Variational Variable Selection To Assess Experimental Condition Relevance In Event-Related fMRI

Christine Bakhous, Florence Forbes, Thomas Vincent, Michel Dojat, Philippe Ciuciu

VARIATIONAL VARIABLE SELECTION TO ASSESS EXPERIMENTAL CONDITION RELEVANCE IN EVENT-RELATED FMRI

Christine Bakhous[1,3]*, Florence Forbes*[1]*, Thomas Vincent*[1]*, Michel Dojat*[3]*, Philippe Ciuciu*[2]

[1] INRIA, MISTIS, Grenoble University, LJK, Grenoble, France
[2] CEA/NeuroSpin & INRIA, Parietal, GIF SUR YVETTE, cedex France
[3] INSERM, U836, Grenoble University, GIN, Grenoble, France

ABSTRACT

Brain functional exploration investigates the nature of neural processing following cognitive or sensory stimulation. This goal is not fully accounted for in most functional Magnetic Resonance Imaging (fMRI) analysis which usually assumes that all delivered stimuli possibly generate a BOLD response everywhere in the brain although activation is likely to be induced by only some of them in specific brain regions. Generally, criteria are not available to select the relevant conditions or stimulus types (e.g. visual, auditory, etc.) prior to activation detection and the inclusion of irrelevant events may degrade the results, particularly when the Hemodynamic Response Function (HRF) is jointly estimated. To face this issue, we propose an efficient variational procedure that automatically selects the conditions according to the brain activity they elicit. It follows an improved activation detection and local HRF estimation that we illustrate on synthetic and real fMRI data.

Index Terms— Irrelevance detection, Stimulus type selection, Joint detection-estimation, Bayesian hierarchical modelling, Functional magnetic resonance imaging

1. INTRODUCTION

Event-related functional MRI refers to a technique for detecting the brain's response to brief stimuli. It relies on both (i) a detection step to localize which brain regions are activated by a given stimulus type, and on (ii) an estimation step to recover the underlying Blood Oxygenation Level Dependent (BOLD) signal dynamics. The latter is generally skipped and a predefined (canonical) unique model for the HRF is assumed for the entire brain although such an assumption may not be valid typically in any pathological context. Event-related designs allow different trials or stimuli to be presented in arbitrary sequences, which enables to investigate both distributed and local neural processing, convolved by the HRF, and provides therefore a powerful tool for functional brain exploration. For relevant cognitive interpretation, an important challenge consists of optimally designing such event-related fMRI experiments so as to maximize the accuracy with which the activation probabilities can be evaluated and the event-related hemodynamic response to different stimuli estimated. Once the data has been acquired, a crucial issue concerns the design of the activation model: most often, this question is addressed in the General Linear Model (GLM) context, using a predefined global HRF model, by comparing different model structures (eg sparse design matrix [1]) using Fisher tests between reduced and full models [2]. Following [2], a critical point in such attempts is the risk of discarding relevant activation due to HRF mis-specification. In the alternative Bayesian detection-estimation approach (JDE) of [3], this model definition becomes region or (parcel)-specific so as to account for spatial hemodynamic variability. The JDE enables HRF estimation in addition to activation detection. Then, to optimize the activation probabilities, it makes sense to consider the most sparse model by assessing the relevance to include or not each of the experimental conditions in a specific parcel. Indeed, the activation of interest is likely to be induced by only a subset of these conditions when functional segregation of the parcel can be assumed with respect to a specific cognitive task. The relevant stimulus types may fluctuate across regions making this optimization procedure a complex combinatorial task.

In this paper, we address this issue as a model specification problem within the JDE framework. In contrast to model selection approaches (e.g. [4]) that require the comparison of the performance of several models and select the most appropriate one, we propose a single parsimonious procedure that includes the automatic selection of the experimental conditions that best explain brain activity. This is done by introducing for each stimulus type an additional binary variable as a measure of its relevance (in terms of evoked activity). In a regression context, the idea of adding such indicator variables is usually referred to as variable selection (see *e.g.* [5]) and has been used in [6, 2] to capture evoked brain activity in a sparse manner. In the JDE framework, this activity detection task is already handled within the model in a more general way. Activated and non-activated voxels are modelled using a two-class Gaussian mixture instead of a Bernoulli-Gaussian prior [6]. Our use of binary variables is then rather

oriented toward the selection of stimulus types which has to be done across the whole set of voxelwise regressions (Section 2). The proposed approach is carried out in a variational EM framework (Section 3), which offers a faster alternative to intensive stochastic procedures as used in [7]. Simulated experiments confirm the ability of our model to select the relevant conditions while real fMRI data illustrate an enhanced determination of activated brain regions (Section 4). The proposed model is referred to as the *parsimonious model* while the model in which all stimulus types are included is referred to as the *complete model*.

2. PARSIMONIOUS JOINT DETECTION ESTIMATION MODEL

A vector is by convention a column vector. The transpose is denoted by $^{\mathrm{t}}$. The Gaussian distribution with mean μ and variance $\mathbf{\Sigma}$ is denoted using $\mathcal{N}(\mu, \mathbf{\Sigma})$.

Following [3, 8], for a given brain parcel γ, the observed data is denoted by $y = \{y_j, j \in \gamma\}$ where y_j is a N-dimensional vector representing the BOLD signal measured at voxel $j \in \gamma$. Activity status are indicated by activation class assignment variables $q = \{q^m, m = 1 : M\}$ where $q^m = \left\{q_j^m, j \in \gamma\right\}$ and $q_j^m = i$ means that voxel j lies in activation class i for the mth experimental condition. Typically the number of classes is 2 for activated $(i = 1)$ and inactivated $(i = 0)$ voxels. The activation amplitudes are modelled by the so-called Neural Response Levels (NRLs) $a = \{a^m, m = 1 : M\}$ with $a^m = \left\{a_j^m, j \in \gamma\right\}$ and $a_j = \left\{a_j^m, m = 1 : M\right\}$. In our parsimonious context, our goal is to account for the fact that only a subset among M experimental conditions (or stimulus types) are necessary to explain the evoked BOLD signal in a given parcel. A stimulus type will be identified as irrelevant for the data under consideration if the average activation amplitude is too small for this condition. In this case, we consider that such evoked activity is artifactual and decide that the stimulus type should be discarded from the model definition. This is different from [7] in which a condition is considered as irrelevant when the number of activated voxels is too small. As confirmed by the experiments below, defining an appropriate relevance threshold appears to be easier and less data dependent for the average activation amplitude criterion. To encode such information, we then consider a set of M binary variables $w = \{w^m, m = 1 : M\}$ where $w^m = 1$ (resp. $w^m = 0$) means that the mth stimulus type is relevant (resp. irrelevant). The unknown HRF function denoted by $h = [h_0, h_{\Delta t}, , \ldots, h_{D\Delta t}]^{\mathrm{t}}$ is then a $(D + 1)$-real valued vector with Δt the sampling period of the HRF. Physiological artifacts are modelled by $\boldsymbol{P}\ell$ where $\boldsymbol{P}$ is a low frequency orthogonal $N{\times}L$ matrix and $\ell = \{\ell_j, j \in \gamma\}$ the set of low frequency drifts.

The observed signals are then explained by the following generative model implying additional parameters to be estimated or fixed as specified below:

$$\forall j \in \gamma, \quad y_j = \sum_{m=1}^{M} w^m a_j^m \boldsymbol{X}^m h + \boldsymbol{P}\ell_j + \varepsilon_j, \tag{1}$$

where $\boldsymbol{X}^m$ denotes the $N{\times}(D+1)$ binary matrix that codes the arrival times of the events of type m which are approximated to fit a Δt-sampled grid, ε_j's stand for the independent and normally distributed noise, $\varepsilon_j \sim \mathcal{N}(\mathbf{0}, \sigma_j^2 \boldsymbol{I}_N)$, ($\boldsymbol{I}_N$is the $N \times N$ identity matrix).

Under standard additional assumptions (see in [3, 8]) and omitting the dependence over parameters, the joint distribution $p(y, w, a, h, q)$ reads $p(y \,|\, w, a, h)\; p(a \,|\, w, q)\; p(w)\; p(h)\; p(q)$ and can be further specified as follows.

The $p(y \,|\, w, a, h)$ **term** follows from (1):
$p(y \,|\, w, a, h) = \prod_{j\in\gamma} p(y_j \,|\, w, a_j, h; \sigma_j^2)$ with
$(y_j \,|\, w, a_j, h; \sigma_j^2) \sim \mathcal{N}\big(\sum_{m=1}^{M} w^m a_j^m \boldsymbol{X}^m h + \boldsymbol{P}\ell_j, \sigma_j^2 \boldsymbol{I}_N\big)$.

In **the** $p(a \,|\, w, q)$ **term**, the assignment variables q_j^m are introduced to segregate activated from non-activated voxels.

The NRLs are assumed to be independent conditionally to the q_j^m's and w^m's: $p(a \,|\, w, q) = \prod_m \prod_j p(a_j^m \,|\, w^m, q_j^m)$. The relevance variable w^m is then accounted for by assuming for $i \in \{0, 1\}$, $\left(a_j^m | w^m = 1, q_j^m = i\right) \sim \mathcal{N}(\mu_i^m, v_i^m)$ and $\left(a_j^m | w^m = 0, q_j^m = i\right) \sim \mathcal{N}(\mu_0^m, v_0^m)$. The Gaussian parameters $\theta_a = \{\mu_1^m, v_1^m, v_0^m, m = 1 : M\}$ need to be estimated but we set $\mu_0^m = 0$ for all m. The idea is that for a relevant stimulus type ($w^m = 1$), the distribution of a_j^m depends on the activation state q_j^m in voxel j while for an irrelevant stimulus type ($w^m = 0$), q_j^m has no influence on a_j^m, which is distributed around 0 to account for the absence of response to stimulus type m.

The $p(w)$ **term** links the relevance variables to the activation class means μ_1^m's. The binary relevance variables w are independent across stimulus types, $p(w) = \prod_{m=1}^{M} p(w^m)$ and for each m, w^m follows a Bernoulli distribution whose probability of success is given via a logit link to μ_1^m, $P(w^m = 1) = \mathcal{F}(\mu_1^m; \tau_1, \tau_2)$,where $\mathcal{F}$ is the sigmoid function $\mathcal{F}(x; \tau_1, \tau_2) = (1 + \exp(-\tau_1(x^2 - \tau_2)))^{-1}$ with τ_1 controlling the slope of the sigmoid and τ_2 the inflection point that can be seen as a relevance threshold above which the stimulus type will be considered as relevant with a high probability. A condition m is considered as irrelevant when the estimated μ_1^m is close to zero reflecting the absence of significant activations for m.

The $p(h)$ **and** $p(q)$ **terms** are then set as in [8] and thus not detailed here. They involve respectively a HRF smoothness parameter σ_h^2 and spatial interaction parameters for each condition $\{\beta^m, m = 1 : M\}$.

The set of parameters θ is then $\theta = \{\sigma_j^2, \mu_1^m, v_1^m, v_0^m, \tau_1, \tau_2, \sigma_h^2, \beta^m, j \in \gamma, m = 1 : M\}$. They will be inferred using the following variational EM framework except for the additional parameters τ_1 and τ_2 involved in the $\mathcal{F}$ logit link which are fixed as indicated in Section 4.

3. ESTIMATION BY VARIATIONAL EM

Our Bayesian model is too complex to be amenable to analytical calculations. Hence, we resort to an iterative variational Expectation-Maximization (EM) procedure as in [8]. Compared to [8], this implies adding in the E-step, M additional stages for the M binary variables w and reporting their impact on the other unobserved variables in the E-step and the unknown parameters in the M-step. Here, of particular interest are the variational estimates of the posterior probabilities $p(w^m \mid y)$ for the stimulus types. At iteration (r), the current estimates of these probabilities are denoted by $\tilde{p}^{(r)}_{w^m}$. We can then derive a *relevance profile* (a M-dimensional vector of 0 and 1 depending on whether a stimulus type is relevant or not). Denoting in addition the current parameter values by $\theta^{(r)}$ and $\mathrm{E}_{\tilde{p}}[.]$ the expectation with respect to some pdf $\tilde{p}$, for each $m = 1 : M$, $\tilde{p}^{(r)}_{w^m}(w^m)$ is proportional to:

$$\exp\left(\mathrm{E}_{\tilde{p}^{(r-1)}_{a,q,h,w^{\setminus m}}}\left[\log p(w^m \mid y, a, q, h, w^{\setminus m}, \theta^{(r)})\right]\right)$$

where $w^{\setminus m} = \{w^{m'}, m' \neq m\}$ and $\tilde{p}^{(r-1)}_{a,q,h,w^{\setminus m}} = \tilde{p}^{(r-1)}_{a}\, \tilde{p}^{(r-1)}_{q}\, \tilde{p}^{(r-1)}_{h}\, \tilde{p}^{(r-1)}_{w^{\setminus m}}$ denotes the current estimated variational pdf over variables a, q, h and $w^{\setminus m}$. It follows that $\tilde{p}^{(r)}_{w^m}(w^m)$ is proportional to the product of three terms involving different groups: $\prod_{j\in\gamma} \exp\left(\mathrm{E}_{\tilde{p}^{(r-1)}_{h,a_j,w^{\setminus m}}}\left[\log p(y_j \mid w, a_j, h; \sigma_j^{2(r)})\right]\right)$, $\prod_{j\in\gamma} \exp\left(\mathrm{E}_{\tilde{p}^{(r-1)}_{a^m_j,q^m_j}}\left[\log p(a^m_j \mid w^m, q^m_j; \theta^{(r)}_a)\right]\right)$ and $p(w^m; \mu_1^{m(r)}, \tau_1, \tau_2)$. More specifically, further calculations show that the probability of w^m is impacted by the estimated value of μ_1^m in the third term above. When $\mu_1^{m2} \geq \tau_2$, the third term tends to favor $w^m = 1$ vs $w^m = 0$. Similarly, the second term depends on the $\tilde{p}^{(r-1)}_{a^m_j}$'s for $j \in \gamma$. When the means of these pdfs tend to 0, $w^m = 0$ is favored. The expression of the first term is more complex but $w^m = 0$ is favored when the y_j's are well explained by the model without type m or equivalently when the current noise model parameters can accommodate the absence of stimulus type m. The E-step can then be completed using similar E-sub-steps obtained by exchanging the role of the model variables (see [8] for details). Regarding the M-step, it also decomposes into four sub-steps involving separately $\{\sigma_j^2, j \in \gamma\}$, σ_h^2, $\{\mu_i^m, v_i^m, i = 0, 1\}$ and $\{\beta^m, m = 1 : M\}$. The first two maximizers admit closed-form expressions close to that in [8] with an additional dependence on the current $\tilde{p}^{(r)}_{w^m}(w^m = 1)$ for the first one. The third maximizer leads to close form expressions for the v_i^m's and uses a half-interval search algorithm for the μ_1^m's. The last M-step requires an iterative maximization procedure to update the β^m's as in [8].

4. RESULTS:

Simulated datasets. Experiments were carried out by simulating, in a 2D slice of 400 voxels, from the model given in Eq.(1) with 2 conditions, one relevant $(w^1 = 1)$ (Fig.1-(a,c)) and the other irrelevant $(w^2 = 0)$ (Fig.1-(b,d)). The noise variance σ_j^2 was set to 1 and the HRF fixed to a non canonical shape (Fig.3-(a)). In the *parsimonious model*, appropriate values are required for the sigmoid parameters τ_1 and τ_2. The τ_2 value was set to 0.5, which means that when $|\mu_1^m| > 0.7$, $w^m = 1$ is favored. Parameter τ_1 can be set to $\frac{1}{\tau_2} \log \frac{1-p_0^m}{p_0^m}$ where $p_0^m = p(w^m = 1; \mu_1^m = 0, \tau_1, \tau_2)$ is the value of the sigmoid at 0 (for $\tau_2 = 0.5$ and $p_0^m = 10^{-5}$, $\tau_1 = 23$). The VEM procedure estimates relevance probabilities $\tilde{p}_{w^m}(w^m = 1)$ close to 1 for relevant condition and to 0 for irrelevant one. An improvement of both detection and estimation is expected by discarding irrelevant conditions. To quantify this potential gain, tests were performed over 100 runs. Using the ground truth NRL's shown in Fig.1-(a,b), the mean square error (MSE) was computed for the NRL's at each voxel for both parsimonious and complete models. The T statistic of the differences between these MSE's are then plotted in Fig.2-(a) for the voxels for which the difference was tested as statistically significant (to a 5% threshold). Fig.2-(a) shows that for most voxels, the error is significantly lower with the parsimonious model (negative differences), especially for activated voxels. This suggests a more acute activity detection for the parsimonious model. Fig.2-(b,c) shows estimated probabilities $\tilde{p}_{Q_j^m}(q_j^m = 1)$ of relevant condition for the parsimonious and complete models. The complete model tends to estimate this probabilities around 0.5 for the irrelevant condition (Fig.2-(d)) causing a high false positive rate ($\frac{263}{400}$ voxels are classified as activated). As shown in Fig.3-(a) the estimated HRFs for both models are close but their differences have been tested as statistically significant ($p < 0.01$).

Real fMRI datasets. We considered real non smoothed data recorded following a fast event-related paradigm [9] comprising 80 auditory, visual and motor stimuli and cognitive tasks such as number processing and language comprehension. It consisted of a single session of $N = 128$ functional scans lasting 2.4s each, yielding 3D volumes composed of $96 \times 96 \times 40$ voxels. The average ISI was of 3.76s with a std of 1.99s.

To better assess the impact of the relevance variables w, we considered a brain region $\gamma$where some of the stimulus types are likely to be irrelevant. We focused on the right motor area (162 voxels) where motor stimuli are likely to induce evoked activity. To derive relevant contrast, the experimental conditions were grouped in four categories: computation, sentence, checkerboard and click (i.e., motor response). In this motor region, it is expected that the checkerboard, sentence and calculation categories are set as irrelevant. Sigmoid parameters were set as before. Inference with the *parsimonious model* concluded that the calculation, sentence and checkerboard conditions were irrelevant, with estimated posterior probabilities $p(w^m = 0 \mid y)$ close to 1. The click stimulus type was estimated as relevant with probability 1. The *parsimonious model* provided larger NRLs and normalized NRLs for the relevant motor condition (click). As an alternative to statistical parametric maps (SPMs) used in

classical inferences, posterior probability maps (PPMs) are images of the probability $p(a_j^m > \alpha|y)$ that a NRL a_j^m exceeds some threshold α, given the data [10]. The threshold α operationally defines the meaning of "activation" and was set to the prior standard deviation of the nonactivated class ($\sqrt{v_0^m}$). We considered that below this value a NRL is not large enough to be considered as activated. As v_0^m is possibly estimated differently for the parsimonious and complete models, we set $\alpha = 5.63$ which is the average of the two estimations. The superimposed histograms of the PPMs (PPM$^\alpha$) values greater than $\rho = 0.5$ for both models are shown in Fig.4-(a). It is then common to threshold the PPMs using a $\rho = 0.95$ or higher threshold, *i.e* to show only the voxels j for which $p(a_j^m > \alpha|y) > 0.95$. Fig.4-(b,c) shows these voxels in red for the parsimonious and complete models. The complete model identifies a smaller number of voxels (69 *vs* 81 for the parsimonious model and it becomes 40 *vs* 81 for $\alpha = 6$) suggesting that it may have missed some activation and that the parsimonious model provides more confidence in the selected activated voxels. Finally, Fig.3-(b) shows the estimated HRFs superimposed to the canonical HRF.

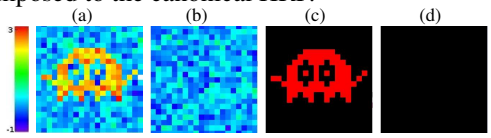

Fig. 1. Simulated NRLs (a,b) and activity status (c,d) resp. for the relevant and irrelevant conditions.

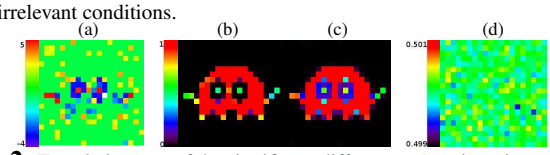

Fig. 2. T-statistic maps of the significant differences (parsimonious vs complete) between NRLs' MSE (a) over 100 runs. $\widetilde{p}_{Q_j^m}(q_j^m = 1)$ maps for the relevant condition in parsimonious (b) and complete (c) models and for the irrelevant condition in the complete model (d).

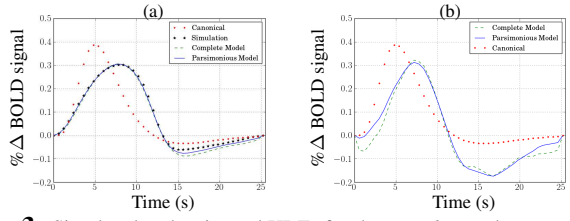

Fig. 3. Simulated and estimated HRFs for the *complete* and *parsimonious* models for the simulated (a) and the real motor region (b) datasets superimposed to the canonical HRF.

5. CONCLUSION AND FUTURE WORK

We proposed to go beyond the standard event-related fMRI data analysis which models all delivered stimuli as effects of interest in the GLM context. Using a Bayesian hierarchical approach we further explored the variable selection principle, used previously to detect evoked brain activity [6], as a tool to perform relevant condition selection. Experiments on synthetic and real data suggested the ability of our model to accurately select and exploit the most relevant stimulus types. Our

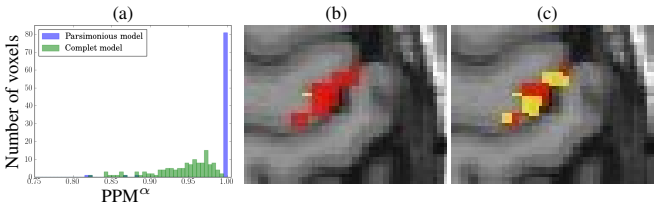

Fig. 4. PPM$^\alpha$ for the relevant condition (click) in the right motor region using an activation threshold $\alpha = 5$: (a) superimposed histograms of the PPM$^\alpha$ (the darkest color shows the overlap). (b,c) Thresholded PPM$^\alpha$ (1 in red and 0 in yellow) for a confidence level of 95% superimposed on the anatomical image, for parsimonious (b) and complete (c) models. Neurological convention: right is right.

parsimonious model improves both detection and estimation compared to the *complete model*. An advantage over the selection made in [7] is an easier choice for sigmoid parameters τ_1 and τ_2 that can be set in practice independently of the parcel size and the number of activated voxels, allowing in particular to detect small clusters with a high activation. Eventually, further real data analysis would be necessary for an extended study with a particular emphasis on the group-level impact of parcel-wise adaptive definition of parsimonious models [11].

6. REFERENCES

[1] H. Luo and S. Puthusserypady, "A Sparse Bayesian Method for Determination of Flexible Design Matrix for fMRI Data Analysis," *IEEE Trans. on Circuits and systems–I: regular papers*, vol. 52, no. 12, pp. 2699–2706, 2005.

[2] V.P. Oikonomou, K. Blekas, and L. Astrakas, "A Sparse and Spatially Constrained Generative Regression Model for fMRI Data Analysis," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 1, pp. 58–67, 2012.

[3] 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, pp. 1059–1074, 2010.

[4] S. Donnet, M. Lavielle, P. Ciuciu, and J.-B. Poline, "Selection of temporal models for event-related fMRI," in *2th Proc. ISBI*, Arlington, VA, 2004, pp. 992–995.

[5] R. B. O. Hara and M. J. Sillanpaa, "A review of Bayesian variable selection methods: What, how and which," *Bayesian Analysis*, vol. 4, no. 1, pp. 85–118, 2009.

[6] M. Smith and L. Fahrmeir, "Spatial Bayesian variable selection with application to functional MRI," *J. Stat. Amer. Assoc.*, vol. 102, no. 478, pp. 417–431, 2007.

[7] C. Bakhous, F. Forbes, T. Vincent, L. Chaari, M. Dojat, and P. Ciuciu, "Adaptive experimental condition selection in event-related fMRI," in *8th Proc. ISBI*, Barcelona, Spain, 2012, pp. 1755–1758.

[8] L. Chaari, T. Vincent, F. Forbes, M. Dojat, and P. Ciuciu, "Fast joint detection-estimation of evoked brain activity in event-related fMRI using a variational approach," *IEEE trans. Med. Imag.*, 10.1109/TMI.2012.2225636, to appear.

[9] P. Pinel, B. Thirion, S. Meriaux, 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, pp. 91, 2007.

[10] K. J. Friston, W. Penny, C. Phillips, S. J. Kiebel, G. Hinton, and J. Ashburner, "Classical and Bayesian inference in neuroimaging," *Neuroimage*, vol. 16, no. 2, pp. 465–483, Jun. 2002.

[11] S. Badillo, T. Vincent, and P. Ciuciu, "Impact of the joint detection-estimation approach on random effects group studies in fMRI," in *7th Proc. ISBI*, Chicago, IL, 2011, pp. 376–380.