Web-based Supplementary Materials for "Spatial Bayesian Latent Factor Regression Modeling of Coordinate-based Meta-analysis Data" by Silvia Montagna, Tor Wager, Lisa Feldman Barrett, Timothy D. Johnson, and Thomas E. Nichols

SILVIA MONTAGNA
School of Mathematics, Statistics and Actuarial Science,
University of Kent, Canterbury CT2 7FS, U.K.

TOR WAGER
Department of Psychology and Neuroscience, University of Colorado at Boulder,
Boulder, CO 80309, U.S.A.

LISA FELDMAN BARRETT
Department of Psychology, Northeastern University, Boston, MA 02115, U.S.A.

TIMOTHY D. JOHNSON
Biostatistics Department, University of Michigan, Ann Arbor, MI 48109, U.S.A.

THOMAS E. NICHOLS
Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK

Web Appendix A

Prior specification

In this Section, we specify prior distributions for the model parameters of the spatial Bayesian latent factor regression model. With regard to specifying priors, it is practically important to choose conditionally conjugate priors given the high dimensionality of the application. A prior distribution is defined conjugate when the resulting posterior distribution is in the same family as the prior, that is, when the posterior has the same algebraic form as the prior, but with different (updated) parameter values. Choosing conditionally conjugate priors leads to an efficient posterior computation via a Gibbs sampler. For this reason, we have chosen Gaussian priors (multivariate, when appropriate) for the model parameters in Sections 2.1 and 2.2 of the paper, and Gamma priors on the residual precisions. We begin by providing details on the prior for the factor loading matrix $\mathbf{\Lambda}$.

Careful modeling of $\mathbf{\Lambda}$ is crucial in that the factor loading matrix ultimately controls sparsity and basis selection. Following [6], we adopt a multiplicative gamma process shrinkage (MGPS) prior on the loadings [2]:

$$\lambda_{jh}|\iota_{jh}, \tau_h \sim N(0, \iota_{jh}^{-1}\tau_h^{-1}), \quad \iota_{jh} \sim \text{Gamma}\left(\frac{\rho}{2}, \frac{\rho}{2}\right), \quad \tau_h = \prod_{l=1}^{h} \delta_l \tag{1}$$

$$\delta_1 \sim \text{Gamma}(a_1, 1), \quad \delta_l \sim \text{Gamma}(a_2, 1), \quad l \geq 2 \tag{2}$$

with $j = 1, \ldots, p$ and $h = 1, \ldots, k$. Elements $\{\delta_l\}_{l \geq 1}$ are independent, τ_h is a global shrinkage parameter for the hth column of $\mathbf{\Lambda}$ and the ι_{jh}'s are local shrinkage parameters for the elements in the hth column. Under the choice $a_2 > 1$, the τ_h's are stochastically increasing favoring more shrinkage as the column index increases and preventing the factor splitting problem [2]. The local shrinkage parameters prevent over-shrinking the nonzero loadings. Therefore, the MGPS prior shrinks a subset of the loadings strongly towards zero while retaining a sparse signal. For more details on the properties of this prior, we defer to [2].

To obviate the need for pre-specifying the number of factors, we follow [2] and implement an adaptive algorithm for choosing k. The idea behind this sampler is to strike a balance between missing important factors by choosing k too small and wasting computation on an overly conservative value. At iteration t, one monitors the number of columns of $\mathbf{\Lambda}$ having all elements within some pre-specified small neighbourhood of zero. If the number of such columns drops to zero, then a column is added to $\mathbf{\Lambda}$ by sampling the new loadings from the prior distribution, and otherwise discard the redundant columns in that the contribution of the factors is negligible. The other parameters are also modified accordingly. By proposing an oversized set of basis functions (large p in Equation 2 of the paper), we can ultimately control the complexity of our model by including or excluding a particular basis (Equation 9 in the paper) based on its contribution to the likelihood of the observed data. To guarantee convergence of the chain, the adaptations are designed to satisfy the diminishing adaptation condition in Theorem 5 of [10].

The Bayesian specification of our model is completed by placing a Gamma prior on the residual precisions, $\sigma_m^{-2} \sim \text{Gamma}(a_\sigma, b_\sigma)$, for $m = 1, \ldots, p$, and, if covariate information is available, a Cauchy prior (equivalently, a t-distribution with one degree of freedom) on the matrix of coefficients $\boldsymbol{\beta}$ as follows

$$\boldsymbol{\beta}_l \sim \text{N}(0, \text{Diag}(w_{lj}^{-1})), \quad w_{lj} \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right) \tag{3}$$

for $l = 1, \ldots k$ and $j = 1, \ldots, r$. To ensure conjugacy and speedy update, we expressed the Cauchy as an inverse gamma mixture of a normal distribution. The Cauchy distribution is used here as an alternative to the Gaussian distribution as it is more robust to outliers.

Finally, the diagonal orthant (DO) multinomial probit extension for reverse inference (Section 2.2 in the paper) requires setting priors on α_j and $\boldsymbol{\gamma}_j$, for $j = 1, \ldots, J$. We choose $\alpha_j \sim \text{N}(m_\alpha, v_\alpha)$ and $\boldsymbol{\gamma}_j \sim \text{N}_k(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma)$, and posterior computation can proceed via conditionally conjugate Gibbs steps.

Setting p, b, and hyperparameters

To implement our methodology, one has to choose the hyperparameters for the priors above as well as the number of bases p in Equation 2 of the paper and their bandwidth $b = \frac{1}{2s^2}$ (Equation 3 in the paper), where s is the length-scale of the kernel. Likely, the most daunting task is the choice of the bandwidth b. For the simulations and real data analyses we considered values of b in the interval $[0.00781, 0.00195]$, which corresponds to $s^2 \in [64, 256]$. For a given number of basis functions p, larger values of s^2 ensure smoother surfaces. We recall that we used a 2 × 2 × 2 mm brain mask, thus a value of $s^2 = 64$ corresponds to a standard deviation of $s = 8$ mm for the 3D kernels (equivalently, 4 voxels). In general, one can not obtain rougher surfaces than the resolution determined by the bandwidth, thus the choice of b requires careful sensitivity analysis to identify a value that induces the desired level of smoothness for the intensities. Applications with intensities exhibiting a different degree of smoothness in different areas of the spatial domain would require spatially adaptive smoothness, which can potentially be achieved by choosing a pre-specified finite dictionary of different bandwidths and then allowing the kernels to have varying unknown bandwidths via a griddy Gibbs sampler.

In addition to choosing the bandwidth, one has to choose the number of bases p. For any given bandwidth, a smaller value for p results in sparser bases with non-overlapping kernels (the same holds for a given number of bases and a narrower bandwidth). This produces visually unappealing surfaces that exhibit peaks at those kernels centred in proximity of clusters of foci while dropping off quickly beyond the kernel core. In general, one can include a rich, pre-specified set of basis functions since the model allows automatic shrinkage and effective removal of basis coefficients not needed to characterize any of the intensities under study. In our 3D applications, we considered values of p ranging between 400 and 600. At a given axial slice, knot locations formed a 2D grid of equally spaced kernels. At the successive axial slice, knots were given a δ translation along the x-axis to help entertaining a richer variety of shapes. In practical situations, sensitivity analysis is required to choose these parameters jointly, and changing either one of p or b possibly requires fine tuning the other as necessary to find the best combination to fit the data. We remark that different combinations of p and b, chosen as advised above, did not impact on the predictive performance of our model in the simulations and real data analyses, but only affected the qualitative representation of the estimated intensities.

With regard to specifying hyperparameter values, one needs to set hypeparameters ρ, a_1, and a_2 for the MGPS prior in Equation (1) and Equation (2). As remarked above, a choice $a_2 > 1$ induces stochastically increasing τ_h in (1), which favors more shrinkage as the column index increases. In the real data analysis (Section 3 in the paper), we set $\rho = 3, a_1 = 2.1$, and $a_2 = 3.1$, and our sensitivity analyses showed robustness to different choices of these hyperparameters. Furthermore, one has to choose a_σ and b_σ, which are the hyperparameters values of the Gamma prior on σ_j^{-2}, $j = 1, \ldots, p$. Our suggestion is to fix a mean and variance for these Gamma priors and solve for the hyperparameters. For example, we assigned a Gamma$(1, 0.3)$ prior distribution with mean $1/3$ to the diagonal elements of $\boldsymbol{\Sigma}^{-1}$ for the real data analysis.

The DO multinominal probit model (Section 2.2) requires setting the hyperprior parameters of priors $\alpha_j \sim \mathrm{N}(m_\alpha, v_\alpha)$ and $\boldsymbol{\gamma}_j \sim \mathrm{N}_k(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma)$, for $j = 1, \ldots, J$. With regard to the emotion meta-analysis dataset, and assuming that all emotion types are equally likely a priori, we can set $m_\alpha = \Phi^{-1}(1/J)$, where J is the number of unique study types and $\Phi(\cdot)$ is the standard normal CDF. Alternatively, we can have the hyperprior mean to be type-specific, $m_{\alpha,j}$, and set $m_{\alpha,j} = \Phi^{-1}(s_j)$, where s_j is the sample proportion of type j studies. Further, we set $v_\alpha = 1$, $\boldsymbol{\mu}_\gamma = \mathbf{0}$, and $\boldsymbol{\Sigma}_\gamma = \boldsymbol{I}$.

In practical applications, there is often little prior information to guide the selection of higher order hyperparameters, therefore it is important to perform sensitivity analysis to investigate robustness of the results to different choices of these values. That is, one should repeat the analysis for different choices of p, b and $\{\rho, a_1, a_2, a_\sigma, b_\sigma, m_\alpha, v_\alpha, \boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma\}$, and see how the posterior distributions of the model parameters (and various statistics) vary. Refer to Algorithm 1 for a sketch of the pseudo-code for sensitivity analysis.

Algorithm 1 Pseudo-code for sensitivity analysis

```
for $s = 1$ : Number of repetitions do
Choose $\{p, b\}$ and hyperparameters $\{\rho, a_1, a_2, a_\sigma, b_\sigma, m_\alpha, v_\alpha, \boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma\}$
Set initial values for the model parameters
    repeat
        Parameters sampling: Gibbs steps + Hamiltonian Monte Carlo
    until Convergence
Export parameters estimates
```

Posterior computation

To facilitate the routine implementation of the proposed method, the Matlab code for the Bayesian spatial latent factor regression model and its extensions to reverse inference are available at the Biometrics website on Wiley Online Library. Here we provide a description of the MCMC algorithm used to update from the posterior distribution of the model parameters in Section 2.1. Posterior computation proceeds via a hybrid Gibbs sampler with a Hamiltonian Monte Carlo (HMC) step [7] to update the basis function coefficients $\boldsymbol{\theta}_i$. The sampler cycles through the following steps:

1. *Update of* $\boldsymbol{\theta}_i$: Conditioning on all other model parameters, the log-posterior is given by

$$\log \pi(\boldsymbol{\theta}_i \mid -) \propto -\int_{\mathcal{B}} \exp\{\mathbf{b}(\mathbf{s})^\top \boldsymbol{\theta}_i\} ds + \sum_{j=1}^{n_i} \mathbf{b}(\mathbf{x}_{ij})^\top \boldsymbol{\theta}_i - \frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\Lambda}\boldsymbol{\eta}_i)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\Lambda}\boldsymbol{\eta}_i),$$

for $i = 1, \dots, n$. We recall that $\{\mathbf{x}_{ij}\}_{j=1}^{n_i}$ denotes the set of foci reported by study i. After discretising to a grid $B \subseteq \mathcal{B}$, the log posterior becomes

$$\log \pi(\boldsymbol{\theta}_i \mid -) \propto -V \sum_{l \in B} \exp\{\mathbf{b}(\mathbf{l})^\top \boldsymbol{\theta}_i\} + \sum_{j=1}^{n_i} \mathbf{b}(\mathbf{x}_{ij})^\top \boldsymbol{\theta}_i - \frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\Lambda}\boldsymbol{\eta}_i)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\Lambda}\boldsymbol{\eta}_i), \quad (4)$$

where V is the volume of each voxel. We resort to HMC [7] to sample from (4) by defining the potential energy function as

$$U(\boldsymbol{\theta}_i) = -\log \pi(\boldsymbol{\theta}_i \mid -),$$

and its derivative with respect to $\boldsymbol{\theta}_i$ corresponds to

$$\frac{\partial U}{\partial \boldsymbol{\theta}_i} = V \sum_{l \in B} \mathbf{b}(\mathbf{l})^\top \exp\{\mathbf{b}(\mathbf{l})^\top \boldsymbol{\theta}_i\} - \sum_{j=1}^{n_i} \mathbf{b}(\mathbf{x}_{ij})^\top + (\boldsymbol{\theta}_i - \boldsymbol{\Lambda}\boldsymbol{\eta}_i)^\top \boldsymbol{\Sigma}^{-1}.$$

The kinetic energy, $K(\tilde{\boldsymbol{\theta}}_i)$, is assumed to have form $K(\tilde{\boldsymbol{\theta}}_i) = \sum_{m=1}^{p} \frac{\tilde{\theta}_{ij}^2}{2}$. A detailed presentation of Hamiltonian dynamics is beyond the scope of this paper. We defer to [7] for a description of the leapfrog method and further details

2. *Update of* $\boldsymbol{\Lambda}$: Sample $\lambda_{jh}, \delta_1, \delta_h, \iota_{jh}$ from the following posteriors:

   (a) Denote the jth row of $\boldsymbol{\Lambda}_{k^*}$ (the loading matrix $\boldsymbol{\Lambda}$ truncated to $k^* \ll p$) by $\boldsymbol{\lambda}_j$; the $\boldsymbol{\lambda}_j$'s have conditionally independent conjugate posteriors given by

$$\pi(\boldsymbol{\lambda}_j \mid -) \sim N_{k^*}((\mathbf{D}_j^{-1} + \sigma_j^{-2}\boldsymbol{\eta}^\top \boldsymbol{\eta})^{-1} \boldsymbol{\eta}^\top \sigma_j^{-2} \boldsymbol{\theta}^{(j)}, (\mathbf{D}_j^{-1} + \sigma_j^{-2}\boldsymbol{\eta}^\top \boldsymbol{\eta})^{-1})$$

   with $\mathbf{D}_j^{-1} = \text{diag}(\iota_{j1}\tau_1, \dots, \iota_{jk^*}\tau_{k^*})$, $\boldsymbol{\eta}^\top = [\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_{k^*}]$ and $\boldsymbol{\theta}^{(j)} = (\theta_{j1}, \dots, \theta_{jn})$, for $j = 1, \dots, p$

   (b) Sample ι_{jh} from

$$\pi(\iota_{jh} \mid -) \sim \text{Gamma}\left(\frac{\rho + 1}{2}, \frac{\rho}{2} + \frac{\tau_h \lambda_{jh}^2}{2}\right)$$

(c) Sample δ_1 from

$$\pi(\delta_1 \mid -) \sim \text{Gamma}\left(a_1 + \frac{pk^*}{2}, 1 + \frac{1}{2}\sum_{l=1}^{k^*} \tau_l^{(1)} \sum_{j=1}^{p} \iota_{jl}\lambda_{jl}^2\right),$$

and for $h \geqslant 2$, sample δ_h from

$$\pi(\delta_h \mid -) \sim \text{Gamma}\left(a_2 + \frac{p}{2}(k^* - h + 1), 1 + \frac{1}{2}\sum_{l=h}^{k^*} \tau_l^{(h)} \sum_{j=1}^{p} \iota_{jl}\lambda_{jl}^2\right),$$

where $\tau_l^{(h)} = \prod_{t=1,t\neq h}^{l} \delta_t$ for $h = 1, \ldots, k^*$

The sampling begins with a very conservative choice of k^*, which is then automatically selected within the adaptive Gibbs sampler as described in [2]

3. *Update of* $\boldsymbol{\eta}_i$: Sample $\boldsymbol{\eta}_i$ from the full conditional posterior

$$\pi(\boldsymbol{\eta}_i \mid -) \sim N\left(\left(\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} + \boldsymbol{I}_k\right)^{-1}(\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}_i + \boldsymbol{\beta}^\top\mathbf{Z}_i), \left(\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} + \boldsymbol{I}_k\right)^{-1}\right)$$

for $i = 1, \ldots, n$. We recall that $\mathbf{Z}_i$ is the $r \times 1$ vector of covariates for study i

4. *Update of* σ_j^2: Denote with $\sigma_j^{-2}, j = 1, \ldots, p$, the diagonal elements of $\boldsymbol{\Sigma}^{-1}$. Assume $\sigma_j^{-2} \sim \text{Gamma}(a_\sigma, b_\sigma)$. Sample σ_j^{-2} from conditionally independent posteriors

$$\pi(\sigma_j^{-2} \mid -) \sim \text{Gamma}\left(a_\sigma + \frac{n}{2}, b_\sigma + \frac{\sum_{i=1}^{n}(\theta_{ij} - \boldsymbol{\lambda_j}\boldsymbol{\eta}_i)^2}{2}\right)$$

where $\boldsymbol{\lambda}_j$ corresponds to the jth row of $\boldsymbol{\Lambda}$

5. *Update of* $\boldsymbol{\beta}$*:* A Cauchy prior is induced on the columns of the $r \times k$ matrix of coefficients as follows

$$\boldsymbol{\beta}_l \sim N(\mathbf{0}, \text{Diag}(w_{lj}^{-1})), \quad \text{with} \quad w_{lj} \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$$

for $l = 1, \ldots, k$ and $j = 1, \ldots r$. The update proceeds by sampling

(a) ω_{lj} from its full conditional posterior

$$\pi(\omega_{lj} \mid -) \sim \text{Gamma}\left(1, \frac{1}{2}\left(1 + \beta_{lj}^2\right)\right)$$

(b) the lth column of $\boldsymbol{\beta}$ from its full conditional posterior

$$\pi(\boldsymbol{\beta}_l \mid -) \sim N\left(\left(\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^\top + \mathbf{E}_l^{-1}\right)^{-1}\tilde{\mathbf{Z}}\boldsymbol{\eta}_{\cdot l}^\top, \left(\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^\top + \mathbf{E}_l^{-1}\right)^{-1}\right),$$

where $\boldsymbol{\eta}_{\cdot l}^\top \sim N(\tilde{\mathbf{Z}}^\top\boldsymbol{\beta}_l, \boldsymbol{I}_n)$ denotes the lth column of the $n \times k$ transpose of the matrix of latent factors $\boldsymbol{\eta}$, $\boldsymbol{I}_n$ is the $n \times n$ identity matrix, and $\tilde{\mathbf{Z}}^\top$ denotes the transpose of the $r \times n$ matrix of predictors $\tilde{\mathbf{Z}}$. Each row i of $\tilde{\mathbf{Z}}^\top$ corresponds to the vector of covariates for study i, for $i = 1, \ldots, n$. Matrix $\mathbf{E}_l$ corresponds to $\mathbf{E}_l = \text{Diag}(\omega_{lj}^{-1})$, for $l = 1, \ldots, k$ and $j = 1, \ldots, r$

The DO multinomial probit extension described in Section 2.2 involves a straightforward modification of the MCMC algorithm described above, which now includes additional steps to sample from the full conditional posterior distributions of the new model parameters. In particular, the update of the latent factors is modified as described in Step (3b) and three steps (Steps 6-8) are added to account for the study-type component model parameters:

(3b) *Update of* $\boldsymbol{\eta}_i$: Sample $\boldsymbol{\eta}_i$ from a multivariate Gaussian full conditional posterior distribution with mean

$$\mathbb{E}(\boldsymbol{\eta}_i|-) = \left(\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} + \boldsymbol{I}_k + \boldsymbol{\gamma}\boldsymbol{\gamma}^\top\right)^{-1}(\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}_i + \boldsymbol{\beta}^\top\mathbf{Z}_i + \boldsymbol{\gamma}(\boldsymbol{\chi}_i - \alpha))$$

and covariance

$$\mathbb{C}\mathrm{ov}(\boldsymbol{\eta}_i|-) = \left(\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} + \boldsymbol{I}_k + \boldsymbol{\gamma}\boldsymbol{\gamma}^\top\right)^{-1},$$

for $i = 1, \ldots, n$, where $\boldsymbol{\gamma}$ denotes the $k \times J$ matrix $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_J]$, $\boldsymbol{\chi}_i = [\chi_{i1}, \ldots, \chi_{iJ}]^\top$, and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_J]^\top$

(6) *Update of* $\boldsymbol{\gamma}_j$: Sample $\boldsymbol{\gamma}_j$ from the following full conditional posterior distribution

$$\pi(\boldsymbol{\gamma}_j|-) \sim N\left((\boldsymbol{\Sigma}_\gamma^{-1} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)^{-1}(\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\mu}_\gamma + \boldsymbol{\eta}(\boldsymbol{\chi}_j - \boldsymbol{\alpha})), (\boldsymbol{\Sigma}_\gamma^{-1} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)^{-1}\right),$$

where $\boldsymbol{\chi}_j = [\chi_{1,j}, \ldots, \chi_{n,j}]^\top$ and $\boldsymbol{\alpha}_j$ is the $n \times 1$ vector of n replicates of α_j, $\boldsymbol{\alpha} = [\alpha_j, \ldots, \alpha_j]_{n\times 1}^\top$, for $j = 1, \ldots, J$

(7) *Update of* α_j: Sample α_j from the following full conditional posterior distribution

$$\pi(\alpha_j|-) \sim N\left(\frac{v_\alpha}{1 + v_\alpha n_j} \times \left(\frac{\mu_\alpha}{v_\alpha} + \sum_{i=1}^{n_j}\left(\chi_{i,j} - \boldsymbol{\eta}_i^\top\boldsymbol{\gamma}_j\right)\right), \frac{v_\alpha}{1 + v_\alpha n_j}\right)$$

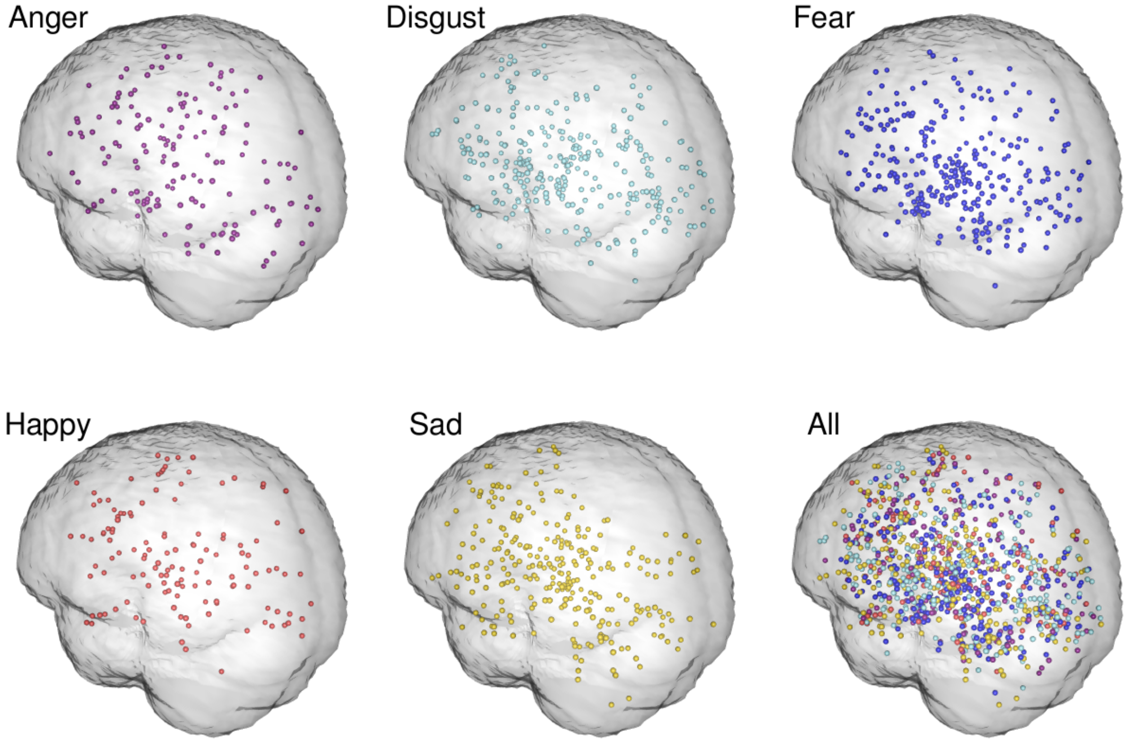
where n_j is the number of studies of type j, for $j = 1, \ldots, J$

(8) *Update of* χ_{ij}: Sample χ_{ij} from the following full conditional posterior distribution

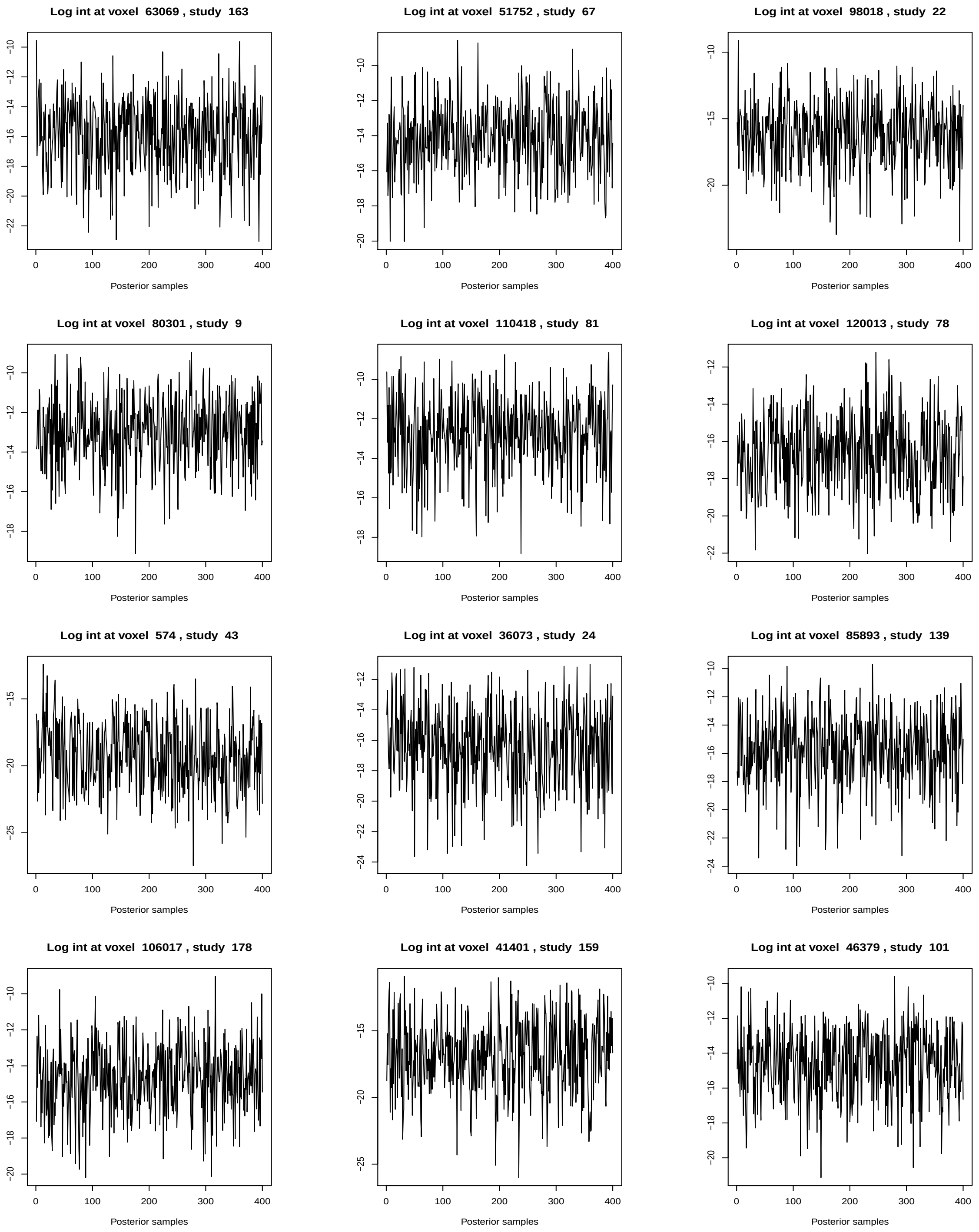
$$\pi(\chi_{ij}|-) = \begin{cases} TN_{[0,\infty)}(\alpha_j + \boldsymbol{\gamma}_j^\top\boldsymbol{\eta}_i, 1) & \text{if } y_i = j \\ TN_{(-\infty,0)}(\alpha_j + \boldsymbol{\gamma}_j^\top\boldsymbol{\eta}_i, 1) & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, n$, where $TN_{[a,b]}$ denotes the truncated normal distribution on the interval $[a, b]$

Web Appendix B
Meta-analysis of emotions (cont.'d Section 3)



Web Figure 1: The foci reported by the 187 studies of five emotions.



Web Figure 2: Traceplots of the posterior log intensity function $\mu_i(\boldsymbol{\nu})$ for a variety of randomly selected studies i and randomly selected voxels $\boldsymbol{\nu}$. The sampler was run for 15,000 iterations, with the first 5,000 samples discarded as a burn-in and collecting every $25th$ sample to thin the chain. Traceplots show the post-burn in and thinned samples.

Web Appendix C

Meta-analysis of emotion and executive control studies

In this Section, we demonstrate the application of our algorithm to real-world data by comparing studies of emotion and cognitive control, using hand-coded activation coordinates from previous meta-analyses [3, 4, 5, 9, 8, 11, 13, 12]. Each domain has been studied extensively using neuroimaging in hundreds of published studies. There is substantial convergence about the systems broadly involved in each, and though they interact, cognitive control and emotion are associated with distinct large-scale networks. In addition, there is substantial converging evidence on the cognitive and emotional functions of homologous systems in invasive animal studies, further validating the functional roles of the systems identified in human neuroimaging. We aim to both characterize the patterns of brain activation that are typical to each type, as well as evaluate the ability of our model to conduct a limited "reverse inference", that is, to predict which topic (cognitive control or emotion) a new study investigates.

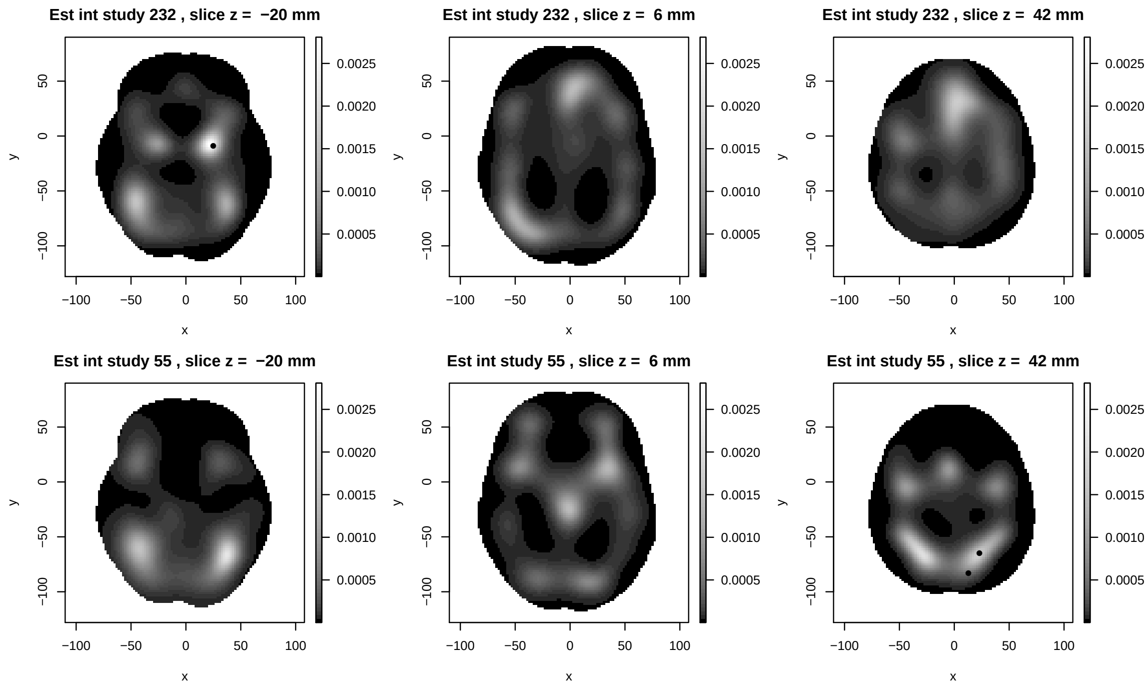
Our dataset consists of 1199 total studies categorized as either "emotion" (860 studies, 6481 foci; [4]) or "executive control" (339 studies, 4332 foci; [3, 9, 11, 8]). Covariate information is not available on these studies. We assigned a Gamma(1, 0.3) prior distribution with mean 1/3 to the diagonal elements of $\mathbf{\Sigma}^{-1}$. The hyperparametes of the MGPS prior were set to $\rho = 3$, $a_1 = 2.1$, $a_2 = 3.1$. In absence of covariate information, we assigned $\boldsymbol{\eta}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{I})$ for $i = 1, \ldots, n$. We chose $p = 352$ Gaussian kernels with bandwidth $b = 0.002$. Kernels were placed on axial slices roughly 8-12 mm apart, at $z = \{-38, -22, -14, -2, 6, 18, 28, 44\}$ mm (about 85% of the foci were located within these axial slices) and, within each slice, were equally spaced by forming a grid of 6×8 knots along the (x, y) direction. We used a standard brain mask with 2 mm^3 voxels and dimensions $91 \times 109 \times 91$. Kernels falling outside this mask were discarded. To update the basis function coefficients via HMC [7], we adopted the leapfrog method for L steps and with a stepsize of ϵ. At each iteration of the MCMC sampler, a new value for L was drawn from Poisson(25) and the stepsize was adapted every 10 iterations during burn-in to benchmark an average acceptance rate of 0.65 over the previous 100 iterations in the Metropolis-Hastings step. The sampler was run for 10,000 iterations, with the first 5,000 samples discarded as a burn-in and collecting every $20th$ sample to thin the chain. We assessed convergence of the chain by multiple runs of the algorithm from over-dispersed starting values and visually inspected the differences in the posterior mean intensity function $\mu_i(\boldsymbol{\nu})$ at a variety of voxels and for different studies. The sampler appeared to converge rapidly and mix efficiently.

Web Figure 3 displays the posterior mean intensity function at three axial slices for two randomly chosen studies. Slices are arranged in columns, the top (bottom) row shows an emotion (executive control) study. The smoothness of the images is due to the choice of the bandwidth parameter. As opposed to the executive control study, the emotion study shows strong activation in two regions at slice $z = -20$ and for $y > -26$ mm. These regions correspond to the amygdalae; almond-shaped structures in the brain of known importance in emotion processing. The executive control study shows stronger activation in more superior slices with a bilateral pattern. Note how we identify active regions even though a study does not have foci at a particular slice or, for example, both the right and left amygdala are active for study 232 even though only one focus is reported around the right amygdala. This a by-product of the borrowing of information across studies in our Bayesian treatment and an essential feature to adequately estimate the intensities over the whole the brain given the sparsity of foci per study.

The top row in Web Figure 4 shows the posterior mean difference between the estimated mean group intensity for the emotion and the executive control studies, respectively. The group intensity at iteration t is obtained by averaging the basis function coefficients for studies that belong to the group, that is,

$$\hat{\mu}_g^t(\boldsymbol{\nu}) = \exp\{\mathbf{b}(\boldsymbol{\nu})^\top \hat{\boldsymbol{\theta}}_g^t\}, \qquad \text{with} \qquad \hat{\boldsymbol{\theta}}_g^t = \frac{1}{\mathrm{Card}(g)} \sum_{i \in g} \hat{\boldsymbol{\theta}}_i^t,$$

where Card(g) is the cardinality of group g. The darker regions in the top row reveal stronger



Web Figure 3: Posterior mean intensity estimates for two randomly chosen studies. Top row: an emotion study; bottom row: an executive control study. Here we only show three axial slices (columns) of the fully 3D results. Black points denote reported foci at the corresponding axial slices.

activation of emotion studies (the amygdalae at slice $z = -20$ mm), whereas the bright regions denote stronger activation of executive control studies. The grey area reveals no difference in activation. We can also obtain standard deviation maps of the difference map as a measure of uncertainty around the point estimate and the corresponding standardized posterior mean difference maps. The clearest of these maps is slice $z = -20$ mm where one can easily identify the amygdalae as significant regions.

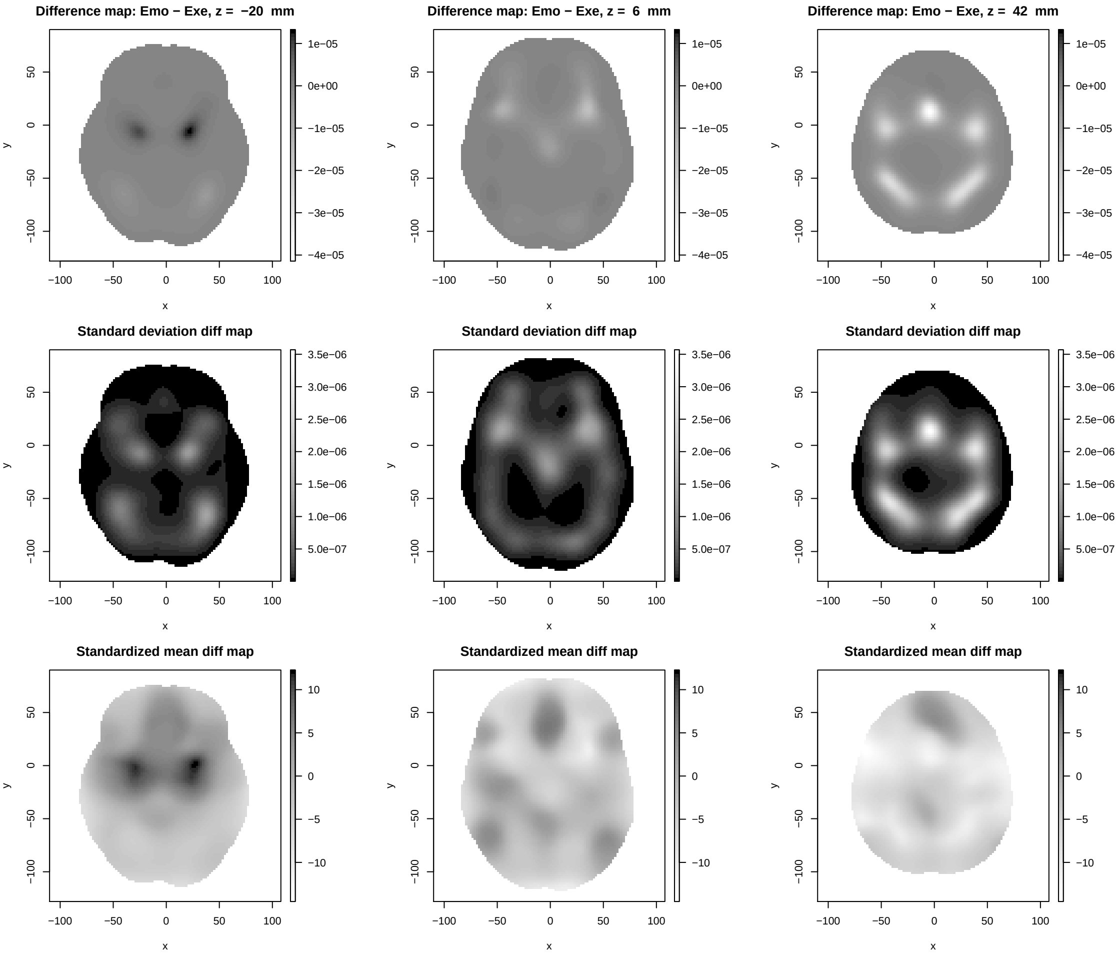
It is also of interest to examine the elements of the dictionary $\{\phi_l\}_{l=1}^k$. The posterior mean number of latent factors is $k = 4$ with 95% credible interval $[3, 6]$. Web Figure 5 shows the first four elements of the dictionary $\{\phi_l\}_{l=1}^4$ (rows) at axial slices $z = -20, 6, 42$ mms (columns). Notice how the magnitude of the learnt bases decreases as k increases, with the first couple of dictionary elements describing the principal patterns of activation and the later elements progressively shrunk toward zero. At every axial slice, the first two dictionary elements combined recover the principal patterns of activation we observed in Web Figures 3-4.

Reverse inference

Our dataset is large enough to split the data into a training set (50%), for which both foci and study type are retained for the analysis, and a testing set, for which the foci only are retained, and we test the predictive accuracy of our model. Let y_i denote the study type, with

$$y_i = \begin{cases} 1 & \text{if study } i \text{ is an emotion study} \\ 0 & \text{if study } i \text{ is an executive control study.} \end{cases}$$

Because the study type can be represented as a binary response, we build a probit model for study type and predict the posterior probability that a new point pattern data arose from



Web Figure 4: Top row: posterior mean of difference map for the emotion studies posterior group intensity vs. executive control studies posterior group intensity; middle row: estimated standard deviation of difference map; bottom row: standardised mean difference map. Here we only show three axial slices (columns) of the fully 3D results. The grey color scale for the middle row has been reversed to improve visibility of the results.

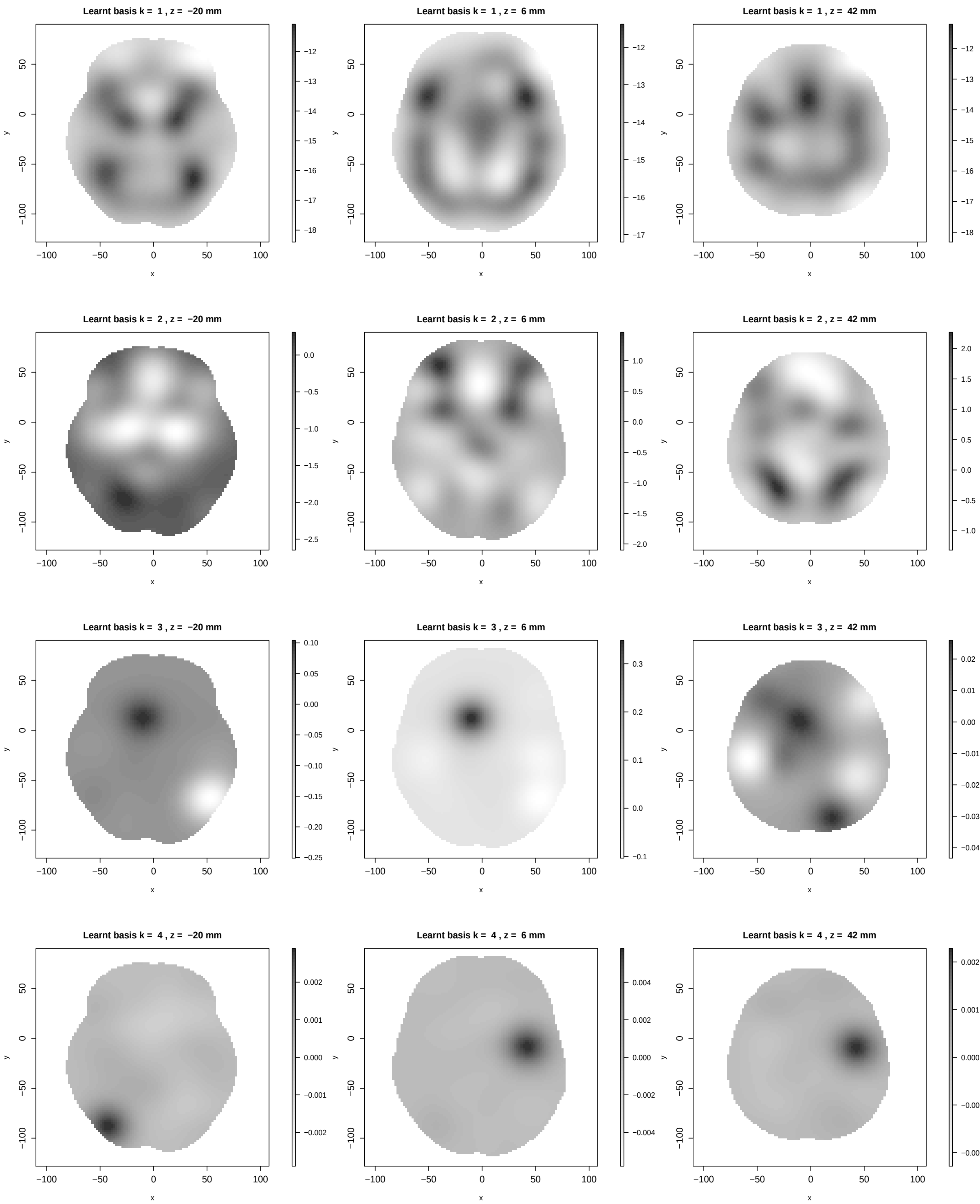
either cognitive process. Specifically, we model $p_{y_i} = \Pr(y_i = 1|\alpha, \boldsymbol{\gamma}, \boldsymbol{\eta}_i) = \Phi(\alpha + \boldsymbol{\gamma}^\top \boldsymbol{\eta}_i)$, where $\Phi(\cdot)$ denotes the standard normal distribution function. Parameter α can be interpreted as the baseline probability that study i is of the emotion type, and $\boldsymbol{\gamma}^\top \boldsymbol{\eta}_i$ accounts for study-specific random deviations.

The intercept α is given a $\mathrm{N}(m_\alpha, v_\alpha)$ prior, with $m_\alpha = \Phi^{-1}(0.50)$ assuming emotion and executive control studies are equally likely a priori, and $\boldsymbol{\gamma}$ is a vector of unknown regression coefficients with conjugate prior distribution $\boldsymbol{\gamma} \sim \mathrm{N}_k(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma)$. The full conditional posterior distributions needed for Gibbs sampling are not automatically available, but we can rely on the data augmentation algorithm of [1] to facilitate the computation. We introduce independent unobservable latent variables $W_1, \ldots, W_n$ and define

$$y_i = \mathbb{1}(W_i > 0), \qquad \text{with} \qquad W_i \sim \mathrm{N}(\alpha + \boldsymbol{\gamma}^\top \boldsymbol{\eta}_i, 1), \tag{5}$$

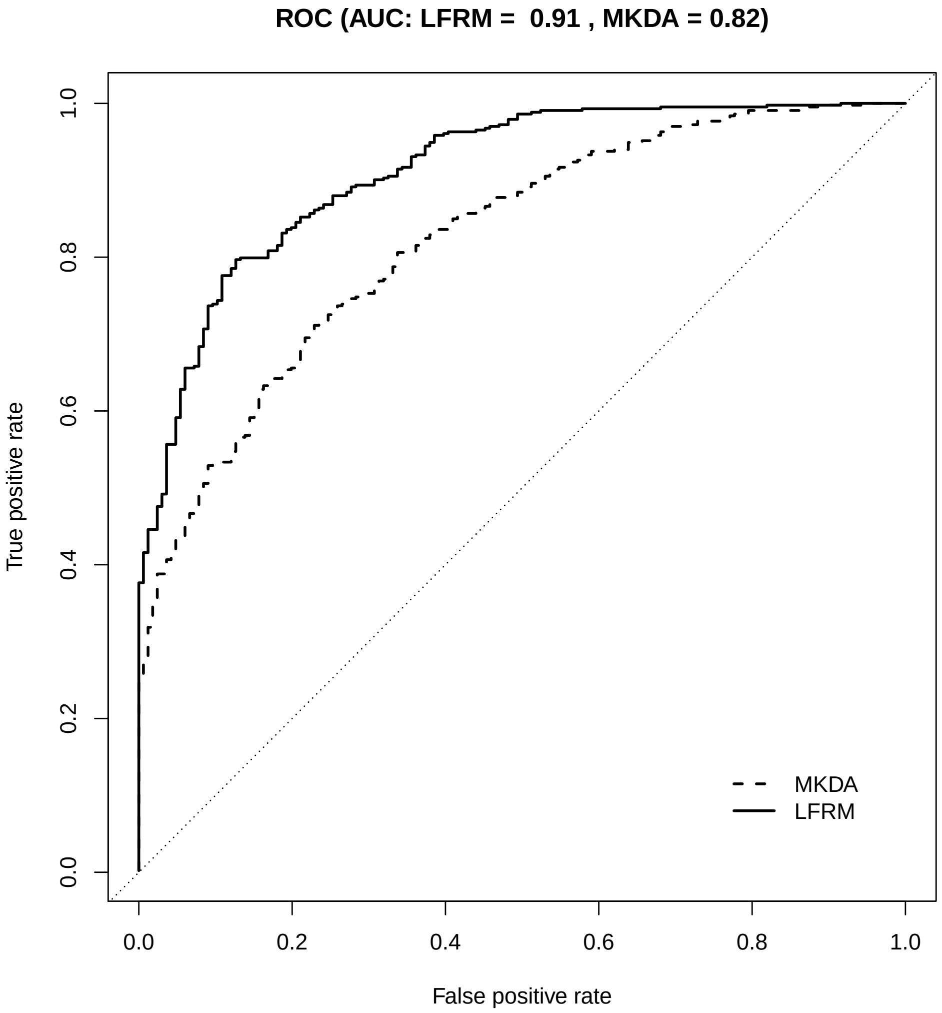
so that $\Pr(y_i = 1|\alpha, \boldsymbol{\gamma}, \boldsymbol{\eta}_i) = \Phi(\alpha + \boldsymbol{\gamma}^\top \boldsymbol{\eta}_i)$ by marginalizing out the latent variable W_i.

We compare our predictive method to previous work that combines MKDA and a naïve Bayesian classifier (NBC) [14]. Web Figure 6 shows the ROC for predicting the study type for studies in the test set with the corresponding area under the curve (AUC). We see that



Web Figure 5: Learnt dictionary elements $\{\phi_l\}_{l=1}^4$ at three axial slices (columns). The estimated posterior mean number of factors is $k = 4$.

our predictive model does a better job that MKDA + NBC at predicting the study type. This result is robust and we could confirm it through additional chains run for sensitivity analysis. Therefore, taking into account the spatial information in the data helps achieving better predictive accuracy, and our Bayesian model captures more sources of variation and conveys the uncertainty in the computation of the predictive probabilities for study type.



Web Figure 6: ROC curve for prediction of study type for studies in the test set. The AUC corresponds to the area under the curve.

Web Appendix D

In this Section, we consider a number of simulation examples to illustrate our approach. In part A, we investigate the sensitivity of our results to different choices for the hyperparameters. In part B, we investigate the performance of our model when studies are functionally unrelated and there is no significant effect. By repeating multiple Coordinate-based Meta-analyses (CBMAs), researchers found there is often little similarity between studies, and some of the published significant results were inadvertently due to software implementation issues. It is therefore of crucial importance to validate our approach under the null hypothesis of no true difference between studies. From the philosophy of CBMA, the result should be null.

Part A: Sensitivity analysis

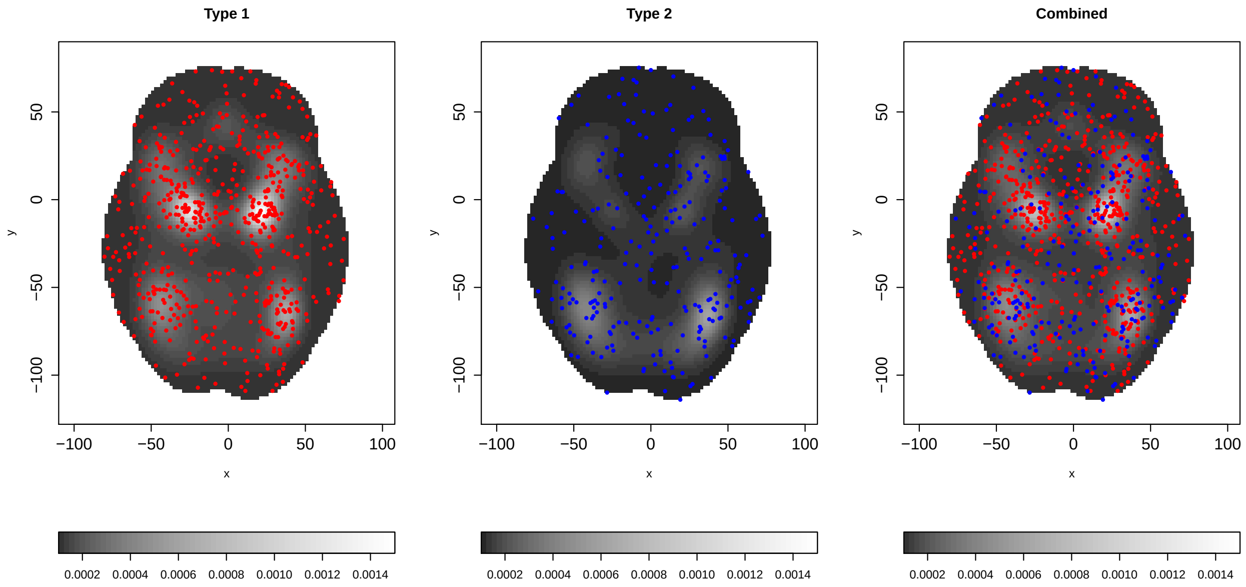
We examine how posterior inference on the intensity functions varies with different specifications of the number of bases p and bandwidth b, and hyperparameters a_σ, b_σ, α, ρ, a_1, and a_2. In particular, we consider nine scenarios of possible combinations of p and b (Web Table 1) and, for one of this scenarios, we also report results on sensitivities to the prior specifications of the remaining parameters.

Scenarios	p	b	a_σ	b_σ	m_α	a_1	a_2	ρ
1	25	1/800	1	0.3	$\Phi^{-1}(0.5)$	2.1	3.1	3
2	25	1/512			$\vdots$			
3	25	1/128			$\vdots$			
4	48	1/800			$\vdots$			
5	48	1/512			$\vdots$			
6	48	1/128			$\vdots$			
7	90	1/800			$\vdots$			
8	90	1/512			$\vdots$			
9	90	1/128	1	0.3	$\Phi^{-1}(0.5)$	2.1	3.1	3
10	90	1/128	1	0.25	$\Phi^{-1}(0.62)$	3.1	2.1	2
11	90	1/128	2	2	$\Phi^{-1}(0.62)$	3.1	2.1	2

Web Table 1: Prior specifications of eleven scenarios for sensitivity analysis.

To generate a synthetic dataset, we randomly selected 200 studies from the real data analysis in Web Appendix C whilst keeping the true proportion of sampled emotion studies equal to that of the real data analysis (70% emotion studies, hereafter called “type 1”). We then retained the estimated posterior mean intensity functions at slice $z = -20$ mm as true intensities for the 2D simulated dataset. Given the true intensities, the foci for each study where generated using the spatstat library in R. Web Figure 7 shows the true group average intensity functions and the simulated data points.

To simulate the posterior distribution, we run each chain for 50,000 iterations with a burn-in of 20,000, and thinned the chain every 20 iterations to reduce the autocorrelation in the posterior samples. To assess the posterior variability of the intensity functions, Web Figure 8 shows the posterior histograms of the intensity function evaluated at voxels with highest intensity values



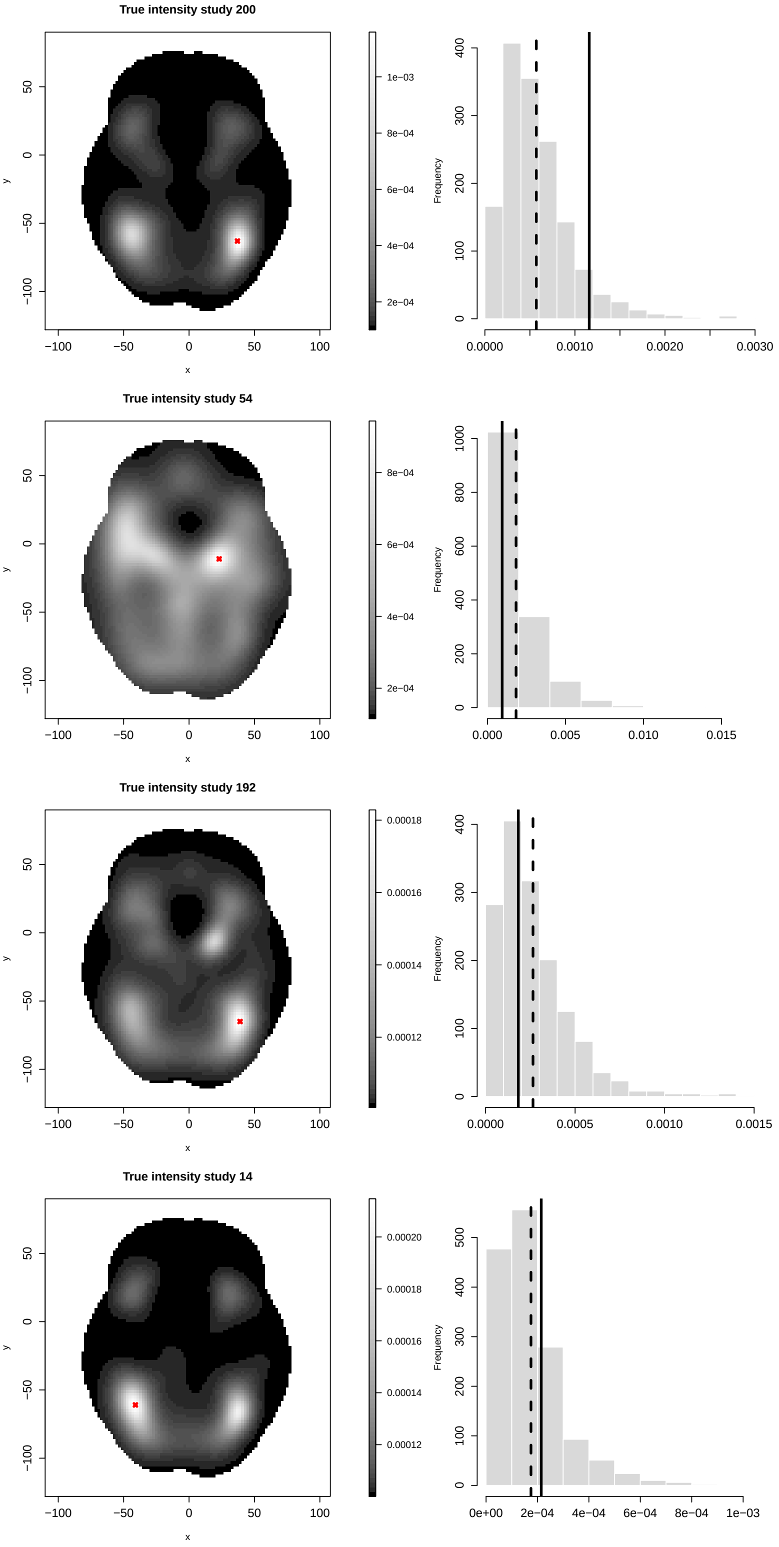
Web Figure 7: True group average intensity functions with data points for the simulation study.

for four studies under case scenario 6. The true values fall in the range of posterior samples; and the posterior mean intensities are close to the true values. We examined a variety of different studies for all scenarios and conclusions were unchanged (Web Figure 9 shows traceplots of the estimated posterior log intensity functions to the truth for a variety of studies and voxels). Therefore, the method provides a good accuracy in estimating the intensities.

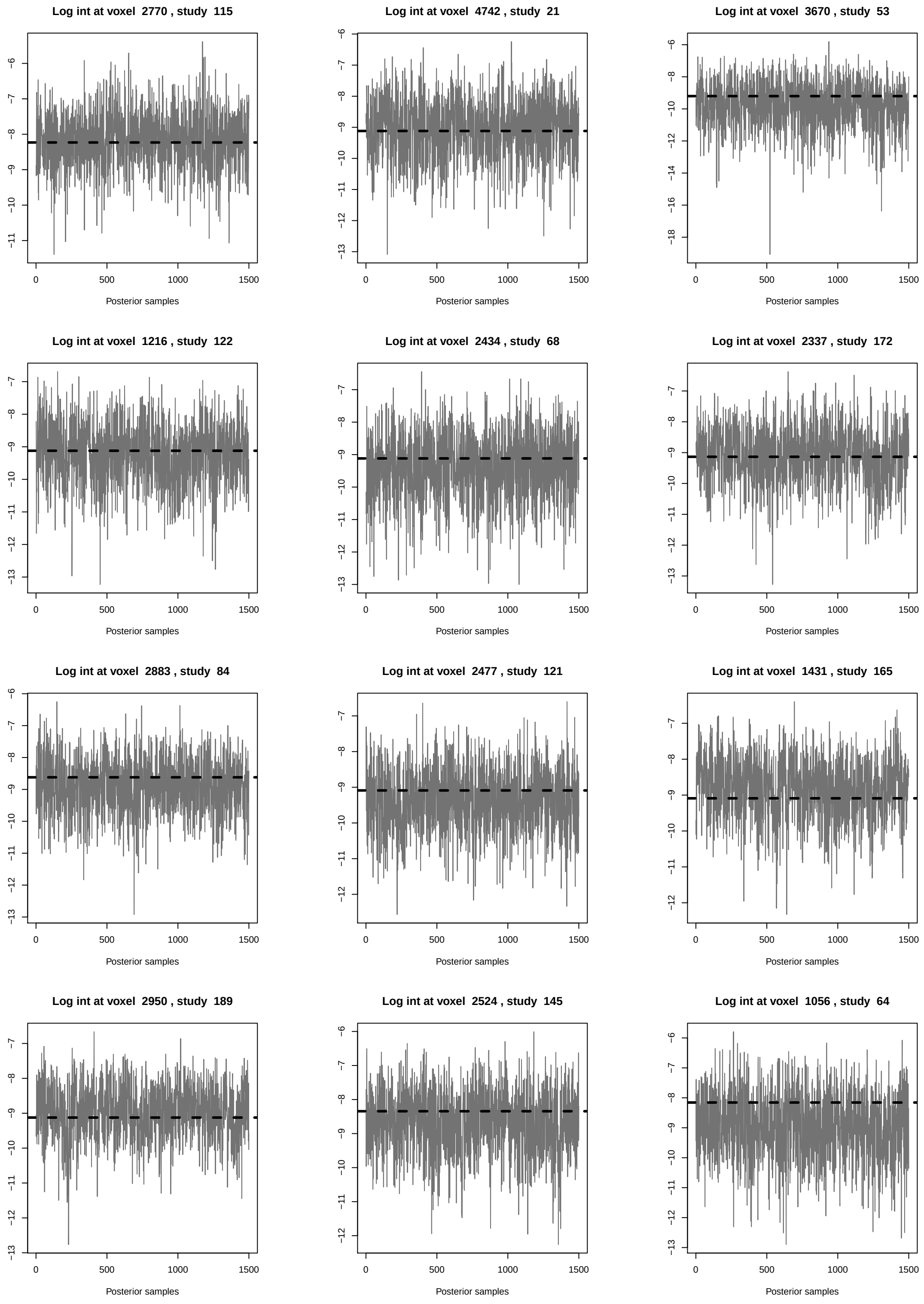
For a numerical comparison across the eleven scenarios, we computed the integrated mean square error (IMSE) between the posterior mean intensity function of for each study and the corresponding true function on axial slice A ($z = -20$ mm),

$$\text{IMSE} = \frac{1}{N}\sum_{i=1}^{N}\int_A \lambda_i(\nu) - \hat{\lambda}_i(\nu)d\nu,$$

where $\hat{\lambda}_i$ is the posterior mean intensity function for study i and λ_i is the true function. Also, we split the data into a training set (50%) and a test set, and computed the AUC to assess the predictive accuracy. Results under the eleven scenarios are reported in Web Table 2. From these results, it appears that a larger number of bases with relatively narrower kernels has to be preferred to be able to capture of large variety of shapes. In general, however, the values for p and b have to be inferred via sensitivity analysis in that the most appropriate values for these parameters always depend on the application at hand. We observe that results under cases 9-11 are very similar. This implies that the intensity estimates are stable and not sensitive to moderate changes of the model hyperparameters.



Web Figure 8: True intensity function for four randomly chosen studies in the simulated dataset. Histograms show the posterior distribution of the intensity function evaluated at voxels of highest intensity values (red "x"). The dashed line is the marginal posterior mean and the solid line is the true intensity.



Web Figure 9: Traceplots of the estimated posterior log intensity function for a variety of synthetic studies and voxels at axial slice $z = -20$ mm. Dashed lines represent true values of the log intensity function. Posterior samples are post burn-in.

Part B: Simulation study with coordinates generated at random

Scenarios	rIMSE	AUC
1	1.181	0.57
2	1.186	0.58
3	1.312	0.49
4	1.147	0.62
5	1.013	0.60
6	1.004	0.65
7	1.233	0.62
8	1.122	0.65
9	**1**	**0.65**
10	1.011	0.65
11	1.061	0.65

Web Table 2: Simulation study results. Comparison of the IMSE and AUC summary measures for our model under the eleven scenarios in Web Table 1. We report the IMSE relative to the value obtained for case scenario 9 (rIMSE).

We conclude this Section by investigating the performance of our model under the null hypothesis of no similarity across studies. We run several simulations on synthetic datasets with coordinates generated at random, and we report here the results of one such simulation in 2D.

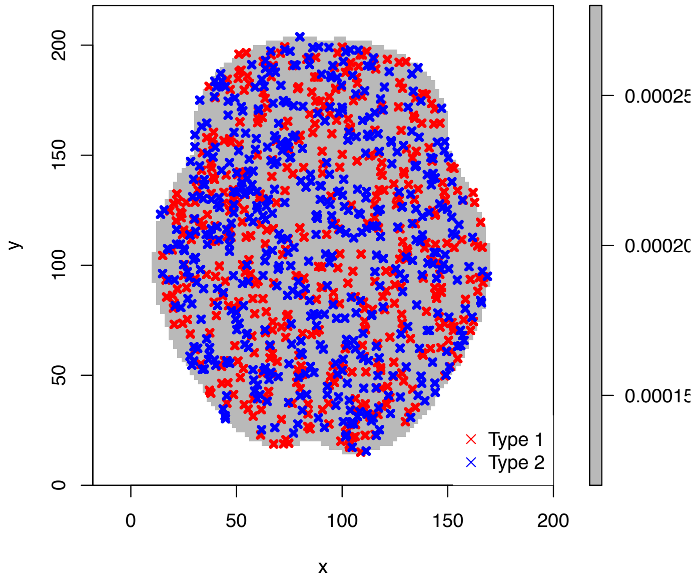
For $N = 200$ synthetic studies, we set a flat intensity function equal to 0.0002 at each and every 2D pixel of a chosen axial slice ($z = -20$ mm), for an expected number of points equal to 4 in the whole region. Given this flat intensity, the spatial point patterns for the $N = 200$ synthetic studies were generated using the spatstat library in R. The number of points per study ranged from a minimum of 1 to a maximum of 12, with an average number of points per study of 4.7 and a total number of points equal to 959. Further, the studies were randomly split between type 1 and type 0, for a total of 99 "type 1" studies and 101 "type 0" studies (Web Figure 10).

To simulate the posterior distribution, we run the chain for 50,000 iterations with a burn-in of 20,000, and thinned the chain every 20 iterations to reduce the autocorrelation in the posterior samples. To assess the posterior variability of the intensity functions, Web Figure 11 shows traceplots of the estimated posterior log intensity function to the truth (red dashed line) for a variety of randomly selected pixels and synthetic studies. The true value of $\log(2e-04)$ falls in the range of posterior samples; and the posterior mean intensities (blue dashed lines) are close to the true value. We examined a variety of other different studies and conclusions were unchanged. Therefore, the method provides a good accuracy in estimating the intensities.

The posterior mean number of latent factors is $k = 7$ with 95% credible interval $[5, 8]$. Therefore, the dimensionality reduction seems slightly less efficient when the studies are unrelated as more elements of the learnt dictionary $\{\tilde{\phi}\}$ are needed to describe the intensities. However, the MGPS prior shrinks progressively the elements of the dictionary (Web Figure 12) and, when plotted on the same color scale, only basis $\tilde{\phi}_1$ does not appear to be shrunk to zero. There is a connection between $\tilde{\phi}_1$ and the intercept of the model. In fact

$$\tilde{\phi}_1(\boldsymbol{\nu}) = \lambda_{11} + \sum_{m=2}^{p} \lambda_{m1} b_m(\boldsymbol{\nu}),$$

where λ_{11} is the first element of the first row of the factor loading matrix $\boldsymbol{\Lambda}$. The first row of $\boldsymbol{\Lambda}$, $\boldsymbol{\lambda}_{1\cdot}$, includes the loadings of θ_{i1}, the basis function coefficient of the intercept, and $\mathbb{M}\text{ean}(\theta_{i1}) =$



Web Figure 10: Data points for the simulation study with coordinates generated at random.

$\boldsymbol{\lambda}_{\mathbf{1}\cdot} \times \boldsymbol{\eta}_i$. By construction, element λ_{11} is the largest (in magnitude) element of the first row of $\boldsymbol{\Lambda}$. No relevant patterns emerge in the learnt dictionary elements.

The studies were split into a training set (75%) and a test set to assess the predictive accuracy. Web Figure 13 shows the ROC for predicting the study type for studies in the test set with the corresponding area under the curve (AUC). The five lines correspond to five replicates of the experiment, where each replicate has studies randomly assigned to either group 0 or group 1 and coordinates generated at random. In general, prediction performance roughly corresponds to random chance under the null hypothesis of truly unrelated studies.

References

[1] JH Albert and Siddhartha Chib. Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American statistical Association*, 88(422):669–679, 1993.

[2] Anirban Bhattacharya and David B Dunson. Sparse Bayesian infinite factor models. *Biometrika*, 98(2):291–306, June 2011.

[3] C. Kim, S. E. Cilles, N. F. Johnson, and B. T. Gold. Domain general and domain preferential brain regions associated with different types of task switching: a meta-analysis. *Human Brain Mapping*, 33(1):130–142, 2012.

[4] H. Kober, L. F. Barrett, J. Joseph, E. Bliss-Moreau, K. Lindquist, and T. D. Wager. Functional grouping and cortical-subcortical interactions in emotion: A meta- analysis of neuroimaging studies. *NeuroImage*, 42:998–1031, 2008.

[5] K. A. Lindquist, T. D. Wager, H. Kober, E. Bliss-Moreau, and L. F. Barrett. The brain basis of emotion: a meta-analytic review. *Behavioral and Brain Sciences*, 35(3):121–143, 2012.

[6] Silvia Montagna, Surya T Tokdar, Brian Neelon, and David B Dunson. Bayesian latent factor regression for functional and longitudinal data. *Biometrics*, 68(4):1064–73, 2012.

[7] Radford M. Neal. MCMC using Hamiltonian dynamics. In Xiao-Li Brooks, Steve; Gelman, Andrew; Jones, Galin; Meng, editor, *Handbook of Markov Chain Monte Carlo*. Boca Raton: Chapman & Hall–CRC Press, 2010.

[8] Derek Evan Nee, Joshua W. Brown, Mary K. Askren, Marc G. Berman, Emre Demiralp, Adam Krawitz, and John Jonides. A meta-analysis of executive components of working memory. *Cerebral Cortex*, 23(2):264–282, 2013.

[9] Derek Evan Nee, Tor D. Wager, and John Jonides. Interference resolution: Insights from a meta-analysis of neuroimaging tasks. *Cognitive, Affective, & Behavioral Neuroscience*, 7(1):1–7, 2007.

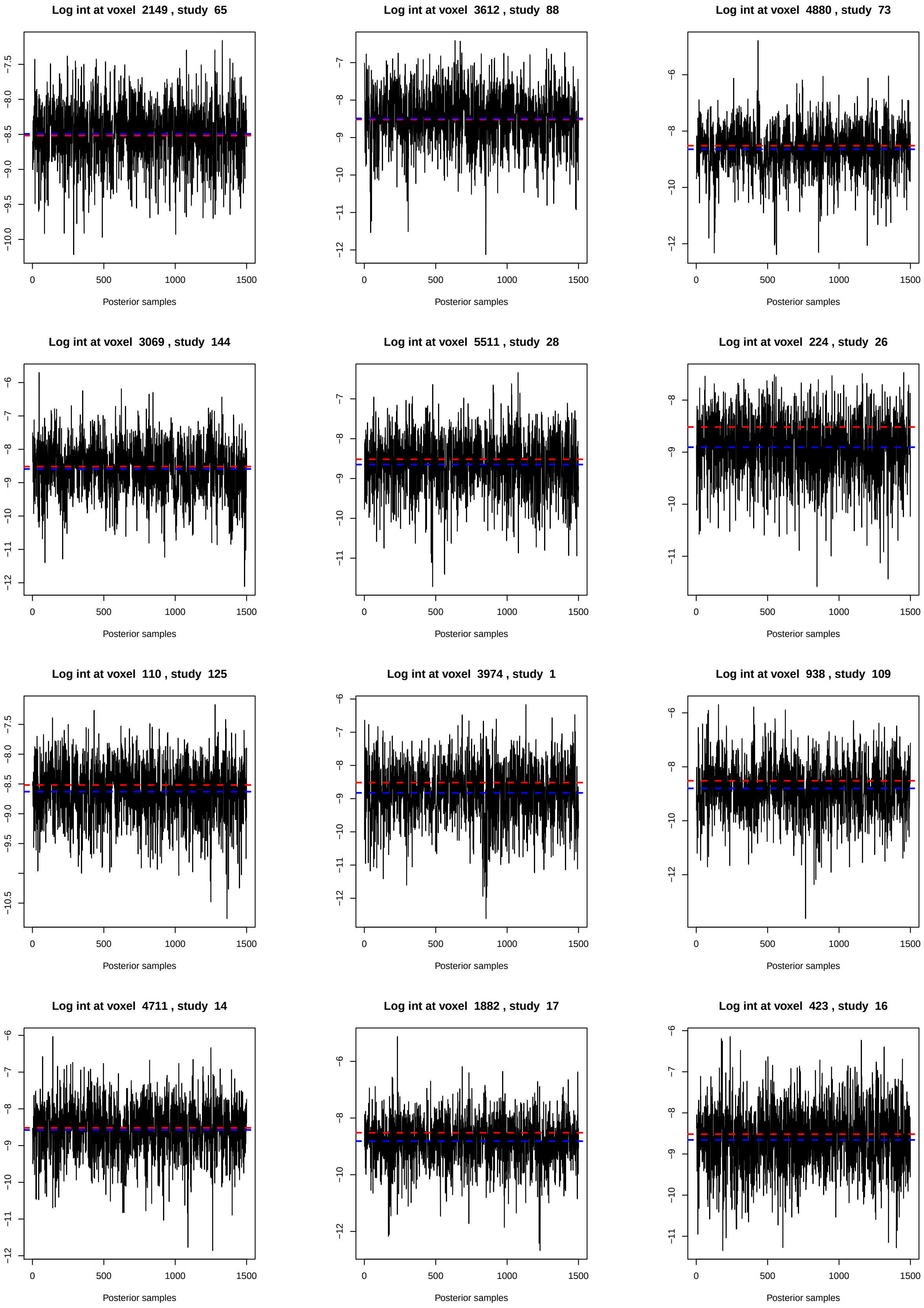
[10] Gareth O. Roberts and Jeffrey S. Rosenthal. Coupling and ergodicity of adaptive Markov Chain Monte Carlo algorithms. *Journal of Applied Probability*, 44(March 2005):458–475, 2007.

[11] C. Rottschy, R. Langner, I. Dogan, K. Reetz, a. R. Laird, J. B. Schulz, P. T. Fox, and S. B. Eickhoff. Modelling neural correlates of working memory: a coordinate-based meta-analysis. *NeuroImage*, 60:830–846, 2012.

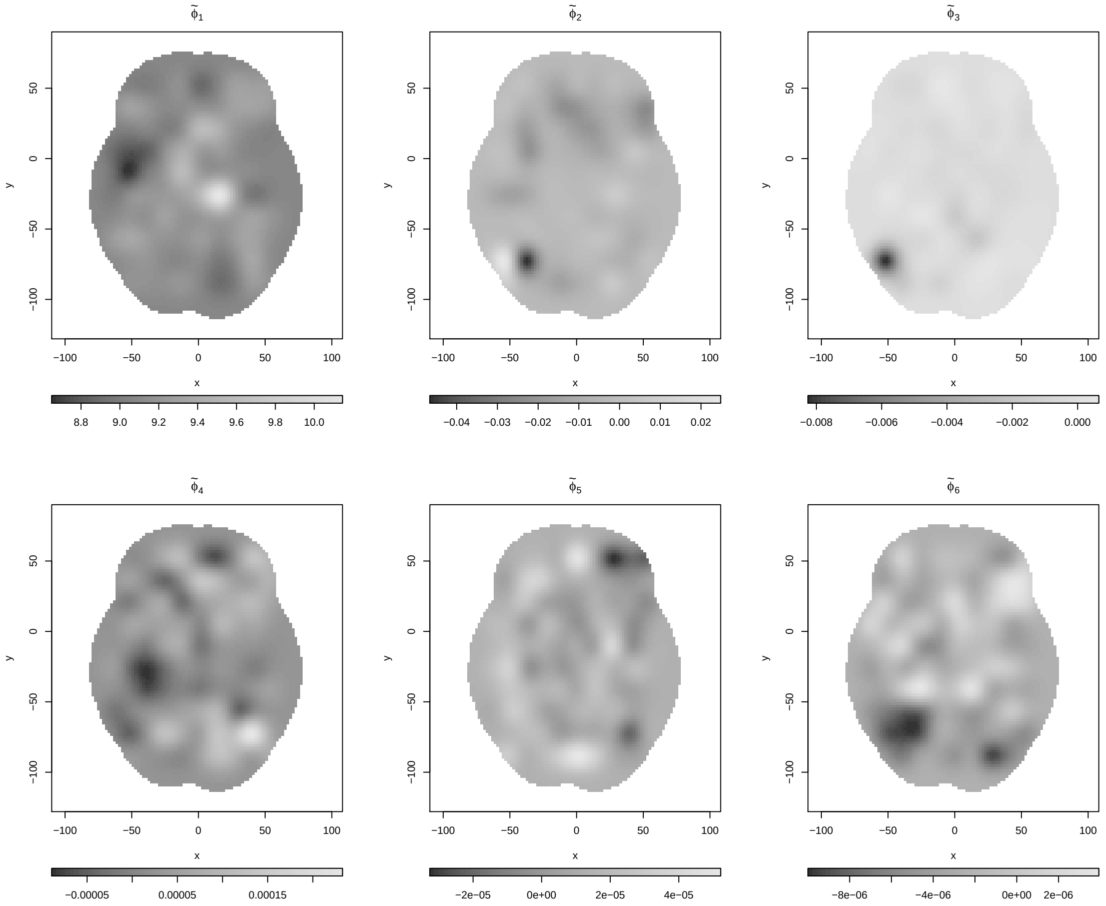
[12] T. D. Wager, J. Jonides, and S. Reading. Neuroimaging studies of shifting attention: a meta-analysis. *NeuroImage*, 22:1679–1693, 2004.

[13] T. D. Wager and E. E. Smith. Neuroimaging studies of working memory. *Cognitive, Affective, and Behavioral Neuroscience*, 3(4):255274, 2003.

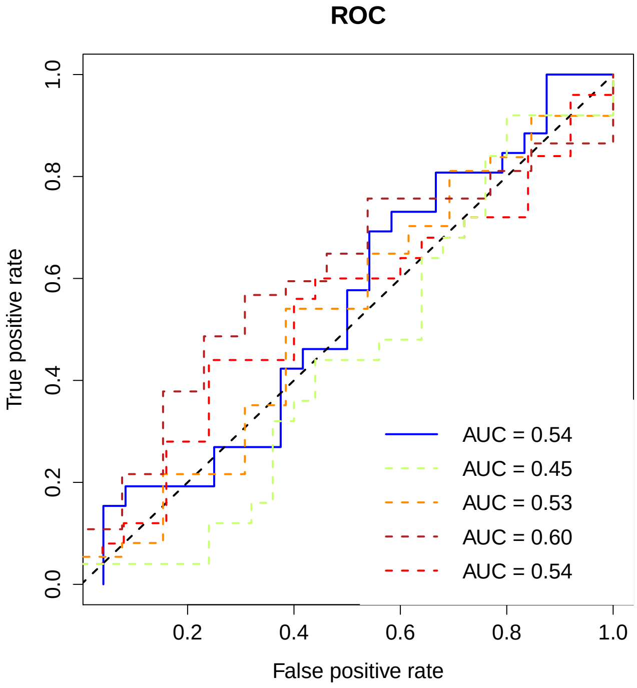
[14] Tal Yarkoni, Russell a Poldrack, Thomas E Nichols, David C Van Essen, and Tor D Wager. Large-scale automated synthesis of human functional neuroimaging data. *Nature methods*, 8(8):665–70, August 2011.



Web Figure 11: Traceplots of the estimated posterior log intensity function to the truth (red dashed line) for a variety of randomly selected pixels and synthetic studies for the simulation with coordinates generated at random. The blue dashed line represents the posterior mean intensity.



Web Figure 12: The first six learnt dictionary elements $\{\phi_l\}_{l=1}^6$ for the simulation study with coordinates generated at random. The estimated posterior mean number of factors is $k = 7$.



Web Figure 13: ROC curve for prediction of study type for studies in the test set. The different lines refer to five independent replicates of the experiment, where studies in each replicate are randomly assigned to either group 0 or group 1 and have coordinates generated at random. The AUC corresponds to the area under the curve.