Fast joint detection-estimation of evoked brain activity in event-related fMRI using a variational approach

Lotfi CHAARI, *Member, IEEE*, Thomas VINCENT, Florence FORBES,
Michel DOJAT, *Senior Member, IEEE* and Philippe CIUCIU, *Senior Member, IEEE*

Abstract*— In standard within-subject analyses of event-related fMRI data, two steps are usually performed separately: detection of brain activity and estimation of the hemodynamic response. Because these two steps are inherently linked, we adopt the so-called region-based Joint Detection-Estimation (JDE) framework that addresses this joint issue using a multivariate inference for detection and estimation. JDE is built by making use of a regional bilinear generative model of the BOLD response and constraining the parameter estimation by physiological priors using temporal and spatial information in a Markovian model. In contrast to previous works that use Markov Chain Monte Carlo (MCMC) techniques to sample the resulting intractable posterior distribution, we recast the JDE into a missing data framework and derive a Variational Expectation-Maximization (VEM) algorithm for its inference. A variational approximation is used to approximate the Markovian model in the unsupervised spatially adaptive JDE inference, which allows automatic fine-tuning of spatial regularization parameters. It provides a new algorithm that exhibits interesting properties in terms of estimation error and computational cost compared to the previously used MCMC-based approach. Experiments on artificial and real data show that VEM-JDE is robust to model misspecification and provides computational gain while maintaining good performance in terms of activation detection and hemodynamic shape recovery.

***Index Terms*— functional MRI, Joint Detection-Estimation, Markov random field, EM algorithm, Variational approximation.**

I. Introduction

Functional Magnetic Resonance Imaging (fMRI) is a powerful tool to non-invasively study the relationship between a sensory or cognitive task and the ensuing evoked neural activity through the neurovascular coupling measured by the BOLD signal [1]. Since the 90's, this neuroimaging modality has become widely used in brain mapping as well as in functional connectivity study in order to probe the specialization and integration processes in sensory, motor and cognitive brain regions [2–4]. In this work, we focus on the recovery of localization and dynamics of local evoked activity, thus on specialized cerebral processes. In this setting, the key issue is the modeling of the link between stimulation events and the induced BOLD effect throughout the brain. Physiological non-linear models [5–8] are the most specific approaches to properly describe this link but their computational cost and their identifiability issues limit their use to a restricted number of specific regions and to a few experimental conditions. In contrast, the common approach, being the focus of this paper, rather relies on linear systems which appear more robust and tractable [2, 9]. Here, the link between stimulation and BOLD effect is modelled through a convolutive system where each stimulus event induces a BOLD response, via the convolution of the binary stimulus sequence with the Hemodynamic Response Function (HRF). There are two goals for such BOLD analysis: the detection of where cerebral activity occurs and the estimation of its dynamics through the HRF identification. Commonly, the estimation part is ignored and the HRF is fixed to a canonical shape which has been derived from human primary visual area BOLD response [10, 11]. The detection task is performed by a General Linear Model (GLM), where stimulus-induced components are assumed to be known and only their relative weighting are to be recovered in the form of effect maps [2]. However, spatial intra-subject and between-subject variability of the HRF has been highlighted [12–14], in addition to potential timing fluctuations induced by the paradigm (*e.g.* variations in delay [15]). To take this variability into account, more flexibility can be injected in the GLM framework by adding more regressors. In a parametric setting, this amounts to adding a function basis, such as canonical HRF derivatives, a set of gamma or logistic functions [15, 16]. In a non-parametric setting, all HRF coefficients are explicitly encoded as a Finite Impulse Response (FIR) [17]. The major drawback of these GLM extensions is the multiplicity of regressors for a given condition, so that the detection task becomes more difficult to perform and that statistical power is decreased. Moreover, the more coefficients to recover, the more ill-posed the problem becomes. The alternative approaches that aim at keeping a single regressor per condition and add also a temporal regularization constraint to fix the ill-posedness are the so-called regularized FIR methods [18–20]. Still, they do not overcome the low signal-to-noise ratio inherent to BOLD signals, and they lack robustness especially in non-activated regions. All the issues encountered in the previously mentioned approaches are linked to the sequential treatment of the detection and estimation tasks. Indeed, these two problems are strongly linked: on the one hand, a precise localization of brain activated areas strongly depends on a reliable HRF model; on the other hand, a robust estimation of the HRF

Lotfi CHAARI, Thomas VINCENT and Florence FORBES are with the Mistis team at Inria Grenoble Rhône-Alpes, 655 avenue de l'Europe, Montbonnot, 38334 Saint Ismier Cedex, France, and University Joseph Fourier, Grenoble, France. Lotfi CHAARI and Thomas VINCENT are also affiliated to CEA/DSV/I^{2}BM/Neurospin, CEA Saclay, Bât. 145, Point Courrier 156, 91191 Gif-sur-Yvette cedex, France. E-mail: {firstname.lastname}@inria.fr. Philippe CIUCIU is with CEA/DSV/I^{2}BM/Neurospin, CEA Saclay, Bât. 145, Point Courrier 156, 91191 Gif-sur-Yvette cedex, France. E-mail: {philippe.ciuciu}@cea.fr. Michel DOJAT is with INSERM, U836, GIN and University Joseph Fourier, Grenoble, France. E-mail: Michel.Dojat@ujf-grenoble.fr.

is only possible in activated areas where enough relevant signal is measured [21]. This interdependence and retroactivity has motivated the idea to jointly perform these two tasks [22–24] (detection and estimation) in a Joint Detection-Estimation (JDE) framework [25] which is the basis of the approach developed in this paper. To improve the estimation robustness, a gain in HRF reproducibility is performed by spatially aggregating signals so that a constant HRF shape is locally considered across a small group of voxels, *i.e.* a region or a parcel. The procedure then implies a partitioning of the data into functionally homogeneous parcels, in the form of a cerebral parcellation [26]. As will be recalled in more detail in Section II, the JDE approach rests upon three main elements: *i.)* a non-parametric or FIR parcel-level modeling of the HRF shape; *ii.)* prior information about the temporal smoothness of the HRF to guarantee its physiologically plausible shape; and *iii.)* the modeling of spatial correlation between the response magnitudes of neighboring voxels within each parcel using condition-specific discrete hidden Markov fields. In [22, 23, 25], posterior inference is carried out in a Bayesian setting using a computationally intensive Markov Chain Monte Carlo (MCMC) method, which is computationally intensive and requires fine tuning of several parameters.

In this paper, we reformulate the complete JDE framework [25] as a missing data problem and propose a simplification of its estimation procedure. We resort to a variational approximation using a Variational Expectation Maximization (VEM) algorithm in order to derive estimates of the HRF and stimulus-related activity. Variational approximations have been widely and successfully employed in the context of fMRI data analysis: *i.)* to model auto-regressive noise in the context of a Bayesian GLM [27]; *ii.)* to characterize cerebral hierarchical dynamic models [28]; *iii.)* to model transient neuronal signals in a Bayesian dynamical system [29] or *iv.)* to perform inference of spatial mixture models for the segmentation of GLM effect maps [30]. As in our study, the primary goal of resorting to variational approximations is to alleviate the computational burden associated with stochastic MCMC approaches. Akin to [30], we aim at comparing the stochastic and variational-based inference schemes, but on the more complex matter of detecting activation *and* estimating the HRF whereas [30] treated only a detection problem.

Compared to the JDE MCMC implementation, the proposed approach does not require priors on the model parameters for inference to be carried out. However, such priors may be injected in the adopted model for more robustness and to make the proposed approach fully auto-calibrated. Experiments on artificial and real data demonstrate the good performance of our VEM algorithm. Compared to the MCMC implementation, VEM is more computationally efficient, robust to mis-specification of the parameters, to deviations from the model, and adaptable to various experimental conditions. This increases considerably the potential impact of the JDE framework and makes its application to fMRI studies in cognitive and clinical neuroscience easier and more valuable. This new framework has also the advantage of providing straightforward criteria for model selection.

The rest of this paper is organized as follows. In Section II, we introduce the hierarchical Bayesian model for the JDE framework in the within-subject fMRI context. In Section III, the VEM algorithm based on variational approximations for inference is described. Evaluation on real and artificial fMRI datasets are reported in Section IV and the performance comparison between the MCMC and VEM implementations is carried out in Section V. Finally, Section VI discusses the pros and cons of the proposed approach and some perspectives.

II. BAYESIAN FRAMEWORK FOR THE JOINT DETECTION-ESTIMATION

Matrices and vectors are denoted with bold upper and lower case letters (*e.g.* $\boldsymbol{P}$ and $\boldsymbol{\mu}$). A vector is by convention a column vector. The transpose is denoted by $^{\mathrm{t}}$. Unless stated otherwise, subscripts j, m, i and n are respectively indexes over voxels, stimulus types, mixture components and time points. The Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is denoted by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The main acronyms used in the paper are defined in Table I. Table II gathers definitions of the main variables and parameters. A graphical representation of the model is given in Fig. 1.

A. The parcel-based model

We first recast the parcel-based JDE model proposed in [23, 25] in a missing data framework. Let us assume that the brain is decomposed in $\boldsymbol{\mathcal{P}} = (\mathcal{P}_\gamma)_{\gamma=1:\Upsilon}$ parcels, each of them containing J_γ voxels and having homogeneous hemodynamic properties. The fMRI time series $\boldsymbol{y}_j$ is measured in voxel $j \in \mathcal{P}_\gamma$ at times $(t_n)_{n=1...N}$, where $t_n = nTR$, N being the number of scans and TR, the time of repetition. The number of different stimulus types or experimental conditions is M. For a given parcel $\mathcal{P}_\gamma$ containing a group of connected voxels, a unique BOLD signal model is used in order to link the observed data $\boldsymbol{Y} = \{\boldsymbol{y}_j \in \mathbb{R}^N, j \in \mathcal{P}_\gamma\}$ to the unknown HRF $\boldsymbol{h}_\gamma \in \mathbb{R}^{D+1}$ specific to $\mathcal{P}_\gamma$, and also to the unknown response amplitudes $\boldsymbol{A} = \{\boldsymbol{a}^m, m = 1\ldots M\}$ with $\boldsymbol{a}^m = \{a_j^m, j \in \mathcal{P}_\gamma\}$, a_j^m being the magnitude at voxel j for condition m. More specifically, the observation model at each voxel $j \in \mathcal{P}_\gamma$ is expressed as follows [23]:

$$\boldsymbol{y}_j = \boldsymbol{S}_j \boldsymbol{h}_\gamma + \boldsymbol{P}\boldsymbol{\ell}_j + \boldsymbol{b}_j, \quad \text{with} \quad \boldsymbol{S}_j = \sum_{m=1}^{M} a_j^m \boldsymbol{X}_m \qquad (1)$$

where $\boldsymbol{S}_j \boldsymbol{h}_\gamma$ is the summation of the stimulus-induced components of the BOLD signal. The binary matrix $\boldsymbol{X}_m = \{x_m^{n-d\Delta t}, n = 1\ldots N, d = 0\ldots D\}$ is of size $N \times (D+1)$ and provides information on the stimulus occurrences for the m-th experimental condition, $\Delta t < TR$ being the sampling period of the unknown HRF $\boldsymbol{h}_\gamma = \{h_{d\Delta t}, d = 0\ldots D\}$ in $\mathcal{P}_\gamma$. This hemodynamic response is a consequence of the neuronal excitation which is commonly assumed to occur following stimulation. The scalars a_j^m's are weights that model the response magnitude evoked by the stimuli, whose occurrences are informed by the $\boldsymbol{X}_m$ matrices ($m = 0\ldots M$). They model the transition between stimuli and the vascular response informed by the filter $\boldsymbol{h}_\gamma$. It follows that the a_j^m's are generally referred to as Neural Response Levels (NRL).

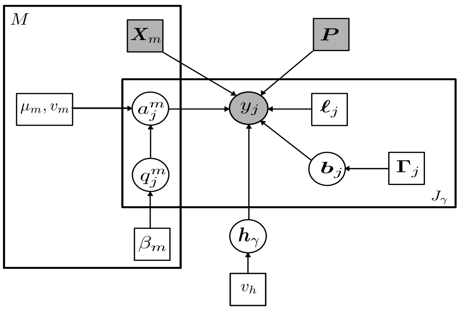

Fig. 1. Graphical model describing dependencies between *latent* and *observed* variables involved in the JDE generative model for a given parcel $\mathcal{P}_\gamma$ with J_γ voxels. Circles and squares indicate random variables and model parameters, respectively. Observed variables and fixed parameters are shaded. We used standard graphical notations where plates represent multiple similar nodes with their number given in the plate.

The rest of the signal is made of matrix $\boldsymbol{P}$, which corresponds to physiological artifacts accounted for via a low frequency orthonormal function basis of size $N \times O$. With each voxel j is associated a vector of low frequency drifts $\boldsymbol{\ell}_j \in \mathbb{R}^O$ which has to be estimated. Within parcel $\mathcal{P}_\gamma$, these vectors may be grouped into the same matrix $\boldsymbol{L} = \{\boldsymbol{\ell}_j, j \in \mathcal{P}_\gamma\}$. As regards observation noise, the $\boldsymbol{b}_j$'s are assumed to be independent with $\boldsymbol{b}_j \sim \mathcal{N}(0, \boldsymbol{\Gamma}_j^{-1})$ at voxel j (see Section II-B.1 for more details). The set of all unknown precision matrices (inverse of the covariance matrices) is denoted by $\boldsymbol{\Gamma} = \{\boldsymbol{\Gamma}_j, j \in \mathcal{P}_\gamma\}$. The forward BOLD model expressed in Eq. (1) relies on the classical assumption of a linear and time invariant system which is adopted in the GLM framework [2]. Indeed, it can easily be recast in the same formulation where the response magnitudes a_j^m's and drift coefficient $\boldsymbol{\ell}_j$'s are equivalent to the effects associated with stimulus-induced and low frequency basis regressors, respectively. However, the JDE forward model generalizes the GLM model since the hemodynamics filter is unknown. Finally, detection is handled through the introduction of activation class assignments $\boldsymbol{Q} = \big\{\boldsymbol{q}^m, m = 1 \dots M\big\}$ where $\boldsymbol{q}^m = \big\{q_j^m, j \in \mathcal{P}_\gamma\big\}$ and q_j^m represents the *activation class* at voxel j for condition m. The NRL coefficients will therefore be expressed conditionally to these hidden variables. In other words, the NRL coefficients will depend on the activation status of the voxel j, which itself depends on the activation status of neighbouring voxels thanks to a Markov model used as a spatial prior on $\boldsymbol{Q}$ (*cf* Section II-B.2.c). Without loss of generality, we consider here two activation classes akin to [25] (activated and non-activated voxels). An additional deactivation class may be considered depending on the experiment as proposed within the JDE context in [31]. In the following developments, all provided formulas are general enough to cover this case.

B. A hierarchical Bayesian Model

In a Bayesian framework, we first need to define the likelihood and prior distributions for the model variables $(\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{h}_\gamma, \boldsymbol{Q})$ and parameters $(\boldsymbol{\Theta})$. Using the hierarchical structure between $\boldsymbol{Y}$, $\boldsymbol{A}$, $\boldsymbol{h}_\gamma$, $\boldsymbol{Q}$ and $\boldsymbol{\Theta}$, the complete model is given by the joint distribution of both the observed and unobserved (or missing) data: $p(\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{h}_\gamma, \boldsymbol{Q}; \boldsymbol{\Theta}) = p(\boldsymbol{Y} \,|\, \boldsymbol{A}, \boldsymbol{h}_\gamma; \boldsymbol{\Theta}) \quad p(\boldsymbol{h}_\gamma; \boldsymbol{\Theta}) p(\boldsymbol{A} \,|\, \boldsymbol{Q}; \boldsymbol{\Theta}) \; p(\boldsymbol{Q}; \boldsymbol{\Theta})$. To fully define the hierarchical model, we now specify each term.

TABLE I

ACRONYMS USED IN THE JDE MODEL PRESENTATION AND INFERENCE.

Acronym	Definition
JDE	Joint Detection-Estimation
HRF	Hemodynamic Response Function
FIR	Finite Impulse Response
NRL	Neural Response Level
PPM	Posterior Probability Map
PV	HRF Peak Value: $\max\,\{h_{d\Delta t}\}_{d=0:D}$
TTP	HRF Time-to-Peak: $\Delta t \times \arg\max\limits_{d}\,\{h_{d\Delta t}\}_{d=0:D}$
FWHM	Full Width at Half Maximum: $\Delta t \times (d_2 - d_1)$ such that $h_{d_1\Delta t} = h_{d_2\Delta t} = \mathrm{PV}/2, d_1 < d_2$
TTU	HRF Time-to-Undershoot: $\Delta t \times \arg\min\limits_{d>d_2}\,\{h_{d\Delta t}\}_{d=0:D}$
ISI	Inter Stimuli Interval

1) Likelihood:

The definition of the likelihood depends on the noise model assumptions. In [23, 32], an autoregressive (AR) noise model has been adopted to account for serial correlations in fMRI time series. It has also been shown in [23] that a spatially-varying first-order AR noise model helps control the false positive rate. In the same context, we will assume such a noise model $\boldsymbol{b}_j \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}_j^{-1})$ with $\boldsymbol{\Gamma}_j = \sigma_j^{-2}\boldsymbol{\Lambda}_j$ where $\boldsymbol{\Lambda}_j$ is a tridiagonal symmetric matrix which depends on the AR(1) parameter ρ_j [23]: $(\boldsymbol{\Lambda}_j)_{1,1} = (\boldsymbol{\Lambda}_j)_{N,N} = 1$, $(\boldsymbol{\Lambda}_j)_{n,n} = 1 + \rho_j^2$ for $n = 2 : N - 1$ and $(\boldsymbol{\Lambda}_j)_{n+1,n} = (\boldsymbol{\Lambda}_j)_{n,n+1} = -\rho_j$ for $n = 1 : N - 1$. These parameters are assumed voxel-specific due to their tissue-dependence [33, 34]. Denoting $\boldsymbol{\theta}_0 = (\sigma_j^2, \rho_j)_{1 \le j \le J_\gamma}$ and $\overline{\boldsymbol{y}}_j = \boldsymbol{y}_j - \boldsymbol{P}\boldsymbol{\ell}_j - \boldsymbol{S}_j\boldsymbol{h}_\gamma$, the likelihood can be factorized over voxels as follows:

$$p(\boldsymbol{Y} \,|\, \boldsymbol{A}, \boldsymbol{h}_\gamma; \boldsymbol{L}, \boldsymbol{\theta}_0) \propto \prod_{j=1}^{J_\gamma} \det\boldsymbol{\Lambda}_{\mathrm{j}}^{1/2}\sigma_{\mathrm{j}}^{-\mathrm{N}} \exp\big(-\frac{\overline{\boldsymbol{y}}_{\mathrm{j}}^{\mathrm{t}}\boldsymbol{\Lambda}_{\mathrm{j}}\overline{\boldsymbol{y}}_{\mathrm{j}}}{2\sigma_{\mathrm{j}}^2}\big). \tag{2}$$

2) Model priors:

a) Hemodynamic response function:

Akin to [23, 25], we introduce constraints in the HRF prior that favor smooth variations in $\boldsymbol{h}_\gamma$ by controlling its second order derivative: $\boldsymbol{h}_\gamma \sim \mathcal{N}(0, v_{\boldsymbol{h}}\boldsymbol{R})$ with $\boldsymbol{R} = (\Delta t)^4 \; (\boldsymbol{D}_2^{\mathrm{t}}\boldsymbol{D}_2)^{-1}$ where $\boldsymbol{D}_2$ is the second-order finite difference matrix and $v_{\boldsymbol{h}}$ is a parameter to be estimated. Moreover, boundary constraints have also been fixed on $\boldsymbol{h}_\gamma$ as in [23, 25] so that $h_0 = h_{D\Delta t} = 0$. The prior assumption expressed on the HRF amounts to a smooth FIR model introduced in [18] and is flexible enough to recover *any* HRF shape.

b) Neural response levels:

Akin to [23, 25], the NRLs a_j^m are assumed to be statistically independent across conditions: $p(\boldsymbol{A}; \boldsymbol{\theta}_{\boldsymbol{a}}) = \prod_{m=1}^{M} p(\boldsymbol{a}^m; \boldsymbol{\theta}_m)$ where $\boldsymbol{\theta}_{\boldsymbol{a}} = \{\boldsymbol{\theta}_m, m = 1 \dots M\}$ and $\boldsymbol{\theta}_m$ gathers the parameters for the m-th condition. A mixture model is then adopted by using the assignment variables q_j^m to segregate non-activated voxels ($q_j^m = 0$) from activated ones ($q_j^m = 1$). For the m-th condition, and conditionally to the assignment variables $\boldsymbol{q}^m$, the NRLs are assumed to be independent: $p(\boldsymbol{a}^m \,|\, \boldsymbol{q}^m; \boldsymbol{\theta}_m) = \prod_{j \in \mathcal{P}_\gamma} p(a_j^m \,|\, q_j^m; \boldsymbol{\theta}_m)$. If $q_j^m = i$ then $p(a_j^m \,|\, q_j^m = i; \boldsymbol{\theta}_m) \sim \mathcal{N}(\mu_{im}, v_{im})$. It is worth noting that the Gaussianity assumed for a given NRL is similar to the assumption of Gaussian effects in the classical GLM context [10]. The Gaussian parameters $\boldsymbol{\theta}_m =$

TABLE II

NOTATIONS FOR VARIABLES AND PARAMETERS USED IN THE MODEL FOR A GIVEN PARCEL $\mathcal{P}_\gamma$ WITH J_γ VOXELS.

	Notation	Definition	Dimension
Variables	$\boldsymbol{y}_j \in \mathbb{R}^N$	Observed BOLD signal at voxel j	N
	$\boldsymbol{b}_j \in \mathbb{R}^N$	Acquisition noise vector at voxel j	N
	$\boldsymbol{h}_\gamma = (h_{d\Delta t})_{d=0\dots D} \in \mathbb{R}^{D+1}$	HRF sampled at Δ_t	$D+1$
	$a_j^m \in \mathbb{R}$	NRL at voxel j for condition m	1
	$\boldsymbol{a}^m = \{a_j^m, j \in \mathcal{P}_\gamma\} \in \mathbb{R}^{J_\gamma}$	NRLs for condition m	J_γ
	$q_j^m \in \{0,1\}$	Activation class for voxel j and condition m	1
	$\boldsymbol{q}^m = \{q_j^m, j \in \mathcal{P}_\gamma\} \in \{0,1\}^{J_\gamma}$	Activation classes for condition m	J_γ
Unknowns	$\boldsymbol{\ell}_j \in \mathbb{R}^O$	Low frequency drifts for voxel j	O
	$\boldsymbol{\Gamma}_j \in \mathbb{R}^{N\times N}$	Noise precision matrix for voxel j	$N \times N$
	$\boldsymbol{\mu}_m = \{\mu_{0m}, \mu_{1m}\} \in \mathbb{R}^2$	Mixture model means for NRLs in condition m	2
	$\boldsymbol{v}_m = \{v_{0m}, v_{1m}\} \in \mathbb{R}_+^2$	Mixture model variances for NRLs in condition m	2
	$\beta_m \in \mathbb{R}_+$	Potts regularization parameter for condition m	1
	$v_{\boldsymbol{h}} \in \mathbb{R}_+$	HRF prior parameter	1
Fixed	$\boldsymbol{X}_m \in \{0,1\}^{N\times D+1}$	Binary stimulus occurrence matrix for condition m	$N \times (D+1)$
	$\boldsymbol{P} \in \mathbb{R}^{N\times O}$	Low frequency orthonormal function basis	$N \times O$

$\{\mu_{im}, v_{im}, i=0,1\}$ are unknown. For the sake of conciseness, we rewrite $\boldsymbol{\theta_a} = (\boldsymbol{\mu}, \boldsymbol{v})$ where $\boldsymbol{\mu} = \{\boldsymbol{\mu}_m, m = 1\dots M\}$ with $\boldsymbol{\mu}_m = \{\mu_{0m}, \mu_{1m}\}$ and $\boldsymbol{v} = \{\boldsymbol{v}_m, m = 1\dots M\}$ with $\boldsymbol{v}_m = \{v_{0m}, v_{1m}\}$. More specifically, for non-activated voxels we set for all m, $\mu_{0m} = 0$.

c) Activation classes:

As in [25], we assume prior independence between the M experimental conditions regarding the activation class assignments. It follows that $p(\boldsymbol{Q}) = \prod_{m=1}^{M} p(\boldsymbol{q}^m; \beta_m)$ where we assume in addition that $p(\boldsymbol{q}^m; \beta_m)$ is a Markov random field prior, namely a Potts model. Such prior modeling assumption is consistent with the physiological properties of the fMRI signal where the activity is known to be correlated in space [33, 35]. Here, the prior Potts model with interaction parameter β_m [25] is expressed as:

$$p(\boldsymbol{q}^m; \beta_m) = Z(\beta_m)^{-1} \exp\big(\beta_m \sum_{j\sim k} I(q_j^m = q_k^m)\big) \quad (3)$$

and where $Z(\beta_m)$ is the normalizing constant and for all $(a,b) \in \mathbb{R}^2$, $I(a=b) = 1$ if $a = b$ and 0 otherwise. The notation $j \sim k$ means that the summation is over all neighboring voxels. The unknown parameters are denoted by $\boldsymbol{\beta} = \{\beta_m, m = 1\dots M\}$. In what follows, we will consider a 6-connexity 3D neighboring system.

For the complete model, the whole set of parameters is denoted by $\boldsymbol{\Theta} = \{\boldsymbol{\Gamma}, \boldsymbol{L}, \boldsymbol{\theta_a}, v_{\boldsymbol{h}}, \boldsymbol{\beta}\}$ and belong to a set $\underline{\boldsymbol{\Theta}}$.

III. ESTIMATION BY VARIATIONAL EXPECTATION-MAXIMIZATION

We propose to use an Expectation-Maximization (EM) framework to deal with the missing data namely, $\boldsymbol{A} \in \mathcal{A}$, $\boldsymbol{h}_\gamma \in \mathcal{H}$, $\boldsymbol{Q} \in \mathcal{Q}$.

A. Variational Expectation-Maximization principle

Let $\mathcal{D}$ be the set of all probability distributions on $\mathcal{A} \times \mathcal{H} \times \mathcal{Q}$. EM can be viewed [36] as an alternating maximization procedure of a function $\mathcal{F}$ on $\mathcal{D}$, for all $q \in \mathcal{D}$,

$$\mathcal{F}(q, \boldsymbol{\Theta}) \quad = \quad \mathrm{E}_q\big[\log p(\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{h}_\gamma, \boldsymbol{Q} \,|\, \boldsymbol{\Theta})\big] + \mathcal{G}(q) \quad (4)$$

where $\mathrm{E}_q\big[.\big]$ denotes the expectation with respect to q and $\mathcal{G}(q) = -\mathrm{E}_q\big[\log q(\boldsymbol{A}, \boldsymbol{h}_\gamma, \boldsymbol{Q})\big]$ is the entropy of q. This function is called the free energy functional. It can be equivalently expressed in terms of the log-likelihood as $\mathcal{F}(q, \boldsymbol{\Theta}) = \log p(\boldsymbol{Y} \,|\, \boldsymbol{\Theta}) - KL(q \,||\, p(\boldsymbol{A}, \boldsymbol{h}_\gamma, \boldsymbol{Q} \,|\, \boldsymbol{Y}, \boldsymbol{\Theta}))$ where $KL(q \,||\, p(\boldsymbol{A}, \boldsymbol{h}_\gamma, \boldsymbol{Q} \,|\, \boldsymbol{Y}, \boldsymbol{\Theta}))$ is the Kullback-Leibler (KL) divergence between q and $p(\boldsymbol{A}, \boldsymbol{h}_\gamma, \boldsymbol{Q} \,|\, \boldsymbol{Y}, \boldsymbol{\Theta})$:

$$KL(q \,||\, p(\boldsymbol{A}, \boldsymbol{h}_\gamma, \boldsymbol{Q} \,|\, \boldsymbol{Y}, \boldsymbol{\Theta})) = \int q(\boldsymbol{A}, \boldsymbol{h}_\gamma, \boldsymbol{Q}) \log \frac{q(\boldsymbol{A}, \boldsymbol{h}_\gamma, \boldsymbol{Q})}{p(\boldsymbol{A}, \boldsymbol{h}_\gamma, \boldsymbol{Q} \,|\, \boldsymbol{Y}, \boldsymbol{\Theta})} \, d\boldsymbol{A} \, d\boldsymbol{h}_\gamma \, d\boldsymbol{Q}. \quad (5)$$

Hence maximizing the free energy with respect to q amounts to minimizing the Kullback-Leibler divergence between q and the posterior distribution of interest $p(\boldsymbol{A}, \boldsymbol{h}_\gamma, \boldsymbol{Q} \,|\, \boldsymbol{Y}, \boldsymbol{\Theta})$. Since the KL divergence is always non-negative, and because the KL divergence of the posterior distribution to itself is zero, it follows easily that the maximum free energy over all $q \in \mathcal{D}$ is the log-likelihood. The link to the EM algorithm follows straightforwardly. At iteration (r), denoting the current parameter values by $\boldsymbol{\Theta}^{(r-1)}$, the alternating procedure proceeds as follows:

$$\textbf{E-step: } p^{(r)}_{A,H_\gamma,Q} = \arg\max_{p \in \mathcal{D}} \; \mathcal{F}(p, \boldsymbol{\Theta}^{(r-1)}) \quad (6)$$

$$\textbf{M-step: } \boldsymbol{\Theta}^{(r)} = \arg\max_{\Theta \in \underline{\boldsymbol{\Theta}}} \; \mathcal{F}(p^{(r)}_{A,H_\gamma,Q}, \boldsymbol{\Theta}) \quad (7)$$

However, the optimization step in Eq. (6) leads to $p^{(r)}_{A,H_\gamma,Q} = p(\boldsymbol{A}, \boldsymbol{h}_\gamma, \boldsymbol{Q} \,|\, \boldsymbol{Y}, \boldsymbol{\Theta}^{(r-1)})$, which is intractable for our model. Hence, we resort to a variational EM (VEM) variant in which the intractable posterior is approximated by constraining the space of possible q distributions in order to make the maximization procedure tractable. In that case, the free energy optimal value reached is only a lower bound on the log-likelihood. The most common variational approximation consists of optimizing over the distributions in $\mathcal{D}$ that factorize as a product of three pdfs on $\mathcal{A}$, $\mathcal{H}$ and $\mathcal{Q}$ respectively.

Previous attempts to use variational inference [37, 38] and in particular in fMRI [27, 30] have been successful, with this type of approximations usually validated by assessing its fidelity to its MCMC counterpart. In Section IV, we will also provide such a comparison. The actual consequences of the factorization may vary with the models under study. Some couples of latent variables may capture more dependencies that would then need to be kept whereas others may induce only weak local correlation at the expense of a long-range correlation which to first order can be ignored (see [39] for more details on the consequences of the factorization for particular models). The way it may affect inference is that often variational approximations are shown to lead to underestimated variances and consequently to confidence intervals that are too narrow. Note that [38] suggested that nonparametric bootstrap intervals whenever possible may alleviate this issue. Also, when the concern is the computation of maximum *a posteriori* estimates, the required ingredients for designing an accurate variational approximation lie in the shape of the optimized free energy. All that is needed for our inference to work well is that the optimized free energy have a similar shape (mode and curvature) to the target log-likelihood whenever the likelihood is relatively large. As a matter of fact, there are cases, such as mixtures of distributions from the exponential family, where the variational estimator is asymptotically consistent. The experiments in [38] even report very accurate confidence intervals. Unfortunately no general theoretical results exist that would include our case to guarantee the accuracy of estimates based on the variational approximation. The other few cases for which the variational approach has shown good theoretical properties are to the best of our knowledge simpler than our setting. The fact that the HRF can be equivalently considered as a missing variable or a random parameter induces some similarity between our VEM variant and the Variational Bayesian EM algorithm in [37]. Our framework varies slightly from the case of conjugate exponential models described in [37] and more importantly, our presentation offers the possibility to deal with extra parameters $\mathbf{\Theta}$ for which prior information may not be available. This is done in a maximum likelihood manner and avoids tusing non-informative priors that could be problematic [40, pp. 64-65]. Consequently, the variational Bayesian M-step of [37] is transferred into our E-step while our M-step has no equivalent in the formulation of [37].

B. Variational Joint Detection-Estimation

We propose here to use a EM variant in which the intractable E-step is instead solved over $\widetilde{\mathcal{D}}$, a restricted class of probability distributions chosen as the set of distributions that factorize as $\widetilde{p}_{A,H_\gamma,Q} = \widetilde{p}_A \widetilde{p}_{H_\gamma} \widetilde{p}_Q$ where $\widetilde{p}_A$, $\widetilde{p}_{H_\gamma}$ and $\widetilde{p}_Q$ are probability distributions on $\mathcal{A}$, $\mathcal{H}$ and $\mathcal{Q}$, respectively. It follows then that our E-step becomes an approximate E-step, which can be further decomposed into three stages that consist of updating the three pdfs, $\widetilde{p}_{H_\gamma}$, $\widetilde{p}_A$ and $\widetilde{p}_Q$ in turn using three equivalent expressions of $\mathcal{F}$ when p factorizes as in $\widetilde{\mathcal{D}}$. At iteration (r) with current estimates denoted by $\widetilde{p}^{(r-1)}_H, \widetilde{p}^{(r-1)}_A, \widetilde{p}^{(r-1)}_Q$ and $\mathbf{\Theta}^{(r-1)}$, the updating rules become:

$$\textbf{VE-H:}\ \ \widetilde{p}^{(r)}_{H_\gamma} = \arg\max_{p_{H_\gamma}} \mathcal{F}(\widetilde{p}^{(r-1)}_A \; p_{H_\gamma} \; \widetilde{p}^{(r-1)}_Q, \mathbf{\Theta}^{(r-1)})$$

$$\textbf{VE-A:}\ \ \widetilde{p}^{(r)}_{A} = \arg\max_{p_{A}} \mathcal{F}(p_A \; \widetilde{p}^{(r)}_{H_\gamma} \; \widetilde{p}^{(r-1)}_Q, \mathbf{\Theta}^{(r-1)})$$

$$\textbf{VE-Q:}\ \ \widetilde{p}^{(r)}_{Q} = \arg\max_{p_{Q}} \mathcal{F}(\widetilde{p}^{(r)}_A \; \widetilde{p}^{(r)}_{H_\gamma} \; p_Q, \mathbf{\Theta}^{(r-1)}).$$

In other words, the factorization is used to maximize the free energy by alternately maximizing it with respect to p_{H_γ}, p_A and p_Q while keeping the other distributions fixed. The steps above can then be equivalently written in terms of minimizations of some Kullback-Leibler divergences. The properties of the latter lead to the following solutions (see Appendix A for details):

VE-H: (8)

$$\widetilde{p}^{(r)}_{H_\gamma}(\boldsymbol{h}_\gamma) \propto \exp\left(\mathrm{E}_{\widetilde{p}^{(r-1)}_A \widetilde{p}^{(r-1)}_Q}\left[\log p(\boldsymbol{h}_\gamma \,|\, \boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{Q}; \mathbf{\Theta}^{(r-1)}\right]\right)$$

VE-A: (9)

$$\widetilde{p}^{(r)}_{A}(\boldsymbol{A}) \propto \exp\left(\mathrm{E}_{\widetilde{p}^{(r)}_{H_\gamma} \widetilde{p}^{(r-1)}_Q}\left[\log p(\boldsymbol{A} \,|\, \boldsymbol{Y}, \boldsymbol{h}_\gamma, \boldsymbol{Q}; \mathbf{\Theta}^{(r-1)})\right]\right)$$

VE-Q: (10)

$$\widetilde{p}^{(r)}_{Q}(\boldsymbol{Q}) \propto \exp\left(\mathrm{E}_{\widetilde{p}^{(r)}_A \widetilde{p}^{(r)}_{H_\gamma}}\left[\log p(\boldsymbol{Q} \,|\, \boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{h}_\gamma; \mathbf{\Theta}^{(r-1)})\right]\right) \; .$$

The corresponding **M-step** is (since $\mathbf{\Theta}$ and $\mathcal{G}(p^{(r)}_{A,H_\gamma,Q})$ are independent, see Eq. (4)):

$$\textbf{M:}\ \ \mathbf{\Theta}^{(r)} = \arg\max_{\mathbf{\Theta}} \; \mathrm{E}_{\widetilde{p}^{(r)}_A \widetilde{p}^{(r)}_{H_\gamma} \widetilde{p}^{(r)}_Q}\left[\log p(\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{h}_\gamma, \boldsymbol{Q}; \mathbf{\Theta})\right] \; . \tag{11}$$

These steps lead to explicit calculations for $\widetilde{p}^{(r)}_{H_\gamma}$, $\widetilde{p}^{(r)}_A$, $\widetilde{p}^{(r)}_Q$ and the parameter set $\mathbf{\Theta}^{(r)} = \left\{\mathbf{\Gamma}^{(r)}, \boldsymbol{L}^{(r)}, \boldsymbol{\theta}^{(r)}_{\boldsymbol{a}}, v^{(r)}_{\boldsymbol{h}}, \boldsymbol{\beta}^{(r)}\right\}$. Although the approximation by a factorized distribution may have seemed initially drastic, the equations it leads to are coupled. More specifically, the approximation consists of replacing stochastic dependencies between latent variables by deterministic dependencies between moments of these variables (see Appendices A to C as mentioned below).

- **VE-H** step: From Eq. (8) standard algebra enables to derive that $\widetilde{p}^{(r)}_{H_\gamma}$ is a Gaussian distribution $\widetilde{p}^{(r)}_{H_\gamma} \sim \mathcal{N}(\boldsymbol{m}^{(r)}_{H_\gamma}, \mathbf{\Sigma}^{(r)}_{H_\gamma})$ whose parameters are detailed in Appendix B. The expressions for $\boldsymbol{m}^{(r)}_{H_\gamma}$ and $\mathbf{\Sigma}^{(r)}_{H_\gamma}$ are similar to those derived in the MCMC case [23, Eq. (B.1)] with expressions involving the a^m_j's replaced by their expectations with respect to (wrt) $\widetilde{p}^{(r-1)}_A$.
- **VE-A** step : Using Eq. (9), standard algebra rules allow to identify the Gaussian distribution of $\widetilde{p}^{(r)}_A$ which writes as $\widetilde{p}^{(r)}_A = \prod_{j \in \mathcal{P}_\gamma} \widetilde{p}^{(r)}_{A_j}$ with $\widetilde{p}^{(r)}_{A_j} \sim \mathcal{N}(\boldsymbol{m}^{(r)}_{A_j}, \mathbf{\Sigma}^{(r)}_{A_j})$. More detail about the update of $\widetilde{p}^{(r)}_A$ is given in Appendix C. The relationship with the MCMC update of $\boldsymbol{A}$ is not straightforward. In [23, 25], the a^m_j's are sampled independently and conditionally on the q^m_j's. This is not the case in the VEM framework but some similarity appears if we set the probabilities $\widetilde{p}^{(r-1)}_{Q^m_j}(i)$ either to 0 or 1 and consider only the diagonal part of $\mathbf{\Sigma}^{(r)}_{A_j}$.
- **VE-Q** step: Using the expressions of $p(\boldsymbol{A}|\boldsymbol{Q})$ and $p(\boldsymbol{Q})$ in Section II, Eq. (10) yields $\widetilde{p}^{(r)}_Q(\boldsymbol{Q}) = \prod_{m=1}^{M} \widetilde{p}^{(r)}_{Q^m}(\boldsymbol{q}^m)$ which is intractable due to the Markov random field prior. To overcome this difficulty, a number of approximation techniques are available. To decrease the computational complexity of our VEM algorithm and to avoid introducing additional variables as done in [30], we use a mean-field like algorithm which consists of

fixing the neighbours to their mean value. Following [41], $\widetilde{p}^{(r)}_{Q^m}(\boldsymbol{q}^m)$ can be approximated by a factorized density $\widetilde{p}^{(r)}_{Q^m}(\boldsymbol{q}^m) = \prod_{j \in \mathcal{P}_\gamma} \widetilde{p}^{(r)}_{Q^m_j}(q^m_j)$ such that if $q^m_j = i$, $\widetilde{p}^{(r)}_{Q^m_j}(i) \propto \mathcal{N}(\boldsymbol{m}^{(r)}_{A^m_j}; \mu^{(r-1)}_{im}, v^{(r-1)}_{im})\ f(q^m_j = i \,|\, \tilde{q}^m_{\sim j}; \beta^{(r-1)}_m, \boldsymbol{v}^{(r-1)}_m)$ where $\tilde{\boldsymbol{q}}^m$ is a particular configuration of $\boldsymbol{q}^m$ updated at each iteration according to a specific scheme, $\sim j$ denotes neighbouring voxels to j on the brain volume and $f(q^m_j = i \,|\, \tilde{q}^m_{\sim j}; \beta^{(r-1)}_m, \boldsymbol{v}^{(r-1)}_m) \propto \exp\{\frac{v^{(r)}_{A^m_j A^m_j}}{v^{(r-1)}_{im}} + \beta^{(r-1)}_m \sum_{k \sim j} I(\tilde{q}^m_k = i)\}$. Hereabove, $m^{(r)}_{A^m_j}$ and $v^{(r)}_{A^m_j A^{m'}_j}$ denote the m and (m, m') entries of the mean vector $(\boldsymbol{m}^{(r)}_{A_j})$ and covariance matrix $(\boldsymbol{\Sigma}^{(r)}_{A_j})$, respectively. The Gaussian distribution with mean μ_{im} and variance v_{im} is denoted by $\mathcal{N}(\ .\ ; \mu_{im}, v_{im})$, while $\tilde{q}^m_{\sim j} = \{\tilde{q}^m_k, k \sim j\}$. More details are given in Appendix D.

- **M** step: For this maximization step, we can first rewrite Eq. (11) as

$$
\begin{aligned}
\boldsymbol{\Theta}^{(r)} = \arg\max_{\boldsymbol{\Theta}} \Big[& \mathrm{E}_{\widetilde{p}^{(r)}_A \widetilde{p}^{(r)}_{H_\gamma}} \big[\log p(\boldsymbol{Y} \,|\, \boldsymbol{A}, \boldsymbol{h}_\gamma; \boldsymbol{L}, \boldsymbol{\Gamma})\big] \\
& + \mathrm{E}_{\widetilde{p}^{(r)}_A \widetilde{p}^{(r)}_Q} \big[\log p(\boldsymbol{A} \,|\, \boldsymbol{Q}; \boldsymbol{\mu}, \boldsymbol{v})\big] \qquad (12) \\
& + \mathrm{E}_{\widetilde{p}^{(r)}_{H_\gamma}} \big[\log p(\boldsymbol{h}_\gamma; v_{\boldsymbol{h}})\big] + \mathrm{E}_{\widetilde{p}^{(r)}_Q} \big[\log p(\boldsymbol{Q}; \boldsymbol{\beta})\big] \Big].
\end{aligned}
$$

The M-step can therefore be decoupled into four sub-steps involving separately $(\boldsymbol{L}, \boldsymbol{\Gamma})$, $(\boldsymbol{\mu}, \boldsymbol{v})$, $v_{\boldsymbol{h}}$ and $\boldsymbol{\beta}$. Some of these sub-steps admit closed-form expressions, while some other require resorting to iterative or alternate optimization. For more details about the related calculations, the interested reader can refer to Appendix E.

IV. Validation of the proposed approach

This section aims at validating the proposed variational approach. Synthetic and real contexts are considered respectively in sub-sections IV-A and IV-B. To corroborate the effectiveness of the proposed method, comparisons with its MCMC counterpart, as implemented in [25], will also be conducted throughout the present section. The two approaches have been tuned at best so as to make them as close as possible. This reduces essentially to moderate the effect of the MCMC priors. The hyper-priors have been parameterized by a set of hyper-parameters that have a limited impact on the priors themselves. For instance, for the mean and variance parameters involved in the mixture model $(\boldsymbol{\mu}, \boldsymbol{v})$, conjugate Gaussian $\mathcal{N}(0, w_\mu)$ and inverse-gamma $\mathcal{IG}(a_v, b_v)$ hyper-prior distributions have been considered whose parameters have been tuned by hand so as to make them flat while proper (eg, $w_\mu < +\infty$). For doing so, the relationship between the hyper-parameters (eg, a_v, b_v) and the statistical moments of the distribution have been carefully studied to guarantee a large variance or a large entropy in the latter. Moreover, given the above mentioned flatness of the hyper-priors, the MCMC approach is fairly robust to hyper-parameter setting.

A. Artificial fMRI datasets

In this section, experiments have been conducted on data simulated according to the observation model in Eq. (1) where $\boldsymbol{P}$ has been defined from a cosine transform basis as in [23]. The simulation process is illustrated in Fig. 2.

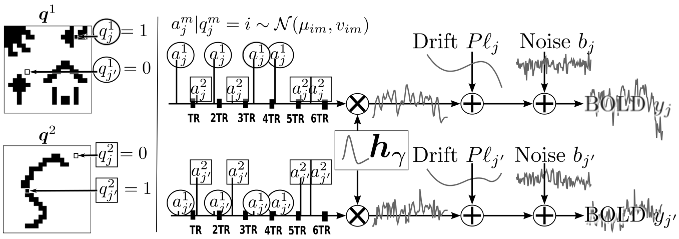

Fig. 2. Single parcel $(\mathcal{P}_\gamma)$ artificial data generation process using two experimental conditions ($M = 2$). From **left to right**: label maps $\boldsymbol{q}^m$ are hand-drawn. Conditionally on them, NRL values $\{a^m_j, j \in \mathcal{P}_\gamma\}$ are drawn from Gaussian distributions with means μ_{im} and variances v_{im} whenever $q^m_j = i$. Then, for any given voxel j, the stimulus sequence of a given experimental condition m is multiplied by the corresponding NRL value a^m_j. The resulting sequence is convolved with a normalized HRF $\boldsymbol{h}_\gamma$ which is common to all conditions and voxels. Nuisance signals are finally added (drift $\boldsymbol{P\ell}_j$ and noise $\boldsymbol{b}_j$) to form the artificial BOLD signal $\boldsymbol{y}_j$.

Different studies have then been conducted in order to validate the detection-estimation performance and robustness. For each of these studies, some simulation parameters have been changed such as the noise or the paradigm properties. Changing these parameters aims at providing for each simulation context a realistic BOLD signal while exploring various situations in terms of Signal-to-Noise Ratio (SNR).

1) Detection-Estimation performance: The first artificial data analysis was conducted on data simulated with a Gaussian white noise $\boldsymbol{\Gamma}^{-1}_j = 1.2\ \boldsymbol{I}_N$ ($\boldsymbol{I}_N$ is the N-dimensional identity matrix). Two experimental conditions have been considered ($M = 2$) while ensuring stimulus-varying Contrast-to-Noise Ratios (for condition m, $\mathrm{CNR}_m = 2(\mu_{1m} - \mu_{0m})^2/(v_{0m} + v_{1m}) = 2\mu^2_{1m}/(v_{0m} + v_{1m})$) achieved by setting $\mu_{11} = 2.8, \mu_{12} = 1.8$, $\mu_{01} = \mu_{02} = 0$, and $v_{11} = v_{12} = v_{01} = v_{02} = 0.5$, so that a higher CNR is simulated for the first experimental condition ($\mathrm{CNR}_{m=1} = 15.68$) compared to the second one ($\mathrm{CNR}_{m=2} = 6.48$). For each of these conditions, the *initial* artificial paradigm comprised 30 stimulus events. The simulation process finally yielded time-series of 268 time-points. Condition-specific activated and non-activated voxels ($\boldsymbol{Q}$ values) were defined on a 20×20 2D slice as shown in Fig. 2[left]. No parcellation was performed. Simulated NRLs and BOLD signal were thus assumed to belong to a single parcel of size 20×20.

The Posterior Probability Maps (PPM) obtained using MCMC and VEM are shown in Fig. 3[middle] and Fig. 3[right]. PPMs here correspond to the activation class assignment probabilities $\widehat{q}^m_j = p(q^m_j = 1 \,|\, \boldsymbol{y}, \boldsymbol{\Theta})$. These figures clearly show the gain in robustness provided by the variational approximation. This gain consists of lower miss-classification error (a lower false positive rate) illustrated by higher PPM values, especially for the experimental condition with the lowest CNR ($m = 2$).

For a quantitative evaluation, the Receiver Operating Characteristic (ROC) curves corresponding to the estimated PPMs using both algorithms were computed. As shown in Fig. 4, they confirm that both algorithms perform well at high CNR

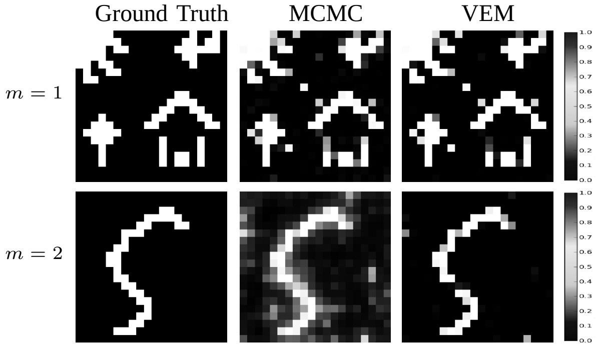

Fig. 3. Detection results for the first artificial data analysis. Ground truth $\boldsymbol{q}^m$ (**left**) and estimated Posterior Probability Maps (PPM) $\widehat{\boldsymbol{q}}^m$ using MCMC (**middle**) and VEM (**right**). Note that condition $m = 2$ (bottom row) is associated with a lower CNR than condition $m = 1$ (top row).

($m = 1$) and that the VEM scheme outperforms the MCMC implementation of the second experimental condition ($m = 2$).

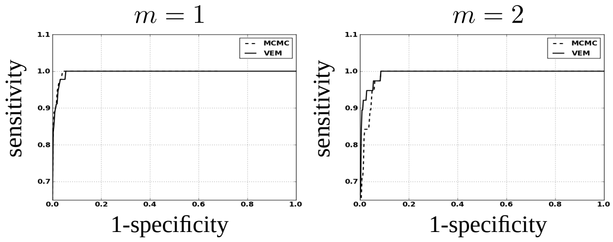

Fig. 4. Detection results for the first artificial data analysis. ROC curves associated with the label posteriors ($\widehat{\boldsymbol{q}}^m$) using VEM and MCMC. Condition $m = 1$ is associated with a higher CNR than condition $m = 2$. Curves are plotted in solid and dashed line for VEM and MCMC, respectively.

Fig. 5 shows the NRL estimates obtained by the two methods. Although some differences are exhibited on the PPMs, both algorithms report similar qualitative results wrt the NRLs. However, the difference between NRL estimates (VEM-MCMC) in Fig. 5[right] points out that regions corresponding to activated areas for the two conditions present positive intensity values, which means that VEM helps retrieving higher NRL values for activated area compared to MCMC.

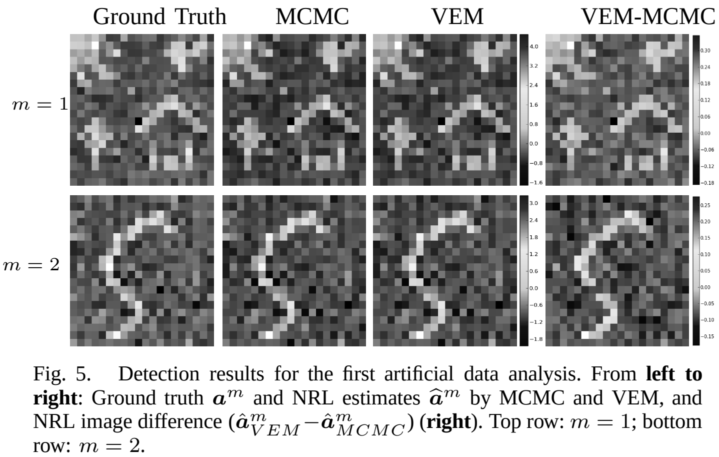

Fig. 5. Detection results for the first artificial data analysis. From **left to right**: Ground truth $\boldsymbol{a}^m$ and NRL estimates $\widehat{\boldsymbol{a}}^m$ by MCMC and VEM, and NRL image difference ($\hat{\boldsymbol{a}}^m_{VEM} - \hat{\boldsymbol{a}}^m_{MCMC}$) (**right**). Top row: $m = 1$; bottom row: $m = 2$.

Quantitatively speaking, the gain in robustness is confirmed by reporting the Sum of Squared Error ($\text{SSE}_{\widehat{\boldsymbol{a}}^m} = \sum_{j=1}^{J_\gamma} (\widehat{a}_j^m - a_j^m)^2 / J_\gamma$) values on NRL estimates which are slightly lower using VEM compared to MCMC for the first experimental condition ($m = 1$: $\text{SSE}_{\text{MCMC}} = 0.012$ *vs.* $\text{SSE}_{\text{VEM}} = 0.010$), as well as for the second experimental condition ($m = 2$: $\text{SSE}_{\text{MCMC}} = 0.010$ *vs.* $\text{SSE}_{\text{VEM}} = 0.009$). These error values indicate that, even though the MCMC algorithm gives the most precise PPMs for the high CNR condition (Fig. 4, $m = 1$), the VEM approach is more robust than its MCMC alternative in terms of estimated NRLs. These values also indicate slightly lower SSE for the second experimental conditions ($m = 2$) compared to the first one ($m = 1$) with higher CNR. This difference is explained by the presence of larger non-activated areas for $m = 2$ where low NRL values are simulated, and for which SSE is very low.

In addition, a test for equality of means has been conducted to test whether the quadratic errors means over voxels obtained with VEM and MCMC were significantly close. Very low p-values of 0.0377 and 0.0015 were obtained respectively for condition $m = 1$ and $m = 2$, which means that the null hypothesis (the two means are equal) is rejected for the usual 5% threshold. In other words, although the obtained error difference is small, it is significant from a statistical viewpoint. Interestingly, As $m = 2$ corresponds to a lower CNR, this highlights the significance of the gain in robustness we got with the VEM version in a degraded CNR context.

As regards HRF estimates, Fig. 6 shows both retrieved shapes using MCMC and VEM. Compared to the ground truth (solid line), the two approaches yield very similar results and preserve the most important features of the original HRF like the peak value (PV), time-to-peak (TTP) and time-to-undershoot (TTU).

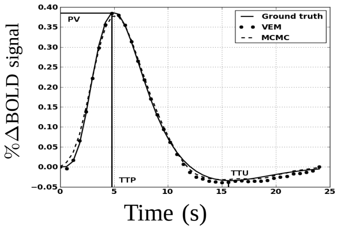

Fig. 6. Estimation results for the first artificial data analysis. Ground truth HRF $\boldsymbol{h}$ and HRF estimates $\widehat{\boldsymbol{h}}$ using the MCMC and VEM algorithms.

2) Estimation robustness: Since estimation errors may be caused by different perturbation sources, in the following, various Monte Carlo analyses have been conducted by varying one-at-a time several simulation parameters, namely the stimulus density, the noise parameters and the amount of spatial regularization. The differences in the obtained estimation errors have all been tested for statistical significance using tests for equality of means for the errors over 100 runs. For all reported comparisons, the obtained p-values were very low ($p < 0.001$) indicating that the performance differences, even when small, were statistically significant.

a) Varying the stimulus density: In this experiment, simulations have been conducted by varying the stimulus density from 5 to 30 stimuli in the artificial paradigm, which leads to decreasing Inter Stimuli Intervals (ISI) (from 47 s to 9 s, respectively). At each stimulus density, 100 realizations of the same artificial dataset have been generated so as to evaluate the estimation bias and variance of NRLs. Here, the stimuli are interleaved between the two conditions so that the above mentioned ISIs correspond to the time interval between two events irrespective of the condition they belong to. A second order autoregressive noise (AR(2)) has also been used for the simulation providing a more realistic BOLD signal [23]. The rest of the simulation process is specified as before. In order to quantitatively evaluate the robustness of the proposed VEM approach to varying input SNR ($\mathrm{SNR} = 10 \log \sum_{j=1}^{J_\gamma} \|\boldsymbol{S}_j \boldsymbol{h}_\gamma\|^2 / \sum_{j=1}^{J_\gamma} \|\boldsymbol{b}_j\|^2$), results (assuming white noise in the model used for estimation for both algorithms) are compared while varying the stimulation rate during the BOLD signal acquisition. Fig. 7 illustrates the error evolution related to the NRL estimates for both experimental conditions wrt the ISI (or equivalently the stimulus density) in the experimental paradigm. Estimation error is illustrated in terms of Mean Squared Error ($\mathrm{MSE}_{\widehat{\boldsymbol{a}}^m} = \mathrm{E}[\|\widehat{\boldsymbol{a}}^m - \boldsymbol{a}^m\|^2] = \mathrm{bias}^2(\widehat{\boldsymbol{a}}^m) + \mathrm{var}(\widehat{\boldsymbol{a}}^m)$), which splits into the sum of the variance (Fig. 7[top]) and squared bias (Fig. 7[bottom]). This figure shows that at low SNR (or high ISI), VEM is more robust in terms of estimation variance to model mis-specification irrespective of the experimental condition. At high SNR or low ISI, the two methods perform similarly and remain quite robust. As regards estimation bias, Fig. 7[bottom] shows less monotonous curves for $m = 2$, which may be linked to the lower CNR of the second experimental condition compared to the first one. However, the two methods still perform well since squared bias values are very low, meaning that the two estimators are not highly biased. As reported in Section IV-A.1, error values on NRL estimates remain comparable for both experimental conditions and all ISI values, although PPM results present some imprecisions for the low CNR condition ($m = 2$).

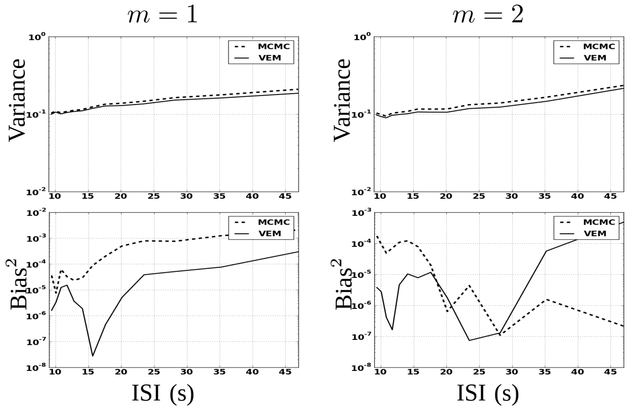

Fig. 7. NRL estimation errors over 100 simulations in a semi-logarithmic scale in terms of variance (top) and squared bias (bottom) wrt ISIs for both experimental conditions $m = 1$ and $m = 2$.

As regards hemodynamic properties, Fig. 8[left] depicts errors on HRF estimates inferred by VEM and MCMC in terms of variance (Fig. 8[left-top]) and squared bias (Fig. 8[left-bottom]) wrt the ISI (or equivalently the stimulus density). The VEM approach outperforms the MCMC scheme over the whole range of ISI values, but the bias and variance remain very low for both methods. When evaluating the estimations of the key HRF features (PV, TTP and TTU), it turns out that the TTP and TTU estimates remain the same irrespective of the inference algorithm, which corroborates the robustness of the developed approach (results not shown). As regards PV estimates, Fig. 8[right] shows the error values wrt the ISIs. The VEM algorithm outperforms MCMC in terms of squared bias over all ISI values. However, the performance in terms of estimation variance remains similar with very low variance values over the explored ISIs range. For more complete comparisons, similar experiments have been conducted while changing the ground truth HRF properties (PV, TTP, TTU), and similar results have been obtained. In contrast to NRL estimations, variance and squared bias are here comparable. We can even note higher squared bias for PV estimates compared to the variance.

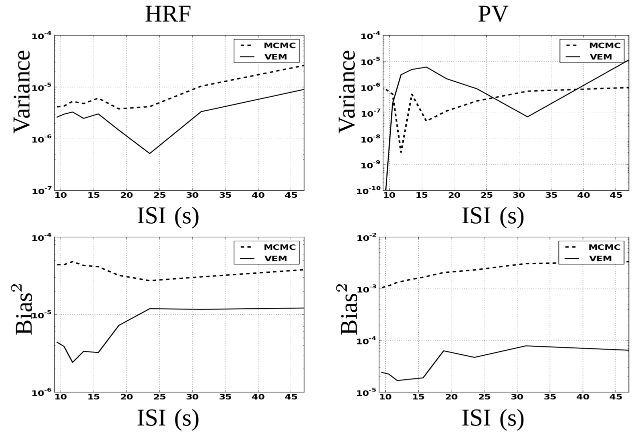

Fig. 8. Estimation errors over 100 simulations in a semi-logarithmic scale in terms of variance (top) and squared bias (bottom) wrt ISIs for both the HRF and its PV.

b) Varying the noise parameters: In this experiment, several simulations have been conducted using an AR(2) noise with varying variance and correlation parameters in order to illustrate the robustness of the proposed VEM approach to noise parameter fluctuation. For each simulation, 100 realizations were generated so as to estimate the estimation bias and variance. Performance of the two methods are evaluated in terms of MSE (splits into the sum of the variance and squared bias). Fig. 9 illustrates in a semi-logarithmic scale for NRLs the variance and squared bias of the two estimators plotted against the input SNR when varying the *noise variance*. This figure clearly shows that the bias introduced by both estimators is very low compared to the variance. Moreover, the bias introduced by VEM is very low compared to MCMC. As regards the variance, our results illustrate that VEM also slightly outperforms MCMC.

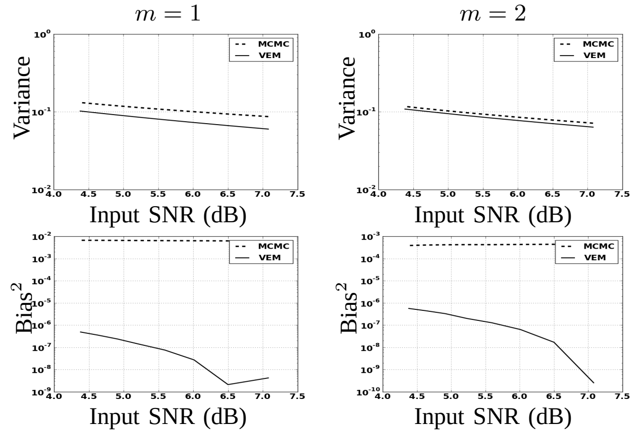

Fig. 9. MSE on NRL estimates split into variance (top) and squared bias (bottom) wrt input SNR by varying the AR(2) noise variance.

Fig. 10 depicts the variance and squared bias plotted in semi-logarithmic scale against input SNR when varying the *noise autocorrelation*. Overall, the same conclusions as for Fig. 9 hold. Moreover, as already observed in [42] at a fixed input SNR value, the impact of a large autocorrelation is stronger than that of a large noise variance irrespective of the inference scheme. Comparing Figs 9 and 10, this property is mainly visible at low input SNR (as usually observed on real BOLD signals). Although a slight advantage is observed for the VEM approach in terms of estimation error and for both experimental conditions, the two methods perform generally well with a relatively low error level. The slightly better performance of VEM compared to MCMC is likely due to a better fit under model mis-specification.

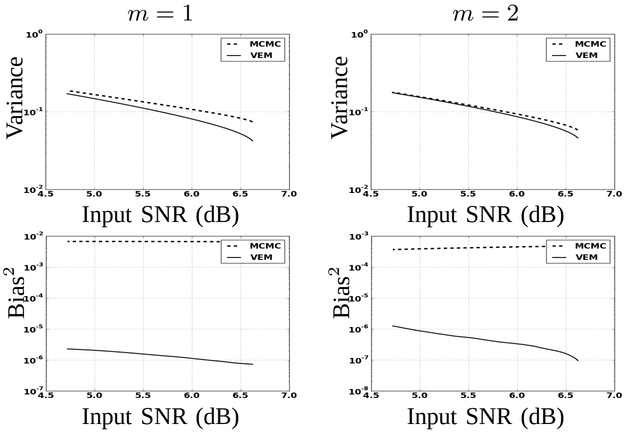

Fig. 10. MSE on NRL estimates split into variance (top) and squared bias (bottom) wrt input SNR (AR(2) noise) by varying the amount of AR(2) noise autocorrelation.

c) Varying the spatial regularization parameter: This section is dedicated to studying the robustness of the spatial regularization parameter estimation. For doing so, the synthetic activation maps of Fig. 2 are replaced with maps obtained as simulations of a 2-class Potts model with interaction parameter β that varies from 0.5 to 1.4. When positive, this parameter (β) favors spatial regularity across adjacent voxels, and hence smoother activation maps. Fig. 11 shows the estimated mean value $\widehat{\boldsymbol{\beta}}$ and standard deviations for $\boldsymbol{\beta} = \{\beta_m, m = 1 \ldots M\}$ over 100 simulations using both algorithms and for the two experimental conditions. Three main regions can be distinguished for both experimental conditions. The first one corresponds to $\beta \leqslant 0.8$, which approximatively matches the phase transition critical value $\beta^c = \log(1 + \sqrt{2}) = 0.88$ for the 2-class Potts model. For this region, Fig. 11 shows that the VEM estimate (green curve) appears to be closer to the Ground truth (black line) than the MCMC one (blue curve). Also, the proposed VEM approach yields more accurate estimation, especially for the first experimental condition having relatively high CNR. The second region corresponds to $\beta \in (0.8, 1.1)$, where MCMC inference becomes more robust than VEM. The third region is identified by $\beta \geqslant 1.1$, where both methods give less robust estimation than for the first two regions. Based on these regions, we conclude that the variational approximation (mean-field) improves the estimation performance up to a given critical value. It turns out that such an approximation is more valid for low β values, which usually correspond to observed β values on real fMRI data.

When comparing estimates for the two conditions, the curves in Fig. 11 show that both methods generally estimate more precise β values for the first experimental condition ($m = 1$) having higher input CNR. For both cases, and across the three regions identified hereabove, the error bars show that the VEM approach generally gives less scattered estimates (lower standard deviations) than the MCMC one, which confirms the gain in robustness induced by the variational approximation.

Note here that estimated β values in the experiment of Section IV-A.1 lie in the first region for the first experimental condition ($\beta_1^{\mathrm{MCMC}} = 0.74$, $\beta_1^{\mathrm{VEM}} = 0.75$). For the second condition, and because low input SNR, no clear conclusion can be made since MCMC and VEM give relatively different values ($\beta_2^{\mathrm{MCMC}} = 0.61$, $\beta_2^{\mathrm{VEM}} = 1.01$) and no ground truth is available since activation maps have been drawn by hand and not simulated according to the Markov model.

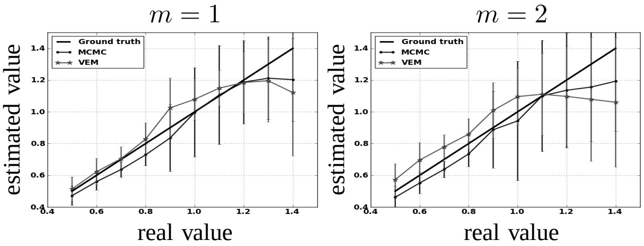

Fig. 11. Reference (diagonal) and estimated mean values of $\boldsymbol{\beta}$ with VEM MCMC for both experimental conditions ($m = 1$ and $m = 2$). Mean values and standard deviations (vertical bars) are computed based on 100 simulations.

B. Real fMRI datasets

This section is dedicated to the experimental validation of the proposed VEM approach in a real context. Experiments were conducted on real fMRI data collected on a single healthy adult subject who gave informed written consent. Data were collected with a 3-Tesla Siemens Trio scanner using a 3D Magnetization Prepared Rapid Acquisition GRadient Echo (MPRAGE) sequence for the anatomical MRI and a Gradient-Echo Echo Planar Imaging (GRE-EPI) sequence for the fMRI experiment. The acquisition parameters for the MPRAGE sequence were set as follows: Time of Echo: TE $= 2.98$ ms; Time of Repetition: TR $= 2300$ ms; sagittal orientation; spatial in-plane resolution: 1×1 mm^2; Field of View: FOV $= 256$ mm^2 and slice thickness: 1.1 mm. Regarding the EPI sequence, we used the following settings: the fMRI session consisted of $N = 128$ EPI scans, each of them being acquired using TR $= 2400$ ms, TE $= 30$ ms, slice thickness: 3 mm, transversal orientation, FOV $= 192$ mm^2 and spatial in-plane resolution was set to 2×2 mm^2. Data was collected using a 32 channel head coil to enable parallel imaging during the EPI acquisition. Parallel SENSE imaging was used to keep a reasonable Time of Repetition (TR) value in the context of high spatial resolution.

The fMRI experiment design was a functional localizer paradigm [43] that enables a quick mapping of cognitive brain functions such as reading, language comprehension and mental calculations as well as primary sensory-motor functions. It consists of a *fast event-related* design comprising sixty auditory, visual and motor stimuli, defined in ten experimental conditions and divided in two presentation modalities (auditory and visual sentences, auditory and visual calculations, left/right aurally and visually induced motor responses, horizontal and vertical checkerboards). The average ISI is 3.75 s including all experimental conditions. Such a paradigm is well suited for simultaneous detection and estimation, in contrast to slow event-related and block paradigms which are considered as optimal for estimation and detection, respectively [44]. After standard pre-processing steps (slice-timing, motion corrections and normalization to the MNI space), the whole brain fMRI data was first parcellated into $\varUpsilon = 600$ functionally homogeneous parcels by resorting to the approach described in [26]. This parcellation method consisted of a hierarchical clustering (Euclidean distance, Ward's linkage) of the experimental condition effects estimated by a GLM analysis. This GLM analysis comprised the temporal and dispersion HRF derivatives as regressors so that the clustering took some HRF variability into account. To enforce parcel connexity, the clustering process was spatially constrained to group only adjacent positions. This parcellation was used as an input of the JDE procedure, together with the fMRI time series. We stress the fact that the latter signals were not spatially smoothed prior to the analysis as opposed to the classical SPM-based fMRI processing. In what follows, we compare the MCMC and VEM versions of JDE with the classical GLM analysis by focusing on two contrasts of interest: i) the **Visual-Auditory (VA)** contrast which evokes positive and negative activity in the primary occipital and temporal cortices, respectively, and ii) the **Computation-Sentences (CS)** contrast which aims at highlighting higher cognitive brain functions. Besides, results on HRF estimates are reported for the two JDE versions and compared to the canonical HRF, as well as maps of regularization factor estimates.

Fig. 12 shows results for the **VA** contrast. High positive values are bilaterally recovered in the occipital region and the overall cluster localizations are consistent for both MCMC and VEM algorithms. The only difference lies in the temporal auditory regions, especially on the right side, where VEM yields rather more negative values than MCMC. Thus VEM seems more sensitive than MCMC. The results obtained by the classical GLM (see Fig. 12[right]) are comparable to those of JDE in the occipital region with roughly the same level of recovered activations. However, in the central region, we observe activations in the white matter that can be interpreted as false positives and that were not exhibited using the JDE formalism. The bottom part of Fig. 12 compares the estimated values of the regularization factors $\widehat{\boldsymbol{\beta}}$ between VEM and MCMC algorithms for two experimental conditions involved in the **VA** contrast. Since these estimates are only relevant in parcels which are activated by at least one condition, a mask was applied to hide non-activated parcels. We used the following criterion to classify a parcel as activated: $\max\{(\widehat{\mu}_{1m})_{1\leq m\leq M}\} \geq 8$ (and non-activated otherwise). These maps of $\widehat{\boldsymbol{\beta}}$ estimates show that VEM yields more contrasted values between the visual and auditory conditions. Table III provides the estimated $\widehat{\boldsymbol{\beta}}$ values in the highlighted parcels of interest. The auditory condition does not elicit evoked activity and yields lower $\widehat{\boldsymbol{\beta}}$ values in both parcels whereas the visual condition is associated with higher values. The latter comment holds for both algorithms but VEM provides much lower values ($\widehat{\boldsymbol{\beta}}^{\text{aud.}}_{\text{VEM}} \approx 0.01$) than MCMC ($\widehat{\boldsymbol{\beta}}^{\text{aud.}}_{\text{MCMC}} \approx 1.07$) for the non-activated condition. For the activated condition, the situation is comparable, with $\widehat{\boldsymbol{\beta}}^{\text{vis.}}_{\text{VEM}} \approx 1.1$ and $\widehat{\boldsymbol{\beta}}^{\text{vis.}}_{\text{MCMC}} \approx 1.25$. This illustrates a noteworthy difference between VEM and MCMC. Probably due to the mean field and variational approximations, the hidden field may not have the same behaviour (different regularization effect) between the two algorithms. Still, this discrepancy is not visible on the NRL maps.

Fig. 12(a-b) depicts HRF estimation results which are rather close for both methods in the two regions under consideration. VEM and MCMC HRF estimates are also consistent with the canonical HRF shape. Indeed, the latter has been precisely calibrated on visual regions [10, 11]. These estimation results explain why JDE does not bring any gain in sensitivity compared to the classical GLM: the canonical HRF is the optimal choice for visual areas. We can note a higher variability in the undershoot part, which can be explained, first, by the fast event-related nature of the paradigm where successive evoked responses are likely to overlap in time so that it is more difficult to disentangle their ends; and second, by the lower signal strength and SNR in the tail of the response. To conclude on the **VA** contrast which focused on well-known sensory regions, VEM provides sensitive results consistent with the MCMC version, both wrt detection and estimation tasks. These results were also validated by a classical GLM analysis which yielded comparable sensitivity in a region where the canonical HRF is known to be valid (see Fig. 12[right]).

Results related to the **Computation-Sentences (CS)** contrast are depicted in Fig. 13. As for **VA**, NRL contrast maps ($\widehat{a}^{\text{comp.}}_j - \widehat{a}^{\text{sent.}}_j$) are roughly equivalent for VEM and MCMC in terms of cluster localizations. Still, in Fig. 13[left-center], we observe that MCMC seems quite less specific than VEM as positive contrast values are exhibited in the white matter for MCMC, and not for VEM (compare especially the middle part of the axial slices). Here, GLM results clearly show lower sensitivity compared to the JDE results. For the estimates of the regularization parameters, the situation is globally almost the same as for the **VA** contrast, with VEM yielding more contrasted $\widehat{\boldsymbol{\beta}}$ maps than MCMC. However, these values are slightly lower than the ones reported for the **VA** contrast.

We first focus on the left frontal cluster, located in the middle frontal gyrus which has consistently been exhibited as involved in mental calculation [45]. HRF estimates in this region are shown in Fig. 13(b) and strongly depart from the canonical version, which explains the weaker sensitivity in the GLM results as the canonical HRF model is not optimal in this region. Especially, the TTP value is much more delayed with JDE (7.5 s), compared to the canonical situation (5 s). The VEM and MCMC shapes are close to each other, except at the beginning of the curves where VEM presents an initial dip. This might be interpreted as a higher temporal regularization

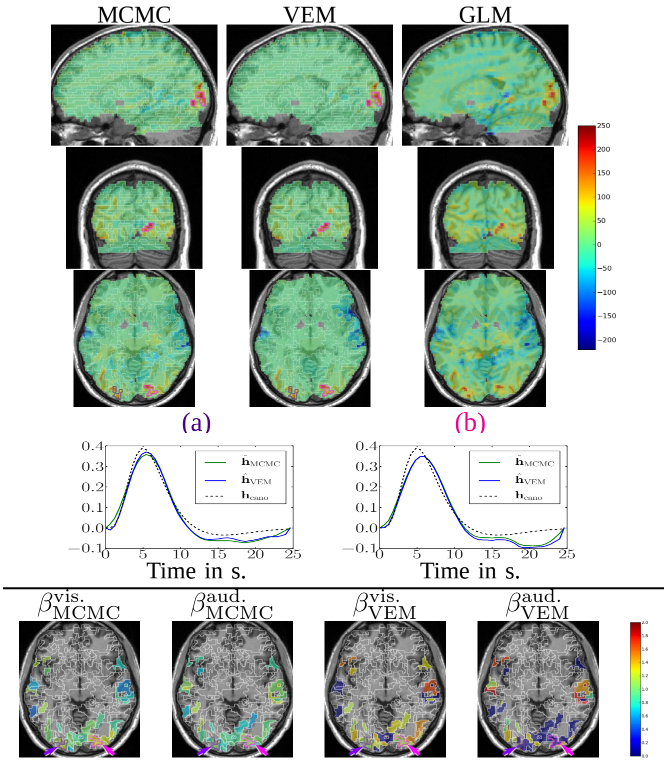
MCMC
VEM
GLM
250
200
150
100
50
0
−50
−100
−150
−200
(a)
(b)
$\hat{\mathbf{h}}_{\mathrm{MCMC}}$
$\hat{\mathbf{h}}_{\mathrm{VEM}}$
$\mathbf{h}_{\mathrm{cano}}$
Time in s.
Time in s.
$\beta^{\mathrm{vis.}}_{\mathrm{MCMC}}$
$\beta^{\mathrm{aud.}}_{\mathrm{MCMC}}$
$\beta^{\mathrm{vis.}}_{\mathrm{VEM}}$
$\beta^{\mathrm{aud.}}_{\mathrm{VEM}}$

Fig. 12. Results for the **Visual-Auditory** contrast obtained by the VEM and MCMC JDE versions, compared to a GLM analysis. Top part, from left to right: NRL contrast maps for MCMC, VEM and GLM with sagittal, coronal and axial views from top to bottom lines (neurological convention: left is left). Middle part: plots of HRF estimates for VEM and MCMC in the two parcels circled in indigo (γ_1) and magenta (γ_2) on the maps: occipital left (a) and right (b), respectively. The canonical HRF shape is depicted in dashed line. Bottom part: axial maps of estimated regularization parameters $\widehat{\boldsymbol{\beta}}$ for the two conditions, auditory (aud.) and visual (vis.), involved in the **VA** contrast. Parcels that are not activated by any condition are hidden. For all contrast maps, the input parcellation is superimposed in white contours.

introduced in the MCMC scheme. Still, the most meaningful HRF features such as the TTP and the Full Width at Half Maximum (FWHM) are very similar.

The second region of interest for the **CS** contrast is located in the inferior parietal lobule and is also consistent with the computation task [45]. Note that the contrast value is lower than the one estimated in the frontal region, whatever the inference scheme. Interestingly, this activation is lost by the GLM fitting procedure. HRF estimates are shown in Fig. 13(a). The statement relative to the previous region holds again: they strongly differ from the canonical version, which explains the discrepancy between the different detection activation results each method retrieved. When comparing MCMC and VEM, even if the global shape and the TTP position are similar, the initial dip in the HRF estimate is still stronger with VEM and the corresponding FWHM is also smaller than for the MCMC version. As previously mentioned, this suggests that MCMC may tend to over-smooth the HRF shape. Results for the regularization parameters, as shown in Table III [4th col.], indicate that $\widehat{\beta}$ for VEM and the Sentence condition is not as low as it is for the other parcels and the non-activated conditions ($\widehat{\beta}^{\gamma_4}_{\mathrm{Sent.}} = 1.19$ against $\widehat{\beta}^{\gamma_3}_{\mathrm{Sent.}} = 0.01$). This is explained by the fact that both Computation and the Sentence conditions yield activations in this parcel as confirmed by the low contrast value.

TABLE III

Estimated regularization parameters $\widehat{\boldsymbol{\beta}}$ obtained with VEM and MCMC JDE for the experimental conditions involved in the studied contrasts: **Visual-Auditory (VA)** and **Computation-Sentences (CS)**. Results are provided for the two highlighted parcels for each contrast (see Figs. 12-13).

	VA contrast				**CS** contrast			
	parcel γ_1		parcel γ_2		parcel γ_3		γ_4	
	$\widehat{\beta}^{\gamma_1}_{\mathrm{Vis.}}$	$\widehat{\beta}^{\gamma_1}_{\mathrm{Aud.}}$	$\widehat{\beta}^{\gamma_2}_{\mathrm{Vis.}}$	$\widehat{\beta}^{\gamma_2}_{\mathrm{Aud.}}$	$\widehat{\beta}^{\gamma_3}_{\mathrm{Comp.}}$	$\widehat{\beta}^{\gamma_3}_{\mathrm{Sent.}}$	$\widehat{\beta}^{\gamma_4}_{\mathrm{Comp.}}$	$\widehat{\beta}^{\gamma_4}_{\mathrm{Sent.}}$
MCMC	1.28	1.08	1.24	1.05	1.08	0.68	1.07	0.64
VEM	1.14	0.01	1.08	0.01	0.91	0.01	0.82	1.19

The studied contrasts represent decreasing CNR situations, with the **VA** contrast being the stronger and **CS** the weaker. From the detection point of view, the contrast maps are very similar for both JDE versions and are only dimly affected by the CNR fluctuation. In contrast, HRF estimation results are much more sensitive to this CNR variation, with stronger discrepancies between the VEM and MCMC versions, especially for the HRF estimates associated with the **CS** parietal cluster. The latter shows the weaker contrast magnitude. Still, both versions provide results in agreement on the TTP and FWHM values. Indeed, the differences mainly concern the heading and tailing parts of the HRF curves.

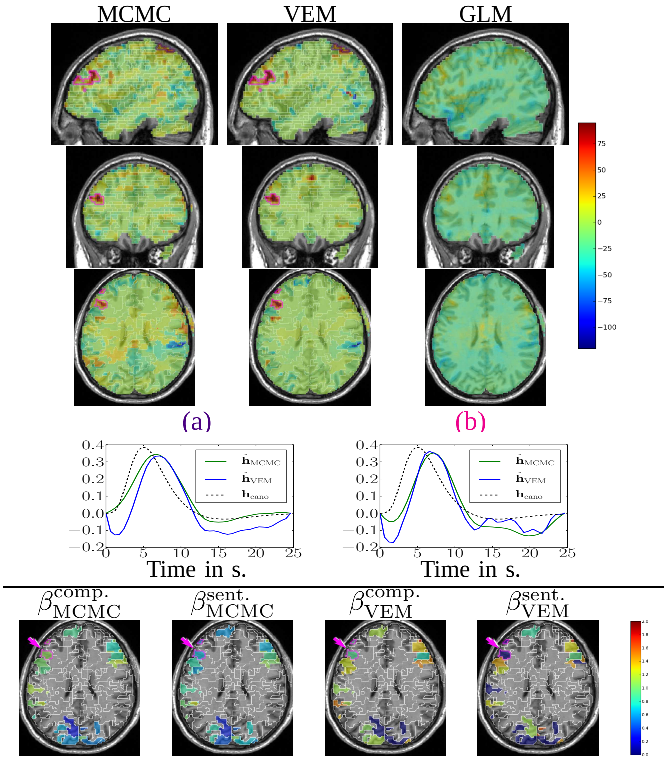
MCMC
VEM
GLM
75
50
25
0
−25
−50
−75
−100
(a)
(b)
$\hat{\mathbf{h}}_{\mathrm{MCMC}}$
$\hat{\mathbf{h}}_{\mathrm{VEM}}$
$\mathbf{h}_{\mathrm{cano}}$
Time in s.
Time in s.
$\beta^{\mathrm{comp.}}_{\mathrm{MCMC}}$
$\beta^{\mathrm{sent.}}_{\mathrm{MCMC}}$
$\beta^{\mathrm{comp.}}_{\mathrm{VEM}}$
$\beta^{\mathrm{sent.}}_{\mathrm{VEM}}$

Fig. 13. Results for the **Computation-Sentences** contrast obtained by the VEM and MCMC JDE versions compared to a GLM analysis. Top part, from left to right: NRL contrast maps for MCMC, VEM, and GLM, with sagittal, coronal and axial views from top to bottom lines (neurological convention: left is left). Middle part: plots of HRF estimates for VEM and MCMC in the two parcels circled in indigo (γ_3) and magenta (γ_4) on the maps: left parietal lobule (a) and left middle frontal gyrus (b), respectively. The canonical HRF shape is depicted in dashed line. Bottom part: axial maps of estimated regularization parameters $\widehat{\boldsymbol{\beta}}$ for the two conditions, computation (comp.) and sentence (sent.), involved in the **CS** contrast. Parcels that are not activated by any condition are hidden. For all contrast maps, the input parcellation is superimposed in white contours.

V. Algorithmic efficiency

In this section, the computational performance of the two approaches are compared on both artificial and real fMRI datasets. Both algorithms were implemented in Python and fully optimized by resorting to the efficient array operations of the Numpy library (`http://numpy.scipy.org`) as well as C-extensions for the computationally intensive parts (*e.g.* NRL sampling in MCMC or the VE-Q step for VEM). Moreover, our implementation handled distributed computing resources as the JDE analysis consists of parcel-wise independent processings which can thus be performed in parallel. This code is available in the PyHRF package (`http://www.pyhrf.org`). For both the VEM and MCMC algorithms, the same stopping criterion was used. This criterion consists of simultaneously evaluating the online relative variation of each estimate. In other words, for instance for the estimated $\widehat{\boldsymbol{h}}_\gamma$, one has to check whether $c_H = \frac{||\widehat{\boldsymbol{h}}_\gamma^{(r+1)} - \widehat{\boldsymbol{h}}_\gamma^{(r)}||_2^2}{||\widehat{\boldsymbol{h}}_\gamma^{(r)}||_2^2} \leq 10^{-5}$. By evaluating a similar criterion c_A for the NRLs estimates, the algorithm is finally stopped once $c_H \leq 10^{-5}$ and $c_A \leq 10^{-5}$. For the MCMC algorithm, this criterion is only computed after the burn-in period, when the samples are assumed to be drawn from the target distribution. The burn-in period has been fixed manually based on several *a posteriori* controls of simulated chains relative to different runs (here 1000 iterations). More sophisticated convergence monitoring techniques [46] should be used to stop the MCMC algorithm, but we chose the same criterion as for the VEM to carry out a more direct comparison.

Considering the artificial dataset presented in Section IV-A.1, Fig. 14 illustrates the evolution of c_H and c_A with respect to the computational time for both algorithms. Only about 18 seconds are enough to reach convergence for the VEM algorithm, while the MCMC alternative needs about 1 minute to converge on the same Intel Core 4 - 3.20 GHz - 4 Gb RAM architecture. The horizontal line in the blue curve relative to the MCMC algorithm corresponds to the burn-in period (1000 iterations). It can also be observed that the gain in computational efficiency the VEM scheme brings is independent of the unknown variables (*e.g.* HRF) we look at.

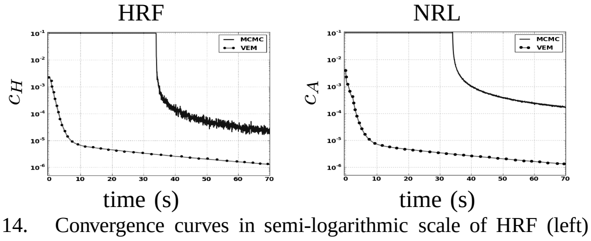

Fig. 14. Convergence curves in semi-logarithmic scale of HRF (left) and NRL (right) estimates using MCMC (blue lines) and VEM (green lines).

To illustrate the impact of the problem dimensions on the computational cost of both methods, Fig. 15 shows the evolution of the computational time of one iteration when varying the number of voxels (left), the number of experimental conditions (center) and the number of scans (right). The three curves show that the computational time increases almost linearly (see the blue and red curves) for both algorithms, but with different slopes. Blue curves (VEM) have steeper slopes than red ones (MCMC) in the three plots showing that the computational time of one iteration increases faster with VEM than with MCMC wrt the problem dimensions.

As regards computational performance on the real fMRI data set presented in Section IV-B and comprising 600 parcels, the VEM also appeared faster as it took 1.5 hours to perform a whole brain analysis whereas the MCMC version took 12 hours. These analysis timings were obtained by a serial processing of all parcels for both approaches. When resorting to the distributed implementation, the analysis durations boiled down to 7 mins for VEM and 20 mins for MCMC (on a 128-core cluster). To go further, we illustrate the computational time difference ($\Delta t = t_{\text{MCMC}} - t_{\text{VEM}}$) between both algorithms in terms of parcel size which ranged from 50 to 580 voxels. As VEM vs. MCMC efficiency appears to be influenced by the level of activity within the parcel, we resorted to the same criterion as in Section IV-B to distinguish non-activated from activated parcels and tag the analysis durations accordingly in Fig. 16. Fig. 16(a) clearly shows that the differential timing Δt between the two algorithms is higher for non-activated parcels (blue dots) and increases with the parcel size, which confirms the utility of the proposed VEM approach especially in low CNR/SNR circumstances. To further investigate the gain in terms of computational time induced by VEM, Fig. 16(b) illustrates the gain factor ($G_f = t_{\text{MCMC}}/t_{\text{VEM}}$) for activated and non-activated parcels. This figure shows that the VEM algorithm always outperforms the MCMC alternative since $G_f \geq 1$ in all parcels (see horizontal line in Fig. 16[b]). Moreover, the gain factor G_f is clearly higher for non-activated parcels for which the input SNRs and CNRs are relatively low, and we found $G_f \in [2.7, 80]$.

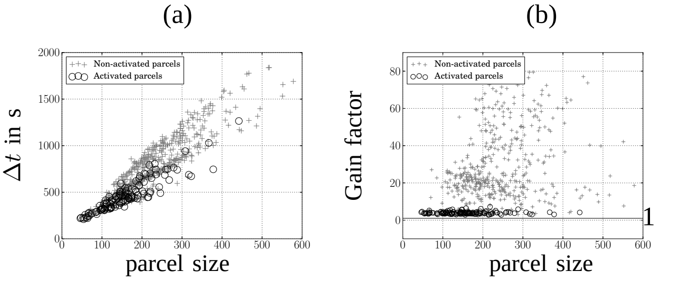

Fig. 16. Comparison of durations for MCMC and VEM analyses in terms of parcel size, each dot coding for a different parcel. **(a)**: differential timing $\Delta t = t_{\text{MCMC}} - t_{\text{VEM}}$. **(b)**: gain factor $G_f = t_{\text{MCMC}}/t_{\text{VEM}}$ of VEM compared to MCMC, horizontal line indicates a gain factor $G_f = 1$. Plus marks $(+)$ indicate parcels estimated as activated, ie $\max\{(\widehat{\mu}_{1m})_{1 \leq m \leq M}\} \geq 8$ and circles (o) indicate parcels estimated as non-activated.

VI. Discussion and conclusion

In this paper, we have proposed a new intra-subject method for parcel-based joint detection-estimation of brain activity from fMRI time series. The proposed method relies on a Variational EM algorithm as an alternative solution to intensive stochastic sampling used in previous work [23, 25]. Compared to the latter formulation, the proposed VEM approach does not require priors on the model parameters for inference to be carried out. However, to achieve gain in robustness and make the proposed approach completely auto-calibrated, the adopted model can be extended by injecting additional priors on some of its parameters as detailed in Appendix E for $\boldsymbol{\beta}$ and $v_{\boldsymbol{h}}$ estimation.

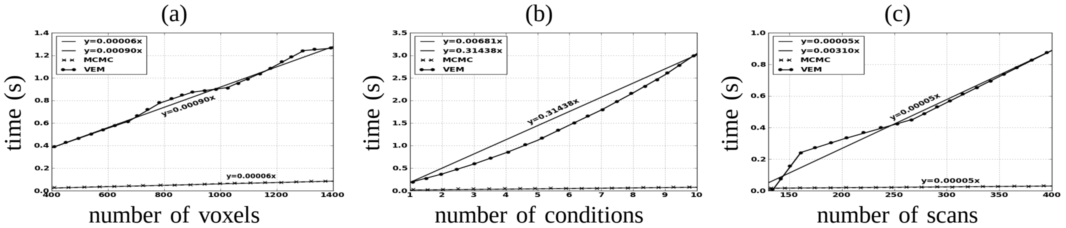

Fig. 15. Evolution of the computational time per iteration using the MCMC and VEM algorithms when varying the problem dimension according to: (a): number of voxels; (b): number of conditions; (c): number of scans.

Illustrations on simulated and real datasets have been deeply conducted in order to assess the robustness of the proposed method compared to its MCMC counterpart in different experimental contexts. Simulations have shown that the proposed VEM algorithm retrieved more accurate activation labels and NRLs especially at low input CNR, while yielding similar performance for HRF estimation. VEM superiority in terms of NRLs and labels estimation has been confirmed by mean equality statistical tests over estimation errors obtained through 100 Monte Carlo runs. Conducted tests showed very low p-values, which means that differences in terms of obtained error means were statistically significant. Simulations have also shown that our approach was more robust to any decrease of stimulus density (or equivalently to any increase of ISI value). Similar conclusions have been drawn wrt noise level and autocorrelation structure. In addition, our VEM approach provided more robust estimation of the spatial regularization parameter and more compact activation maps that are likely to better account for functional homogeneity. These good properties of the VEM approach are obtained faster than using the MCMC implementation. Simulations have also been conducted to study the computational time variation wrt the problem dimensions, which may significantly vary from one experimental context to another.

As a general comment, this performance of VEM compared to MCMC may be counterintuitive. The sampled chain in MCMC is supposed to converge to the true target distribution after the burn-in period. However, since non-informative priors are used in the MCMC model such as the uniform prior for Potts parameters β, imprecisions in the target distribution may occur. In this case, VEM may outperform MCMC as observed in our simulations. Also, in some of the experiments, the model assumes a 2-class Potts model for the activation classes while we used more realistic synthetic images instead. The images we used (*e.g.* the house shape) are more regular and realistic than would be a typical realization of a Potts model. Then, our results show lower MSE (variance and squared bias) for VEM when the noise model is mis-specified and we suspect that VEM is less sensitive to model mis-specification. In a mis-specification context, the variational approach may be favored by the fact that the factorization assumption acts as an extra regularizing term that smoothes out the solutions in a more appropriate manner. There exists a number of other studies in which the variational approach is compared to its MCMC counterpart and provides surprisingly accurate results [30, 38, 47, 48]. It is true that results showing the superior performance of VEM over MCMC are more seldom. For instance, the results reported in [38] show that their variational approach is highly accurate in approximating the posterior distribution. These authors show like us, smaller MSE for the variational approach vs MCMC on simulations and point out a MCMC sensitivity to initialization conditions. On a computational efficiency point of view, most works on variational methods lead to EM-like algorithms in which one iteration consists of updating sufficient statistics (*e.g.* means and variances) that characterize the distribution approximating the target posterior. In our setting the approximating distribution is made of Gaussian parts fully specified by their mean and variance. In contrast, sampling-based methods like MCMC are not focused on sufficient statistics computation, but rather simulate realizations from the full posterior. Approximating a limited number of moments is less complex than approximating a full distribution, which may also be an ingredient that explains improved VEM efficiency and performance.

Regarding real data experiments, VEM and MCMC showed similar results with a higher specificity for the former. Compared to the classical GLM approach, the JDE methodology yields similar results in the visual areas where the canonical HRF is well recovered, whereas in the areas involved in the **CS** contrast, the estimation of more adapted HRFs that strongly differ from the canonical version enables higher sensitivity in the activation maps. These results further emphasize the interest of using VEM for saving a large amount of computational time. From a practical viewpoint, another advantage of the proposed algorithm lies in its simplicity to track convergence even if the latter is local: the VEM algorithm only requires a simple stopping criterion to achieve a local minimizer in contrast to the MCMC implementation [25]. It is also more flexible to account for more complex situations such as those involving higher AR noise order, habituation modeling [49] or considering three instead of two activation classes with an additional deactivation class.

To confirm the impact of the proposed inference, comparisons between the MCMC and VEM approaches should also take place at the group level. The most straightforward way would be to compare the results of random effect analyses (RFX) based on group-level Student t-test on averaged effects, the latter being computed either by a standard individual SPM analysis or by the two VEM and MCMC JDE approaches. In this direction, a preliminary study has been performed in [14] where group results based on JDE MCMC intra-subject analyses provide higher sensitivity than results based on GLM based intra-subject analyses. Ideally, the JDE framework could be extended to perform group-level analysis and yield group-level NRL maps as well as group-level HRFs. To break down the complexity, this extension could operate parcel-wisely by grouping subject-dependent data into group-level functionally homogeneous parcels. This procedure would result in a hierarchical mixed effect model and encode mean group-level values of NRLs and HRFs so that subject-specific NRL and HRF quantities would be modeled as fluctuations around theses means.

Such group-level validations would also shed the light on the impact of the used variational approximation in VEM. In fact, no preliminary spatial smoothing is used in the JDE approach in contrast to standard fMRI analyses where this smoothing helps retrieving clearer activation clusters. In this context, the used mean field approximation especially in the VE-Q step should help getting less noisy activation clusters compared to the MCMC approach. Eventually, akin to [23], the model used in our approach accounts for functional homogeneity at the parcel scale. These parcels are assumed to be an input of the proposed JDE procedure and can be *a priori* provided independently by any parcellation technique [26, 50]. In the present work, parcels have been extracted from functional features we obtained via a classical GLM processing assuming a canonical HRF for the entire brain. This assumption does not bias our HRF local model estimation since a large number of parcels is considered (600 parcels) with an average parcel size of 250 voxels.

However, on real dataset, results may depend on the reliability of the used parcellation technique. A sensitivity analysis has been performed in [51] on real data and for MCMC JDE version, that assesses the reliability of this parcellation against a computationally heavier approach which tends, by randomly sampling the seed positions of the parcels, to identify the parcellation that retrieved the most significant activation maps. Still, it would be of interest to investigate the effect of the parcellation choice in the VEM context, and more generally to extend the present framework to incorporate an automatic online parcellation strategy to better fit the fMRI data while accounting for the HRF variability across regions, subjects, populations and experimental contexts. The current variational framework has the advantage to be easily augmented with parcel identification at the subject-level as an additional layer in the hierarchical model. Automatically identifying parcels raises a model selection problem in the sense of getting sparse parcellation (reduced number of parcels) which guarantees spatial variability in hemodynamic territories while enabling the reproducibility of parcel identification across fMRI datasets. More generally, a model selection approach can be easily carried out within the VEM implementation as variational approximations of standard information criteria based on penalised log-evidence can be efficiently used [52].

Appendix

A. Derivation of the VEM formula:

We show here how to obtain the Variational E-steps using the properties of the Kullback-Leibler divergence without resorting to calculus of variation as usually done. We illustrate the approach for the derivation of the VE-H step (Eq. (8)). The following VE-steps can be derived similarly. Dropping the (r) and $(r-1)$ superscripts, the VE-H step is defined as

$$\widetilde{p}_{H_\gamma} = \arg\max_{p_{H_\gamma}} \mathcal{F}(\widetilde{p}_A \; p_{H_\gamma} \; \widetilde{p}_Q, \boldsymbol{\Theta}) \; .$$

Definition (4) of $\mathcal{F}$ leads to $\mathcal{F}(\widetilde{p}_A p_{H_\gamma} \widetilde{p}_Q, \boldsymbol{\Theta}) = \mathrm{E}_{p_{H_\gamma}}\left[\mathrm{E}_{\widetilde{p}_A \widetilde{p}_Q}\left[\log p(\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{h}_\gamma, \boldsymbol{Q}; \boldsymbol{\Theta})\right]\right] + \mathcal{G}(p_{H_\gamma}) + \mathcal{G}(\widetilde{p}_A \widetilde{p}_Q)$, s.t.:

$$\arg\max_{p_{H_\gamma}} \mathcal{F}(\widetilde{p}_A \; p_{H_\gamma} \; \widetilde{p}_Q, \boldsymbol{\Theta}) =$$
$$\arg\max_{p_{H_\gamma}} \mathrm{E}_{p_{H_\gamma}}\left[\log \exp(\mathrm{E}_{\widetilde{p}_A \; \widetilde{p}_Q}\left[\log p(\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{h}_\gamma, \boldsymbol{Q}; \boldsymbol{\Theta})\right])\right] + \mathcal{G}(p_{H_\gamma}).$$

In the last equality above, we artificially introduced the exponential by taking the logarithm. We then denote by $\widetilde{p}_{H_\gamma}$ the distribution on $\boldsymbol{h}_\gamma$ proportional to $\exp(\mathrm{E}_{\widetilde{p}_A \; \widetilde{p}_Q}\left[\log p(\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{h}_\gamma, \boldsymbol{Q}; \boldsymbol{\Theta})\right])$. The normalizing constant of the latter quantity is by definition independent of $\boldsymbol{h}_\gamma$ so that the above argmax is

$$\arg\max_{p_{H_\gamma}} \mathrm{E}_{p_{H_\gamma}}\left[\log \widetilde{p}_{H_\gamma}\right] + \mathcal{G}(p_{H_\gamma}) = \arg\min_{p_{H_\gamma}} KL(p_{H_\gamma} \,||\, \widetilde{p}_{H_\gamma}),$$

where $KL(p_{H_\gamma} \,||\, \widetilde{p}_{H_\gamma})$ is the Kullback-Leibler divergence between p_{H_γ} and $\widetilde{p}_{H_\gamma}$. From the KL divergence properties, it follows that the optimal p_{H_γ} is $\widetilde{p}_{H_\gamma}$ which provides Eq. (8) as desired.

B. VE-H step:

For the VE-H step, the expressions for $\boldsymbol{m}_{H_\gamma}^{(r)}$ and $\boldsymbol{\Sigma}_{H_\gamma}^{(r)}$ are $\boldsymbol{\Sigma}_{H_\gamma}^{(r)} = \left(1/v_{\boldsymbol{h}}^{(r-1)} \boldsymbol{R}^{-1} + \sum_{j \in \mathcal{P}_\gamma} \left(\sum_{m,m'} v_{A_j^m A_j^{m'}}^{(r-1)} \boldsymbol{X}_m^{\mathrm{t}} \boldsymbol{\Gamma}_j^{(r-1)} \boldsymbol{X}_{m'} + \widetilde{\boldsymbol{S}}_j^{\mathrm{t}} \boldsymbol{\Gamma}_j^{(r-1)} \widetilde{\boldsymbol{S}}_j\right)\right)^{-1}$ and $\boldsymbol{m}_{H_\gamma}^{(r)} = \boldsymbol{\Sigma}_{H_\gamma}^{(r)} \sum_{j \in \mathcal{P}_\gamma} \widetilde{\boldsymbol{S}}_j^{\mathrm{t}} \boldsymbol{\Gamma}_j^{(r-1)} (\boldsymbol{y}_j - \boldsymbol{P}\boldsymbol{\ell}_j^{(r-1)})$ with $\widetilde{\boldsymbol{S}}_j = \sum_{m=1}^{M} m_{A_j^m}^{(r-1)} \boldsymbol{X}_m$, $m_{A_j^m}^{(r-1)}$ and $v_{A_j^m A_j^{m'}}^{(r-1)}$ denoting respectively the m and (m, m') entries of the mean vector ($\boldsymbol{m}_{A_j}^{(r-1)}$) and covariance matrix ($\boldsymbol{\Sigma}_{A_j}^{(r-1)}$) of the current $\widetilde{p}_{A_j}^{(r-1)}$.

C. VE-A step:

The VE-A step also leads to a Gaussian pdf for $\widetilde{p}_A^{(r)}$: $\widetilde{p}_A^{(r)} \sim \prod_{j \in \mathcal{P}_\gamma} \mathcal{N}(\boldsymbol{m}_{A_j}^{(r)}, \boldsymbol{\Sigma}_{A_j}^{(r)})$. The parameters are updated as $\boldsymbol{\Sigma}_{A_j}^{(r)} = \left(\sum_{i=1}^{I} \boldsymbol{\Delta}_{ij} + \widetilde{\boldsymbol{H}}_j\right)^{-1}$ and $\boldsymbol{m}_{A_j}^{(r)} = \boldsymbol{\Sigma}_{A_j}^{(r)} \left(\sum_{i=1}^{I} \boldsymbol{\Delta}_{ij} \boldsymbol{\mu}_i^{(r-1)} + \widetilde{\boldsymbol{G}}^{\mathrm{t}} \boldsymbol{\Gamma}_j^{(r-1)} (\boldsymbol{y}_j - \boldsymbol{P}\boldsymbol{\ell}_j^{(r-1)})\right)$, where a number of intermediate quantities need to be specified. First, $\boldsymbol{\mu}_i^{(r-1)} = \left[\mu_{i1}^{(r-1)}, \ldots, \mu_{iM}^{(r-1)}\right]^{\mathrm{t}}$ and $\widetilde{\boldsymbol{G}} = \mathrm{E}_{\widetilde{p}_{H_\gamma}^{(r)}}\left[\boldsymbol{G}\right]$ where $\boldsymbol{G}$ is the matrix $\boldsymbol{G} = [\boldsymbol{g}_1 \,|\, \ldots \,|\, \boldsymbol{g}_M]$ made of columns $\boldsymbol{g}_m = \boldsymbol{X}_m \boldsymbol{h}_\gamma$. The m-th column of $\widetilde{\boldsymbol{G}}$ is then also denoted by $\widetilde{\boldsymbol{g}}_m = \boldsymbol{X}_m \boldsymbol{m}_{H_\gamma}^{(r)} \in \mathbb{R}^N$. Then, $\boldsymbol{\Delta}_{ij} = \mathrm{diag}_M \left[\widetilde{p}_{Q_j^m}^{(r-1)}(i) / v_{im}^{(r-1)}\right]$ and $\widetilde{\boldsymbol{H}}_j = \mathrm{E}_{\widetilde{p}_{H_\gamma}^{(r)}}\left[\boldsymbol{G}^{\mathrm{t}} \boldsymbol{\Gamma}_j^{(r-1)} \boldsymbol{G}\right]$ is an $M \times M$ matrix whose element (m, m') is given by $\mathrm{E}_{\widetilde{p}_{H_\gamma}^{(r)}}\left[\boldsymbol{g}_m^{\mathrm{t}} \boldsymbol{\Gamma}_j^{(r-1)} \boldsymbol{g}_{m'}\right] = \mathrm{E}_{\widetilde{p}_{H_\gamma}^{(r)}}\left[\boldsymbol{g}_m\right]^{\mathrm{t}} \boldsymbol{\Gamma}_j^{(r-1)} \mathrm{E}_{\widetilde{p}_{H_\gamma}^{(r)}}\left[\boldsymbol{g}_{m'}\right] + \mathrm{trace}\big(\boldsymbol{\Gamma}_j^{(r-1)} cov_{\widetilde{p}_{H_\gamma}^{(r)}}(\boldsymbol{g}_m, \boldsymbol{g}_{m'})\big)$ $= \widetilde{\boldsymbol{g}}_m^{\mathrm{t}} \boldsymbol{\Gamma}_j^{(r-1)} \widetilde{\boldsymbol{g}}_{m'} + \mathrm{trace}\big(\boldsymbol{\Gamma}_j^{(r-1)} \boldsymbol{X}_m \boldsymbol{\Sigma}_{H_\gamma}^{(r)} \boldsymbol{X}_{m'}^{\mathrm{t}}\big)$.

D. VE-Q step:

From $p(\boldsymbol{A}|\boldsymbol{Q})$ and $p(\boldsymbol{Q})$ in Section II, it follows that the $(\boldsymbol{a}^m, \boldsymbol{q}^m)$ couples correspond to independent hidden Potts models with Gaussian class distributions. It follows an approximation that factorizes over conditions: $\widetilde{p}_Q^{(r)}(\boldsymbol{Q}) = \prod_{m=1}^{M} \widetilde{p}_{Q^m}^{(r)}(\boldsymbol{q}^m)$ where $\widetilde{p}_{Q^m}^{(r)}(\boldsymbol{q}^m) = f(\boldsymbol{q}^m | \boldsymbol{a}^m = \boldsymbol{m}_{A^m}^{(r)}; \boldsymbol{\mu}_m^{(r-1)}, \boldsymbol{v}_m^{(r-1)}, \beta_m^{(r-1)})$ is the posterior of $\boldsymbol{q}^m$ in a modified hidden Potts model f, in which the observations a_j^m's are replaced by their mean values $\boldsymbol{m}_{A_j^m}^{(r)}$ and an external field $\{\boldsymbol{\alpha}_j^{m(r)} = -v_{A_j^m A_j^m}^{(r)} [1/v_{0m}^{(r-1)}, 1/v_{1m}^{(r-1)}]^{\mathrm{t}}, j \in \mathcal{P}_\gamma\}$ is added to the prior Potts model $p(\boldsymbol{q}^m; \beta_m^{(r-1)})$. It follows that the defined Potts reads as $f(\boldsymbol{q}^m; \boldsymbol{v}_m^{(r-1)}, \beta_m^{(r-1)}) \propto \exp\Big(\sum_{j \in \mathcal{P}_\gamma} (\boldsymbol{\alpha}_j^{m(r)}(q_j^m) + \frac{1}{2}\beta_m^{(r-1)} \sum_{k \sim j} I(q_j^m = q_k^m))\Big)$. Since the expression hereabove is intractable, and using the mean-field approximation [41], $\widetilde{p}^{(r)}_{Q^m}(\boldsymbol{q}^m)$ is approximated by a factorized density $\widetilde{p}^{(r)}_{Q^m}(\boldsymbol{q}^m) = \prod_{j\in\mathcal{P}_\gamma} \widetilde{p}^{(r)}_{Q^m_j}(q^m_j)$ such that if $q^m_j = i$, $\widetilde{p}^{(r)}_{Q^m_j}(i) \propto \mathcal{N}(\boldsymbol{m}^{(r)}_{A^m_j}; \mu^{(r-1)}_{im}, v^{(r-1)}_{im}) f(q^m_j = i \,|\, \tilde{q}^m_{\sim j}; \beta^{(r-1)}_m, \boldsymbol{v}^{(r-1)}_m)$, where $\tilde{\boldsymbol{q}}^m$ is a particular configuration of $\boldsymbol{q}^m$ updated at each iteration according to a specific scheme and $f(q^m_j \,|\, \tilde{q}^m_{\sim j}; \beta^{(r-1)}_m, \boldsymbol{v}^{(r-1)}_m) \propto \exp\big(\boldsymbol{\alpha}^{m(r)}_j(q^m_j) + \beta^{(r-1)}_m \sum_{k\sim j} I(\tilde{q}^m_k = q^m_j)\big)$.

E. M step:

1) M-$(\boldsymbol{\mu}, \boldsymbol{v})$ step:

By maximizing with respect to $(\boldsymbol{\mu}, \boldsymbol{v})$, Eq. (12) reads:

$$(\boldsymbol{\mu}^{(r)}, \boldsymbol{v}^{(r)}) = \underset{(\boldsymbol{\mu},\boldsymbol{v})}{\arg\max}\ \mathrm{E}_{\widetilde{p}^{(r)}_A \widetilde{p}^{(r)}_Q}\big[\log p(\boldsymbol{A} \,|\, \boldsymbol{Q}; \boldsymbol{\mu}, \boldsymbol{v})\big] \qquad (13)$$

By denoting $\bar{p}^{(r)}_{im} = \sum\limits_{j\in\mathcal{P}_\gamma} \widetilde{p}^{(r)}_{Q^m_j}(i)$, and after deriving wrt μ_{im} and v_{im} for every $i \in \{0,1\}$ and $m \in \{1\ldots M\}$, we get $\mu^{(r)}_{im} = \sum_{j\in\mathcal{P}_\gamma} p^{(r)}_{Q^m_j}(i)\, m^{(r)}_{A^m_j}/\bar{p}^{(r)}_{im}$ and $v^{(r)}_{im} = \sum\limits_{j\in\mathcal{P}_\gamma} \widetilde{p}^{(r)}_{Q^m_j}(i)\big((m^{(r)}_{A^m_j} - \mu^{(r)}_{im})^2 + v^{(r)}_{A^j_m A^j_m}\big)/\bar{p}^{(r)}_{im}$.

2) M-$v_{\boldsymbol{h}}$ step:

By maximizing with respect to $v_{\boldsymbol{h}}$, Eq. (12) reads as $v^{(r)}_{\boldsymbol{h}} = \arg\max_{v_{\boldsymbol{h}}} \mathrm{E}_{\widetilde{p}^{(r)}_{H_\gamma}}\big[\log p(\boldsymbol{h}_\gamma; v_{\boldsymbol{h}})\big]$. Then it follows the closed-form $v^{(r)}_{\boldsymbol{h}} = \mathrm{E}_{\widetilde{p}^{(r)}_{H_\gamma}}\big[\boldsymbol{h}^{\mathrm{t}}_\gamma \boldsymbol{R}^{-1}\boldsymbol{h}_\gamma\big]/(D-1) = \big(\mathrm{trace}(\boldsymbol{\Sigma}^{(r)}_{H_\gamma}\boldsymbol{R}^{-1}) + \boldsymbol{m}^{(r)t}_{H_\gamma}\boldsymbol{R}^{-1}\boldsymbol{m}^{(r)}_{H_\gamma}\big)/(D-1) = \mathrm{trace}\big((\boldsymbol{\Sigma}^{(r)}_{H_\gamma} + \boldsymbol{m}^{(r)}_{H_\gamma}\boldsymbol{m}^{(r)t}_{H_\gamma})\boldsymbol{R}^{-1}\big)/(D-1)$.

For a more accurate estimation of $v_{\boldsymbol{h}}$, one may take advantage of the flexibility of the VEM inference and inject some prior knowledge about this parameter in the model. Being positive, a suitable prior can be an exponential distribution with mean $\lambda^{-1}_{v_{\boldsymbol{h}}}$:

$$p(v_{\boldsymbol{h}}; \lambda_{v_{\boldsymbol{h}}}) = \lambda_{v_{\boldsymbol{h}}} \exp\big(-\lambda_{v_{\boldsymbol{h}}} v_{\boldsymbol{h}}\big). \qquad (14)$$

Accounting for this prior, the new expression of the current estimate $v^{(r)}_{\boldsymbol{h}}$ is: $v^{(r)}_{\boldsymbol{h}} = \frac{(D-1)+\sqrt{8\lambda_{v_{\boldsymbol{h}}} C + (D-1)^2}}{4\lambda_{v_{\boldsymbol{h}}}}$ with $C = \mathrm{trace}\big((\boldsymbol{\Sigma}^{(r)}_{H_\gamma} + \boldsymbol{m}^{(r)}_{H_\gamma}\boldsymbol{m}^{(r)}_{H_\gamma}{}^{t})\boldsymbol{R}^{-1}\big)$

3) M-$\boldsymbol{\beta}$ step:

By maximizing with respect to $\boldsymbol{\beta}$, Eq. (12) reads:

$$\boldsymbol{\beta}^{(r)} = \underset{\boldsymbol{\beta}}{\arg\max}\ \mathrm{E}_{\widetilde{p}^{(r)}_Q}\big[\log p(\boldsymbol{Q}; \boldsymbol{\beta})\big]. \qquad (15)$$

Updating $\boldsymbol{\beta}$ consists of making further use of a mean field-like approximation [41], which leads to a function that can be optimized using a gradient algorithm. To avoid over-estimation of this key parameter for the spatial regularization, one can introduce for each β_m, some prior knowledge $p(\beta_m; \lambda_{\beta_m})$ that penalises high values. As in Eq. (14), an exponential prior with mean $\lambda^{-1}_{\beta_m}$ can be used. The expression to optimize is then given by:

$$\begin{aligned}\beta^{(r)}_m &= \underset{\beta_m}{\arg\max}\, E_{\widetilde{p}^{(r)}_{Q^m}}\,[\log p(\boldsymbol{q}^m; \beta_m) p(\beta_m; \lambda_{\beta_m})] \\ &= \underset{\beta_m}{\arg\max}\big(-\log Z(\beta_m) + \beta_m\big(\sum_{j\sim k} \mathrm{E}_{\widetilde{p}^{(r)}_{Q^m}}\big[I(q^m_j = q^m_k)\big] - \lambda_{\beta_m}\big)\big).\end{aligned}$$

After calculating the derivative wrt β_m, we retrieve the standard equation detailed in [41] in which $\sum\limits_{j\sim k} \mathrm{E}_{\widetilde{p}^{(r)}_{Q^m}}\big[I(q^m_j = q^m_k)\big]$ is replaced by $\sum\limits_{j\sim k} \mathrm{E}_{\widetilde{p}^{(r)}_{Q^m}}\big[I(q^m_j = q^m_k)\big] - \lambda_{\beta_m}$. It can be easily seen that, as expected, subtracting the constant λ_{β_m} helps penalizing high β_m values.

4) M-$(\boldsymbol{L}, \boldsymbol{\Gamma})$ step:

This maximization problem factorizes over voxels so that for each $j \in \mathcal{P}_\gamma$, we need to compute:

$$(\boldsymbol{\ell}^{(r)}_j, \boldsymbol{\Gamma}^{(r)}_j) = \underset{(\boldsymbol{\ell}_j,\boldsymbol{\Gamma}_j)}{\arg\max}\ \mathrm{E}_{\widetilde{p}^{(r)}_{H_\gamma}\widetilde{p}^{(r)}_{A_j}}\big[\log p(\boldsymbol{y}_j \,|\, \boldsymbol{a}_j, \boldsymbol{h}_\gamma; \boldsymbol{\ell}_j, \boldsymbol{\Gamma}_j)\big]\ , \qquad (16)$$

where $\boldsymbol{a}_j = \{a^m_j, m = 1\ldots M\}$. Finding the maximizer wrt $\boldsymbol{\ell}_j$ leads to ($\widetilde{\boldsymbol{G}}$ is defined in the VE-A step):

$$\boldsymbol{\ell}^{(r)}_j = \underset{\boldsymbol{\ell}_j}{\arg\max}\big(2(\widetilde{\boldsymbol{G}}\boldsymbol{m}^{(r)}_{A_j} - \boldsymbol{y}_j)^{\mathrm{t}}\boldsymbol{\Gamma}^{(r)}_j\boldsymbol{P}\boldsymbol{\ell}_j + \boldsymbol{\ell}^{\mathrm{t}}_j\boldsymbol{P}^{\mathrm{t}}\boldsymbol{\Gamma}^{(r)}_j\boldsymbol{P}\boldsymbol{\ell}_j\big). \qquad (17)$$

After calculating the derivative wrt $\boldsymbol{\ell}_j$, we get $\boldsymbol{\ell}^{(r)}_j = (\boldsymbol{P}^{\mathrm{t}}\boldsymbol{\Gamma}^{(r)}_j\boldsymbol{P})^{-1}\boldsymbol{P}^{\mathrm{t}}\boldsymbol{\Gamma}^{(r)}_j(\boldsymbol{y}_j - \widetilde{\boldsymbol{G}}\boldsymbol{m}^{(r)}_{A_j})$. In the AR(1) case, with $\boldsymbol{\Gamma}_j = \sigma^{-2}_j\boldsymbol{\Lambda}_j$, we can then derive the following relationship:

$$\boldsymbol{\ell}^{(r)}_j = \big(\boldsymbol{P}^{\mathrm{t}}\boldsymbol{\Lambda}^{(r)}_j\boldsymbol{P}\big)^{-1}\boldsymbol{P}^{\mathrm{t}}\boldsymbol{\Lambda}^{(r)}_j\big(y_j - \widetilde{\boldsymbol{S}}_j m^{(r)}_{H_\gamma}\big) = F_1(\rho^{(r)}_j), \qquad (18)$$

where F_1 is a function linking the estimates $\boldsymbol{\ell}^{(r)}_j$ and $\rho^{(r)}_j$. The above formula is similar to that in [23, p.965], when replacing $\boldsymbol{h}_\gamma$ by $m^{(r)}_{H_\gamma}$ and $\boldsymbol{a}$ by $m^{(r)}_A$.

Denoting $\boldsymbol{y}^{(r)}_j = \boldsymbol{y}_j - \boldsymbol{P}\boldsymbol{\ell}^{(r)}_j$ and considering the maximization wrt σ^2_j, similar calculations lead to $\sigma^{2(r)}_j = \frac{1}{N}\Big(\mathrm{E}_{\widetilde{p}^{(r)}_{A_j}}\big[\boldsymbol{a}^t_j\widetilde{\boldsymbol{\Lambda}}^{(r)}_j\boldsymbol{a}_j\big] - 2\boldsymbol{m}^{(r)t}_{A_j}\widetilde{\boldsymbol{G}}\boldsymbol{\Lambda}^{(r)}_j\boldsymbol{y}^{(r)}_j + \boldsymbol{y}^{(r)t}_j\boldsymbol{\Lambda}^{(r)}_j\boldsymbol{y}^{(r)}_j\Big) = F_2(\rho^{(r)}_j, \boldsymbol{\ell}^{(r)}_j)$,

where F_2 is a function linking the estimates $\sigma^{2(r)}_j$ with $\boldsymbol{\ell}^{(r)}_j$ and $\rho^{(r)}_j$. Matrix $\widetilde{\boldsymbol{\Lambda}}^{(r)}_j = \mathrm{E}_{\widetilde{p}^{(r)}_{H_\gamma}}\big[\boldsymbol{G}^{\mathrm{t}}\boldsymbol{\Lambda}^{(r)}_j\boldsymbol{G}\big]$ is a $M \times M$ matrix similar to the matrix $\widetilde{\boldsymbol{H}}_j$ introduced in the VE-A step. Its (m, m') entry is given by $\widetilde{\boldsymbol{g}}^{\mathrm{t}}_m\boldsymbol{\Lambda}^{(r)}_j\widetilde{\boldsymbol{g}}_{m'} + \mathrm{trace}\big(\boldsymbol{\Lambda}^{(r)}_j\boldsymbol{X}_m\boldsymbol{\Sigma}^{(r)}_{H_\gamma}\boldsymbol{X}^{\mathrm{t}}_{m'}\big)$.

Eventually, the maximization wrt ρ_j leads to $\rho^{(r)}_j = \underset{\rho_j}{\arg\max}\{\big(\mathrm{trace}(\boldsymbol{U}_1\widetilde{\Lambda_j}) + \mathrm{trace}(\boldsymbol{U}_2\Lambda_j)\big)/\sigma^{2(r)}_j + \log|\Lambda_j|\}$, with $|\Lambda_j| = 1 - \rho^2_j$ and where $\widetilde{\Lambda_j}$ has the same expression as $\widetilde{\boldsymbol{\Lambda}}^{(r)}_j$ without the (r) superscript. Matrices $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are respectively $M \times M$ and $N \times N$ matrices defined as $\boldsymbol{U}_1 = \boldsymbol{\Sigma}^{(r)}_{A_j} + \boldsymbol{m}^{(r)}_{A_j}\boldsymbol{m}^{(r)\mathrm{t}}_{A_j}$ and $\boldsymbol{U}_2 = \boldsymbol{y}^{(r)}_j(\boldsymbol{y}^{(r)}_j - 2\widetilde{\boldsymbol{G}}\boldsymbol{m}^{(r)}_{A_j})^{\mathrm{t}}$. The derivative, denoted by Λ_j of Λ_j wrt ρ_j writes $\Lambda'_j = 2\rho_j\boldsymbol{B} + \boldsymbol{C}$, where the entries of $\boldsymbol{B}$ and $\boldsymbol{C}$ are zero except $(\boldsymbol{B})_{n,n}$ which is 1 for $n = 2 : (N-1)$ and for $(\boldsymbol{C})_{n,n+1}$ and $(\boldsymbol{C})_{n+1,n}$ which are -1 for $n = 1 : (N-1)$. The derivative, denoted by $\widetilde{\Lambda}'_j$, of $\widetilde{\Lambda}_j$ wrt ρ_j can be written as: $\widetilde{\Lambda}'_j = 2\rho_j\widetilde{\boldsymbol{B}} + \widetilde{\boldsymbol{C}}$ where $\widetilde{\boldsymbol{B}}$ and $\widetilde{\boldsymbol{C}}$ are $M \times M$ matrices whose entries (m, m') are respectively $(\widetilde{\boldsymbol{B}})_{m,m'} = \mathrm{trace}\big((\boldsymbol{X}_m\boldsymbol{\Sigma}^{(r)}_{\boldsymbol{H}_\gamma}\boldsymbol{X}^{\mathrm{t}}_{m'} + \widetilde{\boldsymbol{g}}_{m'}\widetilde{\boldsymbol{g}}^{\mathrm{t}}_m)\boldsymbol{B}\big)$ and $(\widetilde{\boldsymbol{C}})_{m,m'} = \mathrm{trace}\big((\boldsymbol{X}_m\boldsymbol{\Sigma}^{(r)}_{\boldsymbol{H}_\gamma}\boldsymbol{X}^{\mathrm{t}}_{m'} + \widetilde{\boldsymbol{g}}_{m'}\widetilde{\boldsymbol{g}}^{\mathrm{t}}_m)\boldsymbol{C}\big)$. Eventually, the derivative wrt ρ_j leads to $\rho^{(r)}_j = \frac{1-\rho^{(r)}_j}{\sigma^{2(r)}_j}\Big(2\rho^{(r)}_j\big(\mathrm{trace}(\boldsymbol{U}_1\widetilde{\boldsymbol{B}}) + \mathrm{trace}(\boldsymbol{U}_2\boldsymbol{B})\big) + \mathrm{trace}(\boldsymbol{U}_1\widetilde{\boldsymbol{C}}) + \mathrm{trace}(\boldsymbol{U}_2\boldsymbol{C})\Big) = F_3(\rho^{(r)}_j, \sigma^{2(r)}_j)$.

Then $\rho^{(r)}_j$ can be estimated as a solution of the fixed point equation $\rho^{(r)}_j = F_3(\rho^{(r)}_j, F_2(\rho^{(r)}_j, F_1(\rho^{(r)}_j)))$.

Note that in the Gaussian noise case, the updating of the noise parameters reduces to the estimation of $\sigma^{2(r)}_j$ which simplifies into $\sigma^{2(r)}_j = \frac{1}{N}\big(\mathrm{E}_{\widetilde{p}^{(r)}_{A_j}}\big[\boldsymbol{a}^t_j\mathrm{E}_{\widetilde{p}^{(r)}_{H_\gamma}}\big[\boldsymbol{G}^{\mathrm{t}}\boldsymbol{G}\big]\boldsymbol{a}_j\big] - 2\boldsymbol{m}^{(r)t}_{A_j}\widetilde{\boldsymbol{G}}\boldsymbol{y}^{(r)}_j + \boldsymbol{y}^{(r)t}_j\boldsymbol{y}^{(r)}_j\big)$.

REFERENCES

[1] S. Ogawa, T. Lee, A. Kay, and D. Tank, "Brain magnetic resonance imaging with contrast dependent on blood oxygenation," *Proc. Natl. Acad. Sci. USA*, vol. 87, no. 24, pp. 9868–9872, 1990.

[2] K. J. Friston, A. P. Holmes, J.-B. Poline, P.J. Grasby, S.C.R. Williams, R.S.J. Frackowiak, and R. Turner, "Analysis of fMRI time-series revisited," *Neuroimage*, vol. 2, no. 2, pp. 45–53, Mar. 1995.

[3] B. Biswal, F. Z. Yetkin, V. M. Haughton, and J. S. Hyde, "Functional connectivity in the motor cortex of resting human brain using echo-planar MRI," *Magn. Reson. Med.*, vol. 34, no. 4, pp. 537–541, 1995.

[4] M. D. Fox, A. Z. Snyder, J. L. Vincent, M. Corbetta, D. C. Van Essen, and M. E. Raichle, "The human brain is intrinsically organized into dynamic, anticorrelated functional networks.," *Proc. Natl. Acad. Sci. USA*, vol. 102, no. 27, pp. 9673–9678, Jul. 2005.

[5] R. B. Buxton and L. Frank, "A model for the coupling between cerebral blood flow and oxygen metabolism during neural stimulation," *J. Cereb. Blood Flow Metab.*, vol. 17, no. 1, pp. 64–72, Jan. 1997.

[6] K. J. Friston, A. Mechelli, R. Turner, and C. J. Price, "Nonlinear responses in fMRI: the balloon model, Volterra kernels, and other hemodynamics," *Neuroimage*, vol. 12, pp. 466–477, June 2000.

[7] R. B. Buxton, K. Uludag, D. J. Dubowitz, and T. T. Liu, "Modeling the hemodynamic response to brain activation," *Neuroimage*, vol. 23, no. Suppl. 1, pp. S220–S233, 2004.

[8] J. J. Riera, J. Watanabe, I. Kazuki, M. Naoki, E. Aubert, T. Ozaki, and R. Kawashima, "A state-space model of the hemodynamic approach: nonlinear filtering of BOLD signal," *Neuroimage*, vol. 21, pp. 547–567, 2004.

[9] D. Lashkari, R. Sridharan, E. Vul, P. J. Hsieh, N. Kanwisher, and P. Golland, "Search for patterns of functional specificity in the brain: a nonparametric hierarchical Bayesian model for group fMRI data," *Neuroimage*, vol. 59, no. 2, pp. 1348–1368, Jan. 2012.

[10] K.J. Friston, P. Jezzard, and R. Turner, "Analysis of functional MRI time-series," *Hum. Brain Mapp.*, vol. 1, pp. 153–171, 1994.

[11] G. M. Boynton, S. A. Engel, G. H. Glover, and D. J. Heeger, "Linear systems analysis of functional magnetic resonance imaging in human V1," *J. Neurosci.*, vol. 16, no. 13, pp. 4207–4221, July 1996.

[12] D. A. Handwerker, J.M. Ollinger, , and M. D'Esposito, "Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses," *Neuroimage*, vol. 21, pp. 1639–1651, Apr. 2004.

[13] P. Ciuciu, T. Vincent, A.-L. Fouque, and A. Roche, "Improved fMRI group studies based on spatially varying non-parametric BOLD signal modeling," in *5th Proc. IEEE ISBI*, Paris, France, May 2008, pp. 1263–1266.

[14] S. Badillo, T. Vincent, and P. Ciuciu, "Impact of the joint detection-estimation approach on random effects group studies in fMRI," in *8th Proc. IEEE ISBI*, Chicago, IL, Apr. 2011, pp. 376–380.

[15] M.A. Lindquist, J. Meng Loh, L.Y. Atlas, and T.D. Wager, "Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling," *Neuroimage*, vol. 45, no. 1 (Suppl), pp. S187–S198, 2009.

[16] G. H. Glover, "Deconvolution of impulse response in event-related BOLD fMRI," *Neuroimage*, vol. 9, no. 4, pp. 416–429, Apr. 1999.

[17] R. Henson, M. Rugg, and K. Friston, "The choice of basis function in event-related fMRI," in *Neuroimage (HBM'01)*, June 2001, vol. 13.

[18] C. Goutte, F. A. Nielsen, and L. K. Hansen, "Modeling the haemodynamic response in fMRI using smooth filters," *IEEE Trans. Med. Imag.*, vol. 19, no. 12, pp. 1188–1201, Dec. 2000.

[19] P. Ciuciu, J.-B. Poline, G. Marrelec, J. Idier, Ch. Pallier, and H. Benali, "Unsupervised robust non-parametric estimation of the hemodynamic response function for any fMRI experiment," *IEEE Trans. Med. Imag.*, vol. 22, no. 10, pp. 1235–1251, Oct. 2003.

[20] G. Marrelec, P. Ciuciu, M. Pélégrini-Issac, and H. Benali, "Estimation of the hemodynamic response function in event-related functional MRI: Bayesian networks as a framework for efficient Bayesian modeling and inference," *IEEE Trans. Med. Imag.*, vol. 23, no. 8, pp. 959–967, 2004.

[21] J. Kershaw, B. A. Ardekani, and I. Kanno, "Application of Bayesian inference to fMRI data analysis," *IEEE Trans. Med. Imag.*, vol. 18, no. 12, pp. 1138–1152, Dec. 1999.

[22] S. Makni, P. Ciuciu, J. Idier, and J.-B. Poline, "Joint detection-estimation of brain activity in functional MRI: a multichannel deconvolution solution," *IEEE Trans. Signal Processing*, vol. 53, no. 9, pp. 3488–3502, Sep. 2005.

[23] S. Makni, J. Idier, T. Vincent, B. Thirion, G. Dehaene-Lambertz, and P. Ciuciu, "A fully Bayesian approach to the parcel-based detection-estimation of brain activity in fMRI," *Neuroimage*, vol. 41, no. 3, pp. 941–969, July 2008.

[24] F. de Pasquale, C. Del Gratta, and G. L. Romani, "Empirical Markov Chain Monte Carlo Bayesian analysis of MRI data," *Neuroimage*, vol. 42, no. 1, pp. 99–111, Aug. 2008.

[25] T. Vincent, L. Risser, and P. Ciuciu, "Spatially adaptive mixture modeling for analysis of within-subject fMRI time series," *IEEE Trans. Med. Imag.*, vol. 29, no. 4, pp. 1059–1074, Apr. 2010.

[26] B. Thirion, G. Flandin, P. Pinel, A. Roche, P. Ciuciu, and J.-B. Poline, "Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets," *Hum. Brain Mapp.*, vol. 27, no. 8, pp. 678–693, Aug. 2006.

[27] W. D. Penny, S. Kiebel, and K. J. Friston, "Variational Bayesian inference for fMRI time series," *Neuroimage*, vol. 19, no. 3, pp. 727–741, July 2003.

[28] K.J. Friston, J. Mattout, N. Trullijo-Barreto, J. Ashburner, and W. Penny, "Variational free energy and the Laplace approximation," *Neuroimage*, vol. 34, no. 1, pp. 220–234, Jan. 2007.

[29] S. Makni, Ch. Beckmann, S. Smith, and M. Woolrich, "Bayesian deconvolution fMRI data using bilinear dynamical systems," *Neuroimage*, vol. 42, no. 4, pp. 1381–1396, Oct. 2008.

[30] M. Woolrich and T. Behrens, "Variational Bayes inference of spatial mixture models for segmentation," *IEEE Trans. Med. Imag.*, vol. 25, no. 10, pp. 1380–1391, Oct. 2006.

[31] L. Risser, T. Vincent, F. Forbes, J. Idier, and P. Ciuciu, "Min-max extrapolation scheme for fast estimation of 3D Potts field partition functions. application to the joint detection-estimation of brain activity in fMRI.," vol. 65, no. 3, pp. 325–338, Dec. 2011.

[32] M. Woolrich, B. Ripley, M. Brady, and S. Smith, "Temporal autocorrelation in univariate linear modelling of fMRI data," *Neuroimage*, vol. 14, no. 6, pp. 1370–1386, Dec. 2001.

[33] M. Woolrich, M. Jenkinson, J. Brady, and S. Smith, "Fully Bayesian spatio-temporal modelling of fMRI data," *IEEE Trans. Med. Imag.*, vol. 23, no. 2, pp. 213–231, Feb. 2004.

[34] W. D. Penny, G. Flandin, and N. Trujillo-Bareto, "Bayesian comparison of spatially regularised general linear models," *Hum. Brain Mapp.*, vol. 28, no. 4, pp. 275–293, Apr. 2007.

[35] W. D. Penny, N. Trujillo-Barreto, and K. J. Friston, "Bayesian fMRI time series analysis with spatial priors," *Neuroimage*, vol. 23, no. 2, pp. 350–362, 2005.

[36] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse and other variants," in *Learning in Graphical Models*, M.I. Jordan, Ed., pp. 355–368. Kluwer Academic Publishers, Dordrecht, Netherlands, 1998.

[37] M. J. Beal and Z. Ghahramani, "The variational Bayesian EM Algorithm for incomplete data: with application to scoring graphical model structures," in *Bayesian Statistics*, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, Eds. 2003, pp. 453–464, Oxford University Press.

[38] J. Goldsmith, M. P. Wand, and C. Crainiceanu, "Functional regression via variational Bayes," *Electronic Journal of Statistics*, vol. 5, pp. 572–602, Jan. 2011.

[39] M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[40] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, *Bayesian data analysis*, Chapmann & Hall, London, UK, Second edition, 2004.

[41] G. Celeux, F. Forbes, and N. Peyrard, "EM procedures using mean field-like approximations for Markov model-based image segmentation," *Pattern Recognition*, vol. 36, no. 1, pp. 131–144, Jan. 2003.

[42] R. Casanova, S. Ryali, J. Serences, L. Yang, R. Kraft, P.J. Laurienti, and J.A. Maldjian, "The impact of temporal regularization on estimates of the BOLD hemodynamic response function: a comparative analysis," *Neuroimage*, vol. 40, no. 4, pp. 1606–1618, May 2008.

[43] P. Pinel, B. Thirion, S. Mériaux, A. Jobert, J. Serres, D. Le Bihan, J.-B. Poline, and S. Dehaene, "Fast reproducible identification and large-scale databasing of individual functional cognitive networks," *BMC Neurosci.*, vol. 8, no. 1, pp. 91, Oct. 2007.

[44] Liu T. T., Frank L. R., Wong E. C., and Buxton R. B., "Detection power, estimation efficiency, and predictability in event-related fMRI," *Neuroimage*, vol. 13, no. 4, pp. 759–773, Apr. 2001.

[45] O. Gruber, P. Indefrey, H. Steinmetz, and A. Kleinschmidt, "Dissociating neural correlates of cognitive components in mental calculation," *Cereb. Cortex*, vol. 11, no. 4, pp. 350–359, Apr. 2001.

[46] A. Gelman and D. B. Rubin, "Inference from iterative simulation using multiple sequences," *Stat. Science*, vol. 7, no. 4, pp. 457–472, 1992.

[47] P. Carbonetto and M. Stephens, "Scalable Variational Inference for Bayesian Variable Selection in Regression, and its Accuracy in Genetic Association Studies," *Bayesian Analysis*, vol. 7, pp. 73–108, 2012.

[48] A. Nummenmaa, T. Auranen, M.S. Hämäläinen, I.P. Jääskeläinen, J. Lampinen, M. Sams, and A. Vehtari, "Hierarchical Bayesian estimates of distributed MEG sources: theoretical aspects and comparison of variational and MCMC methods," *Neuroimage*, vol. 35, no. 2, pp. 669–685, Apr. 2007.

[49] P. Ciuciu, S. Sockeel, T. Vincent, and J. Idier, "Modelling the neurovascular habituation effect on fMRI time series," in *34th Proc. IEEE ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 433–436.

[50] A. Tucholka, B. Thirion, M. Perrot, P. Pinel, J.-F. Mangin, and J.-B. Poline, "Probabilistic anatomo-functional parcellation of the cortex: how many regions?," in *11thProc. MICCAI, LNCS Springer Verlag*, New-York, USA, 2008.

[51] T. Vincent, P. Ciuciu, and B. Thirion, "Sensitivity analysis of parcellation in the joint detection-estimation of brain activity in fMRI," in *5th Proc. IEEE ISBI*, Paris, France, May 2008, pp. 568–571.

[52] F. Forbes and N. Peyrard, "Hidden Markov Random Field model selection criteria based on mean field-like approximations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1089–1101, Sep. 2003.