Bayesian Analysis (2021) **TBA**, Number TBA, pp. 1–28

Spatial 3D Matérn Priors for Fast Whole-Brain fMRI Analysis*

Per Sidén[†,**], Finn Lindgren[‡], David Bolin[§], Anders Eklund[†,¶], and Mattias Villani[†,‖]

Abstract. Bayesian whole-brain functional magnetic resonance imaging (fMRI) analysis with three-dimensional spatial smoothing priors has been shown to produce state-of-the-art activity maps without pre-smoothing the data. The proposed inference algorithms are computationally demanding however, and the spatial priors used have several less appealing properties, such as being improper and having infinite spatial range. We propose a statistical inference framework for whole-brain fMRI analysis based on the class of Matérn covariance functions. The framework uses the Gaussian Markov random field (GMRF) representation of possibly anisotropic spatial Matérn fields via the stochastic partial differential equation (SPDE) approach of Lindgren et al. (2011). This allows for more flexible and interpretable spatial priors, while maintaining the sparsity required for fast inference in the high-dimensional whole-brain setting. We develop an accelerated stochastic gradient descent (SGD) optimization algorithm for empirical Bayes (EB) inference of the spatial hyperparameters. Conditionally on the inferred hyperparameters, we make a fully Bayesian treatment of the brain activity. The Matérn prior is applied to both simulated and experimental task-fMRI data and clearly demonstrates that it is a more reasonable choice than the previously used priors, using comparisons of activity maps, prior simulation and cross-validation.

Keywords: spatial priors, Gaussian Markov random fields, fMRI, spatiotemporal modeling, efficient computation.

1 Introduction

Functional magnetic resonance imaging (fMRI) is a noninvasive technique for making inferences about the location and magnitude of neuronal activity in the living human

*This work was funded by Swedish Research Council (Vetenskapsrådet) grant no 2013-5229 and grant no 2016-04187. Finn Lindgren was funded by the European Union's Horizon 2020 Programme for Research and Innovation, no 640171, EUSTACE. Anders Eklund was funded by Center for Industrial Information Technology (CENIIT) at Linköping University.

[†]Division of Statistics and Machine Learning, Dept. of Computer and Information Science, Linköping University, SE-581 83 Linköping, Sweden, per.siden@liu.se

[‡]School of Mathematics, The University of Edinburgh, James Clerk Maxwell Building, The King's Building, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, United Kingdom, finn.lindgren@ed.ac.uk

[§]CEMSE Division, King Abdullah University of Science and Technology, Saudi Arabia, david.bolin@kaust.edu.sa

[¶]Division of Medical Informatics, Dept. of Biomedical Engineering and Center for Medical Image Science and Visualization (CMIV), Linköping University, SE-581 83 Linköping, Sweden, anders.eklund@liu.se

[‖]Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden, mattias.villani@gmail.com

[**]Corresponding author.

 https://doi.org/10.1214/21-BA1283

brain. The use of fMRI has provided neuroscientists with countless new insights on how the brain operates (Lindquist, 2008). By observing changes in blood oxygenation in a subject during an experiment, a researcher can apply statistical methods such as the general linear model (GLM) (Friston et al., 1995) to draw conclusions regarding task-related brain activations.

One may see fMRI data as a sequence of three-dimensional images collected by a magnetic resonance (MR) scanner over time, where each image can be divided into a large number of voxels. Alternatively, fMRI data can be seen as a collection of time series, one for each voxel, measuring the blood oxygen level dependent (BOLD) signal which is a response to local neural activity (Ogawa et al., 1990). The GLM approach uses a regression model to estimate the linear dependence between the expected BOLD response given the task or condition presented to the scanned subject and the observed BOLD signal in each voxel. Significant regression coefficients determine the activation of a certain voxel for some condition, or, in a Bayesian treatment, the posterior probability of activation of each voxel can be computed and visualised as a posterior probability map (PPM). By designing different tasks, researchers can use this technology to localize different functional regions in the brain, which can for example be used in presurgical planning (Gallen et al., 1994), or to compare the activity patterns between different patients in a group study. A problem with the GLM approach and many of its successors is that the model is mass-univariate, that is, it analyses each voxel independently. This ignores the well-known inherent spatial dependencies in the brain activity between neighboring brain regions, known as functional segregation (Friston and Price, 2011). Instead, this problem is normally addressed in pre- and post-processing as discussed below.

An alternative to the mass-univariate approach is to use Bayesian spatial smoothing priors for the brain activity, and an early example of this is the two-dimensional prior in slice-wise fMRI analysis proposed by Penny et al. (2005). The spatial prior on the activity coefficients reflects the prior knowledge that activated regions are spatially contiguous and locally homogeneous. Penny et al. (2005) use the variational Bayes (VB) approach to approximate the posterior distribution of the activations. Sidén et al. (2017) extend that prior to the 3D case and propose a fast Markov Chain Monte Carlo (MCMC) method and an improved VB approach, that is empirically shown to give negligible error compared to MCMC.

In this paper, we show how the spatial priors used in these previous articles can be seen as special cases of the Gaussian Markov random field (GMRF) representation of Gaussian fields of the Matérn class, using the stochastic partial differential equation (SPDE) approach presented in Lindgren et al. (2011). The Matérn family of covariance functions, attributed to Matérn (1960) and popularized by Handcock and Stein (1993), is seeing increasing use in spatial statistical modeling. It is also a standard choice for Gaussian process (GP) priors in machine learning (Rasmussen and Williams, 2006). In his practical suggestions for prediction of spatial data, Stein (1999) notes that the properties of a spatial field depend strongly on the local behavior of the field and that this behavior is unknown in practice and must be estimated from the data. Moreover, some commonly used covariance functions, for example the Gaussian (also known as the squared exponential), do not provide enough flexibility with regard to this local behavior and Stein summarizes his suggestions with "*Use the Matérn model*". Using the

Matern prior on large-scale 3D data such as fMRI data is computationally challenging, however, in particular with MCMC. We present a fast Bayesian inference framework to make Stein's appeal feasible in practical work.

For fMRI analysis, standard practice has traditionally relied on the Gaussian covariance function rather than the Matérn. The data are pre-smoothed with a Gaussian kernel, with reference to the matched filter theorem (Rosenfield and Kak, 1982), which suggests that the signal-to-noise ratio can be improved by smoothing the data with the same frequency structure as that of the signal. Furthermore, the standard post-correction of multiple hypothesis testing use random field theory (RFT) with an assumed Gaussian covariance function, an assumption partly motivated by the pre-smoothing. However, this approach was shown to lead to spurious results by Eklund et al. (2016), who mention the Gaussian covariance assumption as a principal cause of the invalid results, and demonstrate that the empirical spatial auto-correlation functions of raw fMRI data seem more fat-tailed than a Gaussian, see also Cox et al. (2017). Even though this use of the Gaussian covariance function is different from for example using it in a spatial GP prior for the brain activity, this raises a doubt of its suitability for fMRI data.

A problem with GPs for most commonly used covariance functions, including the Gaussian, is computational. The standard GP formulation results in a dense covariance matrix which becomes too computationally expensive to invert even with only a few thousand voxels. For example, (Groves et al., 2009) use a spatial GP prior with Gaussian covariance and do the analysis slice-wise, due to the computational cost. For this reason, much work on spatial modeling of fMRI data has been using GMRFs instead, see for example Gössl et al. (2001); Woolrich et al. (2004); Penny et al. (2005); Harrison and Green (2010); Sidén et al. (2017). GMRFs have the property of having sparse precision matrices, which make them computationally very fast to use, but do not always correspond to simple covariance functions, especially the intrinsic GMRFs often used as priors, whose precision matrices are not invertible (Rue and Held, 2005).

A different branch of Bayesian spatial models for fMRI has considered selecting active voxels as a variable selection problem, modeling the spatial dependence between the activity indicators rather than between the activity coefficients (see, among others Smith and Fahrmeir, 2007; Vincent et al., 2010; Lee et al., 2014; Zhang et al., 2014; Bezener et al., 2018). These articles mostly use Ising priors for the indicator dependence, which also gives sparsity. However, these priors are rarely defined over the whole brain, but are applied independently to parcels or slices, probably due to computational costs.

The SPDE approach of Lindgren et al. (2011) has been applied to fMRI data using either a slice-wise approach (Yue et al., 2014) or on the sphere after transforming the volumetric data to the cortical surface (Mejia et al., 2020). In both cases integrated nested Laplace approximations (INLA) (Rue et al., 2009) were used for approximating the posterior, which is efficient but presently cumbersome to apply directly to volumetric fMRI data, as the `R-INLA` R-package currently lacks support for three-dimensional data.

Our paper makes a number of contributions. First, we develop a fast Bayesian inference algorithm that allows us to use spatial three-dimensional whole-brain priors of the Matérn class on the activity coefficients, for which previous MCMC and VB approaches are not computationally feasible. The algorithm applies empirical Bayes (EB)

to optimize the hyperparameters of the spatial prior and the parameters of the autoregressive noise model, using an accelerated version of stochastic gradient descent (SGD). The link to the Matérn covariance function gives the spatial hyperparameters a clear interpretation as the range and marginal variance of the corresponding Gaussian field. Given the maximum a posteriori (MAP) values of the optimized hyperparameters, we make a fully Bayesian treatment of the activity coefficients, and compute brain activity PPMs. The convergence of the optimization algorithm is established and the resulting EB posterior is shown to be extremely similar to the exact MCMC posterior for the prior used in Sidén et al. (2017). Second, we develop an anisotropic version of the Matérn 3D prior. The anisotropic prior allows the spatial dependence to vary in the x-, y- and z-direction, and we propose a parameterization such that the new parameters do not the affect the marginal variance of the field. Third, we apply the proposed Matérn priors to both simulated and real experimental fMRI datasets, and compare with the prior used in Sidén et al. (2017) by observing differences in the PPMs, by examining the plausibility of new random samples of the different spatial priors, and by comparing predictive performance, both in terms of point predictions and predictive uncertainty. Collectively, our demonstration strongly suggests that the higher level of smoothness is more reasonable for fMRI data, and also indicates that the second order Matérn prior (see the definition in Section 2.2) is more sensible than its intrinsic counterpart.

The methods in this article are developed for analyzing fMRI data, but can also be applied to other fields with large-scale image data with spatial dependencies, such as diffusion tensor imaging (Gu et al., 2017), microscopy (Barman and Bolin, 2018) or satellite data (Heaton et al., 2019).

The article is organized as follows. Section 2 reviews the model of Penny et al. (2005) and introduces the proposed extension to Matérn priors spatial priors and associated hyperpriors. In Section 3, we derive the optimization algorithm for the EB method, and describe the PPM computation. Experimental and simulation results are shown in Section 4. Section 5 contains conclusions and recommendations for future work. The Supplementary Material to the article (Sidén et al., 2021) provides the derivation of the gradient and approximate Hessian used in the SGD optimization algorithm, and gives the details of the cross-validation (CV) framework used to assess predictive performance.

The new methods in this article have been implemented and added to the BFAST3D extension to the SPM software, available at `http://www.fil.ion.ucl.ac.uk/spm/ext/#BFAST3D`.

2 Model and priors

Our model for fMRI data can be divided into three parts: (i) the measurement model, which consists of a regression model that relates the observed blood oxygen level dependent (BOLD) signal in each voxel to the experimental paradigm and nuisance regressors, and a temporal noise model (Section 2.1), (ii) the spatial prior that models the dependence of the regression parameters between voxels (Sections 2.2 and 2.3), and (iii) the priors on the spatial hyperparameters and noise model parameters (Sections 2.4 and 2.5).

2.1 Measurement model

The single-subject fMRI-data is collected in a $T \times N$ matrix $\mathbf{Y}$, with T denoting the number of volumes collected over time and N the number of voxels. The experimental paradigm is represented by the $T \times K$ design matrix $\mathbf{X}$, with K regressors representing the expected BOLD response for different conditions computed as the hemodynamic response function (HRF) convolved with the binary time series of task events, or for example nuisance regressors to control for head motion artifacts. The model can be written as $\mathbf{Y} = \mathbf{XW} + \mathbf{E}$, where $\mathbf{W}$ is a $K \times N$ matrix of regression coefficients and $\mathbf{E}$ is a $T \times N$ matrix of error terms. We will also work with the equivalent vectorized formulation $\mathbf{y} = \bar{\mathbf{X}}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{y} = \text{vec}(\mathbf{Y}^T)$, $\bar{\mathbf{X}} = \mathbf{X} \otimes \mathbf{I}_N$, $\boldsymbol{\beta} = \text{vec}(\mathbf{W}^T)$ and $\mathbf{e} = \text{vec}(\mathbf{E}^T)$. The error terms are modeled as Gaussian and independent across voxels, possibly following voxel-specific Pth order AR (autoregressive) models, described by the $N \times 1$ vector $\boldsymbol{\lambda}$ of noise precisions and the $P \times N$ matrix $\mathbf{A}$ of AR parameters. For the ease of presentation we will in what follows only consider the special case $P = 0$, that is, error terms that are independent across both time and voxels, and treat the more general case in the supplementary material.

We can divide our parameters into three groups: $\boldsymbol{\beta}$, $\boldsymbol{\theta}_n$ and $\boldsymbol{\theta}_s$. Here, $\boldsymbol{\beta}$ describes the brain activity coefficients which we are mainly interested in, $\boldsymbol{\theta}_n = \{\boldsymbol{\lambda}, \mathbf{A}\}$ are parameters of the noise model, and $\boldsymbol{\theta}_s$ are spatial hyperparameters that will be introduced in the next subsection.

2.2 Spatial prior on activations

We assume spatial, three-dimensional GMRF priors (Rue and Held, 2005; Sidén et al., 2017) for the regression coefficients, which are independent across regressors, that is, we assume $\boldsymbol{\beta}|\boldsymbol{\theta}_s \sim \mathcal{N}\left(\mathbf{0}, \mathbf{Q}^{-1}\right)$. Here $\mathbf{Q}$ is a $KN \times KN$ block diagonal matrix with the $N \times N$ matrix $\mathbf{Q}_k$ as the kth block. The vector $\boldsymbol{\theta}_s = \{\boldsymbol{\theta}_{s,1}, \ldots, \boldsymbol{\theta}_{s,K}\}$ contains the spatial hyperparameters that the different $\mathbf{Q}_k$ depend on. The precision matrices $\mathbf{Q}_k$ may be chosen differently for different k. In this paper, we construct the different $\mathbf{Q}_k$ using the SPDE approach (Lindgren et al., 2011), which allows for sparse GMRF representations of Matérn fields. An overview of the different priors can be seen in Table 1, and are described in more detail below.

Sidén et al. (2017) focus on the unweighted graph Laplacian prior $\mathbf{Q}_k = \tau^2\mathbf{G}$ which we refer to here as the ICAR(1) (first-order intrinsic conditional autoregression) prior. The matrix $\mathbf{G}$ is defined by

$$G_{i,j} = \begin{cases} n_i, & \text{for } i = j, \\ -1, & \text{for } i \sim j, \\ 0, & \text{otherwise,} \end{cases} \tag{2.1}$$

where $i \sim j$ means that i and j are adjacent voxels and n_i is the number of voxels adjacent to voxel i. The ICAR(1) prior can be derived from the local assumption that $x_i - x_j \sim \mathcal{N}(0, \tau^{-2})$, for all unordered pairs of adjacent voxels (i, j), where $\mathbf{x}$ denotes the GMRF (Rue and Held, 2005). Thus, one can see that τ^2 controls how much the

Spatial prior	α	κ	Precision matrix
GS	–	–	$\tau^2\mathbf{I}$
ICAR(1)	1	$=0$	$\tau^2\mathbf{G}$
M(1)	1	>0	$\tau^2\mathbf{K},\ \mathbf{K}=\kappa^2\mathbf{I}+\mathbf{G}$
ICAR(2)	2	$=0$	$\tau^2\mathbf{G}^T\mathbf{G}$
M(2)	2	>0	$\tau^2\mathbf{K}^T\mathbf{K},\ \mathbf{K}=\kappa^2\mathbf{I}+\mathbf{G}$
A-M(2)	2	>0	$\tau^2\mathbf{K}^T\mathbf{K},\ \mathbf{K}=\kappa^2\mathbf{I}+h_x\mathbf{G}_x+h_y\mathbf{G}_y+h_z\mathbf{G}_z$

Table 1: Summary of the spatial priors used and their precision matrices. The global shrinkage (GS) prior is spatially independent, while the intrinsic conditional autoregression (ICAR), Matérn (M) and anisotropic Matérn (A-M) can be seen as GMRF representations of generalized Matérn fields.

field can vary between neighboring voxels, where large values of τ^2 enforce a field that is spatially smooth. The ICAR(1) prior is default in the SPM software for Bayesian fMRI analysis. The second-order ICAR(2) prior is a more smooth alternative, corresponding to a similar local assumption for the second-order differences, and has been used earlier for fMRI analysis in 2D (Penny et al., 2005). The ICAR priors can be extended by adding κ^2 to the diagonal of $\mathbf{G}$ as in the right hand column of Table 1, and when $\kappa > 0$ we refer to these as M(α) (α-order Matérn) priors. The reason for this is the SPDE link established by Lindgren et al. (2011). For example, the M(2) prior can be seen as the solution $\mathbf{u}$ to

$$\tau\left(\kappa^2\mathbf{I}+\mathbf{G}\right)\mathbf{u}\sim\mathcal{N}\left(\mathbf{0},\mathbf{I}\right), \tag{2.2}$$

which can in turn be seen as a numerical finite difference approximation to the SPDE

$$\left(\kappa^2-\Delta\right)^{\alpha/2}\tau u\left(\mathbf{s}\right)=\mathcal{W}\left(\mathbf{s}\right), \tag{2.3}$$

when $\alpha = 2$. Here $\mathbf{s}$ denotes a point in space, α is a smoothness parameter, Δ is the Laplace operator, and $\mathcal{W}(\mathbf{s})$ is spatial white noise. Define also the smoothness parameter $\nu = \alpha - d/2$, where d is the dimension of the domain. For $\nu > 0$ and $\kappa > 0$, it can be shown that a Gaussian field $u(\mathbf{s})$ is a solution to the SPDE in (2.3), when it has the Matérn covariance function (Whittle, 1954, 1963)

$$C(\delta)=\frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}\left(\kappa\delta\right)^{\nu}K_{\nu}\left(\kappa\delta\right), \tag{2.4}$$

where δ is the Euclidean distance between two points in $\mathbb{R}^d$, K_ν is the modified Bessel function of the second kind and

$$\sigma^2=\frac{\Gamma\left(\nu\right)}{\Gamma\left(\nu+d/2\right)\left(4\pi\right)^{d/2}\tau^2\kappa^{2\nu}} \tag{2.5}$$

is the marginal variance of the field $u(\mathbf{s})$. As $d = 3$ in our case, for $\alpha = 2$ we have $\nu = 1/2$ which is a special case where the Matérn covariance function is the same as the

exponential covariance function. In this paper we also consider the SPDE when $\kappa = 0$ or $\nu = -1/2$, in which case the solutions no longer have Matérn covariance, but are still well-defined random measures, and we will refer to them as generalized Matérn fields. We also define $\mathbf{K} = \kappa^2\mathbf{I}+\mathbf{G}$, in which case the solution to (2.2) is $\mathbf{u} \sim \mathcal{N}\left(\mathbf{0}, (\tau^2\mathbf{K}\mathbf{K})^{-1}\right)$, which is largely the same as the solution obtained in Lindgren et al. (2011) using the finite-element method when the triangle basis points are placed at the voxel locations, apart for some minor differences at the boundary.

A benefit of the Matérn model is that its properties can be interpreted from the parameters σ^2, ρ and ν, which has a simple one-to-one relation to the hyperparameters κ^2, τ^2 and α, at least when $\kappa > 0$ and $\nu > 0$. In addition to the marginal variance of the process σ^2, defined in (2.5), we define the range $\rho = \sqrt{8\nu}/\kappa$, which is the distance for which two points in the field have correlation near to 0.13. This reveals an important interpretation of the ICAR(2) prior, since this can be seen as a special case of M(2) with $\kappa = 0 \Leftrightarrow \rho = \infty$, that is, infinite range. Increasing the smoothness parameter ν leads to realizations of the process that appear as more smooth. This can be understood by the property that a random realization of the process is n times differentiable if and only if $n < \nu$, see e.g. Sidén (2020, Figure 2.3) for an illustration.

For values of $\alpha \neq 2$, similar simple discrete solutions of the SPDE are also available. In particular, for $\alpha = 1$ we have $\mathbf{u} \sim \mathcal{N}\left(\mathbf{0}, (\tau^2\mathbf{K})^{-1}\right)$. Extensions to higher integer values of α such as $\alpha = 3, 4, \ldots$ are straightforward in theory (Lindgren et al., 2011), but will result in less sparse precision matrices $\mathbf{Q}_k$ and thereby longer computing times, and more involved gradient expressions for the parameter optimization in Section 3.1.

For each choice of $\mathbf{Q}_k$, we have spatial hyperparameters $\boldsymbol{\theta}_{s,k} = \left\{\tau_k^2, \kappa_k^2\right\}$, which will normally be estimated from data. For regressors not related to the brain activity, that is, head motion regressors and voxel intercepts, we do not use a spatial prior, but instead a global shrinkage (GS) prior with precision matrix $\mathbf{Q}_k = \tau_k^2\mathbf{I}$. We could here infer τ_k^2 from the data, but will normally fix it to some small value, for example $\tau_k^2 = 10^{-12}$, which gives a non-informative prior that provides some numerical stability.

2.3 Anisotropic spatial prior

The SPDE approach makes it possible to fairly easily construct anisotropic priors, for example using a SPDE of the form

$$\left(\kappa^2 - h_x\frac{\partial^2}{\partial x^2} - h_y\frac{\partial^2}{\partial y^2} - h_z\frac{\partial^2}{\partial z^2}\right)^{\alpha/2} \tau u\left(\mathbf{s}\right) = \mathcal{W}\left(\mathbf{s}\right), \tag{2.6}$$

with h_z defined as $h_z = \frac{1}{h_x h_y}$ for identifiability. For $\alpha = 2$, this SPDE has a finite-difference solution with precision matrix $\tau^2\mathbf{K}\mathbf{K}$, with $\mathbf{K}$ now defined as $\mathbf{K} = h_x\mathbf{G}_x + h_y\mathbf{G}_y + h_z\mathbf{G}_z + \kappa^2\mathbf{I}$. Here, $\mathbf{G}_x$ is defined as in (2.1), after redefining the neighbors as being only the adjacent voxels in the x direction. $\mathbf{G}_y$ and $\mathbf{G}_z$ are defined correspondingly, so that $\mathbf{G} = \mathbf{G}_x + \mathbf{G}_y + \mathbf{G}_z$. When using this prior for regressor k we have four parameters, $\boldsymbol{\theta}_{s,k} = \left\{\tau_k^2, \kappa_k^2, h_{x,k}, h_{y,k}\right\}$. The new parameters h_x and h_y allows for different relative length scales of the spatial dependence in the x-, y- and z-direction. This is useful

considering that fMRI data often do not have voxels of equal size in all dimensions and the data collection is normally not symmetric with respect to the three axes. Also, in case the spatial dependence in the underlying activity pattern is different in the different spatial dimensions, this can automatically be inferred from the data, by learning the values of h_x and h_y. Conveniently, $h_x = h_y = 1$ gives the standard isotropic Matérn field defined earlier.

Proposition 2.1. *For $\alpha > d/2$, the anisotropic field u defined in (2.6) on $\mathbb{R}^d$ has the marginal variance defined in (2.5), and the variance thus does not depend on h_x and h_y. Furthermore, $Cov(u(\mathbf{s}), u(\mathbf{t})) = C(\sqrt{(\mathbf{s}-\mathbf{t})^T\mathbf{H}^{-1}(\mathbf{s}-\mathbf{t})})$, where $\mathbf{H}$ is a diagonal matrix with diagonal $(h_x, h_y, 1/(h_x h_y))^T$, and $C(\delta)$ is the isotropic Matérn covariance function defined in (2.4) with $\nu = \alpha - d/2$.*

Proof. We show the covariance formula first, and then the statement about the marginal variance follows as $Cov(u(\mathbf{s}), u(\mathbf{s})) = C(\sqrt{\mathbf{0}^T\mathbf{H}^{-1}\mathbf{0}}) = C(0)$. By using a certain definition of the Fourier transform, the spectral density of u in the anisotropic SPDE in (2.6) is

$$S(\boldsymbol{\omega}) = \frac{1}{(2\pi)^d} \frac{1}{\tau^2 (\kappa^2 + \boldsymbol{\omega}^T\mathbf{H}\boldsymbol{\omega})^\alpha}, \tag{2.7}$$

so the covariance function can be written as

$$Cov(u(\mathbf{s}), u(\mathbf{t})) = \int_{\mathbb{R}^d} \frac{1}{(2\pi)^d} \frac{1}{\tau^2 (\kappa^2 + \boldsymbol{\omega}^T\mathbf{H}\boldsymbol{\omega})^\alpha} e^{-i\boldsymbol{\omega}^T(\mathbf{s}-\mathbf{t})} d\boldsymbol{\omega}. \tag{2.8}$$

An isotropic field v can be written as an anisotropic field with $\mathbf{H} = \mathbf{I}$, so its covariance function for $\delta = \|\mathbf{s}-\mathbf{t}\|_2$ is

$$Cov(v(\mathbf{s}), v(\mathbf{t})) = \int_{\mathbb{R}^d} \frac{1}{(2\pi)^d} \frac{1}{\tau^2 (\kappa^2 + \boldsymbol{\omega}^T\boldsymbol{\omega})^\alpha} e^{-i\boldsymbol{\omega}^T(\mathbf{s}-\mathbf{t})} d\boldsymbol{\omega}. \tag{2.9}$$

On the other hand,

$$\begin{aligned} Cov\left(v(\mathbf{H}^{-1/2}\mathbf{s}), v(\mathbf{H}^{-1/2}\mathbf{t})\right) &= \int_{\mathbb{R}^d} \frac{1}{(2\pi)^d} \frac{1}{\tau^2 (\kappa^2 + \boldsymbol{\omega}^T\boldsymbol{\omega})^\alpha} e^{-i\boldsymbol{\omega}^T(\mathbf{H}^{-1/2}\mathbf{s} - \mathbf{H}^{-1/2}\mathbf{t})} d\boldsymbol{\omega} \\ &= \int_{\mathbb{R}^d} \frac{1}{(2\pi)^d} \frac{1}{\tau^2 (\kappa^2 + \mathbf{z}^T\mathbf{H}\mathbf{z})^\alpha} e^{-i\mathbf{z}^T(\mathbf{s}-\mathbf{t})} \det\left(\mathbf{H}^{1/2}\right) d\mathbf{z}, \end{aligned} \tag{2.10}$$

where the last step used the variable substitution $\boldsymbol{\omega} = \mathbf{H}^{1/2}\mathbf{z}$. Since $\det\left(\mathbf{H}^{1/2}\right) = \sqrt{h_x \cdot h_y \cdot 1/(h_x h_y)} = 1$, the last expression equals that in (2.8). So

$$Cov(u(\mathbf{s}), u(\mathbf{t})) = Cov\left(v(\mathbf{H}^{-1/2}\mathbf{s}), v(\mathbf{H}^{-1/2}\mathbf{t})\right) = C\left(\sqrt{(\mathbf{s}-\mathbf{t})^T\mathbf{H}^{-1}(\mathbf{s}-\mathbf{t})}\right), \tag{2.11}$$

using that $\sqrt{(\mathbf{s}-\mathbf{t})^T\mathbf{H}^{-1}(\mathbf{s}-\mathbf{t})} = \left\|\mathbf{H}^{-1/2}\mathbf{s} - \mathbf{H}^{-1/2}\mathbf{t}\right\|_2$. □

Proposition 2.1 implies that changing h_x or h_y does not affect the marginal variance of the field. This is convenient because it means that the anisotropic parameterization does not change the interpretation of τ^2 and κ^2, apart from that $\rho = \sqrt{8\nu}/\kappa$ will now be the (in some sense) average range in the x-, y- and z-direction. Thus, we can use the same priors for τ^2 and κ^2 as in the isotropic case. By putting log-normal priors on h_x and h_y, as explained in the next subsection, we get priors that are symmetric with respect to the x-, y- and z-direction.

2.4 Hyperparameter priors

We will now specify priors for the spatial hyperparameters $\boldsymbol{\theta}_s = \{\boldsymbol{\theta}_{s,1}, \ldots, \boldsymbol{\theta}_{s,K}\}$, which we assume to be independent across the different regressors k. For brevity, we drop subindexing with respect to k in what follows.

Penalised complexity (PC) priors (Simpson et al., 2017) provide a framework for specifying weakly informative priors that penalize deviation from a simpler base model. Fuglstad et al. (2019) showed the usefulness of PC priors for the hyperparameters of Matérn Gaussian random fields, where the base model is chosen for κ^2 as the intrinsic field $\kappa^2 = 0$ and the base model for $\tau^2|\kappa^2$ is chosen as the model with zero variance, that is $\tau^2 = \infty$ (note that our definition of τ^2 corresponds to τ^{-1} in Fuglstad et al. (2019)). This means exponential priors for $\kappa^{d/2}$ and for $\tau^{-1}|\kappa^2$. The PC prior for M(2) allows the user to be weakly informative about range and standard deviation of the spatial activation coefficient maps, by a priori controlling the lower tail probability for the range $\Pr(\rho < \rho_0) = \xi_1$ and the upper tail probability for the marginal variance $\Pr\left(\sigma^2 > \sigma_0^2\right) = \xi_2$ of the field. By default, we will set $\xi_1 = \xi_2 = 0.05$, ρ_0 to 2 voxel lengths and σ_0^2 corresponding to 5% probability that the marginal standard deviation of the activity coefficients is larger than 2% of the global mean signal. See the supplementary material for full details about the PC prior for the M(2) hyperparameters.

For M(1), PC priors are not straightforward to specify, since the range and marginal variance are not available for $\nu = -1/2$ in the continuous space, so we will instead use log-normal priors for τ^2 and κ^2, as specified in the supplementary material.

For ICAR(1) and ICAR(2) we use the PC prior for τ^2 for Gaussian random effects in Simpson et al. (2017, Section 3.3), and we follow their suggestion for handling the singular precision matrix. Since these spatial priors do not have a finite marginal variance, we let the PC prior control the marginal variances of $\boldsymbol{\beta}|\mathbf{V}^T\boldsymbol{\beta} = \mathbf{0}$ instead, where $\mathbf{V}$ is the nullspace of the prior precision matrix. These nullspaces are known by construction, and the variances measure deviances beyond the addition of a constant to all voxels for ICAR(1), and beyond the addition of constants and linear trends for ICAR(2). The variances are inversely proportional to τ^2, and we numerically computed them through simulation using a typical brain (from the word object experiment described below) to be $\bar{\sigma}^2 = 0.29/\tau^2$ for ICAR(1) and $\bar{\sigma}^2 = 0.76/\tau^2$ for ICAR(2), on average across all voxels. We specify σ_0^2 and ξ_2 so that $\Pr(\bar{\sigma}^2 > \sigma_0^2) = \xi_2$ and use $\xi_2 = 0.05$ and σ_0 corresponding to 2% of the global mean signal.

For the anisotropic priors we use log-normal priors for h_x and h_y as

$$\begin{bmatrix} \log h_x \\ \log h_y \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \sigma_h^2 \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix}\right), \tag{2.12}$$

which means that also $\log\left(1/\left(h_x h_y\right)\right) \sim \mathcal{N}\left(0, \sigma_h^2\right)$ with correlation $-1/2$ with $\log h_x$ and $\log h_y$. The motivation for this prior is that it is centered at the isotropic model $h_x = h_y = 1$, and it is symmetric with respect to the x-, y- and z-direction. We will use $\sigma_h^2 = 0.01$ as default, which roughly corresponds to a $(0.8, 1.2)$ 95%-interval for h_x.

2.5 Noise model priors

We use priors for the noise model parameters $\boldsymbol{\theta}_n = \{\boldsymbol{\lambda}, \mathbf{A}\}$ that are independent across voxels and across AR parameters within the same voxel, with $\lambda_n \sim \Gamma\left(u_1, u_2\right)$ and $A_{p,n} \sim \mathcal{N}(0, 1/\tau_A^2)$, which is the same prior as in Penny et al. (2005). Normally we use $u_1 = 10$ and $u_2 = 0.1$, which are the default values in the SPM software and $\tau_A^2 = 10^{-3}$ which is the value used in Penny et al. (2005). We have seen that the spatial prior for the AR parameters previously used (Penny et al., 2007; Sidén et al., 2017) gives similar results in practice, which is why we use the computationally more simple independent prior for $\mathbf{A}$.

3 Bayesian inference algorithm

The fast MCMC algorithm in Sidén et al. (2017) is not trivially extended to a 3D model with a Matérn prior as the updating step for κ_k^2 conditional on the other parameters requires the computation of log determinants such as $\log\left|\kappa_k^2 \mathbf{I} + \mathbf{G}\right|$ for various κ_k^2. This in general requires the Cholesky decomposition of $\kappa_k^2 \mathbf{I} + \mathbf{G}$ which has overwhelming memory and time requirements for large N and would normally not be feasible for whole-brain analysis. In addition, κ_k^2 would require some proposal density for a Metropolis-within-Gibbs-step, as a conjugate prior is not available. The same problems apply to the MCMC steps for $h_{x,k}$ and $h_{y,k}$ when using the anisotropic model. The lack of conjugate priors also makes the spatial VB (SVB) method in Sidén et al. (2017) more complicated, as the mean-field VB approximate marginal posterior of κ_k^2 will no longer have a simple closed form.

We instead take an EB approach and optimize the spatial and noise model parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_s, \boldsymbol{\theta}_n\}$, for which we are not directly interested in the uncertainty, with respect to the log marginal posterior $p\left(\boldsymbol{\theta}|\mathbf{y}\right)$. Conditional on the posterior mode estimates of $\boldsymbol{\theta}$, we then sample from the joint posterior of the parameters of interest, the activation coefficients in β, from which we construct PPMs of activations. Optimizing $\boldsymbol{\theta}$ is computationally attractive as we can use fast stochastic gradient methods (see Section 3.1) tailored specifically for our problem. We also note that VB tends to underestimate the posterior variance of the hyperparameters (Bishop, 2006; Rue et al., 2009; Sidén et al., 2017). The approximate posterior for $\boldsymbol{\beta}$ in Sidén et al. (2017) only depends on the posterior mean of the hyperparameters, still it gives very small error compared to

MCMC. Thus, if EB is seen as approximating the distribution of each hyperparameter in $\boldsymbol{\theta}$ as a point mass, it might not be much of a restriction compared to VB.

The marginal posterior of $\boldsymbol{\theta}$ can be computed by

$$p(\boldsymbol{\theta}|\mathbf{y}) = \left.\frac{p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\theta})\,p(\boldsymbol{\beta}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{p(\boldsymbol{\beta}|\mathbf{y},\boldsymbol{\theta})\,p(\mathbf{y})}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}, \tag{3.1}$$

for arbitrary value of $\boldsymbol{\beta}^*$, where all involved distributions are known in closed form, apart from $p(\mathbf{y})$, but this disappears when taking the derivative of $\log p(\boldsymbol{\theta}|\mathbf{y})$ with respect to θ_i. In Section 3.1, we comprehensibly present the optimization algorithm, but leave the finer details to the supplementary material. Given the optimal value $\hat{\boldsymbol{\theta}}$, we will study the full joint posterior $\boldsymbol{\beta}|\mathbf{y},\hat{\boldsymbol{\theta}}$ of activity coefficients, which is normally the main interest for task-fMRI analysis. This distribution is a GMRF with mean $\tilde{\boldsymbol{\mu}}$ and precision matrix $\tilde{\mathbf{Q}}$, see details in the supplementary material, and can be used for example to compute PPMs, as described in Section 3.2.

3.1 Parameter optimization

By using the EB approach with SGD optimization, we avoid the costly log determinant computations needed for MCMC, since the computation of the posterior of $\boldsymbol{\theta}$ is no longer needed. Our algorithm instead uses the gradient of $\log p(\boldsymbol{\theta}|\mathbf{y})$ to optimize $\boldsymbol{\theta}$, for which there is a cheap unbiased estimate. We also use an approximation of the Hessian and other techniques to obtain an accelerated SGD algorithm as described below.

The optimization of $\boldsymbol{\theta}$ will be carried out iteratively. At iteration j each θ_i is updated with some step $\Delta\theta_i$ as $\theta_i^{(j)} = \theta_i^{(j-1)} + \Delta\theta_i^{(j)}$. Let $G(\theta_i^{(j-1)}) = \frac{\partial}{\partial\theta_i}\log p(\boldsymbol{\theta}|\mathbf{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(j-1)}}$ denote the gradient and $H(\theta_i^{(j-1)}) = \frac{\partial^2}{\partial\theta_i^2}\log p(\boldsymbol{\theta}|\mathbf{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(j-1)}}$ denote the Hessian for θ_i (note that we here use the term Hessian to describe a single number for each i, rather than the full Hessian matrix for $\boldsymbol{\theta}$ which would be too large to consider). Ideally, one would use the Newton method with $\Delta\theta_i^{(j)} = -G(\theta_i^{(j-1)})/H(\theta_i^{(j-1)})$, or at least some gradient descent method with $\Delta\theta_i^{(j)} = -\eta G(\theta_i^{(j-1)})$, with some learning rate η. It turns out that for our model, this is not computationally feasible in general, since the gradient depends on various traces on the form $\mathrm{tr}(\tilde{\mathbf{Q}}^{-1}\mathbf{T})$ for some matrix $\mathbf{T}$ with similar sparsity structure as $\tilde{\mathbf{Q}}$. For small problems, such traces can be computed exactly by first computing the selected inverse $\tilde{\mathbf{Q}}^{inv}$ of $\tilde{\mathbf{Q}}$ using the Takahashi equations (Takahashi et al., 1973; Rue and Martino, 2007; Sidén et al., 2017), but this is prohibitive for problems of size larger than, say, $KN > 10^5$. However, the Hutchinson estimator (Hutchinson, 1990) gives a stochastic unbiased estimate of the trace as $\mathrm{tr}(\tilde{\mathbf{Q}}^{-1}\mathbf{T}) \approx \frac{1}{N_s}\sum_{j=1}^{N_s}\mathbf{v}_j^T\tilde{\mathbf{Q}}^{-1}\mathbf{T}\mathbf{v}_j$, where each $\mathbf{v}_j$ is a $N\times 1$ vector with independent random elements 1 or -1 with equal probability. This can be computed without computing $\tilde{\mathbf{Q}}^{-1}$, hence, we can obtain an unbiased estimate of the gradient, which enables SGD. Using a learning rate $\eta^{(j)}$ with the decay properties $\sum_j(\eta^{(j)})^2 < \infty$ and $\sum_j \eta^{(j)} = \infty$ guarantees convergence to a local optimum (Robbins and Monro, 1951; Asmussen and Glynn, 2007).

To speed up the convergence of the spatial hyperparameters $\boldsymbol{\theta}_s$, in addition to SGD, we use an approximation of the Hessian $\tilde{H}(\theta_i^{(j-1)}) = E_{\boldsymbol{\beta}|\mathbf{Y},\boldsymbol{\theta}}\left[\frac{\partial^2\log p(\boldsymbol{\theta},\boldsymbol{\beta}|\mathbf{y})}{\partial\theta_i^2}\right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(j-1)}}$,

Algorithm 1 Parameter optimization algorithm.

Require: Initial values $\boldsymbol{\theta}_0$ and $N_{iter}, \gamma_1, \gamma_2, \eta_{mom}, \left\{\eta^{(j)}\right\}_{j=1}^{N_{iter}}, N_{Polyak}, N_s$

1: **for** $j = 1$ to N_{iter} **do**

2: Estimate the gradient $G(\theta_i^{(j-1)}) = \frac{\partial}{\partial \theta_i} \log p(\boldsymbol{\theta}|\mathbf{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(j-1)}}$ for all i

3: Estimate $\tilde{H}(\theta_i^{(j-1)}) = E_{\boldsymbol{\beta}|\mathbf{Y},\boldsymbol{\theta}}\left[\frac{\partial^2 \log p(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\beta})}{\partial \theta_i^2}\right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(j-1)}}$ for all i

4: Average $\bar{G}(\theta_i^{(j-1)}) = \gamma_1 \bar{G}(\theta_i^{(j-2)}) + (1-\gamma_1) G(\theta_i^{(j-1)})$ for all i

5: Average $\bar{H}(\theta_i^{(j-1)}) = \gamma_2 \bar{H}(\theta_i^{(j-2)}) + (1-\gamma_2) \tilde{H}(\theta_i^{(j-1)})$ for all i

6: Compute $\boldsymbol{\theta}_s$ step sizes $\Delta\theta_{s,i}^{(j)} = \eta_{mom} \Delta\theta_{s,i}^{(j-1)} - \frac{\eta^{(j)}}{\bar{H}(\theta_{s,i}^{(j-1)})} \bar{G}(\theta_{s,i}^{(j-1)})$ for all i

7: Compute $\boldsymbol{\theta}_n$ step sizes $\Delta\theta_{n,i}^{(j)} = \eta_n \eta^{(j)} \bar{G}(\theta_{n,i}^{(j-1)})$ for all i

8: Take step $\theta_i^{(j)} = \theta_i^{(j-1)} + \Delta\theta_i^{(j)}$ for all i

9: **end for**

10: Return $\hat{\boldsymbol{\theta}} = \frac{1}{N_{Polyak}} \sum_{i=N_{iter}-N_{Polyak}+1}^{N_{iter}} \boldsymbol{\theta}^{(j)}$

which improves the step length (Lange, 1995; Bolin et al., 2019). This is also stochastically estimated using Hutchinson estimators of various traces, for example $\text{tr}(\mathbf{K}_k^{-1}\mathbf{K}_k^{-1}) \approx \frac{1}{N_s} \sum_{j=1}^{N_s} \mathbf{v}_j^T \mathbf{K}_k^{-1} \mathbf{K}_k^{-1} \mathbf{v}_j$, where $\mathbf{K}_k^{-1}\mathbf{v}_j$ needs only to be computed once for each j. The final optimization algorithm presented in Algorithm 1 also uses: i) averaging over iterations (line 4-5), which gives robustness to the stochasticity in the estimates, ii) momentum (line 6), which gives acceleration in the relevant direction and dampens oscillations, and iii) Polyak averaging (line 10), which reduces the error in the final estimate of $\boldsymbol{\theta}$ by assuming that the last few iterations are just stochastic deviations from the mode. In practice, all parameters are reparametrized to be defined over the whole real line, see the supplementary material for details.

Some practical details about the optimization algorithm follow. Normally, the maximum number of iterations used is $N_{iter} = 200$ the averaging parameters are $\gamma_1 = 0.2$ and $\gamma_2 = 0.9$, the momentum parameter is $\eta_{mom} = 0.5$, the learning rate decreases as $\eta^{(j)} = \frac{0.9}{0.1 \max(0, j-100)+1}$, the learning rate for $\boldsymbol{\theta}_n$ is $\eta_n = 0.001$, we use $N_{Polyak} = 10$ values for the Polyak averaging, and $N_s = 50$ samples for the Hutchinson estimator. These parameter values led to desirable behavior when monitoring the optimization algorithm on different datasets. We initialize the noise parameters by pre-estimating the model without the spatial prior, and the spatial parameters are normally initialized near to the prior mean. We also start the algorithm by running a 5 iterations of SGD with small learning rate. In each iteration, we also check the sign of the approximate Hessian to prevent steps in the direction opposite to the gradient, which could happen due to the stochasticity or local non-convexity, and change the sign if necessary.

The computational bottleneck of the algorithm is the computation of large matrix solves, such as $\tilde{\mathbf{Q}}^{-1}\mathbf{v}_j$, involving the multiplication of the inverse of large sparse precision matrices with a vector. This is carried out using the fast preconditioned conjugate

gradient (PCG) iterative solvers of the corresponding equation system $\tilde{\mathbf{Q}}\mathbf{u} = \mathbf{v}_j$, as described in Sidén et al. (2017), where it is also illustrated that PCG is numerous times faster than directly solving the equation system using the Cholesky decomposition in these models. In addition, since the Hutchinson estimator requires many matrix solves in each iteration, these can be performed in parallel on separate cores, giving great speedup.

3.2 PPM computation

PPMs are used to summarize the posterior information about active voxels. The marginal PPM is computed for each voxel n and contrast vector $\mathbf{c}$ as $P(\mathbf{c}^T\mathbf{W}_{\cdot,n} > \gamma|\mathbf{y}, \hat{\boldsymbol{\theta}})$, for some activity threshold γ, recalling that $\text{vec}(\mathbf{W}^T) = \boldsymbol{\beta}$ are the activity coefficients. Since $\boldsymbol{\beta}|\mathbf{y}, \hat{\boldsymbol{\theta}} \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\mathbf{Q}}^{-1})$ is a GMRF (see the supplementary material), it is clear that $\mathbf{c}^T\mathbf{W}_{\cdot,n}|\mathbf{y}, \hat{\boldsymbol{\theta}}$ is univariate Gaussian and the PPM would be simple to compute for any $\mathbf{c}$ if we only had access to the mean and covariance matrix of $\mathbf{W}_{\cdot,n}|\mathbf{y}, \hat{\boldsymbol{\theta}}$ for every voxel n. The mean is known, but the covariance matrix is non-trivial to compute, since the posterior is parameterized using the precision matrix. We therefore use the simple Rao-Blackwellized Monte Carlo (simple RBMC) estimate in Sidén et al. (2018) to approximate this covariance matrix using

$$\begin{aligned}\text{Var}\left(\mathbf{W}_{\cdot,n}|\mathbf{y}, \hat{\boldsymbol{\theta}}\right) = &\text{E}_{\mathbf{W}_{\cdot,-n}}\left[\text{Var}\left(\mathbf{W}_{\cdot,n}|\mathbf{W}_{\cdot,-n}, \mathbf{y}, \hat{\boldsymbol{\theta}}\right)\right] + \\ &\text{Var}_{\mathbf{W}_{\cdot,-n}}\left[\text{E}\left(\mathbf{W}_{\cdot,n}|\mathbf{W}_{\cdot,-n}, \mathbf{y}, \hat{\boldsymbol{\theta}}\right)\right],\end{aligned} \tag{3.2}$$

where $-n$ denotes all voxels but n. The first term of the right hand side is cheaply computed as the inverse of a $K \times K$ subblock of $\tilde{\mathbf{Q}}$. The second term is approximated by producing N_{RBMC} samples $\mathbf{W}^{(j)}$ from $\mathbf{W}|\mathbf{y}, \hat{\boldsymbol{\theta}}$, computing $\text{E}(\mathbf{W}_{\cdot,n}|\mathbf{W}^{(j)}_{\cdot,-n}, \mathbf{y}, \hat{\boldsymbol{\theta}})$ analytically for each j, and computing the Monte Carlo approximation of the variance. We leave out the details for brevity, but this computation is straightforward due to the Gaussianity and computationally cheap due to the sparsity structure of $\tilde{\mathbf{Q}}$. The PPM computation time will normally be dominated by the GMRF sampling, which is done using the technique invented in Papandreou and Yuille (2010) and summarized in Sidén et al. (2017, Algorithm 2), and requires solving N_{RBMC} equation systems involving $\tilde{\mathbf{Q}}$ using PCG.

4 Results

This section is divided into three subsections. We start by analysing simulated fMRI data, to demonstrate the EB method's capability to estimate the true parameters, and to visualise the differences between the spatial priors in a controlled setting. We then consider real experimental fMRI data from two different experiments, and compare the results when using different spatial priors by: inspecting the posterior activity maps, examining the plausibility of new random samples from the spatial priors, and evaluating the predictive performance using cross-validation. In the last subsection, we evaluate approximation error of the EB method by comparing to full MCMC. All computations are performed using our own Matlab code which is linked to in the end of Section 1.

	True values				A-M(2) estimates			
Condition	ρ	σ	h_x	h_y	ρ	σ	h_x	h_y
Weak	18	1	1	1	15.0	0.96	1.06	1.01
Short range	9	2	1	1	9.1	1.97	0.96	1.02
Long range	60	2	1	1	48.7	1.88	0.90	1.07
Anisotropic	18	1	0.5	2	16.9	1.05	0.60	1.67

Table 2: Spatial hyperparameters of the anisotropic Matérn model (A-M(2)), used for the four conditions when simulating the data, and estimated values for the same data, computed using the EB method. Spatial range $\rho = 2/\kappa$ (in mm), marginal standard deviation σ (see (2.5)) and anisotropic parameters h_x and h_y.

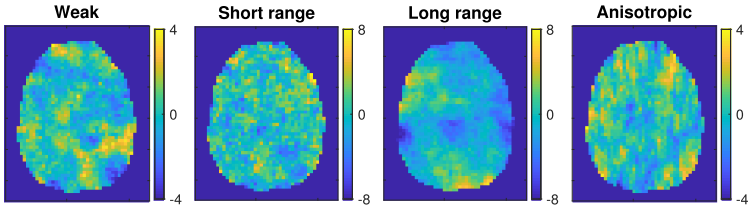


Figure 1: True activity coefficients $\boldsymbol{\beta}$ for the four conditions of the simulated dataset.

4.1 Simulated data

We consider a simulated dataset that is randomly generated using the anisotropic Matérn (A-M(2)) prior with fixed hyperparameters. The size and shape of the brain is taken from the word object dataset, described below. We first simulate four different 3D fields of activity coefficients $\boldsymbol{\beta} = \text{vec}(\mathbf{W}^T)$ using four A-M(2) priors with different hyperparameters. We select the hyperparameters to highlight different spatial characteristics and name the four composed conditions: *Weak* (Small activation magnitude, low σ), *Short range* (Short spatial range ρ), *Long range* (Long spatial range ρ) and *Anisotropic* ($h_x \neq 1$ and $h_y \neq 1$). A summary of the selected hyperparameters can be seen in Table 2, and one slice of the activity coefficient maps are shown in Figure 1.

We then use the simulated $\boldsymbol{\beta}$ coefficients to generate a time series of fMRI volumes. In order to do this, we also borrow the following variables from the word object dataset: the columns of the design matrix $\mathbf{X}$ corresponding to the HRF and intercept, the estimated values for the elements in $\boldsymbol{\beta}$ corresponding to the intercept, and estimated values for the noise variables $\boldsymbol{\lambda}$ and $\mathbf{A}$. The generated dataset has $T = 100$ time points.

We use the EB method to estimate the model with the different spatial priors described in Table 1. The estimated hyperparameters for the A-M(2) can be seen in Table 2. The estimates indicate that the method manages to recover the true parameter values fairly well for the *Short range* condition, when the signal is strong (high σ) and the range is short (small ρ). However, when the signal is weak the estimates are more affected by the noise, and when the range is long there is bias from boundary effects

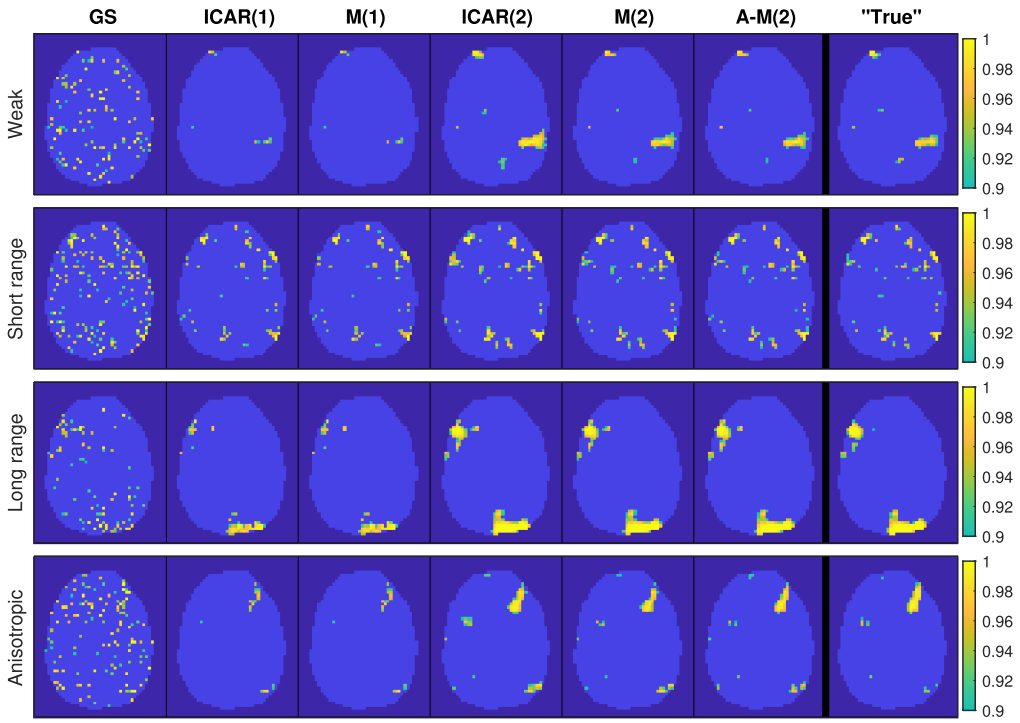


Figure 2: PPMs for the four conditions of the simulated dataset, estimated with different spatial priors. The last column "True" shows the results when the true A-M(2) hyperparameters used to generate the data is used for estimation. The PPMs show probabilities of exceeding 0.2% of the global mean signal, thresholded at 0.9. See the definition of the spatial priors in Table 1. The corresponding posterior means are shown in the supplementary material.

since a long range is harder to infer on a limited domain. The anisotropic parameters h_x and h_y are reasonably well recovered in general, but the anisotropy is somewhat underestimated for the *Anisotropic* condition, due to shrinkage from the prior.

Figure 2 shows the resulting PPMs for the different spatial priors, with hyperparameters estimated by EB, for the same slice as in Figure 1. The last column also shows the "true" PPMs obtained by using the A-M(2) with the hyperparameters used to generate the data. We see how the non-spatial GS prior leads to cluttered PPMs which bear little resemblance with the true activity coefficients. We note that the first-order ICAR(1) and M(1) priors, with smoothness $\alpha = 1$, tend to show smaller activity patterns than the second-order priors with $\alpha = 2$, except for perhaps the *Short range* condition. The differences between the second-order priors ICAR(2), M(2) and A-M(2) are quite subtle, but for the *Weak* and *Anisotropic* conditions ICAR(2) shows some signs of over-smoothing, resulting in slightly larger activity regions compared to the truth. As expected, M(2) and A-M(2) show little discrepancy for the first three isotropic conditions, but for the *Anisotropic* condition A-M(2) is to some degree closer to the truth.

To test the capacity of the spatial priors, we consider an additional simulated dataset, where the activity coefficients $\boldsymbol{\beta}$ were simulated using a spatially independent normal distribution with standard deviation $\sigma \in \{1, 2, 4, 8\}$ for the four different tasks. We observe that the spatial Matérn priors are able to adapt to this situation by for example learning a very short range ρ (shorter than one voxel length) for the M(2) and A-M(2) priors, corresponding to near spatial independence. The PPMs for this simulation study are shown in the supplementary material.

4.2 Experimental data

Description of the data

We evaluate the method on two different real experimental fMRI datasets, the face repetition dataset (Henson et al., 2002) previously examined in Penny et al. (2005); Sidén et al. (2017), and the word object dataset (Duncan et al., 2009). The face repetition dataset is available at SPM's homepage (`http://www.fil.ion.ucl.ac.uk/spm/data/face_rep/`) and the word object dataset is available at OpenNEURO (`https://openneuro.org/datasets/ds000107/versions/00001`) (Poldrack and Gorgolewski, 2017). Both experiments have four conditions or subject tasks. Thus, the design matrix $\mathbf{X}$ for both datasets has $K = 15$ columns, with column $(1, 3, 5, 7)$ corresponding to the standard canonical HRF convolved with the different task paradigms, column $(2, 4, 6, 8)$ corresponding to the HRF derivative, column 9 to 14 corresponding to head motion parameters and the last column corresponding to the intercept.

The face repetition dataset was acquired during an event-related experiment, where greyscale images of non-famous and famous faces were presented to the subject for 500 ms. The four conditions in the dataset correspond to the first and second time a non-famous or famous face was shown. The contrast studied below "mean effect of faces" is the average of the HRF regressors, that is $\mathbf{c}^T\mathbf{W}_{\cdot,n} = (W_{1,n} + W_{3,n} + W_{5,n} + W_{7,n})/4$, and the presented PPMs can therefore be interpreted as showing brain regions involved in face processing. The dataset was preprocessed using the same steps as in Penny et al. (2005) using SPM12 (including motion correction, slice timing correction and normalization to a brain template, but no smoothing), and small isolated clusters with less than 400 voxels were removed from the brain mask. The resulting mask has $N = 57184$ voxels and there are $T = 351$ volumes.

The word object experiment also has conditions that correspond to visual stimuli: written words, pictures of common objects, scrambled pictures of the same objects, and consonant letter strings, which were presented to the subject for 350 ms according to a block-related design. For the word object data, preprocessing consisted only of motion correction and removal of isolated clusters of voxels, as the slice time information was not available. We selected subject 10, which had relatively little head motion, and the resulting brain mask has $N = 41486$ voxels and the number of volumes is $T = 166$.

For both datasets, the voxels are of size $3 \times 3 \times 3$ mm and the global mean signal is computed as the average value across all voxels in the brain mask and all volumes, and the activity threshold γ used in the PPM computation is related to this quantity.

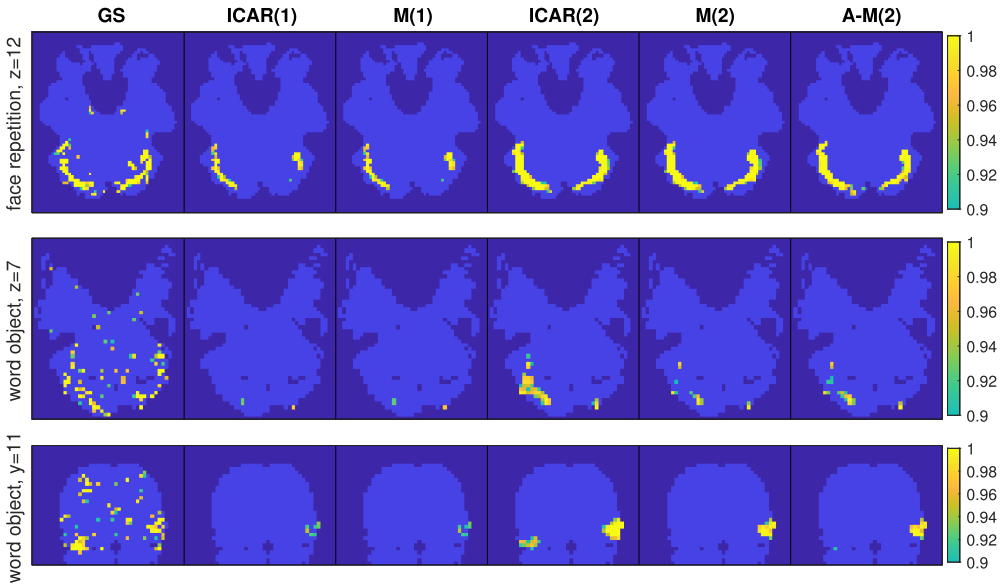


Figure 3: PPMs for the two experimental datasets, when using different spatial priors, thresholded at 0.9. The spatial priors are summarised in Table 1. The top row shows axial slice 12 of the face repetition dataset, and the middle and bottom rows show axial slice 7 and coronal slice 11 of the word object dataset. The face repetition PPMs consider the contrast "mean effect of faces" and show probabilities of effect sizes exceeding 1% of the global mean signal. The word object PPMs consider the first condition "Words" and show probabilities of effect sizes exceeding 0.5% of the global mean signal. The corresponding posterior means and standard deviations are shown in the supplementary material.

Posterior results

We estimate the models with the different spatial priors for the two experimental datasets using the EB method, and present the resulting PPMs in Figure 3. As for the simulated dataset, we observe cluttered PPMs for the non-spatial GS prior, and in general the priors with $\alpha = 1$ (ICAR(1) and M(1)) lead to substantially smaller activity regions compared to the priors with $\alpha = 2$. Given the same α, the differences between the Matérn and ICAR priors do not seem as striking, but for the word object data, the ICAR(2) prior produces an activity region in the left-hand side of the brain that is much smaller for the M(2) and A-M(2) priors.

The use of second order Matérn priors enables simple interpretations of the spatial properties of the inferred activity coefficient fields. We report the estimated hyperparameters when using the A-M(2) prior for the four conditions in respective dataset in Table 3. The results show that the face repetition data activity patterns have longer spatial ranges (higher ρ) and generally larger magnitudes (higher σ), compared to the word object data. The anisotropic parameters indicate stronger dependence in the z-

Face repetition data				
Condition	ρ	σ	h_x	h_y
Non-famous 1	62.9	2.36	0.72	0.75
Non-famous 2	58.7	2.40	0.73	0.73
Famous 1	59.5	2.28	0.70	0.74
Famous 2	47.0	1.98	0.79	0.68

Word object data				
Condition	ρ	σ	h_x	h_y
Words	10.5	1.07	1.21	1.11
Objects	16.0	0.93	1.13	1.08
Scrambled	21.0	1.24	1.18	1.16
Consonant	11.7	2.09	1.14	1.23

Table 3: Estimated spatial hyperparameters for the A-M(2) prior by the EB method, for the different datasets and conditions. Spatial range $\rho = 2/\kappa$ (in mm), marginal standard deviation σ (see (2.5)) and anisotropic parameters h_x and h_y.

direction (between slices) for the face repetition data, while the opposite is true for the word object data.

The observed differences between the datasets could be explained by differences in the studied subjects and tasks, but is likely as well an effect from differences in scanner properties and that the slice timing and normalization preprocessing steps impose some smoothness for the face repetition data. The latter are spatial properties that would preferably be modelled in the noise rather than in the activity patterns, and we view improved, computationally efficient spatial noise models for fMRI data as important future work.

Nevertheless, the ability to flexibly estimate and indicate different spatial properties is indeed a great advantage of the second order Matérn models. Furthermore, these Matérn models correspond to exponential autocorrelation functions, whose fat tails resemble the empirical autocorrelation functions for fMRI data found in Eklund et al. (2016, Supplementary Figure 17) and Cox et al. (2017, Figure 3).

Prior simulation

To better understand the meaning of the different priors in practice, Figure 4 displays samples from the spatial priors using the estimated hyperparameters for the first regressor of the different datasets. The M(1) and ICAR(1) priors produce fields that vary quite rapidly, while the second order priors give realizations that are more smooth. For the word object dataset we note that the short estimated range for M(2) gives a sample with much faster variability than the sample from ICAR(2), which looks unrealistically smooth. This illustrates the problem with using the infinite range ICAR(2) prior for a dataset where the inherent range is much shorter.

Cross-validation

Many studies, including this one, evaluate models for fMRI data by displaying the estimated brain activity maps and deciding whether they look plausible or not. A more scientifically sound approach would be to compare models based on their ability to predict the values of unseen data points, which is the standard procedure in many

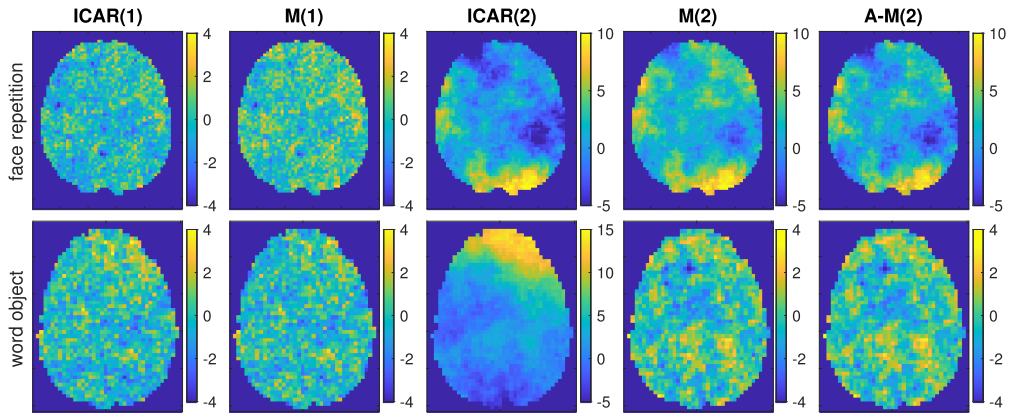


Figure 4: Random samples from the different spatial priors using the estimated hyperparameters for the first regressor of the different datasets. The same seed has been used for the same α and dataset.

other statistical applications. The problem for fMRI data is that the main object of interest, the set of activity coefficients $\mathbf{W}$ corresponding to activity related regressors, is not directly observable, but only indirectly through the observed noisy BOLD signal $\mathbf{Y}$. This makes direct comparison to ground truth activation impossible. We will here attempt to evaluate the performance of the spatial priors for brain activity by measuring the out-of-sample predictive performance by computing various prediction error scores on $\mathbf{Y}$ instead. We cannot, however, expect to find large differences between the different priors, as only a small fraction of the signal is explained by brain activation; most is explained by the intercept and various noise sources.

We compute the out-of-sample fit using CV by repeatedly leaving out 90% of the voxels randomly over the whole brain, and estimating the predictive distribution of the data $\mathbf{Y}$ in the left-out voxels given the data in the remaining 10% of voxels and the original estimates of the hyperparameters $\boldsymbol{\theta}$ based on the whole dataset. The same hyperparameters are used in all repetitions to avoid the computational cost of refitting the model each time, still making the comparison fair across spatial priors. We then compare the estimated predictive distribution and the actual signal $\mathbf{Y}$ in the left-out voxels. In order to focus the comparison on the evaluation of the spatial priors, we must compute the errors in a slightly more cumbersome way than normal, which is explained in the supplementary material, to reduce the impact of the noise model, head motion and intercept regressors.

We use the mean absolute error (MAE) and root mean square error (RMSE) to evaluate the predicted mean of $\mathbf{Y}$ for each prior, and the mean continuous ranked probability score (CRPS), the mean ignorance score (IGN, also known as the logarithmic score) and the mean interval score (INT) to evaluate the whole predictive distribution for $\mathbf{Y}$. All these scores are all examples of proper scoring rules (Gneiting and Raftery, 2007), which encourage the forecaster to be honest and the expected score is maximized

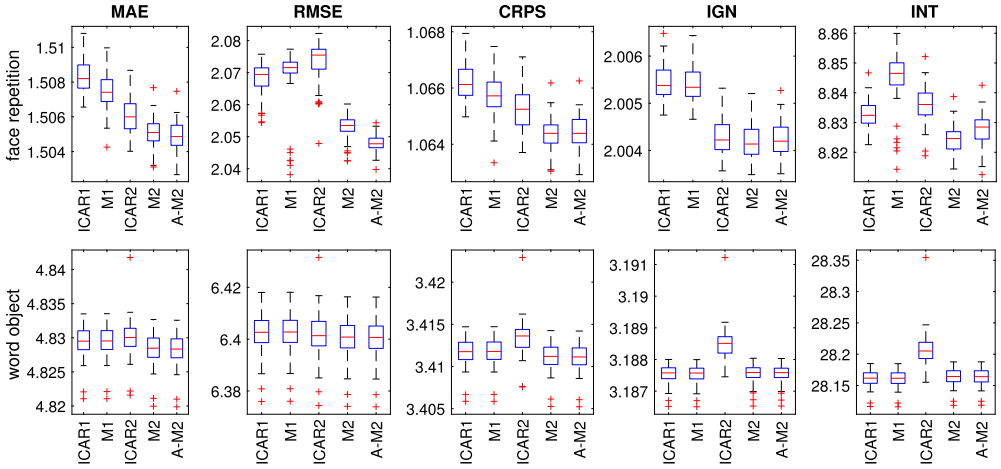


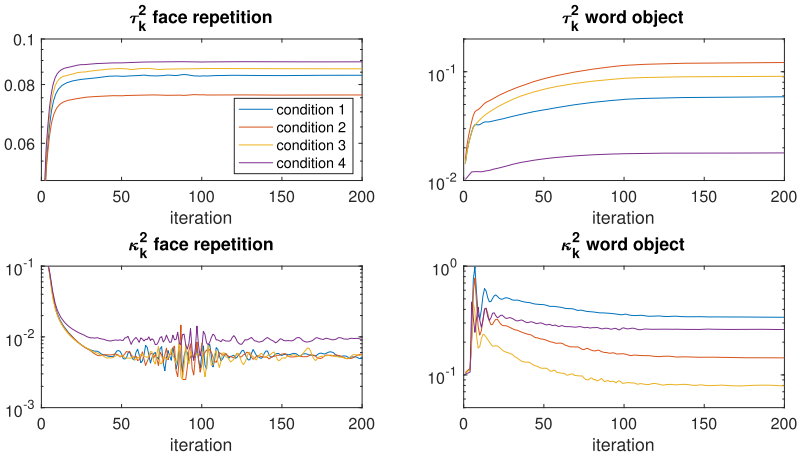
Figure 5: Cross-validation scores computed on 90% left out voxels for the two datasets, comparing the different spatial priors. The scores are computed as means across voxels, and presented in negatively oriented forms, so that smaller values are always better. The boxplots reflect the variation in 50 random sets of left out voxels. Additional results in the supplementary material show these scores also for in-sample fit and 50% left out voxels.

when the predictive distribution equals the generative distribution of the data points. Since the predictive distribution is Gaussian given the hyperparameters, all the scores can be computed using simple formulas, see the supplementary material.

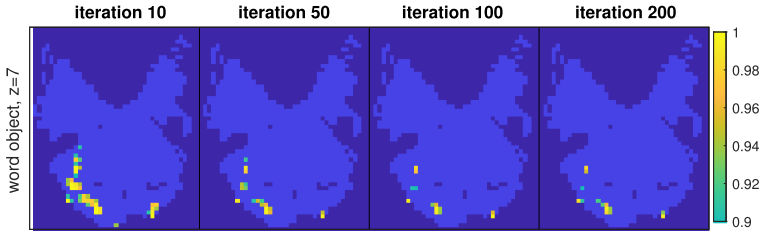
The results can be seen in Figure 5. For the face repetition data, we note that the second order Matérn priors (M(2) and A-M(2)) perform better than the other priors in all cases. For the word object data the differences between different priors are smaller, which can probably be explained by the higher noise level and shorter spatial correlation range in this dataset, but the second order Matérn priors are generally among the best. The absolute differences between the different priors may seem small, but one must remember that most of the error comes from noise that is unrelated to the brain activity, making it hard for a spatial activity prior to substantially reduce the error. The large RMSE for the ICAR(2) prior for the face repetition data indicates that this prior can give relatively large out-of-sample errors, possibly due to over-smoothing.

4.3 Evaluation of the EB method and comparison to MCMC

One of the most challenging aspects with our work has been in the development of the EB method, summarised in Algorithm 1, and in finding optimization parameters that result in stable and fast convergence in the optimization of the spatial hyperparameters. The convergence behaviour for the M(2) prior is depicted in Figure 6. The hyperparameter optimization trajectories in Figure 6a suggest that the parameters reach the right



(a)



(b)

Figure 6: Convergence of the EB method for the M(2) prior. (a) The hyperparameters τ_k^2 and κ_k^2 corresponding to different conditions over the iterations of the Algorithm 1, when using the M(2) prior for the face repetition data (left) and word object data (right). (b) PPM for the word object data after 10, 50, 100 and 200 iterations, where the last is the same as in Figure 3.

GS	ICAR(1)	M(1)	ICAR(2)	M(2)	A-M(2)
0.4	0.9	5.3	1.1	2.0	4.2

Table 4: Computing times (h) for the EB method for different spatial priors.

level in about 100 iterations. The results presented in this paper are all, more conservatively, after 200 iterations of optimization, but future work could include coming up with some automatic convergence criterion, based on the change of some parameters over the iterations. Figure 6b shows how the PPM of the word object data converges. The computing time on a computing cluster, using 16 workers on Intel Xeon Gold 6240 processors at 2.6 GHz, was 2.0h until convergence (100 iterations). The corresponding times for the other spatial priors can be found in Table 4. For comparison, the computing time for MCMC with the ICAR(1) in the analysis below was almost one week.

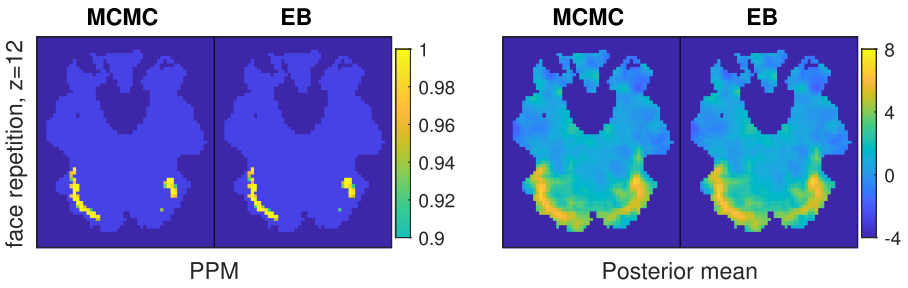


Figure 7: Comparison between MCMC and EB in terms of PPMs (left) and posterior mean of activity coefficients (right) using the ICAR(1) prior for the face repetition data. The presented PPM for EB is the same as in Figure 3.

Due to a maximum time limit on the computing cluster (3 days), the MCMC run was carried out on a different computer with Intel Xeon E5-1620 processors at 3.5 GHz.

To assess how well the EB posterior with optimized hyperparameters approximates the full posterior, we also fit the model with the ICAR(1) prior using MCMC as described in Sidén et al. (2017), using the face repetition data. Figure 7 compares PPMs and posterior mean maps between the two methods, and the differences are practically negligible, and much smaller than, for example, the differences between different spatial priors. The EB estimates for the spatial hyperparameters $\{\tau_k^2\}$ are also very similar to the MCMC posterior mean. For this exercise the same conjugate gamma prior for τ_k^2 as in Sidén et al. (2017) was also for EB. The MCMC method used 10,000 iterations after 1,000 burnin samples and thinning factor 5.

These results support the conjecture made earlier, that the posterior distributions of the hyperparameters $\boldsymbol{\theta}$ are well approximated by point masses when the goal of the analysis is to correctly model the distribution of the activity coefficients $\mathbf{W}$. It would be interesting to do the same comparison for the other spatial priors, and the other hyperparameters (κ^2, h_x and h_y), but to our knowledge there exists no computationally feasible MCMC method to sample these parameters, which lack the conjugacy exploited for τ^2.

5 Conclusions and directions for future research

We propose an efficient Bayesian inference algorithm for whole-brain analysis of fMRI data using the flexible and interpretable Matérn class of spatial priors. We study the empirical properties of the prior on simulated and two experimental fMRI datasets. Based on the experimental data, we conclude that the second order Matérn priors (M(2) or A-M(2)) should be the preferred choice for future studies. The priors with $\alpha = 1$ are clearly inferior in the sense that they do not find the seemingly correct activity patterns that are found by the priors with $\alpha = 2$, they produce new samples that appear too speckled and they perform worse in the cross-validation. The differences between the M(2) and ICAR(2) are less evident, but our paper contains a number of results that are favorable to the M(2) prior: (i) these priors produce somewhat different activity

maps for some datasets while M(2) has better theoretical properties, (ii) new samples from the ICAR(2) look too smooth, (iii) M(2) performed consistently better in the cross-validation, and (iv) the M(2) prior parameters are more easily interpreted.

The introduced anisotropic Matérn prior was shown to perform slightly better than the isotropic Matérn prior in the cross-validation, but overall the differences between the results for the two priors are quite small. Still, A-M(2) has the capacity to model also anisotropic datasets, while containing the M(2) prior as a special case, and could therefore be the best alternative. To get the full potential of the anisotropic model, one should consider non-stationary anisotropic models (Lindgren et al., 2011; Fuglstad et al., 2015), where the anisotropy is allowed to vary locally, capturing different dependence structures in different brain regions.

The optimization algorithm appears satisfactory with relatively fast convergence. Using SGD is an improvement relative to the coordinate descent algorithm employed for SVB in Sidén et al. (2017), because following the gradient is in general the shorter way to reach the optimum and there exists better theoretical guarantees for the convergence. Also, well-known acceleration strategies, such as using momentum or the approximate Hessian information, are easier to adopt to SGD and one can thereby avoid the more ad hoc acceleration strategies used in Sidén et al. (2017, Appendix C).

The EB method is shown to approximate the exact MCMC posterior well empirically, suggesting that properly accounting for the uncertainty in the spatial hyperparameters is of minor importance if the main object is the activity maps.

As the smoothness parameter α appears to be the most important for the resulting activity maps, it would in future research be interesting to estimate it as a non-integer value, which could be addressed using the method in Bolin and Kirchner (2020).

The PPMs reported in this work only contain the marginal probability of activation in each voxel. If instead using joint PPMs (Yue et al., 2014; Mejia et al., 2020) based on excursions sets (Bolin and Lindgren, 2015) to address the multiple comparison problem of classifying active voxels, it is likely to see larger differences between the M(2) and ICAR(2) prior, since the joint PPMs depend more on the spatial correlation. The joint PPMs are easily computed from MCMC output, but harder for the EB method due to the posterior covariance matrix being costly to compute. Future work should address this issue, which could probably be solved by extending the block RBMC method in Sidén et al. (2018).

The estimated spatial hyperparameters for the experimental datasets in Table 3 have strikingly similar values across different HRF regressors. A natural idea is therefore to let these regressors share the same spatial hyperparameters, at least when the tasks in the experiment are similar. Our Bayesian inference algorithm is straightforwardly extended to this setting.

Supplementary Material

Supplementary Material for "Spatial 3D Matérn priors for fast whole-brain fMRI analysis" (DOI: 10.1214/21-BA1283SUPP; .pdf).

References

Asmussen, S. and Glynn, P. W. (2007). Stochastic simulation: algorithms and analysis. In *Stochastic modelling and applied probability*. Springer, New York. MR2331321. 11

Barman, S. and Bolin, D. (2018). A three-dimensional statistical model for imaged microstructures of porous polymer films. *Journal of microscopy*, 269(3):247–258. doi: https://doi.org/10.1111/jmi.12623. 4

Bezener, M., Hughes, J., and Jones, G. (2018). Bayesian spatiotemporal modeling using hierarchical spatial priors with applications to functional magnetic resonance imaging. *Bayesian Analysis*, 13(4):1261–1313. MR3882358. doi: https://doi.org/10.1214/18-BA1108. 3

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer, New York. MR2247587. doi: https://doi.org/10.1007/978-0-387-45528-0. 10

Bolin, D. and Kirchner, K. (2020). The rational SPDE approach for Gaussian random fields with general smoothness. *Journal of Computational and Graphical Statistics*, 29(2):274–285. MR4116041. doi: https://doi.org/10.1080/10618600.2019.1665537. 23

Bolin, D. and Lindgren, F. (2015). Excursion and contour uncertainty regions for latent Gaussian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1):85–106. MR3299400. doi: https://doi.org/10.1111/rssb.12055. 23

Bolin, D., Wallin, J., and Lindgren, F. (2019). Latent Gaussian random field mixture models. *Computational Statistics and Data Analysis*, 130:80–93. MR3860530. doi: https://doi.org/10.1016/j.csda.2018.08.007. 12

Cox, R. W., Chen, G., Glen, D. R., Reynolds, R. C., and Taylor, P. A. (2017). FMRI Clustering in AFNI: False-Positive Rates Redux. *Brain Connectivity*, 7(3):152–171. doi: https://doi.org/10.1089/brain.2016.0475. 3, 18

Duncan, K., Pattamadilok, C., Knierim, I., and Devlin, J. (2009). Consistency and variability in functional localisers. *Neuroimage*, 46(4):1018–1026. doi: https://doi.org/10.1016/j.neuroimage.2009.03.014. 16

Eklund, A., Nichols, T. E., and Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false positive rates. *Proceedings of the National Academy of Sciences*, 113(28):7900–7905. doi: https://doi.org/10.1073/pnas.1602413113. 3, 18

Friston, K. J., Holmes, a. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(4):189–210. 2

Friston, K. J. and Price, C. J. (2011). Modules and brain mapping. *Cognitive neuropsychology*, 28(3-4):241–250. doi: https://doi.org/10.1080/02643294.2011.558835. 2

Fuglstad, G.-A., Lindgren, F., Simpson, D., and Rue, H. (2015). Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy. *Statistica Sinica*, 25(1):115–133. MR3328806. 23

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, 114(525):445–452. MR3941267. doi: https://doi.org/10.1080/01621459.2017.1415907. 9

Gallen, C. C., Bucholz, R., and Sobel, D. F. (1994). Intracranial neurosurgery guided by functional imaging. *Surgical neurology*, 42(6):523–530. doi: https://doi.org/10.1016/0090-3019(94)90083-3. 2

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378. MR2345548. doi: https://doi.org/10.1198/016214506000001437. 19

Gössl, C., Auer, D. P., and Fahrmeir, L. (2001). Bayesian spatiotemporal inference in functional magnetic resonance imaging. *Biometrics*, 57(2):554–562. MR1855691. doi: https://doi.org/10.1111/j.0006-341X.2001.00554.x. 3

Groves, A. R., Chappell, M. A., and Woolrich, M. W. (2009). Combined spatial and non-spatial prior for inference on MRI time-series. *NeuroImage*, 45(3):795–809. doi: https://doi.org/10.1016/j.neuroimage.2008.12.027. 3

Gu, X., Sidén, P., Wegmann, B., Eklund, A., Villani, M., and Knutsson, H. (2017). Bayesian diffusion tensor estimation with spatial priors. *17th international Conference on Computer Analysis of Images and Patterns*. MR3695723. doi: https://doi.org/10.1007/978-3-319-64689-3_30. 4

Handcock, M. S. and Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, 35(4):403–410. doi: https://doi.org/10.1080/00401706.1993.10485354. 2

Harrison, L. M. and Green, G. G. R. (2010). A Bayesian spatiotemporal model for very large data sets. *NeuroImage*, 50(3):1126–1141. doi: https://doi.org/10.1016/j.neuroimage.2009.12.042. 3

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., et al. (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):398–425. MR3996451. doi: https://doi.org/10.1007/s13253-018-00348-w. 4

Henson, R., Shallice, T., Gorno-Tempini, M. L., and Dolan, R. (2002). Face repetition effects in implicit and explicit memory tests as measured by fMRI. *Cerebral Cortex*, 12:178–186. doi: https://doi.org/10.1093/cercor/12.2.178. 16

Hutchinson, M. F. (1990). A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 19(2):433–450. MR1075456. doi: https://doi.org/10.1080/03610919008812864. 11

Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 57(2):425–437. MR1323348. 12

Lee, K.-J., Jones, G. L., Caffo, B. S., and Bassett, S. S. (2014). Spatial Bayesian variable selection models on functional magnetic resonance imaging time-series data. *Bayesian Analysis*, 9(3):699–732. MR3256061. doi: https://doi.org/10.1214/14-BA873. 3

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach. *Journal of the Royal Statistical Society Series B*, 73(4):423–498. MR2853727. doi: https://doi.org/10.1111/j.1467-9868.2011.00777.x. 1, 2, 3, 5, 6, 7, 23

Lindquist, M. (2008). The statistical analysis of fMRI data. *Statistical Science*, 23(4):439–464. MR2530545. doi: https://doi.org/10.1214/09-STS282. 2

Matérn, B. (1960). *Spatial variation*. PhD thesis. MR0169346. 2

Mejia, A. F., Yue, Y. R., Bolin, D., Lindgren, F., and Lindquist, M. A. (2020). A Bayesian general linear modeling approach to cortical surface fMRI data analysis. *Journal of the American Statistical Association*, 115(530):501–520. MR4107654. doi: https://doi.org/10.1080/01621459.2019.1611582. 3, 23

Ogawa, S., Lee, T.-M., Kay, A. R., and Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, 87(24):9868–9872. doi: https://doi.org/10.1073/pnas.87.24.9868. 2

Papandreou, G. and Yuille, A. (2010). Gaussian sampling by local perturbations. *Advances in Neural Information Processing Systems 23*, 90(8):1858–1866. 13

Penny, W. D., Flandin, G., and Trujillo-Barreto, N. J. (2007). Bayesian comparison of spatially regularised general linear models. *Human Brain Mapping*, 28(4):275–293. doi: https://doi.org/10.1002/hbm.20327. 10

Penny, W. D., Trujillo-Barreto, N. J., and Friston, K. J. (2005). Bayesian fMRI time series analysis with spatial priors. *NeuroImage*, 24(2):350–362. doi: https://doi.org/10.1016/j.neuroimage.2004.08.034. 2, 3, 4, 6, 10, 16

Poldrack, R. A. and Gorgolewski, K. J. (2017). OpenfMRI: Open sharing of task fMRI data. *Neuroimage*, 144:259–261. doi: https://doi.org/10.1016/j.neuroimage.2015.05.073. 16

Rasmussen, C. E. and Williams, K. I. (2006). *Gaussian processes for machine learning*. MIT Press. MR2514435. 2

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407. MR0042668. doi: https://doi.org/10.1214/aoms/1177729586. 11

Rosenfield, A. and Kak, A. C. (1982). *Digital Picture Processing*. Academic Press. MR0451925. 3

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC Press. MR2130347. doi: https://doi.org/10.1201/9780203492024. 3, 5

Rue, H. and Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of Statistical Planning and Inference*, 137(10):3177–3192. MR2365120. doi: https://doi.org/10.1016/j.jspi.2006.07.016. 11

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximation. *Journal of the Royal Statistical Society, Series B*, 71(2):319–392. MR2649602. doi: https://doi.org/10.1111/j.1467-9868.2008.00700.x. 3, 10

Sidén, P. (2020). *Scalable Bayesian spatial analysis with Gaussian Markov random fields*. PhD thesis, Linköping University. 7

Sidén, P., Eklund, A., Bolin, D., and Villani, M. (2017). Fast Bayesian whole-brain fMRI analysis with spatial 3D priors. *NeuroImage*, 146:211–225. doi: https://doi.org/10.1016/j.neuroimage.2016.11.040. 2, 3, 4, 5, 10, 11, 13, 16, 22, 23

Sidén, P., Lindgren, F., Bolin, D., Eklund, A., and Villani, M. (2021). "Supplementary Material for "Spatial 3D Matérn priors for fast whole-brain fMRI analysis"." *Bayesian Analysis*. doi: https://doi.org/10.1214/21-BA1283SUPP. 4

Sidén, P., Lindgren, F., Bolin, D., and Villani, M. (2018). Efficient covariance approximations for large sparse precision matrices. *Journal of Computational and Graphical Statistics*, 27(4):898–909. MR3890879. doi: https://doi.org/10.1080/10618600.2018.1473782. 13, 23

Simpson, D. P., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1–28. MR3634300. doi: https://doi.org/10.1214/16-STS576. 9

Smith, M. and Fahrmeir, L. (2007). Spatial Bayesian variable selection with application to functional magnetic resonance imaging. *Journal of the American Statistical Association*, 102:417–431. MR2370843. doi: https://doi.org/10.1198/016214506000001031. 3

Stein, M. L. (1999). *Interpolation of spatial data. Some theory for kriging*. Springer-Verlag, New York. MR1697409. doi: https://doi.org/10.1007/978-1-4612-1494-6. 2

Takahashi, K., Fagan, J., and Chen, M. S. (1973). Formation of a sparse bus impedance matrix and its application to short circuit study. *IEEE Power Industry Computer Applications Conference*, pages 63–69. 11

Vincent, T., Risser, L., and Ciuciu, P. (2010). Spatially adaptive mixture modeling for analysis of fMRI time series. *IEEE transactions on medical imaging*, 29(4):1059–1074. doi: https://doi.org/10.1109/TMI.2010.2042064. 3

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 41:434–449. MR0067450. doi: https://doi.org/10.1093/biomet/41.3-4.434. 6

Whittle, P. (1963). Stochastic processes in several dimensions. *Bulletin of the International Statistical Institute*, 40(2):974–994. MR0173287. 6

Woolrich, M. W., Jenkinson, M., Brady, J. M., and Smith, S. M. (2004). Fully Bayesian spatio-temporal modeling of fMRI data. *IEEE transactions on medical imaging*, 23(2):213–31. doi: https://doi.org/10.1109/TMI.2003.823065. 3

Yue, Y. R., Lindquist, M., Bolin, D., Lindgren, F., Simpson, D., and Rue, H. (2014). A Bayesian general linear modeling approach to slice-wise fMRI data analysis. *Preprint*. URL https://www.researchgate.net/publication/262144219_A_Bayesian_General_Linear_Modeling_Approach_to_Slice-wise_fMRI_Data_Analysis. 3, 23

Zhang, L., Guindani, M., Versace, F., and Vannucci, M. (2014). A spatio-temporal nonparametric Bayesian variable selection model of fMRI data for clustering correlated time courses. *NeuroImage*, 95:162–175. doi: https://doi.org/10.1016/j.neuroimage.2014.03.024. 3