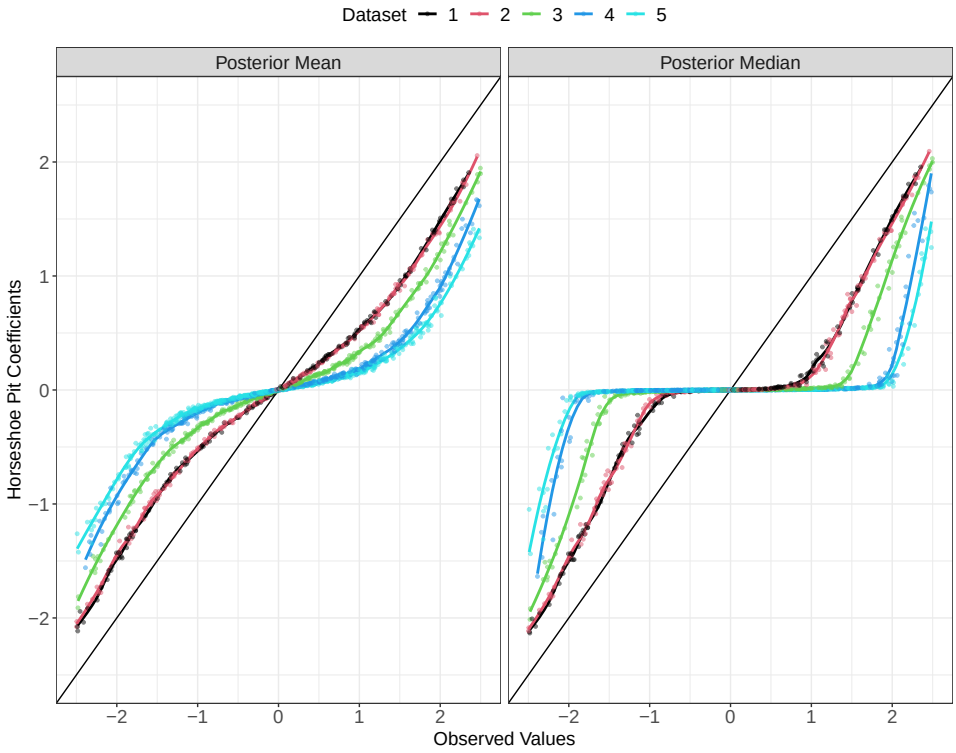

Figure 9: Different profiles of shrinkage affecting the posterior mean and median estimates as the number of irrelevant observations under different scenarios increase.

Appendix E: Additional Figures

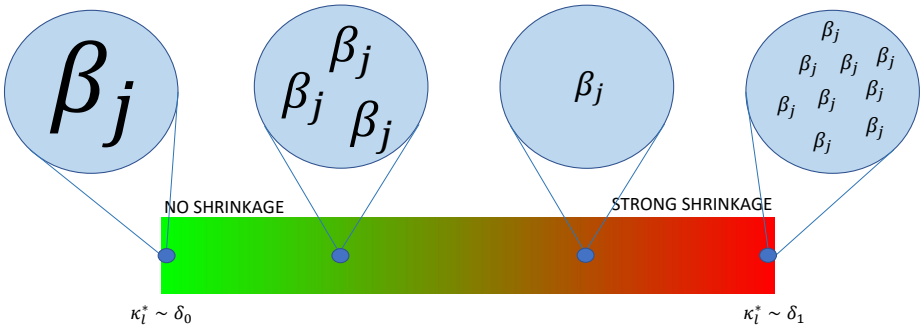

Figure 10: A visual depiction of the shrinkage induced by the HSP model. Clusters containing numerous coefficients with overall low magnitude are regularized the most (right end of the scale). On the contrary, cluster containing few, big coefficients suffer less regularization (left end of the scale).

Multivariate HSP: f-MRI application

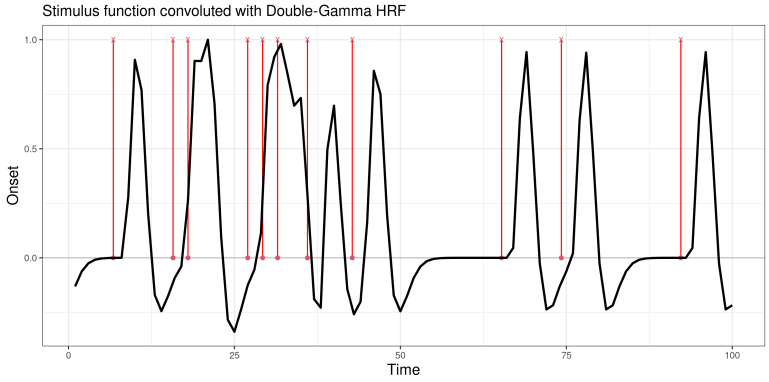

Figure 11: Event-specific stimulus function (red spikes) convoluted with a Double-Gamma HRF adopted for the analysis

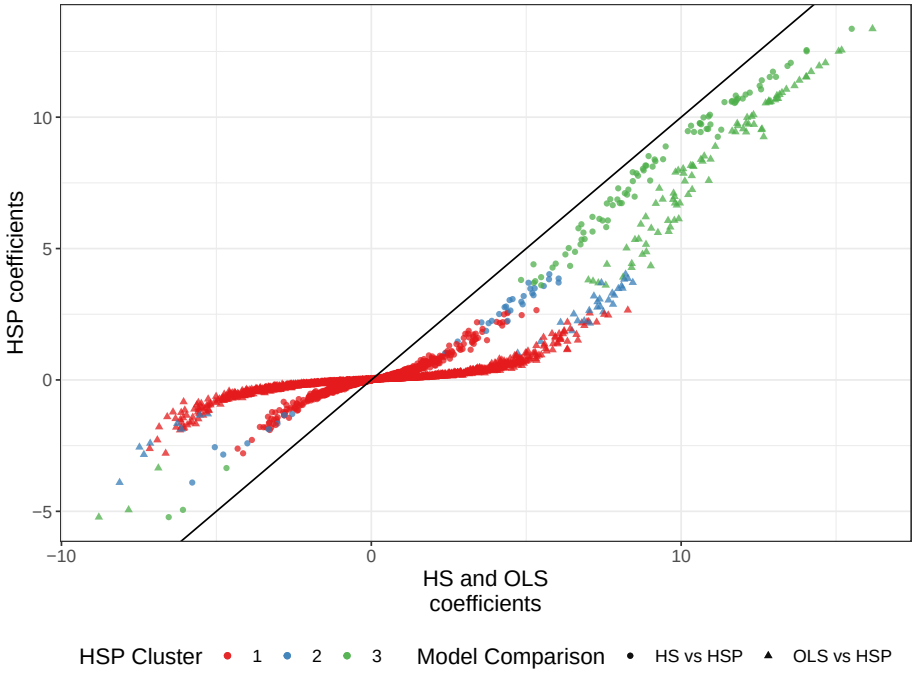

Figure 12: Scatter plot of the estimated coefficients for the simulated f-MRI application. The two different shapes identify the type of estimates (HS or OLS models) plotted against the HSP posterior means. The different colors highlight the induced partition estimated with the HSP model, thresholding the hierarchical clustering solution to find 3 clusters.

Appendix F: Additional details on Whole-Brain Anatomics Application

Detailed description of the experiment

The brain activity is assessed by measuring up-regulation of IEG Npas4, which is unique among IEGs for its high specificity for activity Lin et al. (2008). In our experiment, 14 C57BL/6 (RRID:IMSR_CRL:642) mice were individually housed in the dark for 24h to establish baseline visual activity. Mice were then transferred into a new cage exposed to ambient light. Brains of 6 mice were examined 0-15min after light exposure to serve as the `baseline` group. Brains of another 8 mice were examined 30-120min after light exposure, within the window of Npas4 protein up-regulation (Ramamoorthi et al., 2011). Next, one hemisphere was immunolabeled for Npas4 and rendered optically transparent using iDISCO+ (Renier et al., 2016). Equal numbers of left and right hemispheres were sampled. Hemispheres were then imaged on a Zeiss Z.1 light-sheet microscope with a Mesoscale Imaging System (Translucence Biosystems) at 0.91 µm x 0.91 µm x 6.81 µm resolution (Zeiss).

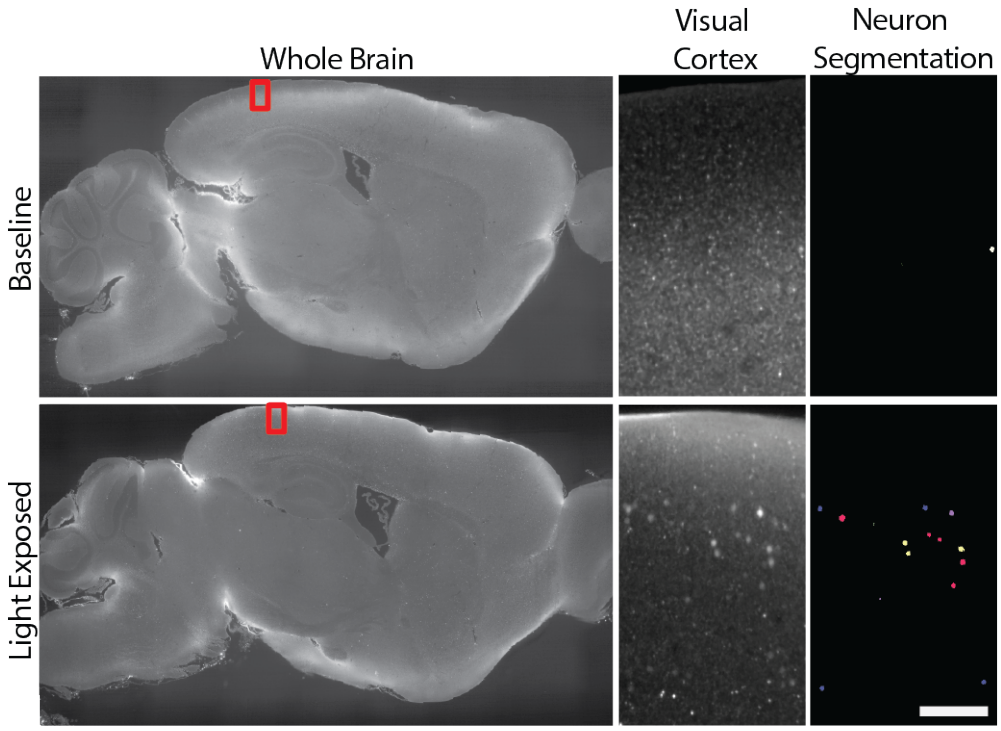

Figure 13: Example LSFM brain images from baseline (top) and light exposed (bottom) mice. Whole coronal section (left) shown with zoom in of visual cortex (red box, middle). Associated neuron segmentation shown for zoom in of visual cortex (right).

Descriptive statistics and boxplots

To obtain a complete picture of the dataset, we report four different sets of descriptive boxplots in Figure 14. Each panel represents different summary statistics computed over each `parent`, stratified by exposure level: the neurons volumes, the number of activated neurons (frequency), the same quantity divided by the volume of the `parent`, and the neuron intensities. As expected, the most noticeable differences between groups appear when the frequencies of neurons are considered 14 (panel 2,3) instead of the raw neurons' intensity (panel 1,4).

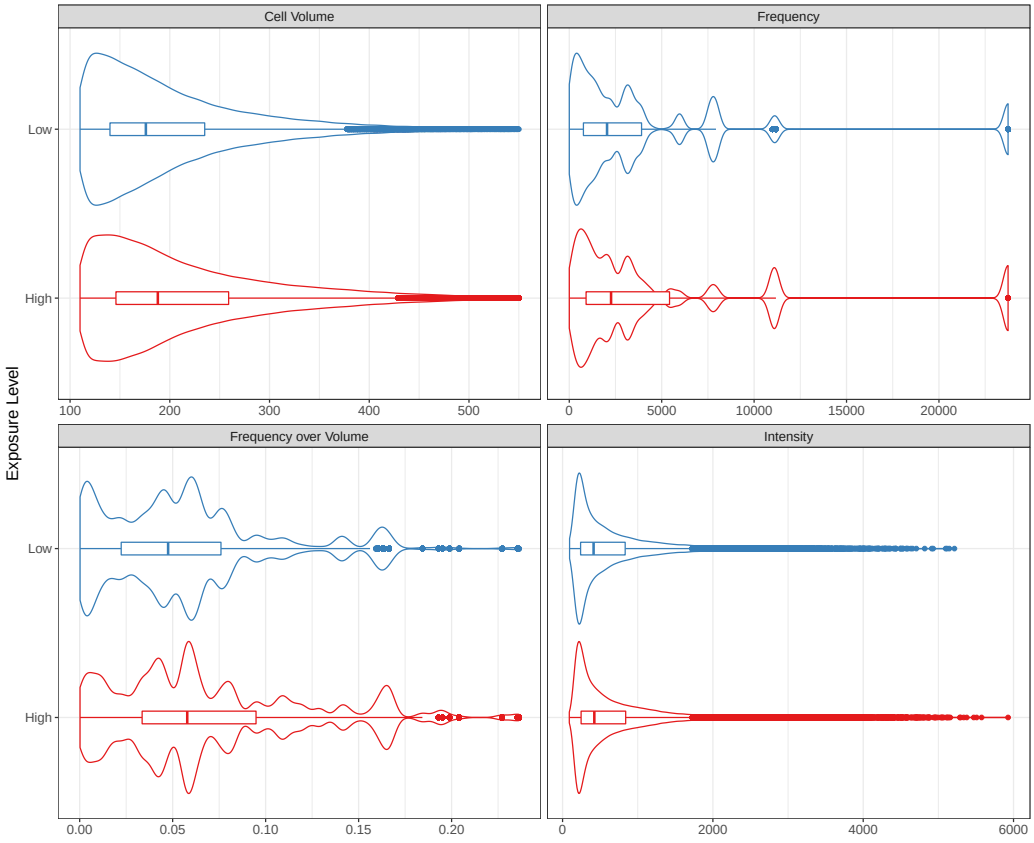

Figure 14: Distribution of descriptive summary statistics computed on the dataset. Each panel corresponds to a different quantity: neuronal volume, frequency per brain region, frequency over volume, and intensity.

Exposure Level	Min	Q25	Median	Mean	Q75	Max
Low	0	21.50	108	376.65	340.50	5507
High	0	34.00	215	1617.80	1946.00	28547

Table 2: Summary statistics of the distribution of neuron frequencies counts in `parents` across all mice, grouped by exposure levels.

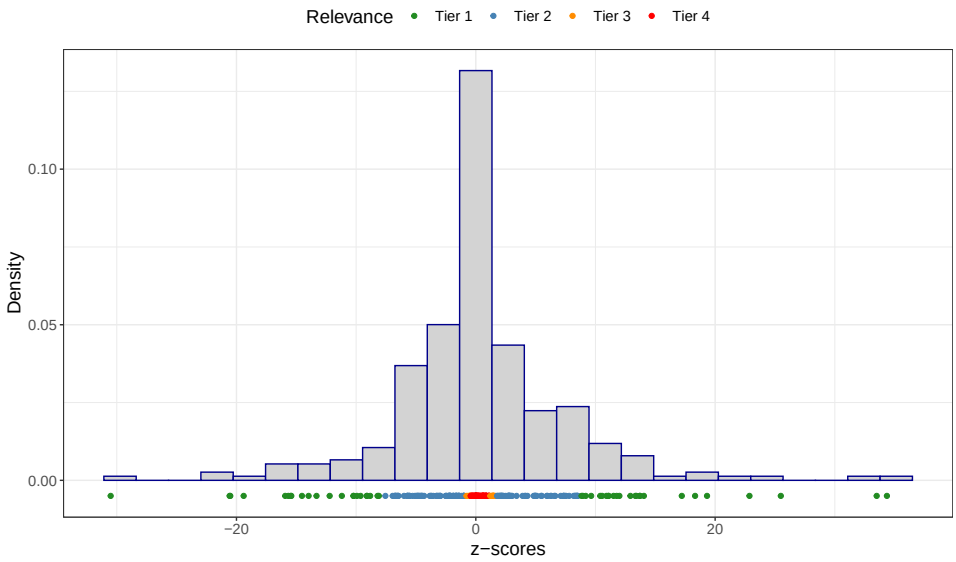

Figure 15: Histogram of the z scores for the 281 brain regions considered. The dots, representing the individual statistics, are colored according to the tiers of relevance.