Physiological Gaussian Process Priors for the Hemodynamics in fMRI Analysis

Josef Wilzén*[a], Anders Eklund[a,b,c], Mattias Villani[a]

[a] *Division of Statistics & Machine Learning, Department of Computer and Information Science, Linköping University, Linköping, Sweden*
[b] *Division of Medical Informatics, Department of Biomedical Engineering, Linköping University, Linköping, Sweden*
[c] *Center for Medical Image Science and Visualization (CMIV), Linköping University, Linköping, Sweden*

Abstract

Inference from fMRI data faces the challenge that the hemodynamic system, that relates the underlying neural activity to the observed BOLD fMRI signal, is not known. We propose a new Bayesian model for task fMRI data with the following features: (i) joint estimation of brain activity and the underlying hemodynamics, (ii) the hemodynamics is modeled nonparametrically with a Gaussian process (GP) prior guided by physiological information and (iii) the predicted BOLD is not necessarily generated by a linear time-invariant (LTI) system. We place a GP prior directly on the predicted BOLD time series, rather than on the hemodynamic response function as in previous literature. This allows us to incorporate physiological information via the GP prior mean in a flexible way. The prior mean function may be generated from a standard LTI system, based on a canonical hemodynamic response function, or a more elaborate physiological model such as the Balloon model. This gives us the nonparametric flexibility of the GP, but allows the posterior to fall back on the physiologically based prior when the data are weak. Results on simulated data show that even with an erroneous prior for the GP, the proposed model is still able to discriminate between active and non-active voxels in a satisfactory way. The proposed model is also applied to real fMRI data, where our Gaussian process model in several cases finds brain activity where previously proposed LTI models, parametric and nonparametric, does not.

Keywords: Bayesian inference, MCMC, fMRI, Hemodynamics, Gaussian processes, mis-specification.

1. Introduction

1.1. Background

Task based fMRI data are typically analyzed using voxel-wise general linear models (GLM), to detect voxels or regions where the blood oxygenation level dependent (BOLD) contrast is correlated with the experimental stimuli paradigm (Friston et al., 1994; Lindquist et al., 2008). BOLD is an indirect measure of neural activation which depends on the hemodynamic response (HR). Understanding the HR is therefore critical in order to correctly infer the brain activity (Handwerker et al., 2004; Lindquist and Wager, 2007; Lindquist et al., 2009). The neurovascular coupling between the neural response triggered by a stimulus and the observed BOLD response in fMRI is not fully understood (Logothetis, 2002, 2003), and the HR has been shown to vary across voxels, brain regions and subjects (Handwerker et al., 2004, 2012). It is common practice in fMRI to model the HR as a linear time invariant system (LTI) (Boynton et al., 2012). Standard GLMs make very strong assumptions about the HR, and since it is unlikely that these models are correct for all voxels and subjects, the inference for the brain activity parameters will be biased (Lindquist and Wager, 2007).

1.2. Joint Detection Estimation framework and Gaussian Processes

The so called joint detection estimation (JDE) framework for the GLM estimates the brain activity jointly with the HR. The JDE approach uses a zero mean Gaussian process prior on filter coefficients, which represent the HR in a LTI context. The filter is often called the hemodynamic response function (HRF) (Goutte et al., 2000; Ciuciu et al., 2003; Marrelec et al., 2003a; Casanova et al., 2008). The activation strength for each voxel is based on a summary statistics on the whole filter. A problem with such voxel-wise approaches is that the filter is unidentified if the specific voxel is inactive. There is also a high risk of overfitting, since a separate HR is estimated in each voxel.

Another approach is to use a bilinear model where both the design matrix and the regression coefficients are estimated jointly. Many models based on the JDE framework use parcellation, see for example Makni et al. (2008); Vincent et al. (2010). Some parameters are constant within each parcel, while other parameters are voxel specific. Parcellation can be done a priori and considered constant (Makni et al., 2008; Vincent et al., 2010), or estimated as a part of the model (Chaari et al., 2012, 2016; Albughdadi et al., 2016). The idea is to restrict the HR to be the same across all voxels in a parcel, but allowing the activation parameters to vary across the voxels within a parcel. In the joint parcellation-detection-estimation (JPDE) frame-

*Corresponding author
Email address: josef.wilzen@liu.se (Josef Wilzén*)

work, the parcellation itself is also learned from the data (Chaari et al., 2012, 2016; Albughdadi et al., 2016). These models use a random effect approach, where the HRF for a given voxel is a random draw from a distribution with a parcel-specific mean HRF. The proposed JDE and JPDE models have been analyzed by variants of Gibbs sampling (Makni et al., 2005, 2006b,a; Ciuciu et al., 2007; Vincent et al., 2007; Makni et al., 2008; Ciuciu et al., 2009; Vincent et al., 2010) and the approximate, but quicker, variational Bayes (VB) method (Chaari et al., 2011, 2012, 2013, 2016; Albughdadi et al., 2016).

Different assumptions have been made regarding the Gaussian process prior for the HRF in the JDE literature. In order to assume a causal filter, the endpoints of the filter are often constrained to be zero (Goutte et al., 2000; Marrelec et al., 2003b; Ciuciu et al., 2003; Makni et al., 2005; Ciuciu et al., 2007; Vincent et al., 2007; Makni et al., 2008; Ciuciu et al., 2009; Vincent et al., 2010; Chaari et al., 2011, 2012, 2013, 2016; Albughdadi et al., 2016). Goutte et al. (2000) and Quirós et al. (2010) use a squared exponential kernel; others use the second order difference matrix as precision matrix (Marrelec et al., 2003b; Ciuciu et al., 2003; Makni et al., 2005, 2006b,a; Ciuciu et al., 2007; Vincent et al., 2007; Makni et al., 2008; Ciuciu et al., 2009; Vincent et al., 2010) and variants thereof (Chaari et al., 2011, 2012, 2013, 2016; Albughdadi et al., 2016). For the approaches that use a fixed and a priori known parcellation, it is assumed that there is *one* HRF per parcel, regardless of the number of stimuli in the experiment. The idea is that a functionally similar region has the same hemodynamic behavior.

1.3. Non-linearity of Predicted BOLD

There is, however, evidence that contradicts the LTI system hypothesis for the HR, see for example Huettel et al. (2004) for a discussion. This has motivated the development of more physiologically realistic models that do not assume an LTI system, and model the predicted BOLD directly (Buxton et al., 1998; Friston et al., 1998, 2000; Buxton et al., 2004; Deneux and Faugeras, 2006; Stephan et al., 2007; Lundengård et al., 2016). Estimation of such nonlinear models are typically more computationally expensive. Non-linear extensions of JDE models that focus on the non-linear habituation effect of repeated stimuli (Ciuciu et al., 2009) are more efficient, but accounts only for a limited class of non-linearities.

1.4. Our approach

In this work, we propose a new model that places a Gaussian Process (GP) prior (Rasmussen and Williams, 2006) directly on the predicted BOLD time series. This is in contrast to earlier work which use a Gaussian process prior on the HRF, and then convolve the posterior HRF with the paradigm. Our approach is therefore not restricted to LTI systems, which means that non-stationary and non-linear properties of the BOLD response can be handled, if supported by the data. Non-stationarity of the BOLD response can for example arise from refractory and adaptation effects (Huettel et al., 2004), or from a participant's failure to perform a task in the MR scanner. Our approach can also implicitly account for the so called stimulus-as-fixed-effect fallacy (Westfall et al., 2016).

A GP prior on the predicted BOLD makes the model very flexible, which can lead to overfitting. Our model therefore incorporates several features to avoid overfitting. First, we use a parcellation approach similar to Makni et al. (2008); Vincent et al. (2010), where the predicted BOLD is restricted to be the same for all voxels in a given parcel, but the activation and other parameters (for example modeling time trends) are voxel-specific. The effect is that the predicted BOLD in a parcel is accurately estimated from data in many voxels. Second, in contrast to the JDE literature, the mean of our GP prior is non-zero and equal to the predicted BOLD from a physiologically motivated model of the hemodynamics, for example the Balloon model proposed by Buxton et al. (1998, 2004). This allows the GP posterior to fall back on the baseline physiological model whenever the data are weak or support the baseline model, while still being able to override the prior mean when the data are incompatible with the baseline model. Third, using a well founded prior mean makes it possible to use relatively tight GP kernels.

The rest of the paper is organized as follow: Section 2 describes the model and the prior on its parameters. The inference procedure is presented in Section 3. Results from simulations and real data are given in Section 4 and 5, respectively. The paper ends with a discussion in Section 6 and conclusions in Section 7.

2. Model, Priors and Posterior Computations

2.1. Notation

Vectors and matrices are denoted with bold lower and upper case letters, respectively. Vectors are assumed to be column vectors. The symbol $^\top$ denotes transpose, $\mathbb{I}_a$ denotes the identity matrix of size $a \times a$, $vec(\cdot)$ is the vectorization operator, $\otimes$ is the Kronecker product and $diag(\boldsymbol{x})$ means a diagonal matrix with vector $\boldsymbol{x}$ as the main diagonal. $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $MN(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$ denote multivariate normal and matrix normal (see Appendix A) distributions, respectively. $InvGamma(a, b)$ denotes the inverse gamma distribution. The following different indices are used:

- j: voxels, $j \in \{1, ..., J\}$, within a parcel
- m: stimulus type, $m \in \{1, \ldots, M\}$
- p: number of nuisance variables, $p \in \{1, \ldots, P\}$
- k: number of parameters in the $AR(k)$ process, $k \in \{1, ..., K\}$.
- t: time, $t \in \mathcal{T}_\star = \{-K+1, -K+2, \ldots, 0, 1, ..., T\}$.

2.2. A physiological Gaussian process prior for predicted BOLD

The fMRI signal will be modeled in the following way: hemodynamic responses are the same for all voxels in a parcel, while task related activations and parameters for the noise process are allowed to vary between voxels in a parcel. The time series contain temporal autocorrelation, which in our case is modeled using an autoregressive (AR) process of order K. We make the usual simplifying assumption in time series analysis that the first K values of the process are known. Let $\mathcal{T}_0 = \{-K+1, -K+2, \ldots, 0\}$ be an initial set of time points where the data values are assumed to be known. Further, define $\mathcal{T} = \{1, \ldots, T\}$ to be the subsequent time points and $\mathcal{T}_\star = \{\mathcal{T}_0, \mathcal{T}\}$ to be the set of all $T_\star = T + K$ time points.

The predicted BOLD is modeled with a Gaussian process (GP) prior (Rasmussen and Williams, 2006) according to

$$f(t) \sim GP(m(t, \boldsymbol{\xi}), \mathtt{k}(t, t', \boldsymbol{\theta})),$$

where $m(t, \boldsymbol{\xi})$ is the mean function and $\mathtt{k}(t, t', \boldsymbol{\theta})$ is the kernel (covariance function) of the process. A sampled value of the GP is denoted f_t and

$$\mathbf{f} = \begin{pmatrix} f_{-k+1} & f_{-k+2} & \cdots & f_T \end{pmatrix}^\top.$$

This prior gives a general framework for modeling the hemodynamic response with a variety of physiological models or constraints. The mean function, defined by the parameters $\boldsymbol{\xi}$, can come from some arbitrary model that can generate the predicted BOLD. These models can be linear (e.g. the HRF used in the SPM software) or nonlinear (e.g. the Balloon model). The kernel controls both the degree of smoothness of $\mathbf{f}$ and the deviation from the mean function. If several stimuli are considered, the total predicted BOLD is considered to be a linear combination of different GPs, which will be denoted $\mathbf{f}_m$, and the $(T+K) \times M$ matrix $\mathbf{F} = \begin{pmatrix} \mathbf{f}_1 & \cdots & \mathbf{f}_M \end{pmatrix}$ gather all the sampled realizations of the different GPs. The parameters $\boldsymbol{\xi}_m$ and $\boldsymbol{\theta}_m$ denote the stimulus specific GP hyperparameters. We use a Matérn kernel with $\frac{5}{2}$ degrees of freedom

$$\mathtt{k}_{\nu=5/2,m}(r) = \omega_m^2 \Big(1 + \tfrac{\sqrt{5}r}{l_m} + \tfrac{5r^2}{l_m^2}\Big) exp\Big(-\tfrac{\sqrt{5}r}{l_m}\Big) \ ,$$

where r is the Euclidean distance between two covariate observations. Let $\boldsymbol{\theta}_m = \begin{pmatrix} l_m & \omega_m^2 \end{pmatrix}$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_M)$. The covariance matrix for all data points from $\mathtt{k}(t, t', \boldsymbol{\theta}_m))$ is denoted $\mathtt{K}(\mathcal{T}_\star, \mathcal{T}_\star)_m$.

The voxel and stimulus specific activations will be represented by the parameter $\beta_{m,j}$ which is contained in the matrix $\mathbf{B}$, of size $M \times J$, where J is the number of voxels. One approach is to let the voxel-wise hemodynamics be modeled as $\mathbf{FB}$. The drawback with this approach is that the parameters enter the likelihood as a product, and are thereby not individually identified since $\mathbf{FB} = \mathbf{F}\boldsymbol{S}^{-1}\boldsymbol{S}\mathbf{B}$, for any invertible matrix $\boldsymbol{S}$ of size $M \times M$. To overcome this problem, we propose an identifying nonlinear transformation for the matrix $\mathbf{F}$. The transformation must fixate the scale of $\mathbf{F}$, and prevent sign flipping as well as linear combinations and permutations. Similar to Pedregosa et al. (2015), we propose the transformation:

$$h(\mathbf{f}_m) = \mathbf{f}_m / \big(||\mathbf{f}_m||_\infty \mathrm{sign}\big(\mathbf{f}_m^\top m(t, \boldsymbol{\xi})\big)\big). \tag{1}$$

This transformation is still sensitive to pure permutations, so in order to identify the column order we introduce a permutation function $\Psi(\cdot)$. Let $\mathbf{F}_0$ be prior mean for $\mathbf{F}$ and $\mathbf{F}_{post}$ be a sample from the posterior for $\mathbf{F}$. The function $\Psi(\mathbf{F}_{post})$ permutes the column in $\mathbf{F}_{post}$ in such a way that

$$\sum_{m=1}^{M} \mathrm{corr}(\mathbf{f}_{post,m} \mathbf{f}_{0,m}),$$

is maximized with respect to the column order of $\mathbf{F}_{post}$. The final transformation is then given by

$$H(\mathbf{F}) = \Psi([h(\boldsymbol{f}_1), \ldots, h(\boldsymbol{f}_M)]). \tag{2}$$

$H(\mathbf{F})$ has its support on a relative scale and is bounded between -1 and 1.

2.3. Multivariate GLM for joint detection and estimation

We use a multivariate regression model for all observed BOLD signals in one parcel, i.e.

$$\mathbf{Y}_\star = H(\mathbf{F})\mathbf{B} + \mathbf{Z}_\star \boldsymbol{\Gamma} + \mathbf{U}_\star, \tag{3}$$

where $\mathbf{Y}_\star = \begin{pmatrix} \mathbf{y}_1 & \cdots & \mathbf{y}_J \end{pmatrix}$ is a $T_\star \times J$ matrix with the observed BOLD signal, where columns represent different voxels and rows represent different time points. The matrix $\mathbf{Z}_\star$, which is of size $T_\star \times P$, contains nuisance covariates, such as time trends and head motion covariates, and the corresponding regression coefficients are stored in the $P \times J$ matrix $\boldsymbol{\Gamma}$. $\mathbf{U}_\star = \begin{pmatrix} \mathbf{u}^{(1)} & \cdots & \mathbf{u}^{(J)} \end{pmatrix}$ is the error matrix, of size $T_\star \times J$, which contains the corresponding errors for all elements in $\mathbf{Y}$. The matrix $\mathbf{F}$ and its transformation $H(\mathbf{F})$ are both of size $T_\star \times M$. The idea is to let $H(\mathbf{F})$ model the dynamics in the predicted BOLD, while $\mathbf{B}$ models the overall response magnitude in each voxel. The separation of the hemodynamics from the activations gives a straight forward measure of voxel activation, that can be used to construct posterior probability maps (PPMs) or t-maps. The hyperparameters for the kernel are for simplicity assumed to be fixed and known, but can in principle be learned in separate updating steps; see the discussion in Section 5.5. The conditional dependency of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ are omitted in the rest of the paper for notational clarity.

Equation (3) can be partitioned with respect to the time observation that corresponds to the pre-sample observations in $\mathcal{T}_0$ and the estimation sample in $\mathcal{T}$ respectively, according to

$$\begin{pmatrix} \mathbf{Y}_0 \\ \mathbf{Y} \end{pmatrix} = H(\mathbf{F})\mathbf{B} + \begin{pmatrix} \mathbf{Z}_0 \\ \mathbf{Z} \end{pmatrix} \boldsymbol{\Gamma} + \begin{pmatrix} \mathbf{U}_0 \\ \mathbf{U} \end{pmatrix}.$$

The upper parts of $\mathbf{Y}_\star$ and $\mathbf{Z}_\star$ have K rows, and the lower parts have T rows. The upper parts will be used as lags in different pre-whitening steps and the lower parts will be used for the inference.

We assume that the noise in each voxel follows an $AR(K)$ process, i.e. for the jth voxel

$$\mathbf{u}^{(j)} = \rho_1 \mathbf{u}_{-1}^{(j)} + \rho_2 \mathbf{u}_{-2}^{(j)} + \ldots + \rho_k \mathbf{u}_{-k}^{(j)} + \boldsymbol{\epsilon}^{(j)},\ j = 1, ...J,$$

where the negative indices denote time lags. We assume that the AR parameters are the same for all voxels in a parcel, but different across parcels. The error terms $\boldsymbol{\epsilon}^{(j)}$ are assumed to be independent across voxels and $\boldsymbol{\epsilon}^{(j)} \sim N\big(\mathbf{0}, \sigma_j^2 \mathbb{I}_T\big)$. Spatial noise dependencies can also be incorporated by replacing $\mathbb{I}_T$ with a matrix $\mathbf{H}$ that models spatial dependences among the elements. The distribution of $\mathbf{U}$ can be expressed as

$$vec(\mathbf{U}) \sim MN(\mathbf{0}, \boldsymbol{\Omega} \otimes \mathbf{M}_{\boldsymbol{\rho}}),$$

where $\boldsymbol{\Omega} = diag\big(\boldsymbol{\sigma}^2\big)$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, ..., \sigma_J^2)$. The matrix $\mathbf{M}_{\boldsymbol{\rho}}$ can be obtained by solving a system of Yule-Walker equations or using the methods of van der Leeuw (1994), but it is not needed explicitly for sampling from the posterior of the model in (3).

The user of our model must decide the parcel sizes. Larger parcels use more data for the estimation of the predicted BOLD, which will be more robust to noise and inactivity, but a drawback is a lower flexibility. Smaller parcels will provide a higher flexibility, at the risk of overfitting.

2.4. Likelihood function and priors

The likelihood function for the model in (3) is of the form

$$L\big(\mathbf{Y}|\mathbf{F}, \mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\sigma}^2, \boldsymbol{\rho}, \mathbf{Z}\big) = (2\pi)^{-TJ/2} |\boldsymbol{\Omega}|^{-T/2} |\mathbf{M}_{\boldsymbol{\rho}}|^{-J/2} \times \exp\Big(-\frac{1}{2} \mathrm{tr}\big[\boldsymbol{\Omega}^{-1} \bar{\mathbf{Y}}^\top \mathbf{M}_{\boldsymbol{\rho}}^{-1} \bar{\mathbf{Y}}\big]\Big),$$

where $\bar{\mathbf{Y}} = \mathbf{Y} - H(\mathbf{F})\mathbf{B} - \mathbf{Z}\boldsymbol{\Gamma}$.

The model (3) has the following parameters and priors:

1. $\mathbf{F}$ has independent Gaussian process priors on each column, i.e.

$$\mathbf{f}_m \sim N(\mathbf{f}_{0,m}, \mathtt{K}(\mathcal{T}_\star, \mathcal{T}_\star)_m) \quad m = 1, \ldots, M,$$

   where $\mathbf{f}_{0,m}$ is the mean function $m(t)$ evaluated at the time points $\mathcal{T}_\star$.
2. The elements of $\boldsymbol{\sigma}^2$ are assumed to be independent apriori and are modeled as

$$\sigma_j^2 \sim \mathrm{InvGamma}(c_{0,j}, d_{0,j}).$$

3. The prior for $\boldsymbol{B}$ is modeled conditional on $\boldsymbol{\Omega} = \mathrm{diag}(\boldsymbol{\sigma}^2)$ as a matrix normal distribution

$$\mathbf{B}|\boldsymbol{\Omega} \sim MN_{M\times \mathrm{J}}\big(\mathbf{B}_0, \boldsymbol{\Omega}, \kappa^{-1}\mathbf{P}^{-1}\big),$$

   where $\mathbf{P}$ is a $M \times M$ positive definite precision matrix over stimuli, $\mathbf{B}_0$ is the $M \times J$ prior mean matrix and κ is a scalar.
4. $\boldsymbol{\Gamma}$ is assigned a matrix normal prior conditional on $\boldsymbol{\Omega}$

$$\boldsymbol{\Gamma}|\boldsymbol{\Omega} \sim MN_{M\times \mathrm{J}}\big(\boldsymbol{\Gamma}_0, \boldsymbol{\Omega}, \tau^{-1}\mathbb{I}_P\big).$$

5. Following Eklund et al. (2016), the prior on the AR process parameters is centered over a stationary AR(1) process:

$$\boldsymbol{\rho} \sim N(\boldsymbol{\rho}_0, \mathbf{A}_0) \cdot I(\boldsymbol{\rho}),$$

   where $\boldsymbol{\rho}_0 = (r, 0, \ldots, 0)$, $\mathbf{A}_0 = \mathrm{diag}(c^2, \frac{c^2}{2^\zeta}, \ldots, \frac{c^2}{K^\zeta})$ and $I(\boldsymbol{\rho})$ is an indicator function for the stationary region

$$I(\boldsymbol{\rho}) = \begin{cases} 1 & \text{if } |\ell_{max}| < 1 \\ 0 & \text{otherwise,} \end{cases}$$

   where ℓ_{max} is the largest absolute (modulus) eigenvalue of the companion matrix

$$\begin{pmatrix} \rho_1 & \rho_2 & \cdots & \rho_{K-1} & \rho_K \\ 1 & 0 & & 0 & 0 \\ 0 & 1 & & & 0 \\ & & \ddots & 0 & \vdots \\ 0 & & 0 & 1 & 0 \end{pmatrix}.$$

   Note that the prior is centered over the noise process $u_t = \rho_1 \cdot u_{t-1} + \epsilon_t$, but assigns probability mass also to higher order AR processes in such a way that longer lags are shrunk more heavily toward zero.

2.5. Posterior computations

The joint posterior of all model parameters in (3) is given by

$$p\big(\mathbf{F}, \mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\sigma}^2, \boldsymbol{\rho}|\mathbf{Y}, \mathbf{Z}\big) \propto L\big(\mathbf{Y}|\mathbf{F}, \boldsymbol{\Gamma}, \mathbf{B}, \boldsymbol{\sigma}^2, \boldsymbol{\rho}\mathbf{Z}\big) \times p(\mathbf{F}) p\big(\mathbf{B}|\boldsymbol{\sigma}^2\big) p\big(\boldsymbol{\Gamma}|\boldsymbol{\sigma}^2\big) p\big(\boldsymbol{\sigma}^2\big) p(\boldsymbol{\rho}). \tag{4}$$

Equation (4) is intractable and cannot be sampled using standard distributions, and we therefore resort to Gibbs sampling. The posterior is sampled in four steps, which are described in Algorithm 1 and detailed formulas are given in the Appendix B.

2.6. Implementation

The proposed estimation of the model in (3) is implemented in the R programming language. Since each parcel is independent, the computation can be parallelized across parcels, which is done with the foreach package (Analytics and Weston, 2015). Functions from the R package neuRosim (Welvaert et al., 2011) are used to simulate data for the simulation study.

Algorithm 1 Schematic description of the Gibbs sampler for the posterior in Equation (4).

1. $\boldsymbol{\rho}$ is sampled from a multivariate Gaussian distribution in Equation (6). $\boldsymbol{\rho}$ is then used to pre-whiten the data.
2. $\boldsymbol{\sigma}^2$ is sampled from an Inverse-gamma distribution in Equation (9).
3. $vec(\boldsymbol{B})$ and $vec(\boldsymbol{\Gamma})$ are sampled from a multivariate normal distribution in Equation (10).
4. $\mathbf{F}$ is sampled with Elliptical slice sampling using the likelihood in (12).

3. Simulations

A simulation study was performed to investigate the proposed model's ability to detect activity, and to estimate the underlying hemodynamics. Two models were used. In the first model, the predicted BOLD is fixated to the prior mean. For the second model, the predicted BOLD is instead estimated from the data using our physiological GP prior. Data were generated from the first model. A total of 32 simulated datasets were generated, for each combination of parameters in the data generating process. In each simulation a single parcel with 100 voxels was used, and 20 of the voxels were active. The following settings were used for each simulation

- A single stimulus was modeled as a block paradigm. The number of time points was set to 150 and the sampling rate (TR) was set to 1 second.
- An AR(3) process was used for the noise process, with autoregressive parameters: $\rho^T = (0.4, 0.1, 0.05)$.
- The contrast to noise ratio (CNR), $b_{m,j}/\sigma$, where $b_{m,j} = 1$ for active voxels and zero otherwise, was set to 5 and 7.
- Constant, linear, quadratic and cubic trends were added to each time series. The coefficients were generated randomly for each voxel and simulation, to reflect realistic trends in fMRI data.

The same priors were used for all simulations, except for $\boldsymbol{\Gamma}$, and were specified to:

1. $\mathbf{F}$: The prior mean for $\mathbf{F}$ was created in such a way that is was either correct, with correlation 1 with the true process, or erroneous with a correlation of 0.615 with the true process. We use two different lengthscales: $l = 2$ or $l = 4$, and $\omega = \sqrt{0.1} \approx 0.316$. The prior mean function was specified to the common double gamma HRF convolved with the paradigm. For the fix model $\mathbf{F}$ is not estimated, but instead fixated to the value of the prior mean.
2. $\boldsymbol{\sigma}^2$: $c_0 = 0$ and $d_0 = 0$, giving a non-informative prior.
3. $\mathbf{B}$: precision scale factor $\kappa = 10^{-10}$, precision matrix $\mathbf{P}$ is diagonal, giving an essentially flat non-informative prior.
4. $\boldsymbol{\Gamma}$: $\tau = 0$, except for the parameter representing the constant, which was given an empirical prior based on voxel mean and four times the voxel variance for each voxel. The nuisance variables are scaled to have zero mean and unit variance.
5. $\boldsymbol{\rho}$: $\boldsymbol{\rho}_0 = (0, 0, 0)$, $\mathbf{A}_0 = \text{diag}(0.5, \frac{0.5}{2^5}, \frac{0.5}{3^5})$

The posterior was sampled 4000 times and 1000 samples were discarded as burn in. The remaining samples were thinned out by a factor of 3, leaving 1000 posterior draws for inference.

There are several ways that voxels can be declared active in an Bayesian model. One way is to construct posterior probabilities of the type $p(b_{m,j} > c|\mathbf{y}) > a$, where c is a user defined effect size and a is a probability threshold. These probabilities can be used to construct PPMs over the brain. However, there is no consensus regarding how to threshold PPMs, and PPMs can be numerically unstable and imprecisely estimated. Instead, we use Bayesian "t-ratios", defined as

$$t = \frac{\mathrm{E}(b|\mathbf{y}) - c}{\sqrt{\mathrm{Var}(b|\mathbf{y})}},$$

which can be easily computed from the posterior samples in each voxel. These ratios have a higher resolution for a fixed amount of posterior samples compared to PPMs. In the simulations $c = 0$ and the test $t > a$ was used, where a is a quantile from a t-distribution. Note that the activation is voxel independent given $\mathbf{F}$ and $\boldsymbol{\rho}$, due to the simulation design. These frequentist calculations are of course not directly transferable to a Bayesian setting, but has the advantage of giving familiar thresholds for fMRI.

A range of values for a were tested and ROC curves were calculated. Figure 1 and 2 show the results for the lengthscales $l = 4$ and $l = 2$, respectively. The results show that our model with a GP prior on predicted BOLD model performs better for a wide range of thresholds, and thus has a much better ability to discriminate between active and non-active voxels. As expected, for very low thresholds the model with estimated predicted BOLD starts to classify non-active voxels as active.

Figure 3 and 4 show the posterior of the predicted BOLD for selected simulations.

4. Real data

4.1. Data

To test our proposed approach on real data, we used open fMRI data from brain tumor patients (Pernet et al., 2016; Gorgolewski et al., 2013), as the hemodynamic response function may be different close to a tumor. A total of 22 patients (9 females) with different types of brain tumors were scanned using both structural (T1, T2, DWI)

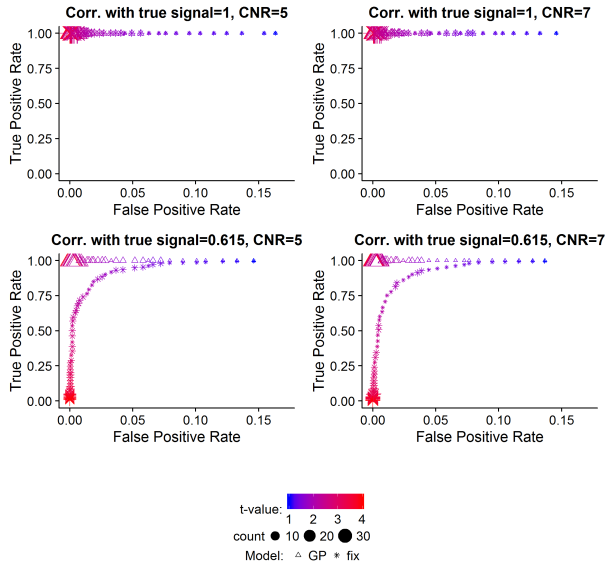

Figure 1: ROC curves for the simulation study. Lengthscale hyperparameter for GP: $l = 4$. Each value of is the average of over 32 simulations. Thresholds for t-values: 60 equidistant values between 1 and 4.

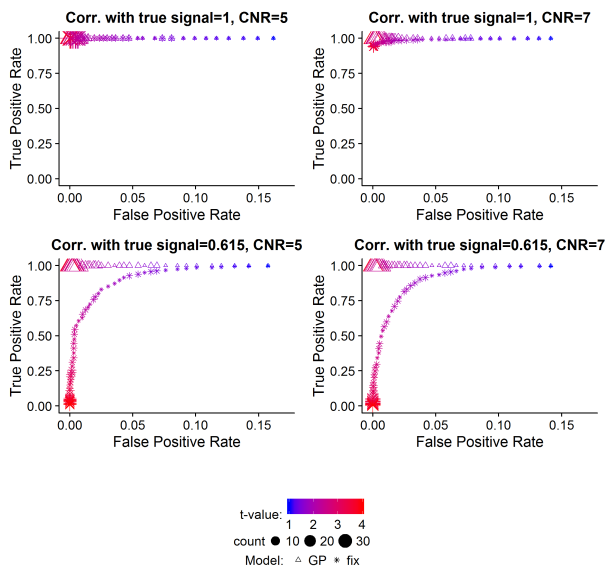

Figure 2: ROC curves for the simulation study. Lengthscale hyperparameter for GP: $l = 2$. Each value of is the average of over 32 simulations. Thresholds for t-values: 60 equidistant values between 1 and 4.

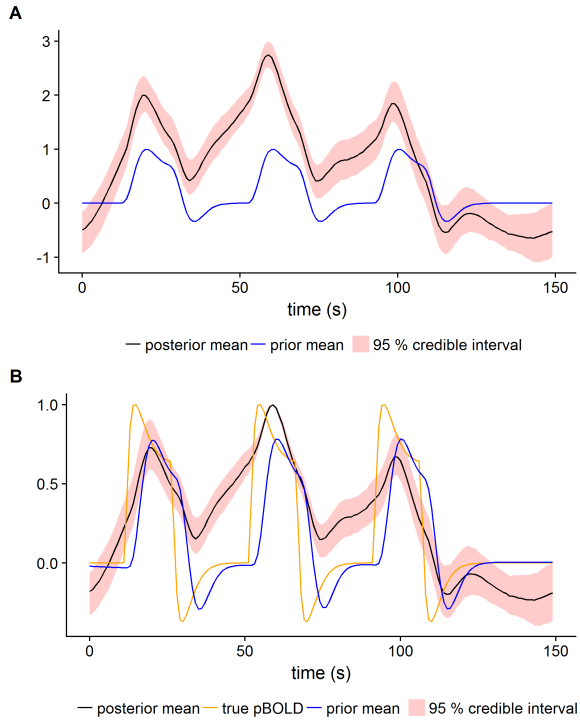

Figure 3: Posterior for predicted BOLD for a simulated parcel. CNR=5, correlation with true signal is 0.615. Lengthscale hyperparameter for GP: $l = 4$. (A) is the Gaussian process $\mathbf{F}$ in 3 and (B) is the transformed Gaussian process $H(\mathbf{F})$, see Equation (1) and (2).

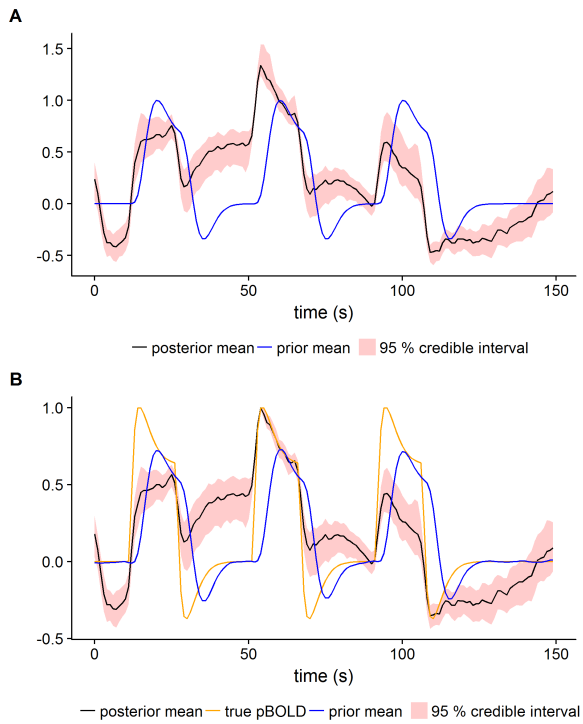

Figure 4: Posterior for predicted BOLD for a simulated parcel. CNR=5, correlation with true signal is 0.615. Lengthscale hyperparameter for GP: $l = 2$. (A) is the Gaussian process $\mathbf{F}$ in 3 and (B) is the transformed Gaussian process $H(\mathbf{F})$, see Equation (1) and (2).

and functional (BOLD T2*) MRI sequences. For the functional scans several tasks were performed: motor, verb generation and word repetition (resting state data are also available). Data were acquired on a General Electric 1.5 Tesla scanner with an 8 channel phased-array head coil. The fMRI data were acquired using a standard EPI sequence with a repetition time of 5.0 seconds (due to sparse sampling for auditory tasks) or 2.5 seconds, and an echo time of 50 milliseconds. Each voxel has a size of 4 x 4 x 4 mm^3, resulting in volumes with 64 x 64 x 30 voxels.

We here focus on the word repetition task. The task is to repeat a given word (overt word repetition), in 6 blocks with 30 seconds of activation and 30 seconds of rest. We can thereby expect activation of the language areas of the brain, parts of the motor cortex that correspond to the mouth and tongue (speech production) and the auditory cortex (listening). Our presented results are for two randomly selected subjects: 18716 and 19628.

4.2. Preprocessing

The fMRI data were preprocessed using motion correction and 6 mm smoothing. The brain parcellation was performed by registering the ADHD 200 parcel atlas[1] (Craddock et al., 2012; Bellec et al., 2017) to EPI space, by combining linear T1-MNI and EPI-T1 transformations using FSL. A brain mask is applied in order to remove voxels outside the brain.

In order to be able to compare effect sizes between voxels, the fMRI data is scaled. We use the scaling

$$\frac{\boldsymbol{y}_j}{sd(\boldsymbol{y}_j)} \cdot \frac{100}{GM},$$

where $\boldsymbol{y}_j$ is observed data for voxel j , $sd(\cdot)$ is the standard deviation function and GM is the global mean of $\boldsymbol{y}_j/sd(\boldsymbol{y}_j)$ over all voxels. This scaling ensures that overall the voxels have an average value close to 100 and the same standard deviation. This makes it possible to use effect sizes in terms of percent of the global mean signal, as done in Penny et al. (2005); Sidén et al. (2017). There are 186 parcels and a total of 19,836 voxels for subject 18716. Corresponding numbers for subject 19628 are 179 and 19,304, respectively. Some example parcels and the distribution of parcel size for subject 19628 are shown in Figure 5.

4.3. Results

Independent models were fitted to each parcel. An autoregressive model of order 3 was used for the noise process. The same priors were used for all parcels. The same prior hyperparameters were used as in the simulation study except for $\mathbf{F}$, which used kernel hyperparameters $l = 4$ and $\omega = 0.1$. The prior mean function for $\mathbf{F}$ was specified to the default HRF in SPM convolved with the paradigm, and was scaled to have zero mean and unit variance.

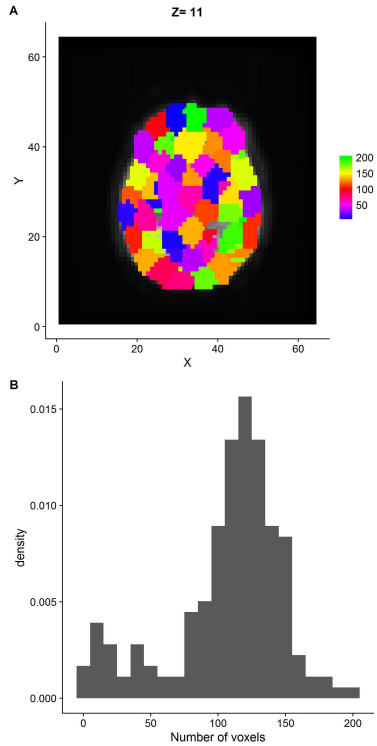

Figure 5: Descriptive statistics for subject 19628. (A) shows all parcels for Z-slice 11. The color specifies parcel belonging. (B) shows the number of voxels in all parcels. Note that our model falls back on the prior mean if there are few active voxels in a parcel, meaning that the model will not break down for small parcels.

Constant, linear, quadratic and cubic trends were included as nuisance variables. Starting values were obtained by using the prior mean for $\mathbf{F}$, and all other parameters were initialized using an Cochrane-Orcutt estimation procedure in each parcel. Group-wise ridge regression, using the R-package glmnet (Friedman et al., 2010), was used to obtain the nuisance and the activation parameters. The autoregressive parameters were estimated for the whole parcel using regularized heteroscedastic regression similar to (6). The two steps were iterated until the difference in mean squared error was less then 0.01.

The proposed GP model is compared with three baseline models. In the first model the predicted BOLD is fixated to the prior mean. The second model also uses the temporal derivative of the prior mean, i.e. two basis functions (which is the most common way to allow for a small time shift of the paradigm). The third model uses a smooth FIR approach (Goutte et al., 2000; Ciuciu et al., 2003; Marrelec et al., 2003a; Makni et al., 2008; Vincent et al., 2010) to model the predicted BOLD (see Appendix C for details). The same priors are used for all other parameters.

The posterior was sampled 9000 times, and 3000 samples were discarded as burn in. The remaining samples

[1] https://www.nitrc.org/frs/?group_id=427

were thinned out by a factor 6, leaving a final sample of 1000 posterior draws for inference. The parameters $\mathbf{B}$, $\mathbf{\Gamma}$, $\boldsymbol{\sigma}^2$ and $\boldsymbol{\rho}$ showed a good mixing with low autocorrelation. In some parcels, the elements in $\mathbf{F}$ had high autocorrelation, but the first and second half of the draws gave similar posteriors.

There can be sizeable differences in activation from using a GP prior on the predicted BOLD, compared to using a fixed predicted BOLD. For example, focusing first on subject 19628, Figures 6 and 7 show that the model that estimates the predicted BOLD finds more activity compared to the two first baseline models. For example, the flexible model detects more brain activity in Broca's language area, which for this subject is close to the brain tumor. According to Figure 8, the flexible GP model also detects stronger brain activity compared to the smooth FIR filter approach.

Figures 9 and 10 show the estimated posterior for the predicted BOLD in the two parcels with most positive activation for subject 19628. Parcel 159 is the yellow cluster in Z-slice 10 and 11 in Figure 6 and has 47 active voxels in total. Parcel 32 is the purple cluster in Z-slice 10 and 11 in Figure 6 and has 47 active voxels in total. Note that the scale of the transformed GP is relative and bounded between -1 and 1, due to the infinity norm. The posteriors have a non-linear behavior, where the amplitude of the peaks and the undershoots vary over time. This form of the posteriors indicate that the data contain important information about the shape of the predicted BOLD, which is not contained in the prior.

Turning now to subject 18716, none of the two first baseline models (using standard basis functions) detected any activity for the given effect size, but our model that estimates the predicted BOLD detected several active voxels, which can be seen in Figure 11. For example, the flexible model detects brain activity in auditory cortex and in motor cortex, not detected by the fix model. As can be seen in Figure 12, the smooth FIR filter approach also detects activity in the motor cortex, but not in the auditory cortex. It should be noted, however, that this activity difference is not caused by a different HR due to the tumor, as the detected activity is on the opposite side of the tumor. The results for the baseline model with two basis functions are not shown.

Figure 13 shows the estimated posterior for the predicted BOLD in one of the parcels with most positive activation for subject 18716. Parcel 32 is the blue cluster in Z-slice 12 in Figure 11 and has 25 active voxels in total. Similar to the posterior predicted BOLD shown for subject 19626, the amplitudes of the peaks and the undershoots are non-stationary.

In order to investigate the effect of the lengthscale hyperparameter, GP models with different lengthscales were estimated. The result is presented in Figure 14. With the shorter lengthscales, the predicted BOLD gets more flexible, and thus finds more activity. Similarly, longer lengthscales restrict the model and the results are more similar to the reference model.

In parcels that lack activity, the posterior predicted BOLD is similar to the prior distribution (not shown).

5. Discussion

5.1. Model the predicted BOLD instead of the HRF

We have proposed and implemented a new way to model the predicted BOLD for task fMRI. The difference from other models is the direct modeling of the predicted BOLD time series, instead of the HRF, combined with a straight forward measure of activity. The simulation study shows that the model has a good ability to discriminate between active and non-active voxels. The proposed model shows robustness to misspecification in the prior mean function for the predicted BOLD. This is a desirable feature, since it is likely that a model with a fix predicted BOLD will not be correct for the whole brain or across subjects. Our proposed model gives the researcher a framework to approach problems related to the HR. For example, for group studies (where all data are transformed to a standard space), the predicted BOLD in one parcel can be compared across subjects. Also, the existence of explicit activity parameters makes it easy to construct PPMs or t-maps.

The non-linear aspect of the hemodynamics can be captured with the GP model. It is interesting to study the properties of the posterior for the predicted BOLD, see Figures 9, 10 and 13. Compared to the prior mean function, the major difference for the posterior is the time-varying amplitude of the peaks and the undershoots. This feature seems to be crucial to find the additional activity compared to the two baseline models and the smooth FIR approach. The described feature is not easily incorporated into traditional GLM approaches. Parametric modulation of the HRF can be used in the LTI context, to obtain hemodynamic features that are non-stationary, but this approach comes with two problems. First, a proper modulator must be chosen. Second, the non-stationarity is assumed to be be known and fix across the brain given the modulator. Our approach handles the non-stationarity in an unsupervised manner, and can of course use a prior mean function that depends on a problem specific modulator.

5.2. Computation

The model is implemented in the R programming language, and the Markov Chain Monte Carlo (MCMC) code is not optimized for speed, except for the CPU parallelization of models across parcels. The computations for our model are therefore rather time consuming, but there are several options to reduce the computational time. One option is to use other inference methods, such as VB inference. Another option is to use GPUs (graphics cards) for parallelization, particularly if there are a relatively large number of parcels.

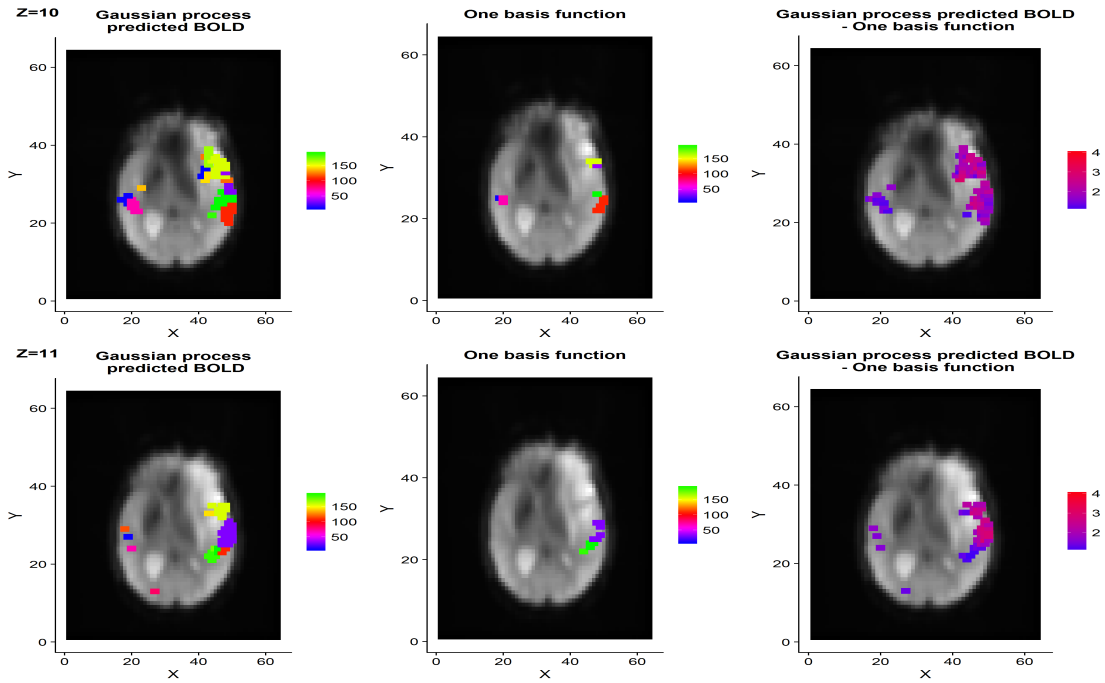

Figure 6: Example slices with Bayesian t-ratios for subject 19836. The activity maps are thresholded at $t \geq 4$ for a test that tests the effect size 0.25. The color specifies parcel belonging for active voxels. The rightmost column shows the differences in t-ratios, thresholded such that only values fulfilling $|t_1 - t_2| > 1$ are shown. Our flexible model clearly detects more activity, compared to the fix predicted BOLD model. Top row: the flexible model detects more brain activity in Broca's language area, which for this subject is close to the brain tumor. Bottom row: the flexible model finds brain activity in the visual cortex and bilateral activation of the auditory cortex, which the fix model struggles to detect.

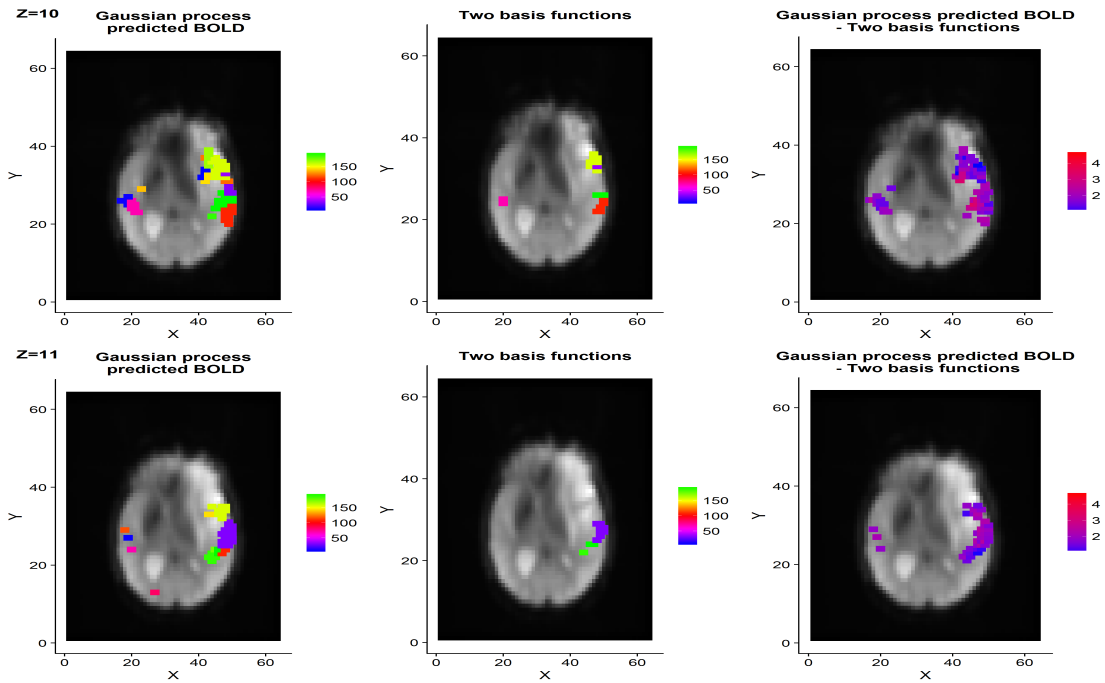

Figure 7: Example slices with Bayesian t-ratios for subject 19836. The activity maps are thresholded at $t \geq 4$ for a test that tests the effect size 0.25. t-ratios for the baseline model is created with b for the first basis. The color specifies parcel belonging for active voxels. The rightmost column shows the differences in t-ratios, thresholded such that only values fulfilling $|t_1 - t_2| > 1$ are shown. The differences between the flexible model and the fix model are very similar to Figure 6, where a single basis function was used.

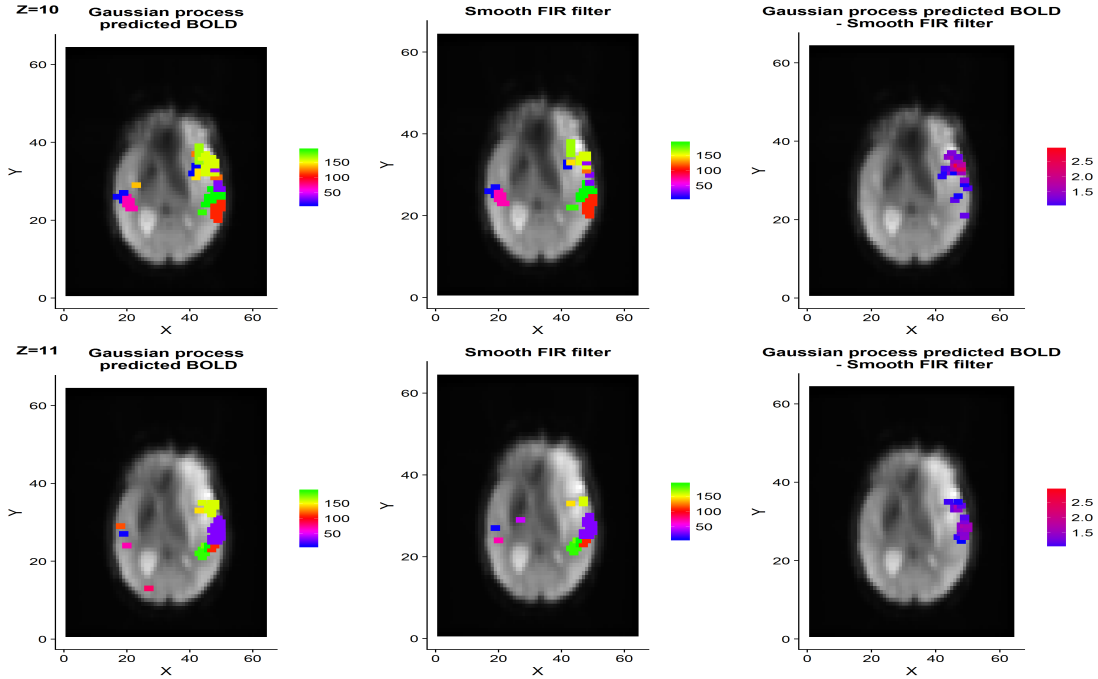

Figure 8: Example slices with Bayesian t-ratios for subject 19836. The activity maps are thresholded at $t \geq 4$ for a test that tests the effect size 0.25. The color specifies parcel belonging for active voxels. The rightmost column shows the differences in t-ratios, thresholded such that only values fulfilling $|t_1 - t_2| > 1$ are shown.

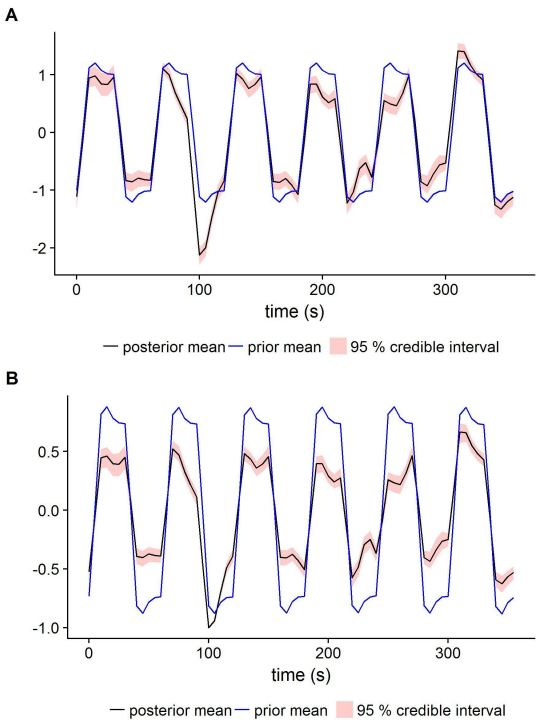

Figure 9: Estimated predicted BOLD for subject 19836 and parcel 159. (A) is the Gaussian process $\mathbf{F}$ in (3) and (B) is the transformed Gaussian process $H(\mathbf{F})$, see Equation (1) and (2).

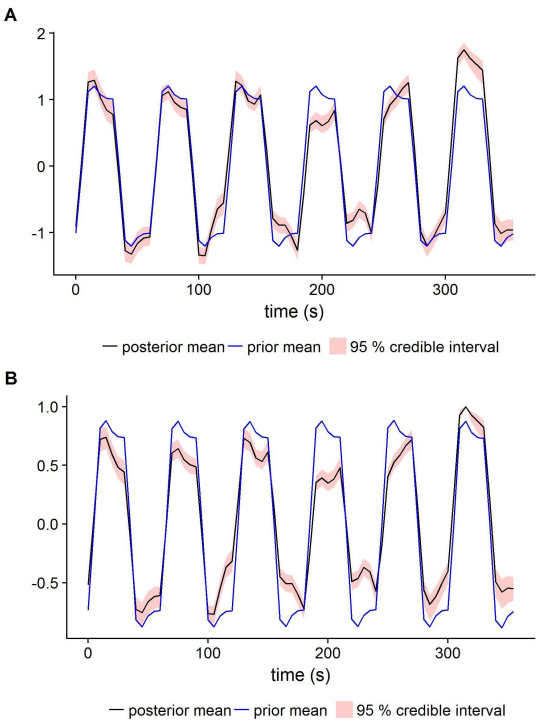

Figure 10: Estimated predicted BOLD for subject 19836 and parcel 32. (A) is the Gaussian process $\mathbf{F}$ in (3) and (B) is the transformed Gaussian process $H(\mathbf{F})$, see Equation (1) and (2).

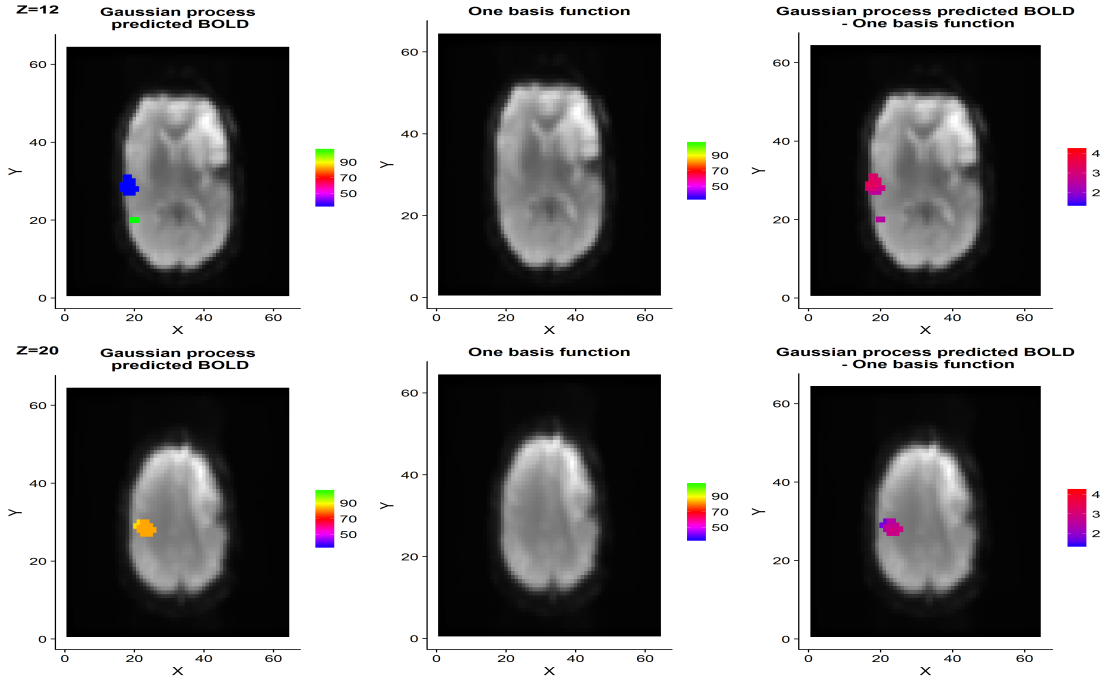

Figure 11: Example slices with Bayesian t-ratios for subject 18716. The activity maps are thresholded at $t \geq 4$ for a test that tests the effect size 0.25. The color specifies parcel belonging for active voxels. The rightmost column shows the differences in t-ratios, thresholded such that only values fulfilling $|t_1 - t_2| > 1$ are shown. Top row: the flexible model finds brain activity in the auditory cortex, which is not found using the fix model. Bottom row: the flexible model finds brain activity in the motor cortex (generated by speech production), which is not found using the fix model.

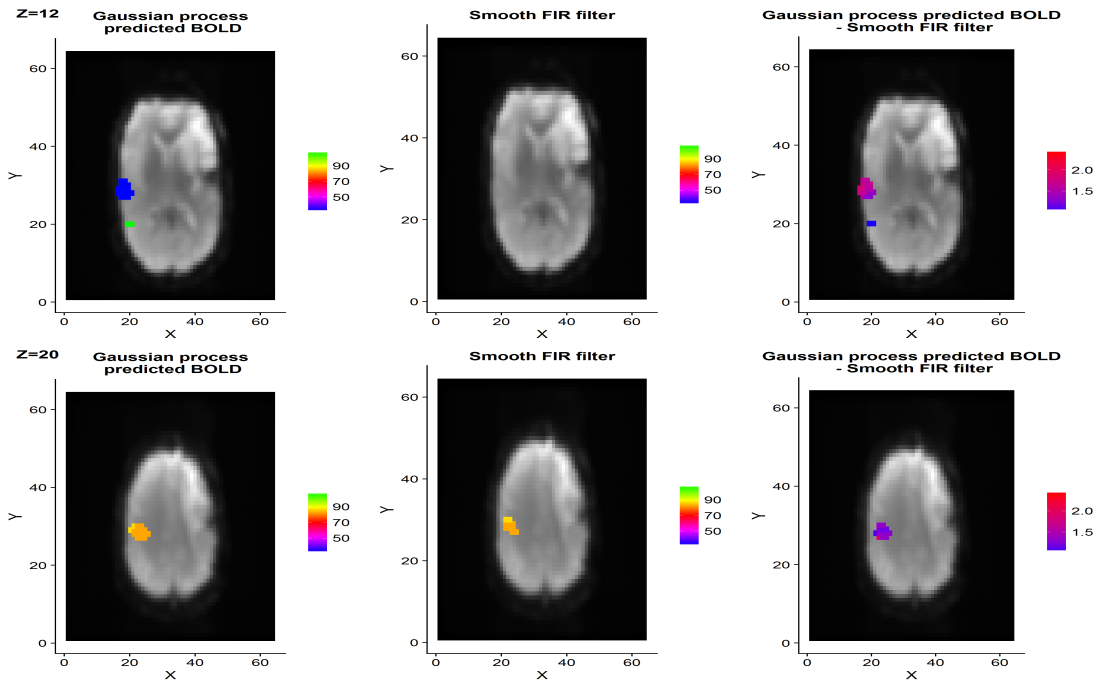

Figure 12: Example slices with Bayesian t-ratios for subject 18716. The activity maps are thresholded at $t \geq 4$ for a test that tests the effect size 0.25. The color specifies parcel belonging for active voxels. The rightmost column shows the differences in t-ratios, thresholded such that only values fulfilling $|t_1 - t_2| > 1$ are shown.

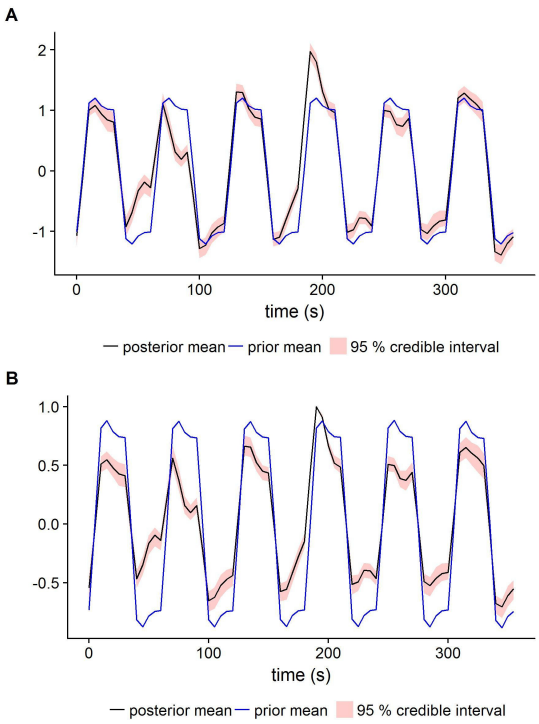

Figure 13: Estimated predicted BOLD for subject 18716 and parcel 32. (A) is the Gaussian process $\mathbf{F}$ in (3) and (B) is the transformed Gaussian process $H(\mathbf{F})$, see Equation (1) and (2).

5.3. Multiple comparisons

In contrast to frequentist methods, there is no consensus in the fMRI field regarding if and how to correct for multiple comparisons for PPMs. In frequentist hypothesis testing the null hypothesis is normally that the parameter representing the brain activity is 0, but using an effect size threshold of 0 for PPMs often leads to activation in a very large portion of the voxels (even for strict probability thresholds for the PPMs). In this paper we have mainly focused on differences between fix and flexible predicted BOLD models, using voxel inference and an effect size threshold of 0.25. One ad-hoc approach to correct for multiple comparisons is to calculate a Bayesian t- or z-score for each voxel, and then apply existing frequentistic approaches for multiple comparison correction (e.g. Gaussian random field theory). This approach is for example used in the FSL software.

5.4. Applications

There are several possible applications of our proposed model. As demonstrated in this paper, a potential application is in clinical fMRI, where fMRI can be used to map out important brain areas prior to tumor surgery. The HR may be different close to a tumor, and our flexible model can then be used to detect more brain activity, and thereby potentially lead to a better treatment plan. Other cases where the HR may be different include young subjects (Richter and Richter, 2003), subjects with epilepsy (Jacobs et al., 2008) and subjects with stroke (Bonakdarpour et al., 2007). Our model can also be used to automatically handle cases where a subject fails to perform one or several events, or where a subject occasionally struggles with the timing of the experiment (adding a temporal derivative can only account for a global shift in time, and not local time shifts). As mentioned in the introduction, our model can also pick up variations in the strength of the BOLD response, while virtually all other models see the stimulus as a fixed effect (Westfall et al., 2016).

5.5. Future work

A natural extension of the model is to do inference for the GP hyperparameters. Since the model uses a non-Gaussian likelihood this is, however, non-trivial. In the MCMC case the methods presented in Filippone and Girolami (2014); Murray and Graham (2016) could be used, where pseudo marginal inference is employed together with an unbiased estimate of the intractable marginal likelihood for the GP.

The prior mean is a quite severely misspecified in the simulation in Section 3, and the GP model cannot fully capture the true underlying signal, see Figures 3 and 4, even though it captures sufficient variation to be able to distinguish between active and non-active voxels. Other kernels could perform even better in this scenario. For block paradigms, locally periodic kernels are a natural candidate, see for example Duvenaud (2014) for a discussion. Those kernels could introduce a positive periodic correlation that decays with the distance; paradigm blocks near in time will be more correlated compared to blocks further away in time. Another alternative is use the prior mean function $m(t)$ and the derivative $\partial m(t)/\partial t$ as covariates in a smooth kernel. This would give a behavior similar to locally periodic kernel, but can also handle event related data. The number of hyperparameters grows with more complex kernels, and it is hard to manually specify these parameters. To efficiently use more complex kernels, inference for the hyperparameters is important, either using an estimate of the posterior mode or sampling via MCMC. Even more sophisticated kernels could be used, such as spectral mixture kernels (Wilson and Adams, 2013) or deep kernels (Wilson et al., 2016). Such kernels are very expressive and can find complicated patterns given sufficient data. However, the proposed inference methods are adapted for the classical GP regression model $y = f(x)+\epsilon$, where f is a GP, and not as a part of a multivariate time series regression model. The problem boils down to finding *data driven* features that can be used in some kernel, in order to explain the prior correlation in $\mathbf{f}$ in a good way for a given parcel.

Another possible future direction is to use more sophisticated priors on $\mathbf{B}$ and $\mathbf{\Gamma}$. Here variable selection would be appropriate, as it could reduce overfitting and improve the model fit. Interesting selection methods could be spike and slab priors or horseshoe priors. Due to the spatial nature of the data, spatial priors would also be an interest-

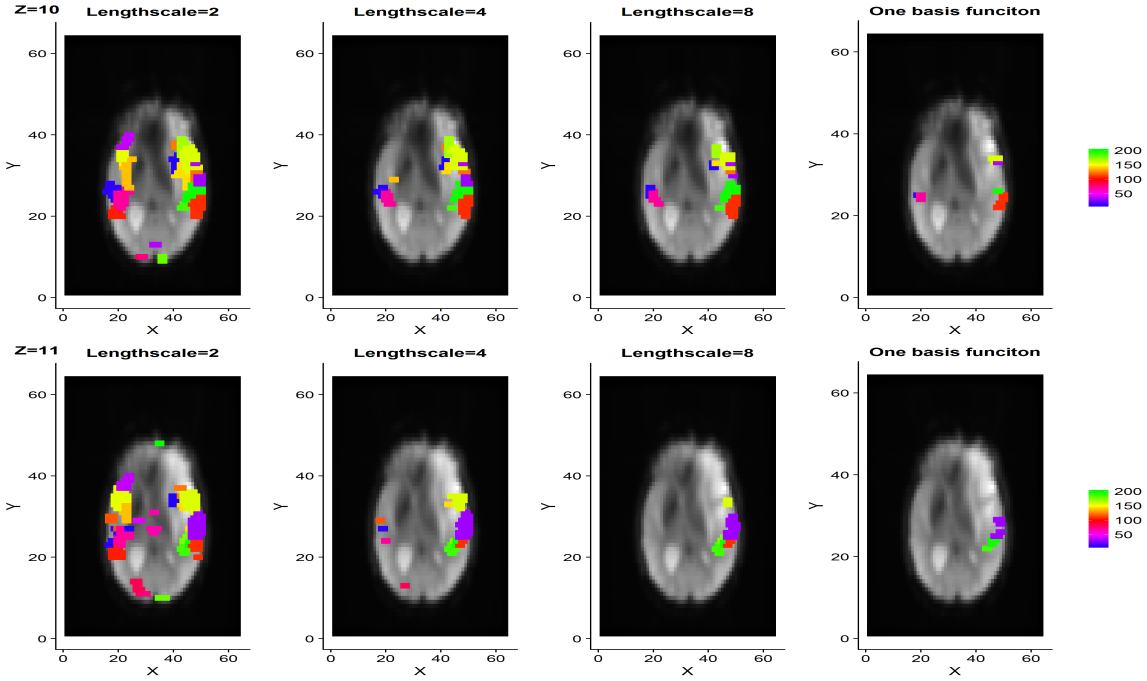

Figure 14: Comparison of different lengthscales for the GP model. Example slices with Bayesian t-ratios for subject 19836. The activity maps are thresholded at $t \geq 4$ for a test that tests the effect size 0.25. The color specifies parcel belonging for active voxels. Clearly, decreasing the lengthscale leads to a more flexible predicted BOLD, which may lead to overfitting.

ing choice. For example, a GP prior could be placed over $\mathbf{B}$ to impose a spatial smoothness, e.g. using a Matérn kernel. This would result in models similar to those in (Luttinen and Ilin, 2009) and (Wilson et al., 2012). Sidén et al. (2017) use sparse precision matrices to model spatial dependencies in whole brain task fMRI data, and derive both fast MCMC and VB methods. Those ideas could be incorporated into the proposed model.

The suggested model is made for single subject data, and in many cases joint inference for many subjects is desirable. A hierarchical model could be used here, with random effects for the $\mathbf{B}$ and $\mathbf{F}$ parameters. Assume that all subjects have been transformed to the same space with the same parcellation. The prior for $\mathbf{f}_m$ for a given parcel could then be expressed as

$$\begin{aligned} \mathbf{f}_m &\sim N(m(t), \mathtt{k}(t,t')) \\ \mathbf{f}_{m,n} &\sim N(\mathbf{f}_m, \varsigma_m \mathbb{I}) \end{aligned},$$

where n is subject index and ς_m is the random effects variance. $\mathbf{B}$ can be modeled in the same way. This construction is similar to the within subject models used in (Chaari et al., 2012, 2016; Albughdadi et al., 2016), but a difference is that they have a random effect on the HRF filter and the hierarchy is over parcels and voxels.

6. Conclusion

We have proposed a novel framework for modeling the hemodynamics in task fMRI. The new model is shown to more accurately detect brain activity compared to traditional parametric and nonparametric LTI models. We model the predicted BOLD directly with a GP prior, as a part of larger time series regression model. We also introduce an identifying transformation that solves the challenging identification problem present in bilinear models in the JDE context. This can be done due to problem specific constraints related to the hemodynamics. Our new framework gives researchers the opportunity to ask new kinds of questions related to hemodynamics, especially with regard to non-linear effects.

Acknowledgments

This work was funded by Swedish Research Council (Vetenskapsrådet) grant no. 2013-5229.

7. References

Mohanad Albughdadi, Lotfi Chaari, Jean-Yves Tourneret, Florence Forbes, and Philippe Ciuciu. Hemodynamic Brain Parcellation Using A Non-Parametric Bayesian Approach. working paper or preprint, February 2016. URL `https://hal.inria.fr/hal-01275622`.

Revolution Analytics and Steve Weston. *foreach: Provides Foreach Looping Construct for R*, 2015. URL `https://CRAN.R-project.org/package=foreach`. R package version 1.4.3.

Pierre Bellec, Carlton Chu, Francois Chouinard-Decorte, Yassine Benhajali, Daniel S Margulies, and R Cameron Craddock. The Neuro Bureau ADHD-200 Preprocessed Repository. *Neuroimage*, 144:275–286, 2017.

B Bonakdarpour, TB Parrish, and CK Thompson. Hemodynamic response function in patients with stroke-induced aphasia: implications for fmri data analysis. *Neuroimage*, 36(2):322–331, 2007.

Geoffrey M Boynton, Stephen A Engel, and David J Heeger. Linear systems analysis of the fMRI signal. *NeuroImage*, 62(2):975–984, 2012.

Richard B Buxton, Eric C Wong, and Lawrence R Frank. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magnetic resonance in medicine*, 39(6):855–864, 1998.

Richard B Buxton, Kâmil Uludağ, David J Dubowitz, and Thomas T Liu. Modeling the hemodynamic response to brain activation. *Neuroimage*, 23:S220–S233, 2004.

Ramon Casanova, Srikanth Ryali, John Serences, Lucie Yang, Robert Kraft, Paul J Laurienti, and Joseph A Maldjian. The impact of temporal regularization on estimates of the BOLD hemodynamic response function: a comparative analysis. *Neuroimage*, 40(4): 1606–1618, 2008.

Lotfi Chaari, Florence Forbes, Thomas Vincent, Michel Dojat, and Philippe Ciuciu. Variational solution to the joint detection estimation of brain activity in fMRI. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011*, pages 260–268, 2011.

Lotfi Chaari, Florence Forbes, Thomas Vincent, and Philippe Ciuciu. Hemodynamic-informed parcellation of fMRI data in a joint detection estimation framework. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, pages 180–188, 2012.

Lotfi Chaari, Thomas Vincent, Florence Forbes, Michel Dojat, and Philippe Ciuciu. Fast joint detection-estimation of evoked brain activity in event-related fMRI using a variational approach. *IEEE transactions on Medical Imaging*, 32(5):821–837, 2013.

Lotfi Chaari, Solveig Badillo, Thomas Vincent, Ghislaine Dehaene-Lambertz, Florence Forbes, and Philippe Ciuciu. Subject-level Joint Parcellation-Detection-Estimation in fMRI. working paper or preprint, January 2016. URL `https://hal.inria.fr/hal-01255465`.

Philippe Ciuciu, J-B Poline, Guillaume Marrelec, Jérôme Idier, Christophe Pallier, and Habib Benali. Unsupervised robust non-parametric estimation of the hemodynamic response function for any fMRI experiment. *IEEE Transactions on medical imaging*, 22(10):1235–1251, 2003.

Philippe Ciuciu, Jérôme Idier, Thomas Veit, and Thomas Vincent. Application du rééchantillonnage stochastique de l'échelle en détection-estimation de l'activité cérébrale par IRMf. In *21° Colloque GRETSI, Troyes, FRA, 11-14 septembre 2007*. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images, 2007.

Philippe Ciuciu, Stéphane Sockeel, Thomas Vincent, and Jérôme Idier. Modelling the neurovascular habituation effect on fMRI time series. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 433–436. IEEE, 2009.

R. Cameron Craddock, G.Andrew James, Paul E. Holtzheimer, Xiaoping P. Hu, and Helen S. Mayberg. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, 33(8):1914–1928, 2012.

Thomas Deneux and Olivier Faugeras. EEG-fMRI fusion of non-triggered data using Kalman filtering. In *Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on*, pages 1068–1071. IEEE, 2006.

David Duvenaud. *Automatic model construction with Gaussian processes*. PhD thesis, University of Cambridge, 2014.

Anders Eklund, Martin A Lindquist, and Mattias Villani. A Bayesian Heteroscedastic GLM with Application to fMRI Data with Motion Spikes. *arXiv preprint arXiv:1612.00690*, 2016.

Maurizio Filippone and Mark Girolami. Pseudo-marginal bayesian inference for gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2214–2226, 2014.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

Karl J Friston, Peter Jezzard, and Robert Turner. Analysis of functional MRI time-series. *Human brain mapping*, 1(2):153–171, 1994.

Karl J Friston, Oliver Josephs, Geraint Rees, and Robert Turner. Nonlinear event-related responses in fMRI. *Magnetic resonance in medicine*, 39(1):41–52, 1998.

Karl J Friston, Andrea Mechelli, Robert Turner, and Cathy J Price. Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. *NeuroImage*, 12(4):466–477, 2000.

Krzysztof J Gorgolewski, Amos Storkey, Mark E Bastin, Ian R Whittle, Joanna M Wardlaw, and Cyril R Pernet. A test-retest fmri dataset for motor, language and spatial attention functions. *GigaScience*, 2(1):6, 2013.

Cyril Goutte, Finn Aarup Nielsen, and KH Hansen. Modeling the hemodynamic response in fMRI using smooth FIR filters. *IEEE transactions on medical imaging*, 19(12):1188–1201, 2000.

Daniel A Handwerker, John M Ollinger, and Mark D'Esposito. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage*, 21 (4):1639–1651, 2004.

Daniel A Handwerker, Javier Gonzalez-Castillo, Mark D'Esposito, and Peter A Bandettini. The continuing challenge of understanding and modeling hemodynamic variation in fMRI. *Neuroimage*, 62(2):1017–1023, 2012.

Scott A Huettel, Allen W Song, and Gregory McCarthy. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, 2004.

Julia Jacobs, Colin Hawco, Eliane Kobayashi, Rainer Boor, Pierre LeVan, Ulrich Stephani, Michael Siniatchkin, and Jean Gotman. Variability of the hemodynamic response as a function of age and frequency of epileptic discharge in children with epilepsy. *Neuroimage*, 40(2):601–614, 2008.

Martin A Lindquist and Tor D Wager. Validity and power in hemodynamic response modeling: a comparison study and a new approach. *Human brain mapping*, 28(8):764–784, 2007.

Martin A Lindquist, Ji Meng Loh, Lauren Y Atlas, and Tor D Wager. Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. *Neuroimage*, 45(1):S187–S198, 2009.

Martin A Lindquist et al. The statistical analysis of fMRI data. *Statistical Science*, 23(4):439–464, 2008.

Nikos K Logothetis. The neural basis of the blood–oxygen–level–dependent functional magnetic resonance imaging signal. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 357(1424):1003–1037, 2002.

Nikos K Logothetis. The underpinnings of the BOLD functional magnetic resonance imaging signal. *The Journal of Neuroscience*, 23(10):3963–3971, 2003.

Karin Lundengård, Gunnar Cedersund, Sebastian Sten, Felix Leong, Alexander Smedberg, Fredrik Elinder, and Maria Engström. Mechanistic mathematical modeling tests hypotheses of the neurovascular coupling in fMRI. *PLoS Comput Biol*, 12(6):e1004971, 2016.

Jaakko Luttinen and Alexander Ilin. Variational gaussian-process factor analysis for modeling spatio-temporal data. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1177–1185. Curran Associates, Inc., 2009. URL `http://papers.nips.cc/paper/3805-variational-gaussian-process-factor-analysis-for-modeling-sp` `pdf`.

Salima Makni, Philippe Ciuciu, Jérôme Idier, and J-B Poline. Joint detection-estimation of brain activity in functional MRI: a multichannel deconvolution solution. *IEEE Transactions on Signal Processing*, 53(9):3488–3502, 2005.

Salima Makni, Philippe Ciuciu, Jérôme Idier, and J-B Poline. Bayesian joint detection-estimation of brain activity using MCMC with a Gamma-Gaussian mixture prior model. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 5, pages V–V. IEEE, 2006a.

Salima Makni, Philippe Ciuciu, Jérôme Idier, and J-B Poline. Joint detection-estimation of brain activity in fMRI using an autoregressive noise model. In *Biomedical Imaging: Nano to Macro,*

2006. 3rd IEEE International Symposium on, pages 1048–1051. IEEE, 2006b.

Salima Makni, Jérôme Idier, Thomas Vincent, Bertrand Thirion, Ghislaine Dehaene-Lambertz, and Philippe Ciuciu. A fully Bayesian approach to the parcel-based detection-estimation of brain activity in fMRI. *Neuroimage*, 41(3):941–969, 2008.

Guillaume Marrelec, Habib Benali, Philippe Ciuciu, Mélanie Pélégrini-Issac, and Jean-Baptiste Poline. Robust Bayesian estimation of the hemodynamic response function in event-related BOLD fMRI using basic physiological information. *Human Brain Mapping*, 19(1):1–17, 2003a.

Guillaume Marrelec, Habib Benali, Philippe Ciuciu, Mélanie Pélégrini-Issac, and Jean-Baptiste Poline. Robust Bayesian estimation of the hemodynamic response function in event-related BOLD fMRI using basic physiological information. *Human Brain Mapping*, 19(1):1–17, 2003b.

Iain Murray and Matthew Graham. Pseudo-marginal slice sampling. In *Artificial Intelligence and Statistics*, pages 911–919, 2016.

Iain Murray, Ryan Prescott Adams, and David JC MacKay. Elliptical slice sampling. In *AISTATS*, volume 13, pages 541–548, 2010.

Fabian Pedregosa, Michael Eickenberg, Philippe Ciuciu, Bertrand Thirion, and Alexandre Gramfort. Data-driven HRF estimation for encoding and decoding models. *NeuroImage*, 104:209–220, 2015.

William D Penny, Nelson J Trujillo-Barreto, and Karl J Friston. Bayesian fmri time series analysis with spatial priors. *NeuroImage*, 24(2):350–362, 2005.

Cyril R. Pernet, Krzysztof J. Gorgolewski, Dominic Job, David Rodrigues, Ian Whittle, and Joanna Wardlaw. A structural and functional magnetic resonance imaging dataset of brain tumour patients. *Scientific Data*, 3:Article number 160003, 2016.

S James Press. Applied multivariate analysis: Using bayesian and frequenist methods of inference, 1982.

Alicia Quirós, Raquel Montes Diez, and Simon P Wilson. Bayesian spatiotemporal model of fMRI data using transfer functions. *NeuroImage*, 52(3):995–1004, 2010.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, 2006.

Wolfgang Richter and Marlene Richter. The shape of the fmri bold response in children and adults changes systematically with age. *NeuroImage*, 20(2):1122–1131, 2003.

Per Sidén, Anders Eklund, David Bolin, and Mattias Villani. Fast bayesian whole-brain fmri analysis with spatial 3d priors. *NeuroImage*, 146:211–225, 2017.

Klaas Enno Stephan, Nikolaus Weiskopf, Peter M Drysdale, Peter A Robinson, and Karl J Friston. Comparing hemodynamic models with DCM. *Neuroimage*, 38(3):387–401, 2007.

Jan van der Leeuw. The covariance matrix of ARMA errors in closed form. *Journal of econometrics*, 63(2):397–405, 1994.

Thomas Vincent, Philippe Ciuciu, and Jérôme Idier. Spatial mixture modelling for the joint detection-estimation of brain activity in fMRI. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 1, pages I–325. IEEE, 2007.

Thomas Vincent, Laurent Risser, and Philippe Ciuciu. Spatially adaptive mixture modeling for analysis of fMRI time series. *IEEE Transactions on Medical Imaging*, 29(4):1059–1074, 2010.

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

Jacob Westfall, Thomas Nichols, and Tal Yarkoni. Fixing the stimulus-as-fixed-effect fallacy in task fMRI. *bioRxiv*, page 077131, 2016.

Andrew Wilson and Ryan Adams. Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1067–1075, 2013.

Andrew G. Wilson, David A. Knowles, and Zoubin Ghahramani. Gaussian process regression networks. In J. Langford and J. Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML)*, Edinburgh, June 2012. Omnipress.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 370–378, 2016.

Appendix A: Distributions

The density function for a matrix normal distribution

$$\mathbf{Y} \sim MN_{n\times p}(\mathbf{M}, \mathbf{U}, \mathbf{V})$$

is of the form

$$p(\mathbf{Y}|\mathbf{M}, \mathbf{U}, \mathbf{V}) = \frac{exp\Big(-\frac{1}{2}tr\Big[\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{M})^\top \mathbf{U}^{-1}(\mathbf{Y}-\mathbf{M})\Big]\Big)}{(2\pi)^{np/2}|\mathbf{V}|^{n/2}|\mathbf{U}|^{p/2}}$$

Parameters:

- $\mathbf{M}$: location, real $n \times p$ matrix
- $\mathbf{U}$: scale, positive-definite real $n \times n$ matrix (dependencies over observations)
- $\mathbf{V}$: scale, positive-definite real $p \times p$ matrix (dependencies over variables)

The matrix normal distribution is related to the multivariate normal distribution in the following way:

$$vec(\mathbf{Y}) \sim N(vec(\mathbf{M}), \mathbf{V} \otimes \mathbf{U})$$

Appendix B: Gibbs sampling

Sampling $\boldsymbol{\rho}$

The autoregressive parameters are sampled using the following formulation

$$\mathbf{u} = \rho_1 \mathbf{u}_{-1} + \rho_2 \mathbf{u}_{-2} + \ldots + \rho_k \mathbf{u}_{-k} + \boldsymbol{\epsilon} \Longleftrightarrow$$

$$\mathbf{u} = \begin{pmatrix} \mathbf{u}_{-1} & \cdots & \mathbf{u}_{-k} \end{pmatrix} \begin{pmatrix} \rho_1 \\ \vdots \\ \rho_k \end{pmatrix} + \boldsymbol{\epsilon} = \mathbf{D}\boldsymbol{\rho} + \boldsymbol{\epsilon}, \quad (5)$$

where $\boldsymbol{u}_\star = vec(\mathbf{Y}_\star - \mathbf{F}_\star \mathbf{B} - \mathbf{Z}_\star \boldsymbol{\Gamma})$ is used to calculate $\mathbf{u}$, given that $\mathbf{B}$, $\mathbf{F}_\star$ and $\boldsymbol{\Gamma}$ are known, $\mathbf{D}$ is a matrix of size $JT \times k$. $\mathbf{u}$ and $\mathbf{D}$ must be updated in every iteration since $\mathbf{B}$, $\mathbf{F}$ and $\boldsymbol{\Gamma}$ also are updated in every iteration. Let $\boldsymbol{\Sigma} = \boldsymbol{\Omega} \otimes \mathbb{I}_T$, then $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. It is clear from (5) that the full conditional posterior for $\boldsymbol{\rho}$ can be obtained using standard formulas for univariate regression with heteroscedastic variance, and the full conditional posterior is given by

$$
\begin{aligned}
\boldsymbol{\rho}_n|\boldsymbol{\Sigma} &\sim N(\boldsymbol{\rho}_n, \mathbf{A}_n) \\
\boldsymbol{\rho}_n &= \mathbf{A}_n\big(\mathbf{D}^\top\boldsymbol{\Sigma}^{-1}\mathbf{D}\hat{\boldsymbol{\rho}} + \mathbf{A}_0\boldsymbol{\rho}_0\big) \\
\mathbf{A}_n &= \big(\mathbf{D}^\top\boldsymbol{\Sigma}^{-1}\mathbf{D} + \mathbf{A}_0\big)^{-1} \\
\hat{\boldsymbol{\rho}} &= \big(\mathbf{D}^\top\boldsymbol{\Sigma}^{-1}\mathbf{D}\big)^{-1}\mathbf{D}^\top\boldsymbol{\Sigma}^{-1}\mathbf{u}.
\end{aligned} \tag{6}
$$

To make sure that $\boldsymbol{\rho}$ is in the stationary region, draws from $\boldsymbol{\rho}_n|\boldsymbol{\Sigma} \sim MN(\boldsymbol{\rho}_n, \mathbf{A}_n)$ are discarded until a draw that is inside the stationary region is obtained.

Sampling $\mathbf{B}$, $\boldsymbol{\Gamma}$ *and* $\boldsymbol{\sigma}^2$

In order to simplify sampling of $\mathbf{B}$ and $\boldsymbol{\Gamma}$ in model (3) the following formulation is used

$$
\begin{aligned}
\mathbf{Y}_\star &= H(\mathbf{F})\mathbf{B} + \mathbf{Z}_\star\boldsymbol{\Gamma} + \mathbf{U}_\star \Leftrightarrow \\
\mathbf{Y}_\star &= \begin{pmatrix} H(\mathbf{F}) & \mathbf{Z}_\star \end{pmatrix}\begin{pmatrix} \mathbf{B} \\ \boldsymbol{\Gamma} \end{pmatrix} + \mathbf{U}_\star \\
\mathbf{Y}_\star &= \mathbf{X}_\star\mathbf{Q} + \mathbf{U}_\star,
\end{aligned} \tag{7}
$$

where $\mathbf{X}_\star$ is of the size $T_\star \times (M+P)$ and $\mathbf{Q}$ is of the size $(M+P)\times J$. In order to use the standard multivariate regression formulas in (7), pre-whitening is used. Let $\Phi_C(L)$ be the column-wise lag polynomial from time series analysis, i.e. $\Phi_C(L) = 1 - \rho_1 L - \rho_2 L^2 - \ldots - \rho_k L^k$. The first K constant observations in $\mathbf{Y}_0$, $\mathbf{F}_0$ and $\mathbf{Z}_0$ are used for obtaining the first K observations of $\mathcal{T}$. This results in a new regression formulation

$$
\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\mathbf{Q} + \mathbf{E}, \tag{8}
$$

where $\tilde{\mathbf{Y}} = \Phi_C(L)\mathbf{Y}$, $\tilde{\mathbf{X}} = \Phi_C(L)\mathbf{X}$ and $\mathbf{E} = \tilde{\mathbf{U}} = \Phi_C(L)\mathbf{U}$.

Given the parcel constant parameters $\mathbf{F}$, $\boldsymbol{\rho}$, and the regression parameters $\mathbf{Q}$, the likelihood $p\Big(\tilde{\mathbf{Y}}, \tilde{\mathbf{X}}|\boldsymbol{\sigma}^2, \mathbf{Q}\Big)$ is independent over voxels, which implies that inference for each element in $\boldsymbol{\sigma}^2$ is performed by regressions of the form: $\tilde{\mathbf{y}}_j = \tilde{\mathbf{X}}\mathbf{q}_j + \boldsymbol{\epsilon}_j$, where $\tilde{\mathbf{y}}_j$ is the j:th column of $\tilde{\mathbf{Y}}$ and $\mathbf{q}_j$ is the j:th column of $\mathbf{Q}$. The full conditional posterior for σ_j^2 is an Inverse-gamma distribution, which is easily obtained from standard formulas for univariate regression

$$
p\Big(\sigma_j^2|\tilde{\mathbf{y}}_j, \tilde{\mathbf{X}}, \mathbf{Q}\Big) \sim \text{InvGamma}(c_n, d_j) \quad j = 1, \ldots, J, \tag{9}
$$

where

$$
\begin{aligned}
c_n &= c_0 + T/2 \\
d_j &= d_0 + \tfrac{1}{2}\Big(\tilde{\mathbf{y}}_j - \tilde{\mathbf{X}}\mathbf{q}_j\Big)^\top\Big(\tilde{\mathbf{y}}_j - \tilde{\mathbf{X}}\mathbf{q}_j\Big) \qquad j = 1, \ldots, J
\end{aligned}
$$

Using standard formulas for multivariate regression with conjugate priors, the full conditional posterior for $vec(\mathbf{Q})$ is a multivariate normal distribution. The likelihood function for (8) is described by Press (1982). Let $\hat{\mathbf{Q}} = \Big(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\Big)^{-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{Y}}$, $\mathbf{S} = \Big(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\mathbf{Q}}\Big)^\top\Big(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\mathbf{Q}}\Big)/T$ and $vec(\mathbf{Q}) = \boldsymbol{q}$. Using standard manipulations, the likelihood can be written as

$$
\begin{aligned}
&p\Big(\tilde{\mathbf{Y}}|\tilde{\mathbf{X}}, \mathbf{Q}, \boldsymbol{\Omega}\Big) \propto \\
&|2\pi\boldsymbol{\Omega}|^{-T/2} exp\big(-\tfrac{1}{2}tr\boldsymbol{\Omega}^{-1}T\cdot\boldsymbol{S}\big)\cdot \\
&exp\Big\{-\tfrac{1}{2}(\mathbf{q}-\hat{\mathbf{q}})^\top\Big[\boldsymbol{\Omega}^{-1}\otimes\Big(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\Big)\Big](\mathbf{q}-\hat{\mathbf{q}})\Big\}.
\end{aligned}
$$

The prior for $\mathbf{Q}$ is now expressed as

$$
\begin{aligned}
\mathbf{Q}|\boldsymbol{\Omega} &\sim MN_{M+p,J}\Big(\mathbf{Q}_0, \boldsymbol{\Omega}\otimes\mathbf{P}_Q^{-1}\Big) \\
\mathbf{Q}_0 = \begin{pmatrix} \mathbf{B}_0 \\ \boldsymbol{\Gamma}_0 \end{pmatrix} &\quad \mathbf{P}_Q = \begin{pmatrix} \kappa\mathbf{P} & \mathbf{0} \\ \mathbf{0} & \tau\mathbb{I} \end{pmatrix},
\end{aligned}
$$

where $\mathbf{Q}_0$ has the same size as $\mathbf{Q}$ and $\mathbf{P}_Q$ has size $(M+P)\times(M+P)$. Using standard formulas for multivariate regression, the full conditional posterior for $\boldsymbol{Q}$ is then given by

$$
\boldsymbol{q}|\boldsymbol{\Omega}, \tilde{\mathbf{Y}}, \tilde{\mathbf{X}} \sim N\Big[\bar{\mathbf{q}}, \boldsymbol{\Omega}\otimes\Big(\mathbf{P}_Q + \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\Big)^{-1}\Big], \tag{10}
$$

where

$$
\bar{\mathbf{q}} = \Big[\boldsymbol{\Omega}\otimes\Big(\mathbf{P}_Q + \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\Big)^{-1}\Big]vec\Big[\Big(\tilde{\mathbf{X}}^\top\tilde{\mathbf{Y}} + \mathbf{P}_Q\mathbf{Q}_0\Big)\boldsymbol{\Omega}^{-1}\Big].
$$

Sampling $\mathbf{F}$

To simplify the sampling of $\mathbf{F}$, we reformulate the model as

$$
\begin{aligned}
\mathbf{g}_\star &= \text{vec}(\mathbf{Y}_\star - \mathbf{Z}_\star\boldsymbol{\Gamma}) \\
&= \text{vec}(\mathbb{I}_{T_\star}H(\mathbf{F})\mathbf{B}) + vec(\mathbf{U}_\star) \\
&= \mathbf{W}\mathbf{f}_H + \boldsymbol{u}_\star,
\end{aligned} \tag{11}
$$

where $\mathbf{W} = \big(\mathbf{B}^\top\otimes\mathbb{I}_{T_\star}\big)$ is of size $JT_\star \times T_\star M$ and $\mathbf{f}_H = \text{vec}(H(\mathbf{F}))$ is of size $(JT_\star)\times 1$. Now, Equation (11) is transformed with a lag polynomial $\Phi_R(L)$ in a row-wise manner. $\Phi_R(L)$ has the same functional form as $\Phi_C(L)$, but operates independently on each voxel time series. The first $T_\star$ rows will first be transformed, followed by transformation of the next $T_\star$ rows, until all rows have been transformed. The transformation results in

$$
\tilde{\mathbf{g}} = \tilde{\mathbf{W}}\mathbf{f}_H + \boldsymbol{\epsilon},
$$

where $\tilde{\mathbf{W}}$ is of size $JT\times T_\star M$, $\tilde{\mathbf{g}}$ and $\boldsymbol{\epsilon}$ are both of size $(JT)\times 1$. The likelihood for $\mathbf{f}_H$ is given by

$$
\begin{aligned}
&p\Big(\tilde{\mathbf{g}}|\tilde{\mathbf{W}}, \boldsymbol{\Omega}, \mathbf{f}_H, \boldsymbol{\rho}\Big) = \\
&\tfrac{1}{\sqrt{(2\pi)^{t+k}|\boldsymbol{\Sigma}|}}exp\Big(-\tfrac{1}{2}\Big(\tilde{\mathbf{g}} - \tilde{\mathbf{W}}\mathbf{f}_H\Big)^\top\boldsymbol{\Sigma}^{-1}\Big(\tilde{\mathbf{g}} - \tilde{\mathbf{W}}\mathbf{f}_H\Big)\Big).
\end{aligned} \tag{12}
$$

Note that $\mathbf{F}$ enters the Gaussian likelihood in a non-linear way, which means that the full conditional posterior is not available in closed form. We use elliptical slice sampling Murray et al. (2010) to sample from the posterior of $\mathbf{F}$. Elliptical slice sampling is a slice sampling technique which is particularly suitable for Gaussian process models with non-Gaussian likelihoods.

Appendix C: Smooth FIR model for predicted BOLD

The proposed model is also compared with a smooth FIR model. In order to make the comparison as fair as possible, the FIR model is formulated as

$$\mathbf{Y}_\star = H(\mathbf{X}_{FIR}\boldsymbol{h})\mathbf{B} + \mathbf{Z}_\star\boldsymbol{\Gamma} + \mathbf{U}_\star, \tag{13}$$

where $\mathbf{X}_{FIR}$ is the standard FIR design matrix (where stimuli are organized column wise) with $K \times M$ columns. K is the filter length, and $\boldsymbol{h}$ is the filters for all stimuli stacked in one vector of size $KM \times 1$. $\boldsymbol{h}$ has independent and identical GP priors for each stimuli. Since $\boldsymbol{h}$ contains much fewer parameters than $\mathbf{F}$, the scale factor of the kernel is specified to a higher value compared to $\mathbf{F}$, in order to increase flexibility. The filters in $\boldsymbol{h}$ is estimated using elliptical slice sampling, in the same way as $\mathbf{F}$ is estimated in our model. The inference for the other parameters is not changed. The prior mean function for $\boldsymbol{h}$ was specified to the common double gamma HRF, and the kernel hyperparameters were specified to $l = 3$ and $\omega = \sqrt{0.5} \approx 0.707$. The endpoints of the filter were constrained to the corresponding prior mean values.