


# A HORSESHOE MIXTURE MODEL FOR BAYESIAN SCREENING WITH AN APPLICATION TO LIGHT SHEET FLUORESCENCE MICROSCOPY IN BRAIN IMAGING

BY FRANCESCO DENTI<sup>1,a</sup> , RICARDO AZEVEDO<sup>2,b</sup>, CHELSIE LO<sup>2,c</sup>,  
DAMIAN G. WHEELER<sup>3,e</sup>, SUNIL P. GANDHI<sup>2,d</sup>, MICHELE GUINDANI<sup>4,f</sup> AND  
BABAK SHAHBABA<sup>5,g</sup>

<sup>1</sup>*Department of Statistics, Università Cattolica del Sacro Cuore, [francesco.denti@unicatt.it](mailto:francesco.denti@unicatt.it)*

<sup>2</sup>*Department of Neurobiology and Behavior, University of California, Irvine, [azevedor@uci.edu](mailto:azevedor@uci.edu), [lochelsie@gmail.com](mailto:lochelsie@gmail.com),  
[sunil.gandhi@uci.edu](mailto:sunil.gandhi@uci.edu)*

<sup>3</sup>*Translucence Biosystems, Inc, [damian@translucencebio.com](mailto:damian@translucencebio.com)*

<sup>4</sup>*Department of Biostatistics, University of California, Los Angeles, [mguindani@g.ucla.edu](mailto:mguindani@g.ucla.edu)*

<sup>5</sup>*Department of Statistics, University of California, Irvine, [babaks@uci.edu](mailto:babaks@uci.edu)*

In this paper we focus on identifying differentially activated brain regions using a light sheet fluorescence microscopy—a recently developed technique for whole-brain imaging. Most existing statistical methods solve this problem by partitioning the brain regions into two classes: significantly and nonsignificantly activated. However, for the brain imaging problem at the center of our study, such binary grouping may provide overly simplistic discoveries by filtering out weak but important signals that are typically adulterated by the noise present in the data. To overcome this limitation, we introduce a new Bayesian approach that allows classifying the brain regions into several tiers with varying degrees of *relevance*. Our approach is based on a combination of shrinkage priors, widely used in regression and multiple hypothesis testing problems, and mixture models, commonly used in model-based clustering. In contrast to the existing regularizing prior distributions, which use either the spike-and-slab prior or continuous scale mixtures, our class of priors is based on a *discrete mixture of continuous scale mixtures* and devises a cluster shrinkage version of the horseshoe prior. As a result, our approach provides a more general setting for Bayesian sparse estimation, drastically reduces the number of shrinkage parameters needed, and creates a framework for sharing information across units of interest. We show that this approach leads to more biologically meaningful and interpretable results in our brain imaging problem, since it allows the discrimination between active and inactive regions, while at the same time ranking the discoveries into clusters representing tiers of similar importance.

**1. Introduction.** A central goal of many neuroscience studies is to detect regional patterns of brain activation associated with an activity, preferably at cellular resolution. A recent strategy to accomplish this goal involves using thin-section microscopy. This technique allows to detect immediate-early gene (IEG) activation, that is, the coordinate activation of genes for which the transcription is fast in response to external stimuli. IEG activation is thus closely related to changes in neurons' activity (Sheng and Greenberg (1990)). By using fluorescent antibodies for labeling IEG proteins along with advanced optical tissue clearing techniques and light sheet fluorescence microscopy (LSFM), we can obtain high-resolution, three-dimensional snapshots of activity in individual neurons across the entire brain (Richardson and Lichtman (2015), Renier et al. (2016)). Using a specific IEG, the Neuronal Per-Arnt-Sim Domain Protein 4 (Npsa4—Lin et al. (2008), Sun and Lin (2016)), our

---

Received January 2022; revised January 2023.

*Key words and phrases.* Bayesian inference, variable selection, mixture models, neuroscience.

goal in this paper is to detect differentially activated brain regions in response to light exposure. Statistical methods for assessing regional differences in activity across the whole brain using IEGs are currently in their infancy. The screening procedure proposed in this paper is a first step to improve statistical inference for quickly emerging high-content imaging techniques such as LSM.

Existing statistical methods for multiple hypothesis testing and variable selection typically group the individual estimates across the regions into two classes: *significant* and *nonsignificant*. This approach, however, oversimplifies the overall objective of such studies as the noise in the data may affect the discovery process. In particular, by using arbitrary cutoffs, the binary partition can also dismiss (i.e., classify as nonsignificant) many weak but biologically relevant signals. The limitations imposed by a dichotomous, symmetric screening are well known, and proposals to improve the decision problem date back at least to [Tukey \(1993\)](#). In a recent report, the U.S. National Academies recommended the consideration of alternatives to binary decision rules (e.g., to reject or not to reject a null hypothesis) as one way to improve the replicability of scientific results ([National Academies of Sciences, Engineering, and Medicine \(2019\)](#)); see also [Wasserstein, Schirm and Lazar \(2019\)](#) and [McShane et al. \(2019\)](#) for more discussion of this concept. Here we propose an alternative to classical binary discrimination with a method that can partition the potential findings into multiple tiers with varying degrees of *relevance*—a term we use instead of significance to distinguish our approach from other hypothesis testing methods (see, for similar usage in Bayesian variable selection, [Tadesse and Vannucci \(2021\)](#)). By allowing the sharing of information across the different regularization profiles and shrinking the noise to zero, our proposed model can better discriminate between signal and noise. Furthermore, this approach allows scientists to rank and classify brain regions without resorting to arbitrary cutoffs or prespecifying the grouping. Thus, investigators can identify interesting activation pathways to consider in their follow-up studies. To achieve these goals, our method combines shrinkage priors with mixture models.

**1.1. One- and two-group based screening.** Screening procedures play a central role in many statistical inference problems involving high-throughput scientific studies. Whether presented as a multiple comparisons problem within a hypothesis testing framework or a variable selection problem within a regression framework, they typically involve inference regarding a set of  $n$  parameters, say  $\beta = \{\beta_i\}_{i=1}^n$ . In a Bayesian framework, many methodologies have been proposed based on *regularization*—or *shrinkage*—of these parameters by using either the spike-and-slab (two-group) models ([McCulloch and George \(1993\)](#), [Mitchell and Beauchamp \(1988\)](#)) or the continuous scale mixture (one-group) models ([Polson et al. \(2012\)](#)).

The first approach treats the prior over the parameters as a discrete mixture of 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 units. [Ročková and George \(2018\)](#) have recently proposed an extension of the Bayesian Lasso ([Park and Casella \(2008\)](#)), called the spike-and-slab Lasso, where the two competing densities are assumed to be from the Laplace family.

The second approach places hierarchical priors on the scale parameter of a given kernel distribution, typically Gaussian (see [Bhadra et al. \(2019\)](#), for a review). The scale parameter is often decoupled into the product of global (i.e., shared across all the regression coefficients) and local shrinkage parameters (i.e., specific to each unit). This framework includes the Bayesian Lasso ([Park and Casella \(2008\)](#)), the Normal-Gamma ([Griffin and Brown \(2010\)](#)), the horseshoe ([Carvalho, Polson and Scott \(2010\)](#)), the horseshoe+, and the Dirichlet–Laplace ([Bhattacharya et al. \(2015\)](#)) priors, all based on Gaussian kernels. Due to their continuous shrinkage profile, the selection between relevant and irrelevant variables

needs to be done through post hoc analysis, usually by thresholding a proxy of the posterior probability  $\mathbb{P}[\beta_i \neq 0|\text{data}]$ .

In this paper we propose a discrete mixture of continuous scale mixtures that bridges the gap between those two alternatives and provides a unified framework. As carefully highlighted in [Hahn and Carvalho \(2015\)](#), the idea of adopting mixtures to model the scale parameters can be traced back to the seminal paper by [Ishwaran and Rao \(2005\)](#), where the authors discuss it within the context of bimodal mixtures. By building on that idea, our contribution allows combining the regularization effect typical of continuous shrinkage priors while inducing a grouping of the coefficients similarly to the spike-and-slab case. In our application to thin-section microscopy, our approach leads to automated model-based detection of groups of brain regions driven by different sparsity levels, imposing an adaptive regularization within each group. With our model we can also rank the discoveries into blocks of increasing relevance, facilitating the interpretation of the results. The discrete mixture also greatly reduces the complexity of the model, avoiding the usual specification of a local shrinkage parameter for each variable and enabling, at the same time, sharing of information across the parameters. From a multiple comparison perspective, the induced clustering goes beyond the classical “significant vs. nonsignificant” paradigm and allows to capture signals that would be otherwise lost within the canonical binary framework. In summary, our approach: (i) provides a more general setting for Bayesian sparse estimation without resorting to arbitrary cutoffs, (ii) drastically reduces the number of shrinkage parameters needed, and (iii) creates a framework for sharing information across units of interest without prespecifying any grouping of the units. The combination of model-based clustering and shrinkage is important for our application since it allows the discrimination between a group of inactive regions (whose effect is aggressively shrunk to zero) and a group of active ones. Moreover, the group of discoveries is partitioned into tiers, characterized by a similar amount of signal, providing neuroscientists with a ranking that can be invaluable for prioritizing further investigation.

Our approach is related to several other methods using mixture models to improve the efficacy of the variable selection and shrinkage processes and models for hypothesis testing. Our proposed method is also related to—but different than—the scale mixture of Gaussian distributions for relevance determination of [Shahbaba and Johnson \(2013\)](#) and the Dirichlet- $t$  distribution of [Finegold and Drton \(2011\)](#), [Finegold and Drton \(2014\)](#). We further elaborate on the connections between our model and the literature in Section A of the Supplementary Material ([Denti et al. \(2023\)](#)).

In the next section, we describe our study and the preprocessing steps required for preparing the raw data for analysis. We also present some preliminary results based on commonly-used methods whose limitations led to the development of our model. We introduce our methodology in Section 3 and the derivation of the corresponding posterior inference in Section 4. Section 5 is devoted to applying our method to the whole-brain imaging data (discussed above) using light sheet fluorescence microscopy to detect degrees of activation across brain regions. Then, in Section 6 we evaluate our model and confirm its validity using several simulation studies. Finally, in Section 7 we summarize the advantages and the shortcomings of our proposed method and discuss future directions.

**2. Thin-section microscopy: Experimental setup, preprocessing pipeline, and preliminary results.** Figure 1 shows a visual representation of the experimental setting in our case study, along with sample images obtained from two representative mice. More specifically, 14 mice were individually housed in the dark for 24 hours to establish baseline visual activity. Mice were then transferred into a new cage exposed to ambient light. The brains of six mice were examined zero to 15 minutes after light exposure to serve as the baseline group.

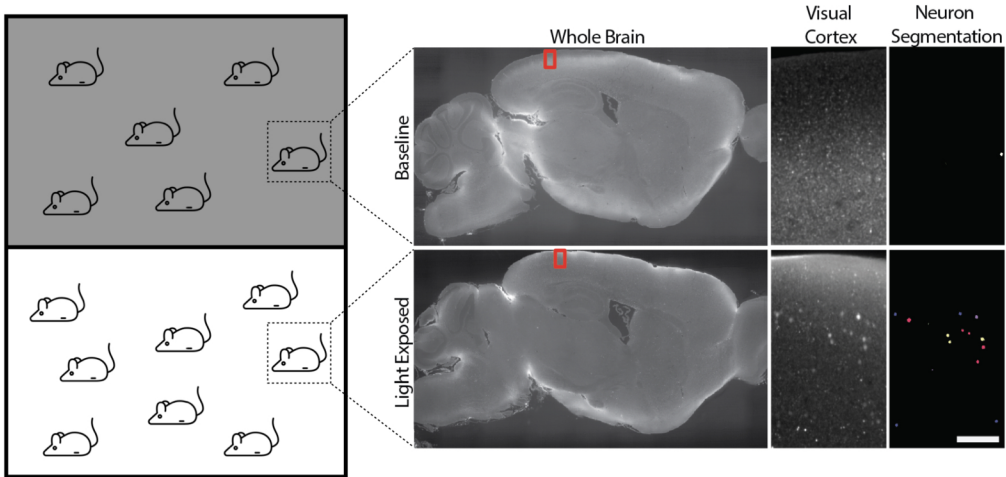


FIG. 1. Visual representation of the experiment setting and images obtained with the LSM technology. For both rows the brain areas highlighted in the rectangles are reported in the right four panels to show the high level of resolution achievable with the LSM.

The brains of another eight mice were examined 30–120 minutes after light exposure, within the window of Npas4 protein up-regulation (Ramamoorthi et al. (2011)). Equal numbers of left and right hemispheres were sampled. The goal was to assess differences in brain activation by comparing the baseline and light-exposed groups. We expect that light exposure induces widespread, visually evoked activity in terms of fluorescence intensity. Through this experiment we measured the location of almost 300,000 active neurons within a common three-dimensional reference space and extracted their intensity and volume with remarkable precision. The neurons are classified into regions according to the *Allen Brain Atlas* (Sunkin et al. (2013)), the anatomic reference atlas commonly used in studies involving brain structures of mice.

Figure 2 displays the three-dimensional images of brain cells measured in two representative mice under the two different experimental conditions: baseline and light-exposed. The intensity per unit of volume,  $i_{ov}$ , is the primary variable of interest in this study. However, before starting our analysis, Figure 2 reveals an important feature of the data: the frequency of observed neurons is strongly affected by the light exposure level. This effect can also be

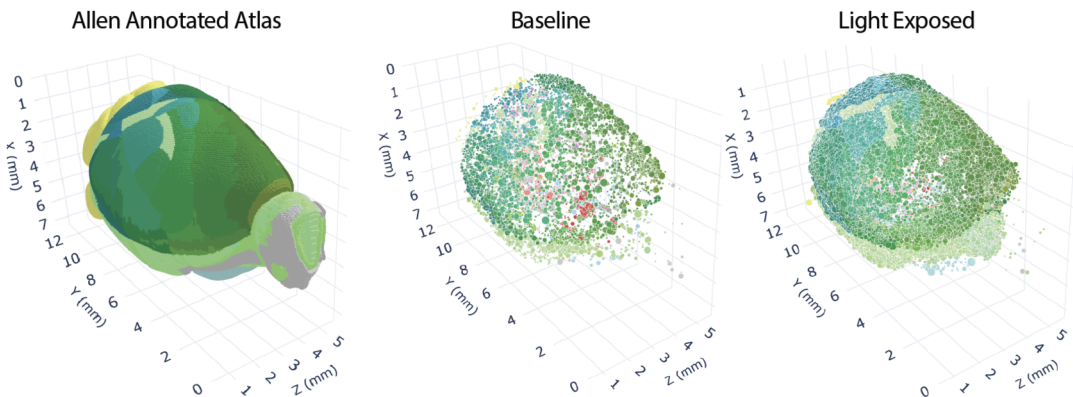


FIG. 2. Comparison between detected Npas4 expressing neurons in brains of two representative mice exposed to different experimental conditions (Allen annotated atlas—left, baseline—middle, light-exposed—right). The points represent the detected neurons. The size of each dot corresponds to the neuron's volume. As we can see, the activated neuron count is higher in the light-exposed group of mice.

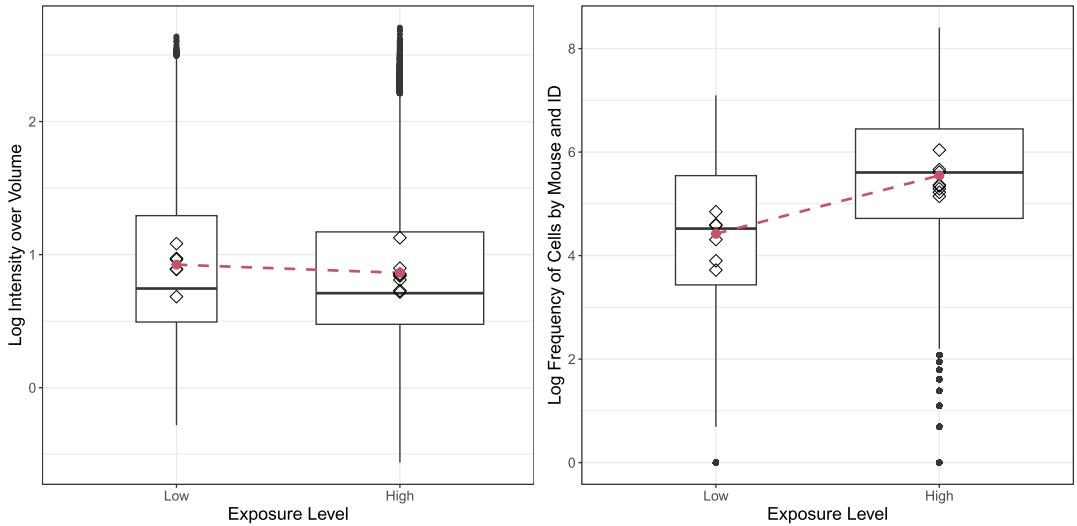


FIG. 3. *Boxplots representing the distributions of  $\log(i_{ov})$  and the log-frequency of cells in each region and mouse stratified by level of exposure. The widths of the boxplots are proportional to the square roots of the number of observations under the two experimental conditions (baseline  $\approx 55,000$  neurons; light-exposed  $\approx 235,000$  neurons). The diamonds represent the mean for each mouse, while the dashed lines connect the overall means across the two subpopulations.*

observed in Figure 3, which reports two boxplots comparing the distributions of the logarithm of  $i_{ov}$  (left panel) and the logarithmic frequency of cells detected in the brain regions of each mouse (right panel), under the two experimental conditions. The log scale is chosen to enhance the visual representation. Basing the entire analysis on  $i_{ov}$  alone would be insufficient and could lead to misleading results, as no clear difference emerges between the two experimental groups. However, the right panel suggests a positive association between the exposure level and the number of activated cells. In other words a proper definition of “activation” needs to incorporate the number of detected cells. Therefore, we will base our analysis on a score derived as a combination of frequency and intensity.

To compare the regions under the two different exposure levels, we need to adjust for the effects of possible confounders. An important source of information provided by our data is the multiresolution, hierarchical organization of the brain regions. Each neuron is assigned to a terminal region, and different terminal regions are connected to a shared, higher-level parent region. This mechanism goes on until all the regions are assigned to a common region, called *root*. We aim to remove the potential distortion in the intensity, given by specific mouse effects, and the possible influence of parent areas (i.e., the closest ancestors). In fact, certain areas may have higher intensity because of the dimension and overall intensity of their parents which, in turn, may blur the activation measures. Therefore, we regress the variable  $i_{ov}$  on all possible interactions between the mouse identifiers and the ancestor identifiers. We denote the resulting residual for each neuron as  $r_{i,c}$ , highlighting the membership of the cell  $c$  to the brain region  $i$ . Let  $m_i$  indicate the number of cells found in brain region  $i$ , with  $i = 1, \dots, n$ . To take into account the frequency distribution of the neurons, we multiply  $r_{i,c}$  by the density of neurons per unit of parent volume. This way, we obtain a new variable of interest:  $\tilde{r}_{i,c} = r_{i,c} \times m_i^* / \text{Vol}_i^*$ , where with  $m_i^*$  and  $\text{Vol}_i^*$  we indicate the frequency of cells and the volume of the parent of region  $i$ , respectively. Finally, we retain all brain regions with at least 15 ( $m_i \geq 15$ ) neurons, leaving  $n = 281$  regions for our analysis.

In a typical analysis, neuroscientists would consider a standard two-sided Welch t-test to detect differential activation of brain regions, comparing the averages of the vector

$\tilde{r}_i = (\tilde{r}_{i,1}, \dots, \tilde{r}_{i,m_i})$  under the baseline vs. light-exposed conditions. In the following we show how this typical approach may fail to identify important regions of interest. We obtain the t-statistics  $\mathbf{t} = \{t_i\}_{i=1}^n$ , the degrees of freedom estimated by the Welch–Satterthwaite equation  $\mathbf{d} = \{d_i\}_{i=1}^n$ , and the corresponding p-values  $\mathbf{p} = \{p_i\}_{i=1}^n$  for each brain region. The p-values are post-processed following [Benjamini and Hochberg \(BH, 1995\)](#) and thresholded at 5% to detect the activated regions. This result provides a first benchmark for later comparisons. We also use Efron’s empirical Bayes two-group model ([Efron \(2007\)](#)) computing the local false discover rate (IFDR). To do so, we first transform the t-statistics to z-scores:  $z_i = \Phi^{-1}(F_{T_{d_i}}(t_i)) \forall i$ , where  $\Phi$  and  $F_{T_d}$  denote the cumulative distribution function (c.d.f.) of the standard normal distribution and a Student-t distribution with  $d$  degrees of freedom, respectively. Then we threshold the resulting IFDR at 0.20, as suggested in the literature.

Within this setting the BH discovers 142 regions. In contrast, the IFDR method flags only 38 brain regions as important, missing many pertinent regions known to be associated with the visual task. On the one hand, such a difference in the results suggests that some brain regions may be active but show weaker signals than others. These regions are the ones that are likely missed by IFDR. On the other hand, it is known that the BH method struggles in cases where the z-scores distribution departs from the theoretical null. This rigidity may explain the large number of regions identified as significant, potentially due to false discoveries. Nonetheless, the discrepancy between the numbers of findings of the two methods highlights the shortcomings of the classical binary hypothesis partition (e.g., significant vs. nonsignificant). Indeed, the model we present in the next section can provide more insights by ranking the signals into several tiers with varying degrees of relevance which identify several levels of biological importance. As shown in [Section 5](#), this ranking allows scientists to examine groups of brain regions from the highest degree of relevance to the lowest degree without setting an arbitrary cutoff. More details about the discoveries can be found in [Section G](#) of the [Supplementary Material](#).

**3. Methodology: A discrete mixture of continuous scale mixtures.** For analyzing the whole-brain imaging data (and potentially similar high-throughput studies), we propose a novel discrete mixture model to cluster the brain regions into several tiers of varying relevance with respect to their activation levels. More specifically, we consider the following model:

$$(1) \quad \mathbf{z} = \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma}),$$

where  $\mathbf{z} = \{z_i\}_{i=1}^n$  is an outcome vector (e.g., z-scores) of length  $n$ ,  $\boldsymbol{\beta} = \{\beta_i\}_{i=1}^n$  is the mean vector, and  $\boldsymbol{\varepsilon}$  is the noise term. We assume homoscedastic and uncorrelated errors, that is,  $\boldsymbol{\Sigma} = \sigma^2 \mathbb{I}_n$ , for simplicity; this assumption seems to hold for our data, but our approach also can be readily generalized for more complex structures. In what follows,  $\mathcal{N}_k(\mathbf{a}, \mathbf{A})$  indicates a multivariate Normal distribution of dimension  $k$  with mean vector  $\mathbf{a}$ , covariance matrix  $\mathbf{A}$ , and density function  $\phi_k(\mathbf{a}, \mathbf{A})$ . In the univariate case, we let  $\mathcal{N}_1 \equiv \mathcal{N}$  and  $\phi_1 \equiv \phi$ .

Our main focus is the specification of suitable prior distributions for the coefficients  $\boldsymbol{\beta}$ . 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, that is,  $\beta_i | \tau, \lambda_i, \sigma^2 \sim \mathcal{N}(0, \sigma^2 \cdot \tau^2 \cdot \lambda_i^2) \forall i = 1, \dots, n$ , with  $\lambda_i$  assumed to be stochastic. Here  $\tau \in \mathbb{R}^+$  denotes a *global* shrinkage parameter, while the vector  $\boldsymbol{\lambda}_n = \{\lambda_i\}_{i=1}^n$ ,  $\lambda_i \in \mathbb{R}^+$  contains all the *local* shrinkage parameters. Conditioning on the variance of the data,  $\sigma^2$ , guarantees a unimodal posterior ([Park and Casella \(2008\)](#)).

We extend this framework and consider 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

$$(2) \quad \beta_i | \tau, \lambda_L, \boldsymbol{\pi}, \sigma^2 \sim \sum_{l=1}^L \pi_l \phi(\beta_i; 0, \sigma^2 \cdot \tau^2 \cdot \lambda_l^2), \quad i = 1, \dots, n,$$

where  $\boldsymbol{\pi}$  is the  $L$ -dimensional vector of mixture weights and the elements of the vector  $\lambda_L = \{\lambda_l\}_{l=1}^L$  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 continuous spike-and-slab framework of George and McCulloch (1997); while when  $L = n$  and  $\pi_l = \delta_i(l) \forall l, i$  (i.e., inducing  $n$  different singleton clusters), we recover the continuous shrinkage framework. In our application the mixture model allows to identify signals characterized by similar levels of shrinkage.

The mean estimation scenario is often considered for hypothesis testing, where the task is to identify the test statistics that depart from the standard Gaussian distribution specified under the null hypothesis (e.g.,  $H_{0,i} : \beta_i = 0$ ). Adopting the classical global-local shrinkage prior for  $\boldsymbol{\beta}$  to induce sparsity and setting  $\sigma^2 = 1$ , one can easily show that  $z_i | \lambda_i, \tau \sim \mathcal{N}(0, 1 + \tau^2 \lambda_i^2)$ . In our discrete mixture of continuous scale mixture model, the induced sampling distribution is itself a mixture,

$$(3) \quad z_i | \tau, \lambda_L, \boldsymbol{\pi} \sim \sum_{l=1}^L \pi_l \phi(z_i; 0, 1 + \tau^2 \lambda_l^2).$$

In the multiple comparison setting, we can see  $\mathbf{z}$  as a vector of  $n$  properly standardized test statistics corresponding to  $n$  different null hypotheses. Thus, model (3) can be interpreted as a *multigroup* extension of the classical two-group model (Efron (2007)). This connection is crucial, since it reveals the limitations of well-established multiple hypothesis testing methods when applied to our neuroscience data, as highlighted in Section 2; see Section B of the Supplementary Material for the derivation of (3).

The interpretation of (3) as a multigroup version of the model presented in Efron (2007) provides an additional justification for the use of continuous scale mixtures of Gaussians. Without loss of generality, let us assume that the first mixture component is characterized by the smallest scale parameter  $\lambda_{(\min)} = \min_l \lambda_l$ . One can impose this constraint a priori or identify the mixture component with the smallest scale parameter after model estimation. Whenever the product  $\tau \lambda_{(\min)} \approx 0$ , the corresponding mixture component can be interpreted as the null distribution, resembling the theoretical standard Gaussian. At the same time, the product  $\tau \lambda_{(\min)}$  is allowed to be different from zero to reflect a departure from the theoretical null, leading to the estimation of the so-called *empirical* null which could capture, for example, unexplained correlations among brain regions (Efron (2004)). The remaining mixture components describe the alternative distribution which can be decomposed into degrees of relevance according to the magnitude of the remaining parameters,  $\lambda \setminus \lambda_{(\min)}$ .

Finally, we highlight that our proposal can be extended to a more generic regression problem. The linear regression case can be obtained by simply substituting  $\mathbf{X}\boldsymbol{\beta}$  as the mean term of model (1), where  $\mathbf{X}$  is a  $n \times p$  covariate matrix and  $\boldsymbol{\beta} = \{\beta_j\}_{j=1}^p$  the corresponding regression coefficients. We explore the performance of our prior specification in such a scenario with a simulation study, reported in Section 6.

3.1. *Mixture and shrinkage: The horseshoe mix.* Whether we are adopting our model to perform variable selection or hypothesis testing, we need to specify prior distributions for the remaining parameters to complete the Bayesian specification. In addition, we can also specify a distribution for the global shrinkage parameter  $\tau$ . A common choice for the prior

distribution of the error variance  $\sigma^2$  is the Jeffreys prior  $\pi(\sigma^2) \propto 1/\sigma^2$ . The prior distribution for the weights changes if we assume a finite or infinite number of mixture components. If we assume  $L$  to be finite, we can simply set  $\boldsymbol{\pi} \sim \text{Dirichlet}(a_1, \dots, a_L)$ . Notice that even  $L > n$  is a viable option since one has to distinguish between mixture components and active components, that is, the actual clusters found in the dataset; see [Malsiner-Walli, Frühwirth-Schnatter and Grün \(2016\)](#) for more discussion on the use of sparse finite mixture (SFM) models. Setting the hyperparameters  $a_l = \epsilon \forall l$  with  $\epsilon$  small ( $\leq 0.05$ ) allows the model to parsimoniously select the number of active components needed to describe the data. Another option is to specify a nonparametric model via a Dirichlet Process (DP) mixture model,

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

where  $DP(\alpha, H)$  indicates a Dirichlet Process with concentration parameter  $\alpha$  and base measure  $H$ . Adopting the Stick Breaking (SB) representation of [Sethuraman \(1994\)](#), model (4) becomes

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

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  $u_l \sim \text{Beta}(1, \alpha)$  for  $l \geq 1$ .

We have introduced multiple mixture specifications (both parametric and nonparametric) to present a general working framework that can be adapted beyond our specific application. Depending on the problem at hand, specific priors can be used to incorporate our domain knowledge about the possible number of tiers. The Bayesian nonparametric approach is preferable if the number of clusters ( $L$ ) is expected to increase with the number of tests (i.e., regions). In our application a higher resolution brain atlas would lead to a larger number of tests and possibly the identification of new activation profiles of brain subregions. In contrast, using a sparse finite mixture implies that the number of clusters has the upper bound  $L$ . Nevertheless, as we will show in the simulation study of Section 6, the two approaches achieve very similar results if  $L$  is set to a sufficiently large number to ensure that many superfluous mixture components are not assigned any observation a posteriori. This rule of thumb is based on the posterior behavior of overfitted mixtures ([Rousseau and Mengersen \(2011\)](#)). In our experience the two methods usually provide similar results for all practical purposes when  $L > 30$ .

To summarize, the introduction of mixture component shrinkage parameters is beneficial for several reasons. 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 use a relatively small number of shrinkage parameters to borrow information across all the units and self-adapt to the different degrees of sparsity characterizing subsets of the coefficients. In our application this feature would help compound the signal in each tier and thus differentiate between pure noise—effectively shrunk to zero—and weak signals. Also, the model-based clustering nature of our approach enables the ranking of groups of coefficients into several *shrinkage profiles*, improving on commonly-used binary solutions (i.e., significant vs. nonsignificant) by providing more flexibility and insight for decision making.

In what follows, we will adopt a half-Cauchy prior for the mixture component shrinkage parameters:  $\lambda_l \sim \mathcal{C}^+(0, 1)$ ,  $\forall l$ . The half-Cauchy has been successfully employed in sparse mean estimation tasks, and its aggressive shrinkage property is ideal for our discovery problem. Henceforth, we refer to this model as horseshoe mix (HSM), in the spirit of the horseshoe (HS) prior introduced by [Carvalho, Polson and Scott \(2010\)](#).

Finally, we point out that, although the two models involve similar distributions, our model is fundamentally different from the Dirichlet–Laplace (DL) prior of [Bhattacharya et al.](#)



(2015). Under the DL prior, the conditional distribution of each coefficient is  $\beta_i | \pi_i^*, \lambda_i, \tau \sim \mathcal{N}(0, \lambda_i^2 \pi_i^{*2} \tau^2)$ , where  $\pi^* \sim \text{Dirichlet}(a_1, \dots, a_n)$ . Thus, the model by Bhattacharya et al. (2015) assumes that the Dirichlet random vector  $\pi^*$  lies on the  $(n - 1)$ -dimensional simplex; that is, its dimension is tied to the sample size. In our model the vector of mixture weights has only  $L$  entries. More importantly, our likelihood is the convex combination of  $L$  different kernels which is different from the single-parameter kernel structure assumed under the DL distribution.

3.2. *Mixture component and cluster shrinkage.* Consider now the Normal mean estimation framework, and define  $\kappa_i = 1/(1 + \tau^2 \lambda_i^2) \in (0, 1)$ . It follows that  $\mathbb{E}[\beta_i | z_i] = (1 - \mathbb{E}[\kappa_i | z_i]) \cdot z_i$  and  $\mathbb{E}[\beta_i | \lambda_i, \tau, z_i] = \frac{\tau^2 \lambda_i^2}{1 + \tau^2 \lambda_i^2} \cdot z_i$ , where  $\kappa_i$  is known as the *shrinkage factor* for observation  $i$ , which can be interpreted as a proxy of the complement of the posterior probability of relevance in the two-group model (Carvalho, Polson and Scott (2010)). It is interesting to see how these key quantities change under our model specification. For the conditional model, the posterior expected values of the coefficients become

$$(6) \quad \begin{aligned} \mathbb{E}[\beta_i | \boldsymbol{\pi}, \mathbf{z}] &= \sum_{l=1}^L \mathbb{E}[r_l(z_i)(1 - \kappa_l^*) | \mathbf{z}] \cdot z_i, \\ \mathbb{E}[\beta_i | \tau, \boldsymbol{\lambda}_L, \boldsymbol{\pi}, \mathbf{z}] &= \left( \sum_{l=1}^L r_l(z_i)(1 - \kappa_l^*) \right) \cdot z_i = (1 - \tilde{\kappa}_i) \cdot z_i, \end{aligned}$$

where  $r_l(z_i) = \frac{\pi_l \phi(z_i; 0, 1 + \tau^2 \lambda_l^2)}{\sum_{l=1}^L \pi_l \phi(z_i; 0, 1 + \tau^2 \lambda_l^2)}$ ; see Section B of the Supplementary Material 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 parameter)  $\tilde{\kappa}_i = \sum_{l=1}^L r_l(z_i) \kappa_l^*$ . Each CSF is a function of a convex combination of the  $L$  MCSFs and directly controls the amount of shrinkage that affects each parameter  $\beta_i$ . Simultaneously, the weights of the convex combination depend on the components of the marginal sampling distribution  $\phi(z_i; 0, 1 + \tau^2 \lambda_l^2)$ . It becomes clear how the model structure takes advantage of the sharing of statistical strength across parameters. Indeed, the posterior mean for  $\beta_i$  is the result of two effects. Given its mixture nature, the shrinkage is affected by all the other mixture component 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  $\mathbf{z}$ .

4. **Posterior inference.** To conduct posterior inference, we rely on Markov chain Monte Carlo (MCMC) algorithms because the posterior distribution is not directly available in closed form. To simplify posterior simulation, we augment model (2) with the latent membership labels  $\boldsymbol{\zeta} = \{\zeta_i\}_{i=1}^n$ , where  $\zeta_i \in \{1, \dots, L\}$ , linking each coefficient with a cluster; that is,  $\zeta_i = l$  if the  $i$ th coefficients has been assigned to the  $l$ th cluster. We obtain

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

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, Scott and Windle (2014)). The details of the Gibbs sampler are deferred to Section C of the Supplementary Material. In Section D of the Supplementary Material, we also comment on additional insights that the data augmentation procedure (7) provides about the model.

**4.1. Postprocessing of the results.** Once the posterior samples have been collected, we can estimate the cluster shrinkage factors from the membership labels. We map each coefficient  $\beta_i$  to the assigned local shrinkage parameter via  $\zeta_i$ , constructing the vector  $(\lambda_{\zeta_1}, \dots, \lambda_{\zeta_n})$ . It is then straightforward to compute  $\hat{\kappa}_i = 1/(1 + \tau^2 \lambda_{\zeta_i}^2)$ . One of the main advantages of our model is that, once the MCMC samples of size  $T$  are collected, it allows the estimation of the best partition that groups the coefficients into classes of similar magnitude. Let  $\zeta^{(t)} = \{\zeta_1^{(t)}, \dots, \zeta_n^{(t)}\}$  be the realization of the membership labels at iteration  $t = 1, \dots, T$ . With this information we can estimate the posterior probability coclustering (PPC) matrix, whose entries are defined as  $\widehat{\text{PPC}}_{i,i'} = \sum_{t=1}^T \mathbb{1}_{(\zeta_i^{(t)} = \zeta_{i'}^{(t)})} / T$ , for  $i, i' = 1, \dots, n$ .

In other words,  $\widehat{\text{PPC}}_{i,i'}$  estimates the proportion of times that coefficients  $i$  and  $i'$  have been assigned to the same cluster along the MCMC iterations. Hierarchical clustering can be applied directly to the  $\widehat{\text{PPC}}$  matrix for fast solutions as in Medvedovic, Yeung and Bumgarner (2004). The choice of the number of tiers can be driven by the simultaneous inspection of the dendrogram obtained from the hierarchical clustering approach and exogenous knowledge from domain experts. When the latter is unavailable, we recommend thresholding the resulting dendrogram using a moderate value of potential tiers (e.g., ranging between 2 and 6) and avoiding partitions with clusters containing only a negligible fraction of the observations. We elaborate more on this point in Section H of the Supplementary Material.

The resulting partition is easy to interpret. The HSM prior allows for a model-based clustering driven by the cluster shrinkage parameter vector  $\lambda_L$ . Therefore, the clusters in the solution specified by the optimal partition  $\hat{\zeta}$  can be described as classes of different magnitudes. Therefore, we can explicitly identify the subgroup of coefficients characterized by the smallest magnitude that can be deemed as irrelevant, similarly to the null component in the two-group model. In a linear regression framework, this means that we are able to identify the set of indices that indicate the least relevant covariates, say  $\mathcal{B}_0 = \{i \in \{1, \dots, n\} : \beta_i = 0\}$ , 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.

In the next section, we apply the HSM model to the light-sheet fluorescence microscopy data presented in Section 2. We emphasize that, despite being tailored to the differential activation detection problem, our HSM model represents a viable alternative to shrinkage priors in a wide range of problems. In Section 6 we compare HSM to well-established and state-of-the-art methods for variable selection and multiple hypothesis testing.

**5. Application: Segmenting brain regions into activation tiers.** In Section 2 we presented the preprocessing steps along with the results obtained from IFDR (38 discoveries) and BH procedures (142 discoveries).

To present additional benchmarks, here we estimate the posterior mean of the vector of  $z$ -statistics, using the spike and slab (SnS) and horseshoe (HS) models, after proper centering. Under the former model, we deem a region as relevant if its inclusion probability over the MCMC samples is over 5%, given the high level of sparsity induced by the SnS model in our data. Under the latter model, we select a brain region as significant if the credible set for the corresponding mean does not contain zero (van der Pas, Szabó and van der Vaart (2017)).

Next, we apply the HSM model directly to the centered  $z$ -scores,

$$(8) \quad z_i | \beta_i, \sigma^2 \sim \mathcal{N}(\beta_i, \sigma^2), \quad \beta_i | \lambda, \tau, \sigma \sim \sum_{l \geq 1} \pi_l \phi(0, \lambda_l^2 \tau^2 \sigma^2), \quad i = 1, \dots, n.$$

As mentioned previously, the expression in (8) can be regarded as a multigroup 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, which are ranked in increasing order, represent different degrees of relevance. To fit model (8), we use a Bayesian nonparametric approach with a DP stick-breaking representation over the mixture weights and adopt an Inverse Gamma distribution for  $\tau^2$ . A sparse finite mixture would also suffice, as the experiments we conduct in Section 6 show that the two specifications appear to provide similar results. The hyperprior on  $\tau^2$  was chosen to ensure good mixing of the global shrinkage parameter. We ran 10,000 iterations as burn-in period and used the next 10,000 samples for inference. Then we postprocessed the resulting posterior coclustering matrix with the Medvedovic approach. Although the flexibility of our model allows estimating the number of tiers through inspection of the nonempty components of the posterior distribution, for practical and inferential purposes a choice often needs to be made post-MCMC. For this application the inspection of the postprocessing results and the insight of our collaborators led to partitioning the z-scores into four tiers of relevance, ranging from no activation (Tier 4) to clear activation (Tier 1).

Figure 4 presents the posterior means (circles) and posterior medians (crosses) for different quantities. The elements in both panels are represented according to the tier to which they are assigned. The top-left panel shows the estimated coefficients. We can see how the model groups the scores according to their magnitude. A scatter plot of the z-scores vs. the posterior estimates is displayed in the bottom-left panel. The axes are cropped to showcase the shrinking effect of the HSM model on the z-scores for Tiers 3 and 4. Finally, the right panel presents the posterior probabilities of relevance  $1 - \tilde{\kappa}_i$ , for  $i = 1, \dots, n$ . This plot helps

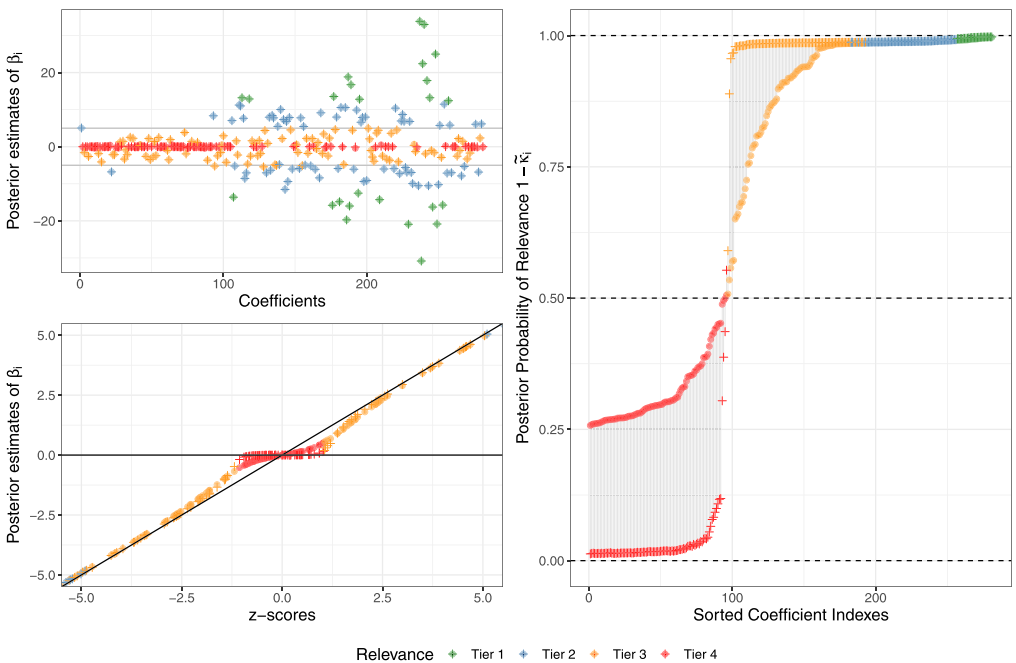


FIG. 4. All the panels show posterior means (circles) and posterior medians (crosses) for different quantities. Top-left panel: Estimates for  $\beta$  stratified according to the retrieved segmentation. Bottom-left panel: Posterior estimates for  $\beta$  plotted against the z-scores. The plot is cropped between  $(-5, 5)$  on both axes to show the shrinkage induced on the z-scores belonging to the low tiers of relevance. Right panel: 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 variability in the posterior distributions.

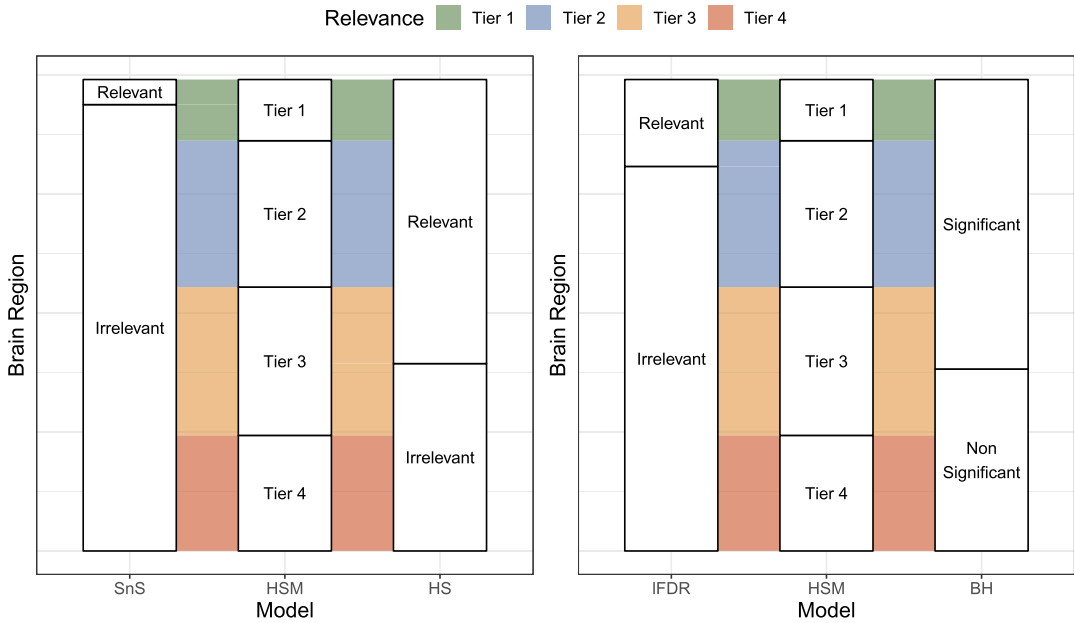


FIG. 5. Alluvial plots displaying how the partitions of the 281 regions, obtained with different models (SnS, HS, IFDR, and BH), relate with the HSM partition into tiers.

interpret the tiers of relevance: we notice the shift from Tier 4 to Tier 3 in both posterior estimates 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 filtering out the irrelevant units, it partitions the remaining ones into different sets with increasing levels of importance, capturing more information from the z-scores.

We now compare the results obtained by the five alternative methods discussed in this paper: BH, IFDR, HS, SnS, and HSM. Figure 5 juxtaposes the different results with an alluvial plot. Each column represents a model, and the horizontal lines (the brain regions) display how HSM tiers are associated with the results of the other models. A contingency table is also provided in Section F of the Supplementary Material. The SnS and IFDR methods are the most conservative, detecting only nine and 38 regions, respectively. All these selected regions are part of HSM's top two tiers. The results from HS and BH are also similar: the two methods detect 138 and 142 regions, respectively. The HSM model places 25 regions in the top tier, 68 regions in the second tier, 49 in the third, and deems 96 regions as irrelevant.

We next sought to identify the biological relevance of these findings. We expect that the introduction of animals to light will drive Npas4 expression in neurons within the laminar subregions (e.g., layers) of different visual cortex areas (Hübener (2003), Andermann et al. (2011)). Previous studies have shown that neurons in the primary visual (V1) area of the cortex respond to light exposure by expressing Npas4 mRNA (Hrvatín et al. (2018)). Our results align with the literature, capturing the activation of the V1 laminae due to light exposure in terms of increased Npas4 protein expression. Other cortex regions are expected to exhibit visually evoked activity, such as the lateral, posteromedial, anterolateral, and anteromedial visual areas. According to our results, all of these regions show Npas4 expression following light exposure (Andermann et al. (2011)). However, across the 20 brain laminae, comprised by these regions, five of them fall into Tiers 3 and 4, reflecting their lack of activation. We report the list of the two top-tier areas in Table 1. The complete list of findings for the four models we considered is reported in Section G of the Supplementary Material. From that list we can appreciate how the HSM model provides a more articulate solution, mediating

TABLE 1  
*Lists of brain regions assigned to Tier 1 and Tier 2 of activation by the HSM model*

Tier 1	Tier 2
Agranular insular area, ventral part: Layer 1	Agranular insular area, dorsal part: Layer 1, 6a
Anterior cingulate area, dorsal part: Layer 1, 5	Anterior cingulate area, ventral part: Layer 1, 2/3, 5
Anterolateral visual area: Layer 1	Anterior olfactory nucleus
Dorsal auditory area: Layer 1, 2/3, 5, 6a	Anteromedial visual area: Layer 1
Ectorhinal area: Layer 2/3	Central amygdalar nucleus
Lateral visual area: Layer 1, 5	Ectorhinal area: Layer 1, 5, 6a
Posteromedial visual area, 4, 5	Fiber tracts
Postpiriform transition area	Hippocampal formation
Primary auditory area: Layer 2/3, 5	Infralimbic area: Layer 1, 5
Primary visual area: ILayer 1, 4, 5, 6a	Main olfactory bulb
Subiculum	Nucleus accumbens
Taenia tecta	Orbital area, medial part: Layer 1, 5
Temporal association areas: Layer 2/3, 6a	Orbital area, ventrolateral part: Layer 1
Ventral auditory area: Layer 2/3	Parasubiculum
.	Perirhinal area: Layer 1, 2/3, 5, 6a
.	Piriform area
.	Piriform-amygdalar area
.	Posterior auditory area: Layer 2/3
.	Posterolateral visual area: Layer 2/3, 5
.	posteromedial visual area: Layer 1
.	Postsubiculum
.	Prelimbic area: Layer 1, 5
.	Primary auditory area: Layer 1, 6a
.	Primary motor area: Layer 1, 5
.	Primary somat. area, barrel field: Layer 1, 2/3, 5, 6a
.	Primary somat. area, lower limb: Layer 2/3, 5
.	Primary somat. area, mouth: Layer 2/3
.	Primary somat. area, nose, 2/3
.	Primary somat. area, trunk: Layer 2/3, 4, 5
.	Primary somat. area, upper limb: Layer 1, 2/3
.	Primary visual area, 2/3
.	Retrosplenial area, lateral agranular part, 1
.	Retrosplenial area, lateral agranular part: Layer 5
.	Retrosplenial area, ventral part: Layer 1, 6a
.	Secondary motor area: Layer 2/3, 5, 6a
.	Suppl. somatosensory area: Layer 1, 5, 6a
.	Temporal association areas: Layer 1, 5
.	Third ventricle
.	Unlabeled
.	Ventral auditory area: Layer 5, 6a

between the more conservative IFDR and SnS methods and the numerous discoveries of the BH and Horseshoe models.

The tiering results, obtained from our method, allow us to stratify our findings by the level of activation without resorting to successive manual and arbitrary p-value cutoffs. We can utilize this approach to identify laminar activity patterns. For example, neurons in layer 2/3 of the primary visual cortex are known to exhibit lower activity than those in other V1 laminae (Niell and Stryker (2010)). HSM identifies this by placing layer 2/3 in tier 2, while all other V1 laminae are placed in tier 1. Interestingly, layers 2/3 of different visual cortex areas are assigned one tier below the other laminae, suggesting lower activity in layer 2/3 may be a common feature throughout the visual cortex. To our knowledge, these results are

novel and have not been previously reported in mice. Hence, these findings warrant further investigation. These results illustrate the additional insights provided by HSM when applied to high-throughput studies.

**6. Numerical experiments validating the HSM model.** In this section we illustrate our method and evaluate its performance, in terms of generic mean estimation and linear regression, to establish its competitiveness with commonly-used and the-state-of-the-art statistical methodologies.

**6.1. Illustrative example.** As a simple example, consider a sample of 500 observations generated from a linear regression model with true vector of coefficients  $\beta$ , 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  $\sigma^2 = 0.5$ . Given this dataset, we estimate the HSM model with a nonparametric specification of the mixture weights and fix  $\tau^2 = 0.001$ . The two panels of Figure 6 show the estimated posterior coclustering matrix  $\widehat{\text{PPC}}$  (left) and the posterior mean for each  $\beta_i$  (right), transformed as  $\log|\hat{\beta}_i|$  to emphasize the differences in terms of magnitude. On top of the PPC matrix, we highlighted three blocks representing the true clusters present in the data which are also represented by the different shapes in the right plot. From the right panel, we can see that the model can effectively group the parameters in terms of magnitude. The accuracy of the classification is 0.89, with an Adjusted Rand Index of 0.72.

**6.2. Performance in mean estimation.** Next, we investigate the performance of the HSM model in terms of mean estimation. To this end, we generate random vectors from a multivariate Gaussian distribution with mean  $\beta$ . The elements in  $\beta = \{\beta_i\}_{i=1}^n$  are organized into three different blocks:  $\beta_i^{(1)} \sim \mathcal{N}(0, 100)$  for  $i = 1, \dots, q$ ,  $\beta_i^{(2)} \sim \mathcal{N}(0, 1)$  for  $i = q + 1, \dots, 2q$ , and  $\beta_i^{(3)} \sim \delta_0$  for  $i = 2q + 1, \dots, n$ . We consider four scenarios (S1–S4), varying according to the values assumed by  $q$  and  $n$ . Specifically, for S1 and S2 we set  $n = 500$ , while for S3 and S4  $n = 1000$ . Moreover, we set  $q = 50$  for S1 and S3, while  $q = 100$  for S2 and S4.

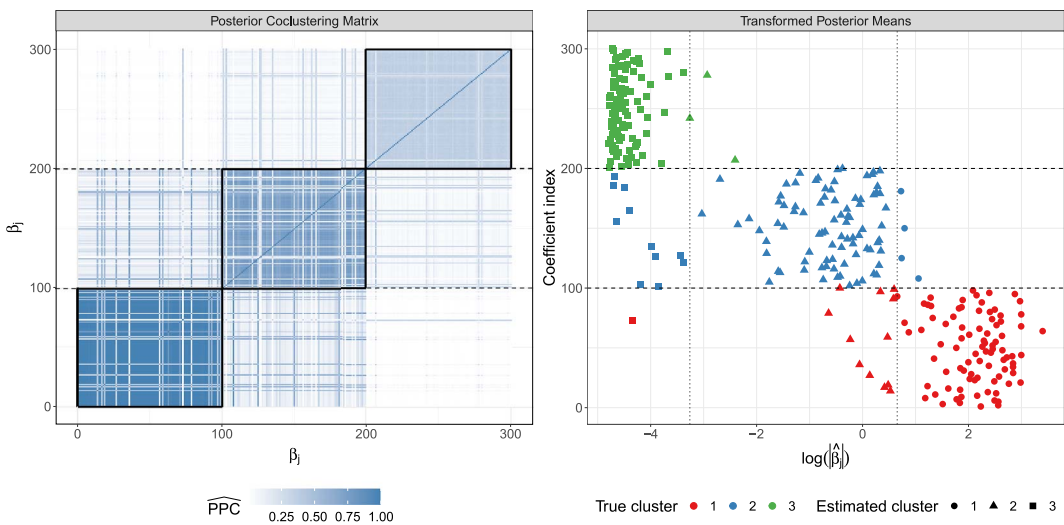


FIG. 6. Left panel: The estimated posterior coclustering matrix where the actual clusters are superimposed with solid lines. Right panel: A scatterplot presenting the estimated posterior mean for each coefficient transformed as  $\hat{\beta}_i \tau \sigma$ . The horizontal lines separate the true clusters while the vertical lines highlight the estimated partition. Note how the magnitude of the coefficients leads the estimation of the clusters.

TABLE 2  
 Summary of the different HSM model specifications used in our first simulation experiment

	HSM1	HSM2	HSM3	HSM4	HSM5	HSM6
Mixture	SFM	SFM	SMF	BNP	BNP	BNP
Global shrinkage parameter	0.0001	$\tau^2 \sim IG$	$\tau \sim \mathcal{C}^+$	0.0001	$\tau^2 \sim IG$	$\tau \sim \mathcal{C}^+$

We compare the results obtained from three models: HSM, horseshoe (HS), and spike and slab (SnS). Results for the latter two models were obtained via the R packages `horseshoe` (v0.2.0) and `BoomSpikeSlab` (v1.2.5), respectively. We consider six different specifications for the HSM model by varying mixture type (sparse finite mixture, SFM, and Dirichlet process mixture, BNP in all cases, we set  $L = 50$ ) and distribution for the global shrinkage parameter (fixed, Inverse Gamma, and half-Cauchy). We summarize the specifications and the corresponding acronyms in Table 2. We also consider two specifications for the HS model, where  $\tau \sim \mathcal{C}^+(0, 1)$  (HS1) or  $\tau \sim \delta_{0.0001}$  (HS2).

To quantify the performance of the models, for each simulated dataset  $k = 1, \dots, K$ , we compute the mean squared error between the posterior mean  $\hat{\beta}_k$  and the ground truth, defined as  $MSE(\beta_k, \hat{\beta}_k) = \sum_{i=1}^n (\beta_{i,k} - \hat{\beta}_{i,k})^2/n$ . We also stratify the same quantity across the three different parameter blocks to understand which magnitude group contributes the most to the error. All the results are averaged over  $K = 30$  replicates. For each replicate we ran 10,000 MCMC iterations, discarding the first half as burn-in. The outcome of the first and second scenarios are displayed in the bar plots with error bars (representing the standard errors) in Figure 7. The table containing the values used to draw the bar plots is reported in Section F of the Supplementary Material. The same table also includes the results for S3 and S4 for which the MSE values are very similar to the ones we display here. Lastly, in Section E of the Supplementary Material we also include an alternative version of Figure 7 without the SnS output to ease the visual comparison of the remaining models.

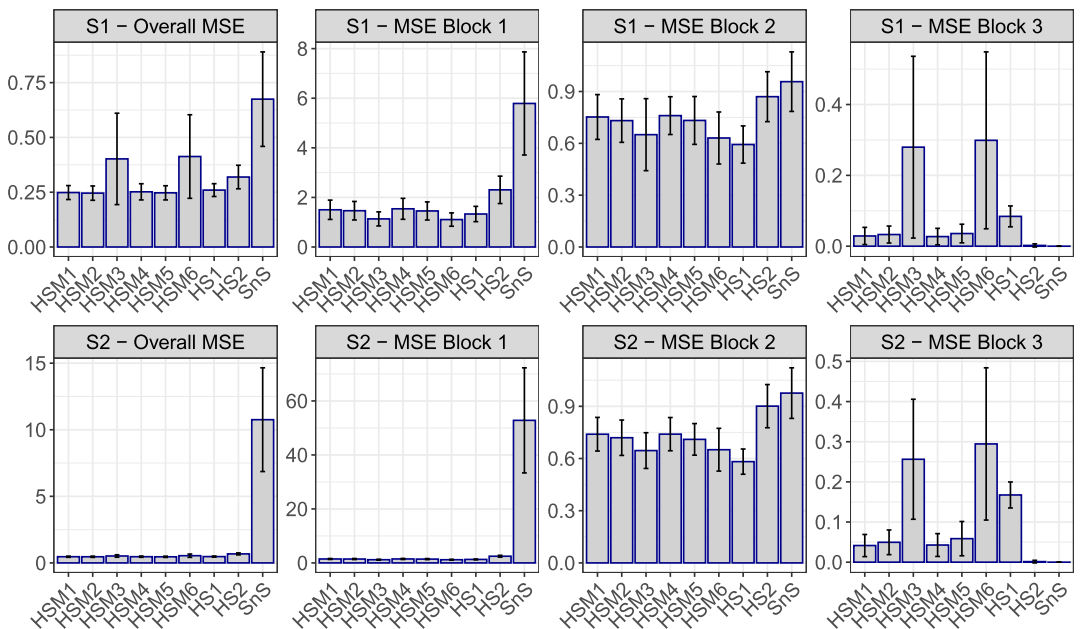


FIG. 7. Bar plots of the average overall and stratified MSEs and corresponding standard errors (error bars) for simulation scenarios S1 and S2.

The HSM obtains very competitive results both in terms of overall and stratified MSE. Overall, the HSM specifications with constant or Inverse-Gamma global shrinkage parameters attain low MSE combined with small standard errors across the replicates. Instead, whenever a half-Cauchy prior is used, the average overall MSE increases but so does the variability of the results. The same happens for the HS model. Also, we do not observe any clear result that favors one mixture type over the other. The stratification of the results into the parameter blocks shows an interesting trade-off between the precision of the parameter estimation across the different magnitudes. As expected, the SnS perfectly captures the true zeros, but it tends to overshrink the nonzero  $\beta$ 's. In the cases where  $\tau \sim \mathcal{C}^+$  (HSM3 and HSM6), the model better captures the nonzero parameters, while the remaining specifications perform very good regularization. The same rationale applies to the HS models.

*6.3. Performance in the estimation of regression coefficients.* Here we consider a linear regression framework with  $n$  observations and  $p$  covariates and compare the estimation performance of the HSM model with both well-established and recent Bayesian shrinkage models, including Bayesian Lasso, horseshoe, and horseshoe+. To estimate these models under different regularizing prior specifications, we use the R package `bayesreg` (v1.2.0).

Our experiment consists of three scenarios, characterized by different values of the ratio  $n/p$ , describing the proportion between the sample size and the number of variables. Specifically, we consider the following three ratios:  $n/p \in \{(500, 250) = 2, (500, 500) = 1, (500, 750) = 0.667\}$ . Under each scenario we generate  $K = 30$  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  $\mathbf{X}_k$ ,  $k = 1, \dots, K$ . Then we sample the regression coefficients  $\beta_k$ , organized in three different blocks:  $\beta_{j_1,k}^{(1)} \sim \mathcal{N}(0, 100)$  for  $j_1 = 1, \dots, 100$ ,  $\beta_{j_2,k}^{(2)} \sim \mathcal{N}(0, 1)$  for  $j_2 = 1, \dots, 100$ , and  $\beta_{j_3,k}^{(3)} \sim \delta_0$  for  $j_3 = 1, \dots, p - 200$ . That is, for a fixed number of covariates  $p > 100$ , we generate 100 coefficients of high magnitude ( $\sigma^{(1)} = 10$ ), 100 coefficients of low magnitude ( $\sigma^{(2)} = 1$ ), and  $p - 100$  coefficients identically equal to zero. Finally, we set  $\mathbf{y}_k = 5 + \mathbf{X}_k \beta_k + \epsilon_k$ , with  $\epsilon_k \sim \mathcal{N}_n(0, \mathbb{I}_n)$ .

For the mixture weights, we adopt a sparse mixture specification, using  $L = 50$  mixture components and  $a = 0.05$ , while we fix  $\tau^2 = 0.0001$ . As in the previous case, we ran 5000 iterations as burn-in period and retained 5000 iterations (thinning every five steps) for posterior inference.

The average and standard errors of the overall MSE, obtained by each model over the 30 replicates, are reported in Table 3. Each row corresponds to a simulation scenario. In general, all models obtain very good performance. Indeed, in the first scenario the average MSEs obtained by different models are very similar. As the number of covariates increases, we see how the sharing of information across the HSM parameters leads to lower MSE, followed, in order, by the horseshoe+, the horseshoe, and the Bayesian Lasso.

We decompose the overall MSE replicates into different magnitude blocks and report the results in Figure 8. The boxplots describe the distributions of the results obtained over 30

TABLE 3  
Overall average MSE (and relative standard errors) for the four different models under the three simulation scenarios

Scenario	HSM		HS		HS+		Lasso	
1	0.0031	(4e-04)	0.0035	(4e-04)	0.0033	(4e-04)	0.0039	(0.0005)
2	0.0019	(3e-04)	0.0044	(6e-04)	0.0031	(4e-04)	0.0287	(0.0059)
3	0.0014	(3e-04)	0.0039	(6e-04)	0.0028	(5e-04)	0.3075	(0.0707)



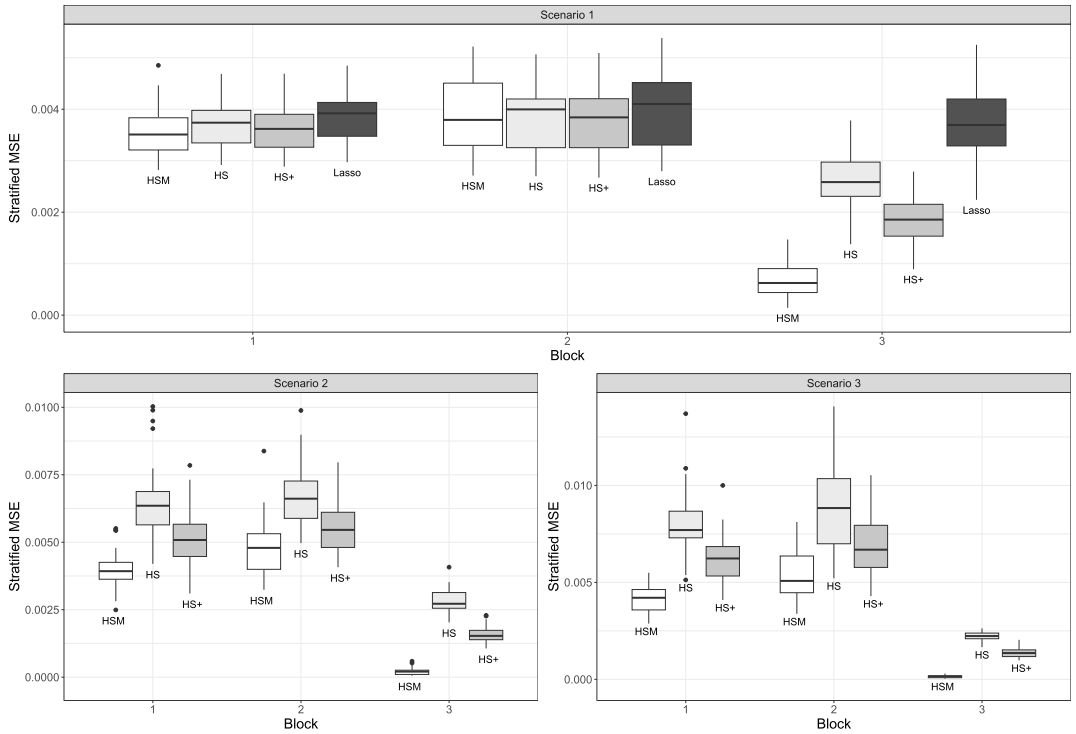


FIG. 8. *Boxplots of the stratified MSE obtained over 30 replicates by the different models. Each panel corresponds to a simulation scenario, with boxplots organized by blocks of parameters. The results obtained with the Lasso are omitted in the center and bottom panels to ease the visual comparison.*

replicates. Each panel depicts a simulation scenario. In the center and bottom panels, we removed the Bayesian Lasso’s results to facilitate the visual comparison since its MSE is much larger than the ones obtained with HSM, horseshoe, and horseshoe+. The complete figure is reported in Section E of the Supplementary Material. First, we notice how the HSM model obtains better performance in block 3, regardless of the scenario. Here the gain in MSE reflects the ability of the model to target and shrink the true zeros effectively. Second, it is interesting to see how HSM obtains lower MSE than its competitors as the ratio  $n/p$  decreases. While the boxplots for blocks 1 and 2 in the top panel are almost equivalent, MSE gains start to manifest in the other two panels.

**7. Discussion.** In this paper we have developed a novel class of priors that for multiple hypothesis testing and variable selection problems. Our proposed method groups the units of interest and their corresponding parameters into several tiers with varying degrees of relevance. This feature proved to be particularly valuable for the specific application discussed in this paper: discovering differentially activated brain regions from data collected via state-of-the-art brain imaging technology. Specifically, our approach involves adopting a discrete mixture model, reminiscent of the two-group models, where each mixture component is itself a continuous scale mixture distribution. We then assumed half-Cauchy priors for the shrinkage parameters, mimicking the horseshoe model. This way, we can retain the strong shrinkage properties of the continuous mixtures while performing model-based clustering typical of the discrete mixtures. Furthermore, the clustering detects irrelevant units and potentially segments the relevant ones into tiered classes, according to each coefficient’s magnitude. With respect to the specific application discussed in this paper, the enhanced stratification of coefficients induced by our model detects important regions that were left out by more conservative

methods. Further, these regions are ranked into groups of varying importance, avoiding arbitrary decisions in screening. Notably, our model ranks in Tier 2 regions that are related to the visual task, such as the primary visual area (layer 2/3) and various posterolateral visual areas, missed by the lFDR and SnS models. Combining the two discrete and continuous mixtures for shrinkage shows promising results, especially in targeting and regularizing the null coefficients via the clustering-induced shrinkage structure. We have showcased the potential of our approach using simulation studies.

The results presented in this paper can pave the way for many future research directions. For example, the data analysis can be enriched by including the hierarchical structure of brain regions—each brain region can be divided into progressively smaller subregions. In our pre-processing workflow, we only considered the relationship between the areas at the highest resolution and their parents to account for potential correlations across regions. However, including information regarding the whole-brain structure could add substantial information since it is usually unclear at which hierarchical level a relevant differential effect emerges. Therefore, we are currently developing a two-group model that incorporates such information directly in the Bayesian model, studying how the activation probability is partitioned across the regions' ancestors. From a methodological point of view, one can consider different continuous scale mixture types to improve the HSM model. For example, a possibility would be to take into consideration a Laplace distribution generalizing the model in Ročková and George (2018) or more refined horseshoe distributions, such as the horseshoe-like distribution (Bhadra et al. (2021)) or regularized horseshoe (Piironen and Vehtari (2017a)). Moreover, the prior specification for the global shrinkage parameter,  $\tau$ , could be improved in the context of our clustering approach, for example, by using the method of Piironen and Vehtari (2017b). Inspired by the idea of splitting the error rate introduced by Tukey (1993), it may be possible to develop a mixture of shrinkage priors that behaves asymmetrically around the origin. The approach would estimate the probability that each z-score is assigned to a particular relevance tier while allowing for different tail behaviors in over- and underexpressed brain regions. Lastly, the scalability of our method could also be improved using more efficient MCMC alternatives, such as the two algorithms for horseshoe estimation recently proposed in Johndrow, Orenstein and Bhattacharya (2020). Alternatively, one can adopt an approximate inference method such as mean-field variational Bayes (Neville, Ormerod and Wand (2014)).

**Acknowledgments.** The authors are grateful to the Editor-in-Chief and the editorial board for insightful comments that helped improve the manuscript. This work was supported by NIH award R01-MH115697. Ricardo Azevedo, Chelsie Lo, and Sunil P. Gandhi are affiliated with Translucence Biosystems, Inc, Irvine. Additionally, Sunil P. Gandhi is affiliated with the Center for the Neurobiology of Learning and Memory, University of California, Irvine.

## SUPPLEMENTARY MATERIAL

**Supplement to “A horseshoe mixture model for Bayesian screening with an application to light sheet fluorescence microscopy in brain imaging”.** (DOI: [10.1214/23-AOAS1736SUPPA](https://doi.org/10.1214/23-AOAS1736SUPPA); .pdf). We provide additional materials to support the results in this paper. These include the details of the Gibbs sampler algorithm, its implementation, additional figures, and a discussion of the proposed prior and models in the context of robust Bayesian statistics.

**Code** (DOI: [10.1214/23-AOAS1736SUPPB](https://doi.org/10.1214/23-AOAS1736SUPPB); .zip). The R and C++ scripts used for the simulation studies presented in this paper are available as in online Supplement (doi: [10.1214/23-AOAS1736SUPPB](https://doi.org/10.1214/23-AOAS1736SUPPB)). In addition, one can find the latest version of the software at the [Github](https://github.com/Fradenti/Horseshoe_Mix) repository Fradenti/Horseshoe\_Mix.

## REFERENCES

- ANDERMANN, M. L., KERLIN, A. M., ROUMIS, D. K., GLICKFELD, L. L. and REID, R. C. (2011). Functional specialization of mouse higher visual cortical areas. *Neuron* **72** 1025–1039. <https://doi.org/10.1016/j.neuron.2011.11.013>
- BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. [MR1325392](https://doi.org/10.1111/j.1467-9868.1995.tb01701.x)
- BHADRA, A., DATTA, J., POLSON, N. G. and WILLARD, B. (2019). Lasso meets horseshoe: A survey. *Statist. Sci.* **34** 405–427. [MR4017521](https://doi.org/10.1214/19-STS700)
- BHADRA, A., DATTA, J., POLSON, N. G. and WILLARD, B. T. (2021). The horseshoe-like regularization for feature subset selection. *Sankhya B* **83** 185–214. [MR4256316](https://doi.org/10.1007/s13571-019-00217-7)
- BHATTACHARYA, A., PATI, D., PILLAI, N. S. and DUNSON, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *J. Amer. Statist. Assoc.* **110** 1479–1490. [MR3449048](https://doi.org/10.1080/01621459.2014.960967)
- CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. [MR2650751](https://doi.org/10.1093/biomet/asq017)
- DENTI, F., AZEVEDO, R., LO, C., WHEELER, D. G., GANDHI, S. P., GUINDANI, M. and SHAHBABA, B. (2023). Supplement to “A horseshoe mixture model for Bayesian screening with an application to light sheet fluorescence microscopy in brain imaging.” <https://doi.org/10.1214/23-AOAS1736SUPPA>, <https://doi.org/10.1214/23-AOAS1736SUPPB>
- EFRON, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104. [MR2054289](https://doi.org/10.1198/016214504000000089)
- EFRON, B. (2007). Size, power and false discovery rates. *Ann. Statist.* **35** 1351–1377. [MR2351089](https://doi.org/10.1214/009053606000001460)
- FINEGOLD, M. and DRTON, M. (2011). Robust graphical modeling of gene networks using classical and alternative  $t$ -distributions. *Ann. Appl. Stat.* **5** 1057–1080. [MR2840186](https://doi.org/10.1214/10-AOAS410)
- FINEGOLD, M. and DRTON, M. (2014). Robust Bayesian graphical modeling using Dirichlet  $t$ -distributions. *Bayesian Anal.* **9** 521–550. [MR3256052](https://doi.org/10.1214/13-BA856)
- GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–373.
- GRIFFIN, J. E. and BROWN, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Anal.* **5** 171–188. [MR2596440](https://doi.org/10.1214/10-BA507)
- HAHN, P. R. and CARVALHO, C. M. (2015). Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective. *J. Amer. Statist. Assoc.* **110** 435–448. [MR3338514](https://doi.org/10.1080/01621459.2014.993077)
- HRVATIN, S., HOCHBAUM, D. R., NAGY, M. A., CICONET, M., ROBERTSON, K., CHEADLE, L., ZILIONIS, R., RATNER, A., BORGES-MONROY, R. et al. (2018). Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat. Neurosci.* **21** 120–129.
- HÜBENER, M. (2003). Mouse visual cortex. *Curr. Opin. Neurobiol.* **13** 413–420. [https://doi.org/10.1016/S0959-4388\(03\)00102-8](https://doi.org/10.1016/S0959-4388(03)00102-8)
- ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. [MR2163158](https://doi.org/10.1214/009053604000001147)
- JOHNDROW, J., ORENSTEIN, P. and BHATTACHARYA, A. (2020). Scalable approximate MCMC algorithms for the horseshoe prior. *J. Mach. Learn. Res.* **21** 73. [MR4095352](https://doi.org/10.48550/jmlr.2020.21.1)
- LIN, Y., BLOODGOOD, B. L., HAUSER, J. L., LAPAN, A. D., KOON, A. C., KIM, T. K., HU, L. S., MALLIK, A. N. and GREENBERG, M. E. (2008). Activity-dependent regulation of inhibitory synapse development by Npas4. *Nature* **455** 1198–1204.
- MAKALIC, E. and SCHMIDT, D. F. (2016). High-dimensional Bayesian regularised regression with the BayesReg package. ArXiv Preprint 1–18.
- MALSINER-WALLI, G., FRÜHWIRTH-SCHNATTER, S. and GRÜN, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Stat. Comput.* **26** 303–324. [MR3439375](https://doi.org/10.1007/s11222-014-9500-2)
- MCCULLOCH, R. E. and GEORGE, E. I. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.
- MC SHANE, B. B., GAL, D., GELMAN, A., ROBERT, C. and TACKETT, J. L. (2019). Abandon statistical significance. *Amer. Statist.* **73** 235–245. [MR3925729](https://doi.org/10.1080/00031305.2018.1527253)
- NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE (2019). *Reproducibility and Replicability in Science*. The National Academies Press, Washington, DC.
- MEDVEDOVIC, M., YEUNG, K. Y. and BUMGARNER, R. E. (2004). Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics* **20** 1222–1232. <https://doi.org/10.1093/bioinformatics/bth068>

- MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. [MR0997578](#)
- NEVILLE, S. E., ORMEROD, J. T. and WAND, M. P. (2014). Mean field variational Bayes for continuous sparse signal shrinkage: Pitfalls and remedies. *Electron. J. Stat.* **8** 1113–1151. [MR3263115](#) <https://doi.org/10.1214/14-EJS910>
- NIELL, C. M. and STRYKER, M. P. (2010). Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron* **65** 472–479.
- PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. [MR2524001](#) <https://doi.org/10.1198/016214508000000337>
- PIIRONEN, J. and VEHTARI, A. (2017a). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electron. J. Stat.* **11** 5018–5051. [MR3738204](#) <https://doi.org/10.1214/17-EJS1337SI>
- PIIRONEN, J. and VEHTARI, A. (2017b). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017* 1–9.
- POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2014). The Bayesian bridge. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 713–733. [MR3248673](#) <https://doi.org/10.1111/rssb.12042>
- POLSON, N. G., SCOTT, J. G., CLARKE, B. and SEVERINSKI, C. (2012). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Stat.* **9** 1–30.
- RAMAMOORTHI, K., FROPF, R., BELFORT, G. M., FITZMAURICE, H. L., MCKINNEY, R. M., NEVE, R. L., OTTO, T. and LIN, Y. (2011). Npas4 regulates a transcriptional program in CA3 required for contextual memory formation. *Science* **334** 1669–1675.
- RENIER, N., ADAMS, E. L., KIRST, C., WU, Z., AZEVEDO, R., KOHL, J., AUTRY, A. E., KADIRI, L., UMADEVI VENKATARAJU, K. et al. (2016). Mapping of brain activity by automated volume analysis of immediate early genes. *Cell* **165** 1789–1802.
- RICHARDSON, D. S. and LICHTMAN, J. W. (2015). Clarifying tissue clearing. *Cell* **162** 246–257. <https://doi.org/10.1016/j.cell.2015.06.067>
- ROČKOVÁ, V. and GEORGE, E. I. (2018). The spike-and-slab LASSO. *J. Amer. Statist. Assoc.* **113** 431–444. [MR3803476](#) <https://doi.org/10.1080/01621459.2016.1260469>
- ROUSSEAU, J. and MENGENSEN, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 689–710. [MR2867454](#) <https://doi.org/10.1111/j.1467-9868.2011.00781.x>
- SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. [MR1309433](#)
- SHAHBABA, B. and JOHNSON, W. O. (2013). Bayesian nonparametric variable selection as an exploratory tool for discovering differentially expressed genes. *Stat. Med.* **32** 2114–2126. [MR3067360](#) <https://doi.org/10.1002/sim.5680>
- SHENG, M. and GREENBERG, M. E. (1990). The regulation and function of c-fos and other immediate early genes in the nervous system. *Neuron* **4** 477–485. [https://doi.org/10.1016/0896-6273\(90\)90106-p](https://doi.org/10.1016/0896-6273(90)90106-p)
- SUN, X. and LIN, Y. (2016). Npas4: Linking neuronal activity to memory. *Trends Neurosci.* **39** 264–275.
- SUNKIN, S. M., NG, L., LAU, C., DOLBEARE, T., GILBERT, T. L., THOMPSON, C. L., HAWRYLYCZ, M. and DANG, C. (2013). Allen brain atlas: An integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res.* **41** 996–1008.
- TADESSE, M. and VANNUCCI, M. (2021). *Handbook of Bayesian Variable Selection*. CRC Press/CRC, New York.
- TUKEY, J. W. (1993). Tightening the clinical trial. *Control. Clin. Trials* **14** 266–285. [https://doi.org/10.1016/0197-2456\(93\)90225-3](https://doi.org/10.1016/0197-2456(93)90225-3)
- VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Anal.* **12** 1221–1274. [MR3724985](#) <https://doi.org/10.1214/17-BA1065>
- WASSERSTEIN, R. L., SCHIRM, A. L. and LAZAR, N. A. (2019). Moving to a world beyond “ $p < 0.05$ ”. *Amer. Statist.* **73** 1–19.