HAL
open science

Unsupervised robust nonparametric estimation of the hemodynamic response function for any fMRI experiment.

Philippe Ciuciu, Jean-Baptiste Poline, Guillaume Marrelec, Jérôme Idier, Christophe Pallier, Habib Benali

To cite this version:

Philippe Ciuciu, Jean-Baptiste Poline, Guillaume Marrelec, Jérôme Idier, Christophe Pallier, et al.. Unsupervised robust nonparametric estimation of the hemodynamic response function for any fMRI experiment.. IEEE Transactions on Medical Imaging, 2003, 22 (10), pp.1235-51. 10.1109/TMI.2003.817759 . cea-00333694

HAL Id: cea-00333694

https://cea.hal.science/cea-00333694v1

Submitted on 23 Oct 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Unsupervised robust non-parametric estimation of the hemodynamic response function for any fMRI experiment

Philippe Ciuciu, *Member, IEEE*,
Jean-Baptiste Poline, Guillaume Marrelec, Jérôme Idier, Christophe Pallier and Habib Benali

***Abstract*— This paper deals with the estimation of the Blood Oxygen Level-Dependent (BOLD) response to a stimulus, as measured in Functional Magnetic Resonance Imaging (fMRI) data. A precise estimation is essential for a better understanding of cerebral activations. The most recent works have used a non-parametric framework for this estimation, considering each brain region as a system characterized by its impulse response, the so-called *Hemodynamic Response Function* (HRF). However, the use of these techniques has remained limited since they are not well-adapted to real fMRI data. Here, we develop a threefold extension to previous works. We consider *asynchronous* event-related paradigms, account for different trial types and integrate several fMRI sessions into the estimation.**

These generalizations are simultaneously addressed through a badly-conditioned observation model. Bayesian formalism is used to model temporal prior information of the underlying physiological process of the brain hemodynamic response. By this way, the HRF estimate results from a tradeoff between information brought by the data and by our prior knowledge. This tradeoff is modeled with hyperparameters that are set to the maximum-likelihood estimate using an Expectation Conditional Maximization (ECM) algorithm. The proposed unsupervised approach is validated on both synthetic and real fMRI data, the latter originating from a speech perception experiment.

***Index Terms*— event-related fMRI paradigm; HRF modeling; Bayesian estimation; ECM algorithm.**

I. Introduction

Manuscript received April 03, 2002; revised April 10, 2003.

Philippe Ciuciu, Jean–Baptiste Poline and Christophe Pallier are with SHFJ/CEA/INSERM U562, 91401 Orsay cedex, France (emails: name@shfj.cea.fr). Guillaume Marrelec and Habib Benali are with INSERM U494, CHU Pitié-Salpêtrière, 91 Bd de l'Hôpital, 75634 Paris cedex 13, France (emails: Firstname.Lastname@imed.jussieu.fr). Guillaume Marrelec is supported by the Fondation pour la Recherche Médicale. Jérôme Idier is with IRCCyN (CNRS), École Centrale de Nantes, 1 rue de la Noë, BP 92101, 44321 Nantes cedex 3 (email: Jérôme.Idier@irccyn.ec-nantes.fr). All authors apart J. Idier are with the Institut Fédératif de Recherche 49, Paris, France.

DYNAMIC brain functional imaging was born in the last decade with fMRI [1]. For one subject, an fMRI experiment consists of the acquisition of a large number (100 to 1500) of 3D volumes (for instance 64x64x32 voxels *i.e.*, volume element) measuring in each voxel the BOLD contrast [2], which is related to the total amount of deoxygenated hemoglobin present in *the* voxel. The subject is submitted to an experimental paradigm consisting of different conditions designed to study a particular brain system (*e.g.* memory, language, vision), while a continuous acquisition of brain volumes is performed. This is called a *run* or *session* and lasts for approximately 5 to 10 minutes. A session is generally repeated several times — typically between three and seven — for a given subject.

This technique has allowed to detect and localize dynamic brain processes for various stimulations or tasks [3] with a high spatial resolution (of the order of a mm), but a poor time resolution and a low signal or contrast-to-noise ratio (CNR) so far. This makes the use of well-designed data acquisition protocols necessary. Two classes of protocols can be distinguished : block-designed and event-related experiments. The latter has emerged as a means of observing the fMRI time course in response to a single, very short stimulus (a trial) [1], [4], while the former has better CNR and may require averaging over many trials presented in close succession. Despite their lower CNR, event-related paradigms are often inevitable, for instance to avoid habituation effects. In such paradigms, random intermixing of trial types is used to eliminate habituation, anticipation or other strategy effects [1] that might occur in *deterministic* paradigms. Actual experiments consist of either *synchronous* or *asynchronous* paradigms, depending whether the onsets of the conditions are synchronized with the data acquisition rate or not. Paradigms are often asynchronous because the onset of the response can be given by the subject himself (response after a variable reaction time).

The end goal of activation detection in brain functional imaging experiments is to retrieve as much as possible

of the neuronal activity in response to cognitive or behavioral tasks [3]. However, the relation between this activity and the BOLD response [2] is not completely understood and still under study [5]–[8]. The partially known mechanisms of coupling neuronal (synaptic) activity to the vascular system produces significant blurring and delay to the original neuronal response over time, indicating that the BOLD sequence is heavily low-pass filtered [5], [8]. As a consequence, hemodynamic events have time scales of a few seconds, whereas neuronal events have time constants of milliseconds. An accurate and robust estimate of the brain hemodynamic response to a stimulus may be a first step towards a better quantification of the brain neuronal activity. In each region, the brain hemodynamic response can be characterized in first instance by the response function to a very short stimulus (the transfer function if the system is linear), denoted as the *Hemodynamic Response Function* (HRF).

Modeling the HRF has become an intensive topic of research for many reasons. First, a precise modeling should lead to a better understanding of cerebral activations. Second, within a region, the signal variations between conditions or stimuli (such as the magnitude of the response, but also its delay or width) can only be studied with an accurate estimate of the response and of its variability. Third, estimation of the HRF can be done at every position in the brain in order to investigate its spatial variability. Lastly, some recent technological progress gives access to simultaneous recordings of electrical (Electroencephalography) and metabolic (fMRI) activities. For these reasons, the HRF has been the subject of many studies that usually assume that the brain system is linear and time invariant (LTI) [9]–[16]. Although the question whether the brain response can be considered linear is not yet fully answered, it has been shown that this assumption is a tenable and useful approximation [4], [10], [14], and thus holds in the present work.

Parametric methods for estimating the HRF as a transfer function of a LTI system appeared first in the literature [9], [11]–[13]. These approaches impose the shape of the HRF by choosing a particular function (*e.g.*, Gamma or Gaussian density). The *nonlinear* parameters of this function are fitted to the data to take variations of the delay and blurring effects of the HRF into account. Parametric models may introduce some bias on the HRF, since it is unlikely that they capture the shape variations of the HRF within the brain.

By contrast, recent works have introduced temporal prior information on the underlying physiological process of the brain hemodynamic response to accurately estimate the HRF in a Bayesian framework. Such priors compensate for the lack of information provided by the data [15], [17], [18]. These techniques only apply to periodic or *synchronous* event-related paradigms and are devoted to the estimation of *one* HRF in response to one condition or stimulus. They also deal with each session separately and average the HRF estimates *a posteriori* without taking fluctuations of physiological factors across sessions into account. Hence, their use has remained limited since they are not well-adapted to actual fMRI data.

The aim of this paper is to propose a threefold efficient generalization of [15], [18]. First, we derive a temporally regularized estimator of the HRF when shorter and jittered Inter-Stimulus Intervals (ISIs) are used, such as in asynchronous paradigms. Second, we propose a simple extension that is able to cope with *mixed task* paradigms, in which mixed trial stimuli are presented in a random order and in a rapid succession to one another. This extension consists in estimating a HRF per trial type. Third, we develop an estimation procedure that is able to simultaneoulsy process all fMRI time series (all sessions or runs) recorded for a subject in a given region. The specific treatment of each session is important because noise characteristics (low frequencies) may be different between sessions. We will show that this leads to more accurate estimates and relevant error bars provided that the drift terms are modeled with few parameters per session (typically three or four). Because these extensions require a significant computational effort, we develop a powerful optimization scheme that makes the computation fast enough (typically one second for a single time series of two hundreds samples) for the analysis of real data in an imaging center environment.

The rest of the paper is organized as follows. Section II starts with the introduction of the LTI model of the HRF for *single trial* asynchronous paradigms. Successive generalizations for *asynchronous multitrial multisession* paradigms are then taken into account in a more complex badly-conditioned observation model. In Section III, we motivate our modeling choices and derive the selected estimator for the HRF within the Bayesian formalism. Since such an estimate depends on a few hyperparameters, Section IV adresses the problem of their tuning according to the maximum-likelihood estimator. Section V illustrates the performances of our *unsupervised* approach with synthetic data. The method is applied to a language comprehension fMRI study in Section VI. In Section VII, we discuss the limits of applicability and possible extensions of the proposed method.

II. Modeling the cerebral hemodynamic response

A. Notations

Throughout the paper, random variables and realizations of thereof are respectively denoted by uppercase (*e.g.*, $\boldsymbol{H}$) and corresponding lowercase (*e.g.*, $\boldsymbol{h}$) symbols; in addition, notations such as $p(\boldsymbol{y}\,|\,\boldsymbol{h})$ are employed as shorthands for $p_{\boldsymbol{Y}\,|\,\boldsymbol{H}}(\boldsymbol{y}\,|\,\boldsymbol{h})$, whenever unambiguous. Furthermore, $p(\boldsymbol{y}\,|\,\boldsymbol{h}\;;\,\boldsymbol{\theta})$ means that the parameter vector $\boldsymbol{\theta}$ is assumed unknown but deterministic. The main notations used in the following are summarized in Table I.

[Table 1 about here.]

B. LTI system for asynchronous paradigms

In event-related protocols with synchronous ISI, the BOLD fMRI time course $(y_n)_{1\leq n\leq N}$ is measured in any voxel of the brain at times $(t_n = n\,\mathrm{TR})_{1\leq n\leq N}$, TR being the Time of Repetition, while stimuli occur with a fixed-delayed impulse signal $(x_n)_{1\leq n\leq N}$. This means that the sampling period is equal to TR when the stimuli occur only at times of acquisition. The HRF is then modeled as the convolution kernel of a LTI system [9], [15], [18].

In *asynchronous* experiments, the presented stimuli occur at any time during scanning. In such cases, we propose to put the data and the trials on a finer temporal grid, which has to be defined such that *(i)* the time occurrences of the stimuli $(x_t)_{t\geqslant t_1}$ are defined on this grid, and that *(ii)* two stimuli do not occur at the same time. Let Δt be the sampling period of this grid. Our strategy consists in approximating the true onsets by their closer neighbor on this grid. This very simple procedure can be seen as a zero-order interpolation and generates what is called in the following an *instant-matching* error. Let $(x_t)_{t\geqslant t_1}$ the time occurrences of the stimuli on this finer grid. Accordingly, the HRF has to be estimated with the same temporal resolution:

$$\begin{aligned} y_{t_n} &= \sum_{k=0}^{K} h_{k\Delta t}\, x_{t_n - k\Delta t} + b_{t_n} \\ &= \boldsymbol{x}_{t_n}^{\mathrm{t}}\,\boldsymbol{h} + b_{t_n}, \text{ for } t_n = t_1, \ldots, t_N, \end{aligned} \tag{1}$$

with $\boldsymbol{x}_{t_n} = [x_{t_n}, x_{t_n-\Delta t}, \ldots, x_{t_n - K\Delta t}]^{\mathrm{t}} \in \mathbb{R}^{K+1}$ and $\boldsymbol{h} = [h_0, h_{\Delta t}, \ldots, h_{K\Delta t}]^{\mathrm{t}}$. Note first that the number of unknowns, *i.e.* $K+1$, may be dramatically larger than its counterpart in the synchronous case and second, that oversampling of the data is not required. b_{t_n} is the nth sample of a zero-mean Gaussian white noise process $\boldsymbol{b}$ of unknown variance $r_b > 0$, independent of $\boldsymbol{h}$. Such a hypothesis may seem restrictive since it is well-known that fMRI time series are correlated in time [19]. Nonetheless, as shown in [18] various noise correlation structures have little influence on the performances of the estimation. The same result has also been emphasized in [20], where the authors analyze the influence of the colored nature of fMRI noise on the *average bias* of the HRF estimate. Alternatively, one could estimate the temporal covariance structure of the noise with an autoregressive model, as done in [21].

In real neuroimaging experiments, the fMRI raw data are contaminated by a low-frequency drift mainly due to physiological artifacts [22]: breathing and cardiac pulses are aliased since the sampling frequency of the data is below Nyquist's bound. Thus, these physiological factors introduce some low frequency fluctuations. A high-pass filter is generally used to remove those trends before estimating the HRF. In this study, we simultaneously estimate the HRF and the trend with the following model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{h} + \boldsymbol{P}\boldsymbol{l} + \boldsymbol{b}, \tag{2}$$

where $\boldsymbol{X} = [\boldsymbol{x}_{t_1}, \ldots, \boldsymbol{x}_{t_N}]^{\mathrm{t}}$ defines the binary onsets matrix. Matrix $\boldsymbol{P} = [P_1, \ldots, P_Q]$ consists of an orthonormal basis of functions $P_q = [P_q(t_1), \ldots, P_q(t_N)]^{\mathrm{t}}$ modeling the low frequencies (*e.g.* a one dimensional discrete cosine transform). The number Q of basis functions depends on the lowest frequency $f_{\min}$ attributable to the drift term and can be defined as $Q = [2N\,\mathrm{TR}\,f_{\min}] + 1$, where "+1" stands for the mean (constant term) and $[\cdot]$ is the integer part operator. Matrix $\boldsymbol{P}$ can also take any covariate of no interest into account, supposed to influence the signal intensity in a linear way. Vector $\boldsymbol{l} \in \mathbb{R}^Q$ defines the unknown weighting coefficients of the basis functions, called *nuisance variables* in the following.

C. Asynchronous multitask paradigms

We further extend (2) to allow for a different HRF estimate for different trial types (*e.g.*, different stimuli or conditions). Let $\left(\boldsymbol{X}^{(m)}\right)_{1\leqslant m\leqslant M}$ be the different trial-dependent matrices, each of them being defined as the previous $\boldsymbol{X}$ matrix, and then suppose that the HRFs $\boldsymbol{h}_m$ add in a *linear* way. Such an extension requires to correctly define the oversampling period Δt as the smallest sampling interval that allows to separate the two closest events, whatever their type. For the sake of simplicity, let us define

$$\begin{aligned} \boldsymbol{X} &= [\boldsymbol{X}^{(1)}|\ldots|\boldsymbol{X}^{(M)}] \in \mathbb{R}^N \times \mathbb{R}^{M(K+1)} \\ \boldsymbol{h} &= [\boldsymbol{h}_1^{\mathrm{t}}, \ldots, \boldsymbol{h}_M^{\mathrm{t}}]^{\mathrm{t}} \in \mathbb{R}^{M(K+1)} \end{aligned}$$

from which model (2) is able to cope with *asynchronous multitask* paradigms.

D. Multisession likelihood

As previously mentioned, the experimental paradigm is repeated several times for a given subject, leading to a few sessions of about 200 to 1000 data each.

It is generally assumed that the HRF remains approximately constant provided that all the exogenous parameters (voxel, task, subject) are fixed. Accordingly, the same vector $\boldsymbol{h}$ is sought from the available fMRI times series, say $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_I$, of respective length N_i, so that a specific matrix $\boldsymbol{X}_i$ is involved for each session. In addition, it seems relevant to select a session-dependent value of $\boldsymbol{l}_i$ and possibly Q_i for the definition of $\boldsymbol{P}_i \in \mathbb{R}^{N_i} \times \mathbb{R}^{Q_i}$ since the physiological factors (breathing and cardiac rates) fluctuate throughout the sessions. Hence, the multisession extension of (2) is given by

$$(\mathcal{M}): \qquad \boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{h} + \boldsymbol{P}_i \boldsymbol{l}_i + \boldsymbol{b}_i, \qquad \text{for } i \in \mathbb{N}_I^*.$$

Model $(\mathcal{M})$ calls for two comments.

First, it relies on the following assumptions about noises $\boldsymbol{b}_i$:

- the mean of $\boldsymbol{b}_i$ (*i.e.*, the baseline of $\boldsymbol{h}$) may vary across sessions. This variation is captured by the constant column of matrix $\boldsymbol{P}_i$.
- the variance of $\boldsymbol{b}_i$ is supposed constant across the sessions for the sake of simplicity. In Section V, it will be shown that the proposed HRF estimate remains robust with regard to departures from this hypothesis. There is no theoretical limitation preventing us from introducing a specific variance for each session. Nonetheless, we advocate the use of the same unknown variance for all sessions on the same subject since we observed on our data that the major fluctations occuring in real data are rather due to physiological variations (modeled by session-dependent trends) than to some modification of the noise scaling.

Second, model $(\mathcal{M})$ allows to introduce more information than model (2), through the introduction of new independent data.

To estimate and make inferences about the hemodynamic response, we first need to take model $(\mathcal{M})$ into account through the definition of the likelihood function. For the sake of conciseness, let us introduce the following notations:

$$\mathbb{y} = \left[\boldsymbol{y}_1^{\mathrm{t}}, \ldots, \boldsymbol{y}_I^{\mathrm{t}}\right]^{\mathrm{t}}, \qquad \mathbb{X} = \left[\boldsymbol{X}_1^{\mathrm{t}} \mid \ldots \mid \boldsymbol{X}_I^{\mathrm{t}}\right]^{\mathrm{t}},$$
$$\mathbb{P} = \operatorname{diag}\left[\boldsymbol{P}_1, \ldots, \boldsymbol{P}_I\right], \qquad \mathbb{l} = \left[\boldsymbol{l}_1^{\mathrm{t}} \mid \ldots \mid \boldsymbol{l}_I^{\mathrm{t}}\right]^{\mathrm{t}},$$

where the diag operator is used to define $\mathbb{P}$ as a block-diagonal matrix. From these assumptions, the likelihood of the asynchronous multitrials multisession model $(\mathcal{M})$ is given by

$$\begin{aligned} p(\mathbb{y} \mid \boldsymbol{h}, \mathcal{M}\ ; r_b, \mathbb{l}) &= \prod_{i=1}^{I} p(\boldsymbol{y}_i \mid \boldsymbol{h}, \mathcal{M}\ ; r_b, \boldsymbol{l}_i) \\ &= (2\pi r_b)^{-N/2} \exp\left(-\sum_{i=1}^{I} \left\|\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{h} - \boldsymbol{P}_i \boldsymbol{l}_i\right\|^2 / 2r_b\right) \\ &= (2\pi r_b)^{-N/2} \exp\left(-\left\|\mathbb{y} - \mathbb{X}\boldsymbol{h} - \mathbb{P}\mathbb{l}\right\|^2 / 2r_b\right) \end{aligned} \tag{3}$$

where $N = \sum_{i=1}^{I} N_i$.

The number of parameters still remains large so that least squares estimation is unreliable when $[\mathbb{X} \mid \mathbb{P}]^{\mathrm{t}}[\mathbb{X} \mid \mathbb{P}]$ is ill-conditioned (the variance of the fitted parameters is too large). A straight application of the inversion lemma for block matrices provides a necessary and sufficient condition: this block-matrix is invertible and well-conditioned if and only if the inverse of $\mathbb{X}^{\mathrm{t}}(\boldsymbol{I}_M - \mathbb{P}(\mathbb{P}^{\mathrm{t}}\mathbb{P})^{-1}\mathbb{P}^{\mathrm{t}})\mathbb{X} = \mathbb{X}^{\mathrm{t}}(\boldsymbol{I}_M - \mathbb{P}\mathbb{P}^{\mathrm{t}})\mathbb{X}$ exists and is well-conditionned[1]. Note that such a matrix can be ill-conditionned even if $\mathbb{X}^{\mathrm{t}}\mathbb{X}$ is well-conditionned. The limiting case, corresponding to the underdetermination of model $(\mathcal{M})$, can be reached if Δt is too low or M too large.

III. Identification of the HRFs

Since the information provided by the data may not be sufficient to derive a robust HRF estimate, we introduce some constraints on the temporal structure of the HRFs that correspond to some available physiological prior knowledge. The proposed HRF estimate will result from an appropriate tradeoff of both types of information (data-driven and prior).

A. Prior information

Following [15], [18], we introduce temporal prior information within the Bayesian framework. As physiologically advocated in [5], each HRF is characterized by the following features:

(a) Its amplitude is close to zero at the first and end points. As a matter of fact, the HRF is causal, hence $h_{m,k\Delta t}$ should be zero for $k \leq 0$ and $m \in \mathbb{N}_M^*$. This means that the stimulus at time k should only have influence for $k > 0$. Similarly, the influence of an activation should vanish in the past, implying that the filter parameters should tend to zero for large delays ($h_{m,K\Delta t} = 0$).

(b) Its variations are smooth.

[1]Let $\boldsymbol{A} \in \mathbb{R}^K \times \mathbb{R}^K$ be an invertible matrix, then $\operatorname{Cond}(\boldsymbol{A}) = \operatorname{Cond}(\boldsymbol{A}^{-1})$ if Cond stands for the condition number [23].

(c) Prior statistical independence is supposed to hold between stimuli and thus between HRFs. In addition, each HRF may be related to an underlying physiological process having its own dynamics (specific time-to-peak and dispersion of response).

Condition *(a)* is easily introduced by redefining vectors $\boldsymbol{h}_m$ and $\boldsymbol{x}_{i,t_n}^{(m)}$ for the ith session and all $n \in \mathbb{N}_N^*$, setting the first and last parameters of $\boldsymbol{h}_m$ to zero:

$$\begin{aligned}
\boldsymbol{h}_m &= [h_{m,\Delta t}, \ldots, h_{m,(K-1)\Delta t}]^{\mathrm{t}} \in \mathbb{R}^{K-1}, \\
\boldsymbol{x}_{i,t_n}^{(m)} &= [x_{i,t_n-\Delta t}^{(m)}, \ldots, x_{i,t_n-(K-1)\Delta t}^{(m)}]^{\mathrm{t}} \in \mathbb{R}^{K-1}, \\
\boldsymbol{X}_i^{(m)} &= [\boldsymbol{x}_{i,t_1}^{(m)}, \ldots, \boldsymbol{x}_{i,t_N}^{(m)}]^{\mathrm{t}} \in \mathbb{R}^{N_i} \times \mathbb{R}^{K-1}.
\end{aligned}$$

The likelihood function (3) remains unchanged with $\boldsymbol{h} = \left[\boldsymbol{h}_1^{\mathrm{t}}, \ldots, \boldsymbol{h}_M^{\mathrm{t}}\right]^{\mathrm{t}} \in \mathbb{R}^{M(K-1)}$ and $\boldsymbol{X}_i = \left[\boldsymbol{X}_i^{(1)} | \ldots | \boldsymbol{X}_i^{(M)}\right] \in \mathbb{R}^{N_i \times M(K-1)}$ for the ith session.

Quantification of condition *(b)* is achieved by setting a Gaussian probability density function (pdf) $\mathcal{N}(0, \tau_m \boldsymbol{R})$ for $p(\boldsymbol{h}_m; \boldsymbol{R}, \tau_m)$. We have chosen $\boldsymbol{R} = (\boldsymbol{D}_2^{\mathrm{t}} \boldsymbol{D}_2)^{-1}$ for the prior covariance where $\boldsymbol{D}_2$ is the truncated second-order finite difference matrix in order to fullfil constraint*(a)*:

$$\boldsymbol{D}_2 = \begin{bmatrix} -2 & 1 & 0 & \cdots & \cdots & 0 \\ 1 & -2 & 1 & 0 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 0 & 1 & -2 & 1 \\ 0 & \cdots & \cdots & 0 & 1 & -2 \end{bmatrix}.$$

Since $\boldsymbol{R}^{-1}$ is of full rank, $p(\boldsymbol{h}_m; \boldsymbol{R}, \tau_m)$ defines a proper prior[2]. For model ($\mathcal{M}$), condition *(c)* may be taken into account by the following pdf:

$$\begin{aligned}
p(\boldsymbol{h}; \boldsymbol{R}, \boldsymbol{\theta}_{\boldsymbol{H}}) &= \prod_{m=1}^{M} p(\boldsymbol{h}_m \; ; \boldsymbol{R}, \tau_m) \\
&\propto \frac{\prod_{m=1}^{M} \tau_m^{-(K-1)/2}}{\det(\boldsymbol{R})^{-M/2}} \exp\left(-\frac{\boldsymbol{h}^{\mathrm{t}} \boldsymbol{R}_{\boldsymbol{H}}^{-1} \boldsymbol{h}}{2}\right), \qquad (4)
\end{aligned}$$

where $\boldsymbol{\theta}_{\boldsymbol{H}} = [\tau_1, \ldots, \tau_M] \in \mathbb{R}^M$ stands for the hyperparameters of the prior model and

$$\boldsymbol{R}_{\boldsymbol{H}} = \mathrm{diag}\left[\tau_1 \boldsymbol{R}, \tau_2 \boldsymbol{R}, \ldots, \tau_M \boldsymbol{R}\right] \in \mathbb{R}^{M(K-1)} \times \mathbb{R}^{M(K-1)}.$$

This prior model clearly favors *smooth* responses, since for each condition it amounts to minimizing $||\boldsymbol{D}_2 \boldsymbol{h}_m||^2$ *i.e.*, the discrete approximation of the second-order derivative of $\boldsymbol{h}_m$. We note also that the sampling period Δt is integrated in the definition of the prior variances τ_m and that introduction of different parameters τ_m for different HRFs allows to model specific dynamics for each condition.

[2]The pdf $p(\boldsymbol{h} \; ; \boldsymbol{R}, \tau_m)$ is said to be proper if $\int_{\boldsymbol{h}} p(\boldsymbol{h} \; ; \boldsymbol{R}, \tau_m) \, d\boldsymbol{h} < \infty$.

B. Modeling choices and error bar computation

Our HRF estimation technique is a two-steps operation, which first consists in jointly estimating $\mathbb{l}$ and $\boldsymbol{\theta}$ in the maximum-likelihood sense (see details in Section IV) before substituting the estimates for the true values in $p(\boldsymbol{h} \,|\, \mathbb{y}, \mathcal{M} \; ; \boldsymbol{\theta}, \mathbb{l})$ in order to compute its maximizer. In the first step, we assume that the drift parameters $\mathbb{l}$ as well as hyperparameters $\boldsymbol{\theta} = \left[r_b, \boldsymbol{\theta}_{\boldsymbol{H}}^{\mathrm{t}}\right]^{\mathrm{t}}$ are *deterministic*. We motivate this choice below.

It does not seem to us that considering the drift terms as random variables would allow the addition of significant information because we have rather poor prior knowledge at hand. Therefore, a reasonable prior on those would have been uninformative or even improper. Moreover, in Subsection V-E we argue that with a small number of nuisance variables (*i.e.*, when Q/N is small, typically $Q/N \leqslant 0.02$) and a standard noise level, we are able to accurately estimate these parameters and be reasonably confident in the variability of our HRF estimate. In other words, when we substitute $\widehat{\mathbb{l}}^{\mathrm{ML}}$ for $\mathbb{l}$, the extra source of error on $\widehat{\boldsymbol{h}}^{\mathrm{MAP}}$ still is small. The same holds for hyperparameters $\boldsymbol{\theta}$. It follows that the most important quantity $\boldsymbol{h}$ is estimated from the maximum a posteriori (MAP) *i.e.*, the maximum of the posterior distribution $p(\boldsymbol{h} \,|\, \mathbb{y}, \mathcal{M} \; ; \boldsymbol{\theta}, \mathbb{l})$. The Bayes rule ensures the fusion of the likelihood (3) and the prior (4) into the Gaussian posterior distribution of $\boldsymbol{h}$ given $(\mathbb{y}, \boldsymbol{\theta}, \mathbb{l})$, from which we can derive the MAP estimate:

$$\begin{aligned}
p(\boldsymbol{h} \,|\, \mathbb{y}, \mathcal{M} \; ; \boldsymbol{\theta}, \mathbb{l}) &\sim \mathcal{N}(\widehat{\boldsymbol{h}}^{\mathrm{MAP}}, \boldsymbol{\Sigma}) \qquad (5) \\
\boldsymbol{\Sigma}^{-1} &= \frac{1}{r_b} \sum_{i=1}^{I} \boldsymbol{X}_i^{\mathrm{t}} \boldsymbol{X}_i + \boldsymbol{R}_{\boldsymbol{H}}^{-1}, \\
\widehat{\boldsymbol{h}}^{\mathrm{MAP}} &= \frac{1}{r_b} \boldsymbol{\Sigma} \sum_{i=1}^{I} \boldsymbol{X}_i^{\mathrm{t}} \left(\boldsymbol{y}_i - \boldsymbol{P}_i \boldsymbol{l}_i\right).
\end{aligned}$$

The influence of this choice of model should have limited impact on the estimation of $\boldsymbol{h}$. However, the influence on the error bars of $\widehat{\boldsymbol{h}}^{\mathrm{MAP}}$ might not be negligeable. As we would like to assess the error made on $\widehat{\boldsymbol{h}}_m^{\mathrm{MAP}}$, we focus on the marginal posterior pdf $p(\boldsymbol{h}_m \,|\, \mathbb{y}, \mathcal{M}; \boldsymbol{\theta}, \mathbb{l})$. As expected, this pdf is $\mathcal{N}(\widehat{\boldsymbol{h}}_m^{\mathrm{MAP}}, \boldsymbol{\Sigma}_m)$-distributed with $\boldsymbol{\Sigma}_m$, the mth diagonal block of $\boldsymbol{\Sigma}$. The marginal error bars can then be derived from the standard deviations $\boldsymbol{\sigma}_m = \left[\sigma_{m,1}, \ldots, \sigma_{m,K-1}\right]$, that are given by the square roots of the main diagonal of $\boldsymbol{\Sigma}_m$. Although slightly

underestimated, such error bars provide a good approximation of the range of variation of $\boldsymbol{h}_m$. Note that these error bars are created for pointwise inference and appropriate only for a single preselected time sample since the search over all time samples requires to analyse the posterior covariance matrix $\boldsymbol{\Sigma}_m$, which cannot be easily represented on the same graph as $\widehat{\boldsymbol{h}}_m^{\text{MAP}}$.

C. Equivalence with the marginal MAP estimate

Under technical conditions, we show that our HRF estimate $\widehat{\boldsymbol{h}}^{\text{MAP}}$ is a valid approximation of another Bayesian estimate that takes all fluctuations of $\mathbb{l}$ into account. This solution corresponds to the maximizer $\widehat{\boldsymbol{h}}^{\text{MMAP}}$ of the *marginal* pdf $p(\boldsymbol{h}\,|\,\mathbb{y},\mathcal{M};\boldsymbol{\theta})$:

$$\begin{aligned} p(\boldsymbol{h}\,|\,\mathbb{y},\mathcal{M};\boldsymbol{\theta}) &= \int p(\boldsymbol{h},\mathbb{l}\,|\,\mathbb{y},\mathcal{M};\boldsymbol{\theta})\,d\mathbb{l} \qquad (6)\\ &= \int p(\boldsymbol{h}\,|\,\mathbb{y},\mathbb{l},\mathcal{M};\boldsymbol{\theta})\,p(\mathbb{l}\,|\,\mathbb{y},\mathcal{M};\boldsymbol{\theta})\;d\mathbb{l}, \end{aligned}$$

where $\mathbb{l}$ is assumed to be random. Using Bayes rule, the marginal posterior pdf $p(\mathbb{l}\,|\,\mathbb{y},\mathcal{M}\;;\boldsymbol{\theta})$ rereads

$$p(\mathbb{l}\,|\,\mathbb{y},\mathcal{M}\;;\boldsymbol{\theta}) \propto p(\mathbb{y}\,|\,\mathbb{l},\mathcal{M}\;;r_b)\,p(\mathbb{l}\;;\boldsymbol{\theta}_{\mathbb{L}}). \qquad (7)$$

In the absence of relevant prior information about $\mathbb{l}$, a flat distribution is considered for $p(\mathbb{l}\;;\boldsymbol{\theta}_{\mathbb{L}})$ leading in (7) to $p(\mathbb{l}\,|\,\mathbb{y},\mathcal{M}\;;\boldsymbol{\theta}) \propto p(\mathbb{y}\,|\,\mathbb{l},\mathcal{M}\;;\boldsymbol{\theta})$. Assuming that $p(\mathbb{l}\,|\,\mathbb{y},\mathcal{M}\;;\boldsymbol{\theta})$ is peaked enough, we may write $p(\mathbb{l}\,|\,\mathbb{y},\mathcal{M}\;;\boldsymbol{\theta}) \approx \delta(\mathbb{l}-\widehat{\mathbb{l}}^{\text{MAP}})$, where $\widehat{\mathbb{l}}^{\text{MAP}} = \widehat{\mathbb{l}}^{\text{ML}}$ is the maximum-likelihood solution computed by our approach (see details in Section IV). If this hypothesis is fullfiled, the marginal distribution (6) is thus approximated by

$$\begin{aligned} p(\boldsymbol{h}\,|\,\mathbb{y},\mathcal{M}\;;\boldsymbol{\theta}) &\approx \int p(\boldsymbol{h}\,|\,\mathbb{y},\mathbb{l},\mathcal{M}\;;\boldsymbol{\theta})\;\delta(\mathbb{l}-\widehat{\mathbb{l}}^{\text{ML}})\;d\mathbb{l}\\ &\approx p(\boldsymbol{h}\,|\,\mathbb{y},\mathbb{l}=\widehat{\mathbb{l}}^{\text{ML}},\mathcal{M}\;;\boldsymbol{\theta}). \qquad (8) \end{aligned}$$

Therefore, our approach amounts to computing $\widehat{\boldsymbol{h}}^{\text{MMAP}}$ provided that $p(\mathbb{l}\,|\,\mathbb{y},\mathcal{M}\;;\boldsymbol{\theta}) \approx \delta(\mathbb{l}-\widehat{\mathbb{l}}^{\text{ML}})$. In Subsection V-E, we discuss the validity of this assumption.

IV. Hyperparameters and nuisance variable estimation

This section focuses on the automatic tuning of hyperparameters $\boldsymbol{\theta}$ and nuisance variables $\mathbb{l}$. This part starts with an introduction to the Maximum-Likelihood principle and then examines a well-adapted way to tackle the underlying optimization problem.

A. Maximum-Likelihood principle

Maximum-likelihood estimation for hyperparameters is a very common procedure, which is currently used in various fields of signal and image processing when dealing with a small number of unknown but deterministic hyperparameters $\boldsymbol{\theta}$ (see for instance [24], [25]). The underlying reason that makes this approach feasible and attractive is that there is a large number of data (N) to estimate $(\boldsymbol{\theta},\mathbb{l})$ accurately. Similarly to [15], we select the hyperparameters and nuisance variables by maximization of their likelihood $p(\mathbb{y}|\mathcal{M};\boldsymbol{\theta},\mathbb{l})$, obtained from the joint pdf of $(\boldsymbol{H},\mathbb{Y})$ after integration over the parameters $\boldsymbol{H}$:

$$\begin{aligned} p(\mathbb{y}\,|\,\mathcal{M};\boldsymbol{\theta},\mathbb{l}) &= \int_{\boldsymbol{h}} p(\mathbb{y}\,|\,\boldsymbol{h},\mathcal{M};r_b,\mathbb{l})\;p(\boldsymbol{h};\boldsymbol{R},\boldsymbol{\theta}_{\boldsymbol{H}})\;d\boldsymbol{h}\\ &= \frac{(2\pi)^{-N/2}}{\det\left(\boldsymbol{\Upsilon}\right)^{1/2}}\exp\left(-\frac{(\mathbb{y}-\mathbb{P}\mathbb{l})^{\mathrm{t}}\boldsymbol{\Upsilon}^{-1}(\mathbb{y}-\mathbb{P}\mathbb{l})}{2}\right) \qquad (9) \end{aligned}$$

with[3]

$$\boldsymbol{\Upsilon} = r_b\boldsymbol{I}_N + \mathbb{X}\boldsymbol{R}_{\boldsymbol{H}}^{-1}\mathbb{X}^{\mathrm{t}}.$$

The Maximum-Likelihood (ML) estimator $(\widehat{\boldsymbol{\theta}}^{\text{ML}},\widehat{\mathbb{l}}^{\text{ML}})$ of $(\boldsymbol{\theta},\mathbb{l})$ minimizes

$$\log\det(\boldsymbol{\Upsilon}) + (\mathbb{y}-\mathbb{P}\mathbb{l})^{\mathrm{t}}\boldsymbol{\Upsilon}^{-1}(\mathbb{y}-\mathbb{P}\mathbb{l}). \qquad (10)$$

Computation of $(\widehat{\boldsymbol{\theta}}^{\text{ML}},\widehat{\mathbb{l}}^{\text{ML}})$ is a complicated nonlinear optimization problem of several variables. First, we must resort to an iterative optimization scheme because of the nonquadratic behavior of $\log\det(\boldsymbol{\Upsilon})$. Second, since $\log p(\mathbb{y}\,|\,\mathcal{M}\;;\boldsymbol{\theta},\mathbb{l})$ is not concave with respect to (w.r.t.) $(\boldsymbol{\theta},\mathbb{l})$, there may exist several local minima. Therefore, the hyperparameters and nuisance variables given by any deterministic algorithm depend on the initial values. In the next subsection, we propose to resort to a variant of the Expectation Maximization (EM) algorithm to avoid direct optimization of (10).

B. EM-based strategy

For notation compactness, let us denote $\widetilde{\boldsymbol{\theta}} = \left[\boldsymbol{\theta}^{\mathrm{t}},\mathbb{l}^{\mathrm{t}}\right]^{\mathrm{t}}$. The EM algorithm, introduced by [26], is a general iterative method which ensures the increasing of the likelihood function $p(\mathbb{y}\,|\,\mathcal{M}\;;\widetilde{\boldsymbol{\theta}})$ of a parameter vector $\widetilde{\boldsymbol{\theta}}$ given observations $\mathbb{y}$ at each iteration. Starting from an initial value $\widetilde{\boldsymbol{\theta}}^0$, a series of succesive estimates $\widetilde{\boldsymbol{\theta}}^k$ is

[3]In these definitions, every matrix $\boldsymbol{P}_i$ is supposed to have $Q_i = Q$ columns $\forall\, i \in \mathbb{N}_I^*$. If such a choice turns out to be inappropriate, the definitions of matrix $\mathbb{P}$ and vector $\mathbb{l}$ should be revisited, by zero-padding matrices $\boldsymbol{P}_i$ and vectors $\boldsymbol{l}_i$ that do not have the maximal number of components *i.e.*, $\max_i Q_i$.

generated by alternating the following two steps:

$$\text{Expectation (E):} \quad \text{Evaluate } \mathcal{Q}(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}^k; \mathbb{y}) \tag{11a}$$

$$\text{Maximization } (\mathcal{M}): \quad \widetilde{\boldsymbol{\theta}}^{k+1} = \arg\max_{\widetilde{\boldsymbol{\theta}}} \mathcal{Q}(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}^k; \mathbb{y}), \tag{11b}$$

where function $\mathcal{Q}$ is defined as

$$\begin{aligned}\mathcal{Q}(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}^k; \mathbb{y}) &= \int_{\boldsymbol{h}} \log p(\mathbb{y}, \boldsymbol{h} | \mathcal{M}; \widetilde{\boldsymbol{\theta}}) p(\boldsymbol{h} | \mathbb{y}, \mathcal{M}; \widetilde{\boldsymbol{\theta}}^k) d\boldsymbol{h} \\ &= \mathrm{E}_{\boldsymbol{H}} \left[\log p(\mathbb{y}, \boldsymbol{h} | \mathcal{M} \; ; \widetilde{\boldsymbol{\theta}}) | \mathbb{y}, \mathcal{M}; \widetilde{\boldsymbol{\theta}}^k \right],\end{aligned} \tag{12}$$

$\boldsymbol{H}$ being an auxiliary random variable whose practical role is to make the *complete* likelihood $p(\mathbb{y}, \boldsymbol{h} \,|\, \mathcal{M} \; ; \widetilde{\boldsymbol{\theta}})$ easier to compute than the original one $p(\mathbb{y} \,|\, \mathcal{M} \; ; \widetilde{\boldsymbol{\theta}})$. Following [27], parameter vector $\widetilde{\boldsymbol{\theta}}$ can be partitioned into two subvectors $\boldsymbol{\theta}_{\mathbb{Y} \,|\, \boldsymbol{H}} = [r_b, \mathbb{l}^{\mathrm{t}}]^{\mathrm{t}}$ and $\boldsymbol{\theta}_{\boldsymbol{H}} = [\tau_1, \ldots, \tau_M]^{\mathrm{t}}$ which respectively control the conditional pdfs $p(\mathbb{y} \,|\, \boldsymbol{h}, \mathcal{M} \; ; \boldsymbol{\theta}_{\mathbb{Y} \,|\, \boldsymbol{H}})$ and $p(\boldsymbol{h} \; ; \boldsymbol{R}, \boldsymbol{\theta}_{\boldsymbol{H}})$. Then, as shown in Appendix I, the M-step (11b) can be divided into two operations: maximization of $\mathcal{Q}_{\mathbb{Y} \,|\, \boldsymbol{H}}$ w.r.t. $\boldsymbol{\theta}_{\mathbb{Y} \,|\, \boldsymbol{H}}$ and maximization of $\mathcal{Q}_{\boldsymbol{H}}$ w.r.t. $\boldsymbol{\theta}_{\boldsymbol{H}}$ (see (19) and (20) for the definitions of $\mathcal{Q}_{\mathbb{Y} \,|\, \boldsymbol{H}}$ and $\mathcal{Q}_{\boldsymbol{H}}$, respectively).

In the present case, detailed examination of the maximization of $\mathcal{Q}_{\mathbb{Y} \,|\, \boldsymbol{H}}$ w.r.t. $\boldsymbol{\theta}_{\mathbb{Y} \,|\, \boldsymbol{H}}$ shows a tricky problem: the *joint* maximization w.r.t. r_b and $\mathbb{l}$, required by the definition (11b), cannot be performed since the updating step of r_b strongly depends on $\mathbb{l}$. Nonetheless, the M-step (11b) can be replaced by a *conditional* optimization strategy that reads:

$$\forall \, \beta_l \in \boldsymbol{\beta} = \boldsymbol{\theta}_{\mathbb{Y}|\boldsymbol{H}}, \;\; \beta_l^{k+1} = \arg\max_{\beta_l} \mathcal{Q}_{\mathbb{Y}|\boldsymbol{H}}(\boldsymbol{\beta}_{l-1}^{k+1}, \widetilde{\boldsymbol{\theta}}^k; \mathbb{y})$$

$$\text{with } \boldsymbol{\beta}_{l-1}^{k+1} = \left\{ \beta_1^{k+1}, \ldots, \beta_{l-1}^{k+1}, \beta_l, \beta_{l+1}^k, \ldots, \beta_{Q+1}^k \right\}. \tag{13}$$

The variant (13) is known as the *Expectation Conditional Maximization*, (ECM) [28]. It is actually a subclass of generalized EM algorithms that are more broadly applicable than EM but share its desirable convergence properties[4]. Therefore, we resort to ECM for function $\mathcal{Q}_{\mathbb{Y} \,|\, \boldsymbol{H}}$. Hereafter, we provide the updating equations of ECM for the estimation of $\boldsymbol{\theta}_{\mathbb{Y} \,|\, \boldsymbol{H}}$ and $\boldsymbol{\theta}_{\boldsymbol{H}}$.

C. Re-estimation formulas for $\boldsymbol{\theta}_{\mathbb{Y} \,|\, \boldsymbol{H}}$

The present M step of ECM is performed by searching an optimal parameter of $\boldsymbol{\theta}_{\mathbb{Y} \,|\, \boldsymbol{H}}$ at a time, say β_l^{k+1}, given by (13) or equivalently by the solution of $\partial \mathcal{Q}_{\mathbb{Y} \,|\, \boldsymbol{H}} / \partial \beta_l^{k+1} = 0$. This procedure is repeated for all parameters belonging to $\boldsymbol{\theta}_{\mathbb{Y} \,|\, \boldsymbol{H}}$. As shown in Appendix II, the computation of the optimal nuisance variables $\widehat{\mathbb{l}}$ can be parallelized since

$$\widehat{\boldsymbol{l}}_i = \boldsymbol{P}_i^{\mathrm{t}} \left(\boldsymbol{y}_i - \boldsymbol{X}_i \widehat{\boldsymbol{h}}^{\mathrm{MAP}} \right), \text{ for } i \in \mathbb{N}_I^*. \tag{14}$$

Taking the new estimates $\widehat{\boldsymbol{l}}_i$ and all the available datasets into account[5], we get for the noise variance:

$$\widehat{r}_b = \frac{\sum_{i=1}^I \mathcal{E}_i(\widehat{\boldsymbol{l}}_i)}{N}, \tag{15}$$

where functions $\mathcal{E}_i(\cdot)$ are defined by (24) (see Appendix II). As it appears in (24), updating both $\boldsymbol{l}_i$ and r_b requires the computation of $\widehat{\boldsymbol{h}}^{\mathrm{MAP}}$. Strictly speaking, this estimate is given by the solution of a linear system without explicitely computing matrix $\boldsymbol{\Sigma}$. Nonetheless, the latter matrix is needed for setting $\widehat{r}_b$ and is thus computed before updating $\boldsymbol{l}_i$.

D. Re-estimation formulas for $\boldsymbol{\theta}_{\boldsymbol{H}}$

The M step of the ECM algorithm w.r.t. $\boldsymbol{\theta}_{\boldsymbol{H}}$ is actually identical to a M step of the standard version of EM since all the parameters belonging to $\boldsymbol{\theta}_{\boldsymbol{H}}$ can be simultaneously updated. As previously derived, the optimal parameters $\boldsymbol{\theta}'_{\boldsymbol{H}}$ are given by the solution of $\partial \mathcal{Q}_{\boldsymbol{H}} / \partial \boldsymbol{\theta}'_{\boldsymbol{H}} = 0$ that is (see Appendix III for details)

$$\tau'_m = \begin{cases} \dfrac{\mathrm{tr} \left\{ \left(\widehat{\boldsymbol{h}}^{\mathrm{MAP}} (\widehat{\boldsymbol{h}}^{\mathrm{MAP}})^{\mathrm{t}} + \boldsymbol{\Sigma} \right) \boldsymbol{R}_{\boldsymbol{H}}^{-1} \right\}}{M(K-1)} & \text{if } \tau'_m = \tau', \forall m \\[2ex] \dfrac{\mathrm{tr} \left\{ \left(\widehat{\boldsymbol{h}}_m^{\mathrm{MAP}} (\widehat{\boldsymbol{h}}_m^{\mathrm{MAP}} \right)^{\mathrm{t}} + \boldsymbol{\Sigma}_m \right\} \boldsymbol{R}^{-1}}{K-1}, & \text{otherwise.} \end{cases} \tag{16}$$

Finally, the successive steps of ECM are summarized in Table II.

[Table 2 about here.]

As to numerical implementation of ECM, the following conjunction has been selected as stopping criterion:

$$\frac{\left| \mathcal{Q}(\widetilde{\boldsymbol{\theta}}^k, \widetilde{\boldsymbol{\theta}}^{k-1} \; ; \mathbb{y}, \mathcal{M}) - \mathcal{Q}(\widetilde{\boldsymbol{\theta}}^{k-1}, \widetilde{\boldsymbol{\theta}}^{k-1} \; ; \mathbb{y}, \mathcal{M}) \right|}{\left| \mathcal{Q}(\widetilde{\boldsymbol{\theta}}^k, \widetilde{\boldsymbol{\theta}}^{k-1} \; ; \mathbb{y}, \mathcal{M}) \right|} < \gamma_1$$

$$\sup_i \left[\frac{||\widetilde{\theta}_i^k - \widetilde{\theta}_i^{k-1}||_2}{||\widetilde{\theta}_i^k||_2} \right] < \gamma_2,$$

where $\widetilde{\boldsymbol{\theta}}^k$ denotes the solution at the kth iteration of the minimization stage and the thresholds have been set to $(\gamma_1, \gamma_2) = (10^{-4}, 10^{-5})$.

[4]The EM algorithm converges monotonically to a stationary point of $p(\mathbb{y} \,|\, \mathcal{M} \; ; \widetilde{\boldsymbol{\theta}})$ if $\mathcal{Q}(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}^k \; ; \mathbb{y}, \mathcal{M})$ is continuous in both $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}^k$ [26], [29].

[5]Since a constant noise variance has been assumed throughout the sessions.

Sections II – IV have allowed us to design *unsupervised non-parametric estimates* of the HRF at a given voxel of the brain, depending on the trial type and the subject under study. Moreover, we have provided uncertainty measures on this estimation to be able to quantify the confidence we may have in the results. Hereafter, we demonstrate the accuracy and robustness of the proposed estimates, first on simulated datasets, and finally through an experimental fMRI paradigm.

V. Simulations results

The first part of this section is a Monte-Carlo study that compares the statistical properties of the Maximum-Likelihood (ML) and MAP HRF estimates (summarized in Table III).

In the following simulations, the CNR is defined as the ratio between the l_1-norm of the HRF and the standard deviation $r_b^{1/2}$ of the noise[6]:

$$\text{CNR} \triangleq \frac{\sum_{k=0}^{K} |h_{m,k\Delta t}|}{(K+1)\, r_b^{1/2}},$$

To provide a single CNR value for time series containing several HRFs with different shapes, we simply average the CNR of each response. The CNR is taken in the range of observed CNR in fMRI data.

The rest of this section emphasizes the performances of the MAP estimator and focuses on the following topics, specific to our extensions:

- Effect of *oversampling*, *i.e.*, whether it is relevant to choose a HRF sampling period Δt for $\boldsymbol{h}$ lower than TR, the sampling period of the data (see Subsection V-C). For this topic, we have chosen a high CNR to rigorously quantify the instant-matching error.
- Successive improvements when allowing for trends and when processing several sessions at the estimation stage (see Subsection V-D). For this investigation, we have considered a higher noise level to better highlight the improvements brought by the multisession method on the average bias and variance of estimation of the HRF. In addition, a session-dependent drift term, similar to the one observed in fMRI data, has been added to quantify the gain that we can achieve when modeling these fluctuations.
- Validity of the error bars on the HRF estimates (see Subsection V-E).
- Robustness of the HRF estimate to departure from the hypothesis of equal noise variance across sessions (see Subsection V-F).
- Influence of *overparametrization*, *i.e.*, when too many HRFs have been modeled and estimated compared to the number that are actually present in the voxel under study (see Subsection V-G). For this purpose, favorable experimental conditions have been chosen.

Each of these topics is analyzed using synthetic data. The parameters of the simulation have been chosen to be compatible with experimental conditions usually encountered, as explained now.

[Table 3 about here.]

A. Common procedure for generating simulated datasets

For each session, we have first simulated a *random-intermixed* sequence of indexes coding for two different event types ($M = 2$). Each index corresponds to a specific stimulus. Unless otherwise specified, the timing of the trials is random, since the ISIs between successive trials follow a uniform distribution on [2.5, 3.5]. This might not be optimal for Maximum-Likelihood estimate. However, an optimal design for the estimation of the HRF will not be optimal for signal detection [20] and we have therefore chosen a tradeoff that is generally considered in neuro-imaging experiments. To investigate the influence of the experimental design on the behavior of the ML and MAP HRF estimates, we compare the statistical properties of the solutions computed for 1) an event-related paradigm (low detection efficiency and good HRF estimation [20], [30]) and 2) a block design (high detection efficiency and poor HRF shape estimation [20], [31]) at the same noise level.

While the optimization of the design parameters is out of the scope of this work, these two settings contrast two opposite situations from the HRF estimation/signal detection point of view. The reader interested by the optimization of experimental design may refer to an excellent survey [32] (and references quoted therein), which also introduces a stochastic framework based on genetic algorithms[7] to optimize a fitness measure of the experimental design wrt several parameters (detection efficiency, HRF shape estimation, counterbalancing of events, ...).

The onsets of the trials are put together on the same temporal grid using $\delta t = 0.5$ s for sampling period. This step simply requires to move the onsets to the nearer time points on the $n\delta t$ grid.

[6]The CNR is sometimes defined as the ratio between the magnitude of the peak signal change and the standard deviation of the noise. However, the peak value is no longer a good descriptor of the entire signal when different HRF shapes are considered.

[7]These tools are availbale at `http://www.lsa.umich.edu/psych/research\&labs/jjonides/download.html`

Each binary time series coupled to a stimulus is then convolved with a specific HRF, $\boldsymbol{h}_1$ or $\boldsymbol{h}_2$, whose exact shapes are plotted in Fig. 1 for instance. $\boldsymbol{h}_1$ is the canonical HRF used by the SPM99 software[8], whereas $\boldsymbol{h}_2$ is chosen as an extreme example of a very peaky HRF. The true HRFs used in the following subsections are similar up to a scaling factor, leading to different values of the CNR at a given noise level.

A white Gaussian noise of variance r_b has been added to the data, as well as a session-dependent low-frequency drift, which was generated from a cosine transform basis which coefficients $\boldsymbol{l}_i$ were drawn from a normal distribution. The amount of low frequency signal was tuned to be significant: we have checked for each session that the ratio between the quadratic norm of the drift components $\boldsymbol{P}_i\boldsymbol{l}_i$ and the quadratic norm of the drift-free data $\boldsymbol{X}_i\boldsymbol{h} + \boldsymbol{b}_i$ was no less than 50 %. We stress here that the events are well distributed over time such that collinearity with the low frequency signal is unlikely.

Number Q_i depends on the chosen Cut-off-Period (COP) for the drift term present in session i. For instance, the selected parameters in row number 3 of Table III leads to $Q_i = 4$. The data are then obtained after undersampling the sequences at a $\left[\mathrm{TR}/\delta t\right]$ rate, the inter-scan interval being $\mathrm{TR}=2$ s. Note also that the length of the datasets varies across sessions.

B. Statistical properties of the MAP estimate

We assess the statistical properties of the ML and MAP estimates through a Monte-Carlo study in the finite and asymptotic[9] data cases. The ML solution can be derived from (5) by setting $\lambda = r_b/\tau = 0$. In particular, we first outline the bias-variance tradeoff, which is intrinsic to Bayesian estimation in the finite data case and illustrate the asymptotic convergence of the MAP estimates to the true HRFs.

To obtain an approximation of the bias for the ML and MAP estimates, we have first computed the *average solution* over $J = 100$ realizations of the noise process:

$$\widetilde{\boldsymbol{h}}_J{}^{\mathrm{MAP}} = \frac{1}{J}\sum_{j=1}^{J}\widehat{\boldsymbol{h}}_j{}^{\mathrm{MAP}} \qquad \widetilde{\boldsymbol{h}}_J{}^{\mathrm{ML}} = \frac{1}{J}\sum_{j=1}^{J}\widehat{\boldsymbol{h}}_j{}^{\mathrm{ML}}.$$

[8] `www.fil.ion.ucl.ac.uk/spm/spm99.html`

[9] There are generally two asymptotic situations. The first one occurs when the number of data $N \to +\infty$ whereas the second one amounts to increasing the CNR in order to characterize an estimator with noise-free data. Here, we have investigated the second case. Nonetheless, both situations are equivalent in our approach since hyperparameter $\lambda = r_b/\tau$ is automatically tuned from the data. This means that when CNR $\to +\infty$, $\lambda \to 0$, which amounts to overweighting the likelihood term.

The estimation variance of the MAP and ML solutions is then approximated using

$$\begin{aligned}\mathcal{V}(\widetilde{\boldsymbol{h}}_J{}^{\mathrm{MAP}}) &= \frac{1}{J}\sum_{j=1}^{J}\left(\widehat{\boldsymbol{h}}_j{}^{\mathrm{MAP}} - \widetilde{\boldsymbol{h}}_J{}^{\mathrm{MAP}}\right)^2\\ \mathcal{V}(\widetilde{\boldsymbol{h}}_J{}^{\mathrm{ML}}) &= \frac{1}{J}\sum_{j=1}^{J}\left(\widehat{\boldsymbol{h}}_j{}^{\mathrm{ML}} - \widetilde{\boldsymbol{h}}_J{}^{\mathrm{ML}}\right)^2\end{aligned}$$

It follows that the quantities of interest *i.e.*, the mean square error for any HRF $\boldsymbol{h}_m$ can be computed using

$$MSE(\widehat{\boldsymbol{h}}\ ;\ \boldsymbol{h}) \triangleq \mathrm{E}\left[||\widehat{\boldsymbol{h}} - \mathrm{E}\left[\widehat{\boldsymbol{h}}\right]||^2\right] \approx \mathcal{V}(\widetilde{\boldsymbol{h}}_J) + (\boldsymbol{h} - \widetilde{\boldsymbol{h}}_J)^2.$$

The global MSE (gMSE) is then obtained after averaging over all time points:

$$gMSE(\widehat{\boldsymbol{h}}\ ;\ \boldsymbol{h}) = \sum_{k=1}^{K} MSE(\widehat{h}_k\ ;\ h_k)/K.$$

To remove the *instant-matching error*, we have considered synchronous paradigms. The onsets of the stimuli and the data are defined on the same grid with $\Delta t = \mathrm{TR} = 1$ s. The ML and MAP estimates are now compared for low and high CNR values (see Table III).

1) Low CNR or "finite data" case: Fig. 1 shows the average HRF estimates computed both for the ML and MAP strategies. As it clearly appears on Fig. 1(a), the ML solution is unbiased. The pointwise error bars that are depicted on Fig. 1(a) have been computed as the square roots of the variance of estimation $\mathcal{V}(\widetilde{\boldsymbol{h}}_J{}^{\mathrm{ML}})$. The average solutions depicted in Fig. 1(b)-(c) correspond to the MAP estimates computed either for a *constant* prior model ($\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ have the same prior variance: $\tau_1 = \tau_2$) or for an *adaptative* prior model ($\tau_1 \neq \tau_2$). These average time courses illustrate the well-known intrinsic bias-variance tradeoff that appears in Bayesian solutions in the finite data case: the MAP solutions are biased but less variable than the ML estimate (their error bars, computed from $\mathcal{V}(\widetilde{\boldsymbol{h}}_J{}^{\mathrm{MAP}})$, are lower). To measure the gain brought by Bayesian methodology, we use the MSE and the summarizing index gMSE. We also note that $\boldsymbol{h}_2$ has been taken as an extreme example of HRF since it is much more peaked than the usual response, better represented by $\boldsymbol{h}_1$.

To emphasize the role played by the experimental paradigm in terms of bias-variance tradeoff, the same quantities have been computed for the block design and the results are depicted in Fig. 1(d)-(f). For the ML estimate plotted in Fig. 1(d), choosing a block design rather than an event-related one generates an increase of the variance of estimation and a slight bias. For the MAP estimates (Fig. 1(e)-(f)), the variance remains roughly

constant but the bias significantly increases compared to the event-related situation. To summarize, working with a poor design for HRF estimation amounts to decreasing the CNR both for ML and Bayesian procedures.

[Figure 1 about here.]

Fig. 2(a)-(b) provide the corresponding MSE over all time samples for $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$, respectively. Fig. 2(a) shows that the MSE is always lower for the MAP solutions when dealing with $\boldsymbol{h}_1$. The MAP estimate computed with an adaptative prior model has a larger bias around the peak but a lower gMSE compared to ML as reported in Table IV. By contrast, Fig. 2(b) shows that both MAP solutions have a larger MSE around the peak of $\boldsymbol{h}_2$ even if the adaptative prior model allows to significantly reduce this error. As shown in Table IV, the MAP estimates still remain more attractive than the ML solution since their gMSE is always lower. It is also important to compare both MAP estimates and to note that the adaptative prior model provides better solutions in terms of gMSE at the expense of the computational cost (twice more expensive).

[Figure 2 about here.]

The MSE has also been computed for the block design experiment. As illustrated in Fig. 2(c)-(d) as well as in Table IV, these results show the significant increase of the mean square error both for the ML and MAP estimates, but more importantly for the ML solution. To conclude, the designer of the paradigm should carefully select the objective function to be optimized according to the underlying question (detection *vs* HRF estimation) [32].

[Table 4 about here.]

2) High CNR or "asymptotic" data case: Fig. 3(a) shows the average ML estimates in the nearly noise-free case. As illustrated, the ML estimate remains unbiased and is *asymptotically consistent* since its estimation variance tends to zero. The average MAP estimates computed with a constant prior model[10] are plotted in Fig. 3(b) for comparison. As shown on this figure, there is a strong evidence that the MAP converges to the true HRF too, so that our solution is *asymptotically consistent*. It should be stressed that such a result is a direct consequence of the automatic tuning of the hyperparameters. Indeed, if $\lambda = r_b/\tau$ was kept constant (as in supervised estimation) the MAP estimate would be biased when the noise variance r_b decreases. Since $\lambda \to 0$ when $r_b \to 0$, the weight of the prior model tends to vanish, so that asymptotically the MAP estimate identifies with the ML solution. As reported in Table IV, the gMSE index provides similar results for both estimates.

[10]the MAP solution computed with an adaptative prior model is not reported here since it provides the same average time courses.

[Figure 3 about here.]

C. How to choose the HRF sampling rate ?

The present simulation addresses tradeoff between the bias and computational burden depending on the chosen sampling period Δt given δt and TR. Here, we have considered a sufficiently high CNR to be able to assess the statistical properties from one single realization of the noise process.

Since the onsets of the trials occur with a temporal resolution of δt seconds, they are not synchronised with the data. Our aim is to determine the optimal oversampling period $\widehat{\Delta t}$. This optimal value should be able to control the instant-matching error in the estimate, and help us to best set the tradeoff between bias reduction and potential variance increase.

Here, we have tested the single session asynchronous model (2), in which the drift component has been discarded. In addition, we have imposed identical prior variances ($\tau_1 = \tau_2$).

Fig. 4 shows a qualitative comparison between $\Delta t =$ TR and $\Delta t <$ TR models when dealing with asynchronous paradigms. The influence of the oversampling period appears clearly: the closer Δt is to δt, the less biased the estimate is.

[Figure 4 about here.]

We have quantitatively checked this result measuring the *quadratic error* $\mathcal{E}$ and the *mean standard deviation* $\mathcal{S}$ of the HRF estimates with

$$\begin{aligned} \mathcal{E}(\widehat{\boldsymbol{h}}_m\ ;\ \boldsymbol{h}_m) &= \frac{||\widehat{\boldsymbol{h}}_m - \boldsymbol{h}_m||^2}{K-1}, \quad m = 1,2 \\ \mathcal{S}(\widehat{\boldsymbol{h}}_m) &= \frac{1}{K-1}\sum_{p=1}^{K-1} \sigma_{m,p}. \end{aligned} \tag{17}$$

For $\boldsymbol{h}_1$ that fits well with the prior model, both criteria $\mathcal{E}$ and $\mathcal{S}$ decrease with Δt, even when the sampling period goes below $\widehat{\Delta t}$ (see Table V). By contrast, for $\boldsymbol{h}_2$, $\mathcal{E}$ decreases until $\Delta t = \delta t =$ TR/4 and increases for lower values, as reported in Table V. Therefore, selecting a sampling period under $\widehat{\Delta t} =$ TR/4 does not really improve the accuracy and robustness of the global estimate $\widehat{\boldsymbol{h}}^{\text{MAP}}$. These results are also obtained for lower CNR values.

[Table 5 about here.]

Choosing $\Delta t < \widehat{\Delta t}$ leads to an increase of the computational cost. More precisely, the results of Fig. 4(a)-(d) where the CNR is high have been obtained in 0.03 s, 0.08 s, 0.33 s and 3.1 s respectively on a Pentium IV 1 GHz. Decreasing the CNR induces a slower convergence, but for $\Delta t = \delta t$, the HRF estimate is usuallly computed in about one second for a dataset of two hundreds samples. The computational burden is therefore low enough to make this approach feasible for several hundreds time series (voxels or region of interest).

Moreover, choosing $\Delta t = \widehat{\Delta t}$ compared to larger values leads to a better estimate of the noise variance r_b. This is likely to be due to the fact that some data variance may be introduced by the undersampling step.

Finally, we note that the hyperparameters ($\lambda = r_b/\tau$, if $\tau_m = \tau,\ \forall m \in \mathbb{N}_M^*$, or $\lambda_m = r_b/\tau_m$ otherwise) automatically adapt to the level of discretization. In the simulation performed for different sampling periods, we have checked that $\lambda_K \approx 8\lambda_{2K}$ (or $\lambda_{\Delta t} \approx 8\lambda_{\Delta t/2}$) is approximately satisfied.

D. Improvements brought by successive models

The goal of this section is to classify the performances of the HRF estimates for the models that have been successively designed in Section II, in terms of quadratic error and variance reduction.

i. the influence of modeling drift terms in (2);

ii. the *session-dependent* model (2), where each HRF is estimated from each session (before a possible averaging step over the sessions), compared to the *multisession* model ($\mathcal{M}$), where each HRF admits a single estimate derived from the whole set of sessions.

[Figure 5 about here.]

(i) Comparison of rows 1-3 and 2-4 of Table VI proves that the quadratic error of the HRF estimate diminishes when modeling a drift component (more significantly for $\boldsymbol{h}_2$): the criterion $\mathcal{E}$ is smaller when the nuisance variables are jointly estimated. The variance reduction or the decrease of $\mathcal{S}$ is rather marginal, probably because the number of unknown parameters is larger when a trend is modeled with a constant number of data. These improvements are emphasized when dealing with multisession estimation. Fig. 5(b) illustrates the gain in robustness brought by model ($\mathcal{M}$), which takes a session-dependent drift into account, compared to the results when the drift has not been modeled in Fig. 5(c). Quantitative results reinforce these improvements, as shown in rows 7-8 of Table VI.

(ii) Rows 5-6 of Table VI as well as the comparison of Fig. 5(a)-(c) show that multisession modeling has a greater influence on the accuracy of the HRF estimates than the consideration of nuisance variables in the model. Assuming stability of the HRFs across sessions in a given voxel actually allows to bring more information, and therefore leads to a significant bias reduction and a slight variance decrease. The variance reduction is even greater on data without drift component (results not shown). Nonetheless, the drift embedded in the data being random and session-varying, the $\mathcal{S}$ values reported in Table VI do not go down when a session-dependent drift has not been modeled. Comparison of Fig. 5(a)-(b) demonstrates the relevance of the HRF estimates computed from all available datasets ($I = 4$).

[Table 6 about here.]

E. Validity of the error bars

A concern is to know whether the approximation leading to (8) has a good chance to be satisfied. This approximation is at the origin of the variance underestimation of $\widehat{\boldsymbol{h}}^{\text{MAP}}$. Our argument is that with a small number of drift parameters compared to the number of data (when Q/N is small enough) and a standard CNR our approach provides an accurate estimate of the nuisance variables such that the extra source of error related to these parameters should be negligeable. Such cases are realistic and occur when $Q/N \leqslant 0.02$, for instance when the lowest frequency $f_{\min}$ is 1/120, $N = 180$ and $TR = 2$ s we can choose $Q_i = 4$. Consequently, we are interested in quantifying the bias on the variance estimate provided by our approach. To achieve this goal, we performed several Monte Carlo studies[11] in which we analyzed the influence of the number of nuisance variables ($Q_i = 4$ and $Q_i = 18$) as well as the the noise level (CNR=0.73 as in Subsection V-D and CNR=0.29) on the estimation variance of $\mathbb{l}$. The drift terms were held to the same value for a complete run. First of all, we checked that our estimate $\widehat{\mathbb{l}}^{\text{ML}}$ is unbiased. For a small number of drift terms ($Q_i = 4$, $Q_i/N_i = 0.02$) and a standard CNR (CNR=0.73), the standard deviation of our nuisance variable estimate $\widehat{\mathbb{l}}^{\text{ML}}$ was no larger than 3 % of the maximal amplitude of the drift. This analysis only gives the behavior of our estimate $\widehat{\mathbb{l}}^{\text{ML}}$, but even indirectly, the dispersion of $\widehat{\mathbb{l}}^{\text{ML}}$ is related to the computed error bars on $\widehat{\boldsymbol{h}}^{\text{MAP}}$. In such cases, our approach essentially provides an accurate approximation of the dispersion of $\widehat{\boldsymbol{h}}^{\text{MMAP}}$ at low cost.

When increasing the noise level (CNR=0.29) or multiplying the number of nuisance varia-bles ($Q_i = 18$,

[11] $J = 10^3$ realizations of the noise have been drawn to accurately estimate the dispersion of the nuisance variable solution $\widehat{\mathbb{l}}^{\text{ML}}$.

$Q_i/N_i = 0.1$), we observed that the standard deviation of $\widehat{\mathbb{I}}^{\text{ML}}$ was at least multiplied by three. In these more difficult cases, the solutions $\widehat{\boldsymbol{h}}^{\text{MAP}}$ and $\widehat{\boldsymbol{h}}^{\text{MMAP}}$ are close to each other but the main difference is that the error bars derived on $\widehat{\boldsymbol{h}}^{\text{MAP}}$ will be underestimated since they will not be able to capture the dispersion attributable to the drift terms. In other words, the estimation variance of $\widehat{\mathbb{I}}^{\text{ML}}$ is not taken into account in $\widehat{\boldsymbol{h}}^{\text{MAP}}$. Hence, we should compute $\widehat{\boldsymbol{h}}^{\text{MMAP}}$ and the marginal posterior covariance matrix $\Sigma_{\boldsymbol{H}}$[12] to remove the existing bias on our error bars. These developments are beyond the scope of this paper (see Section VII). We therefore limit the field of applicability of our technique to small number of nuisance variables ($Q_i \leqslant 5$) and reasonable CNR (roughly CNR $\geqslant 0.5$).

F. Robustness to the equal noise variance hypothesis

We are also interested in testing the robustness of the HRF estimate to a departure from the hypothesis that the noise is constant across sessions. For doing so, we have considered two sessions with respective noise variances $r_b^{(1)} = 0.08$ and $r_b^{(2)} = 0.3$. We stress here that $r_b^{(2)} = 0.3$ is an extremely high noise level, not found in actual datasets but used to emphasize the robustness of our approach in a multisession framework.

[Figure 6 about here.]

As shown in Fig. 6(a) and as expected, the HRF estimates are closer to the original ones when the noise level is low. Fig. 6(b) demonstrates that increasing the noise variance provides oversmoothed results and stresses the limits of the prior model for such low CNR values. Moreover, small error bars that are visible in Fig. 6(b) result from overconfidence in the prior. In this extremely noisy situation, the problem discussed in Subsection V-E is amplified. Here, this variance underestimation concerns the hyperparameters and is due to the large uncertainty on the hyperparameter estimates. With a more important computational effort, this uncertainty could be computed using the methodology derived in [33]. The opposite result would be observed in the standard maximum-likelihood framework, where the higher the noise level, the larger the error bars. Note also that the noise variance is better estimated when the CNR is not too low. Finally, the HRFs plotted in Fig. 6(c) have been computed from both available datasets, assuming they have been generated with the same noise variances, since this assumption holds in model ($\mathcal{M}$). Clearly, Fig. 6(c) shows that the proposed technique remains robust with regard to some departures from this hypothesis. In other words, the multisession approach provides better results than the session-dependent technique, even with session-varying noise levels.

[12] given by the left superior block of the joint covariance matrix $\Sigma_{\boldsymbol{H},\mathbb{L}}$ of size $M(K-1)Q \times M(K-1)Q$.

G. Overparametrization of the number of HRFs

Here, we have used two trial types: one generating the peaked HRF $\boldsymbol{h}_2$, depicted in Fig. 5 for instance, and one yielding no response (zero function). Two different HRFs were modeled at the estimation stage. The question addressed here is whether modeling the non-existing function as a HRF influences bias and variance of the estimate of $\boldsymbol{h}_2$. Indeed, it is well known in the maximum-likelihood framework that, the larger the number of unknowns, the higher the variance of estimation. Simulations (not reported here) show that inclusion in the model of a *uncorrelated* and *non-existant* HRF does not have any influence on the estimation of the existing HRF, whatever the noise level.

Note that it is also possible to design a statistical test either for assessing whether an estimated HRF is zero or not ($\boldsymbol{h}_1 = \boldsymbol{0}$?), or for comparing both estimates ($\boldsymbol{h}_1 = \boldsymbol{h}_2$?), since the sum of squares of the difference between $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ follows a χ^2 distribution (see [34] for details).

VI. Experiment

A. MRI parameters

The experiment was performed on a 3-T whole-body system (Bruker, Germany) equipped with a quadrature birdcage radio frequency (RF) coil and a head-gradient coil insert designed for echoplanar imaging. Functional images were obtained with a T2*-weighted gradient echo, echo planar imaging sequence (TR = 3.3s, TE = 30ms, FOV $= 240 \times 240$ mm^2, matrix $= 64 \times 64$). Each image, acquired in 1.3s, comprised twenty-two 4-mm-thick axial slices covering most of the brain. A high-resolution ($1 \times 1 \times 1.2$ mm) anatomical image using a 3-D gradient-echo inversion-recovery sequence, was also acquired for each participant.

B. Description of the paradigm

The method was assessed on real data acquired in a speech discrimination experiment. The experiment consisted of six sessions ($I = 6$) comprising one hundred trials ($N_i = 100$) lasting 3.3 seconds each. In each trial, the participant heard two pseudo-words over headphones. His task was to indicate whether he had perceived or not a difference between the two stimuli. There were three types of trials ($M = 3$): 'Phonological', 'Acoustic', 'Control'. In trials belonging to the 'Control' condition,

the two auditory stimuli in the pair were exactly the same. In the 'Phonological' condition, the stimuli differed along a contrast used to distinguish words in the language of the participant (it was *linguistically relevant*: path *vs* bath in English). In the 'Acoustic' condition, the stimuli also differed but the contrast between the stimuli was not relevant in the language of the participant (*e.g.,* beat *vs* beet in English).

The stimuli pairs were presented during the silent gaps lasting 2 seconds between two succesive acquisitions (the TR was 3.3 s and the time of acquisition of one volume was 1.3 s). The onsets of events were aligned with the start of the second stimulus in a pair (*i.e.,* at 1.65 s), which felt in between two successive acquisitions. $\Delta t = TR/2$ is therefore an appropriate choice for the sampling period of the HRF.

C. Results

The top of Fig. 7 shows t maps (thresholded at $P = 0.001$ corrected for multiple comparisons) superimposed on axial slices of averaged T1-weighted images, computed with SPM99. To obtain such results, one first has to specify an fMRI model that uses a canonical HRF $\boldsymbol{h}_c$. Second, least squares estimation and inference on relevant contrasts of the parameter estimates is performed. The first half of the regressors of the fMRI model is defined by $f_m(t_n) \triangleq (\boldsymbol{x}^{(m)} \star \boldsymbol{h}_c)(t_n)$ for $n = 1, \ldots, N = \sum_i N_i = 600$ and $m = 1, \ldots, M = 3$. The second half derives from the convolution of the first-order derivative of $\boldsymbol{h}_c$ with the onsets. The use of derivatives allows to gain robustness against variations of the maximum amplitude or the delay of this maximum.

The middle row of Fig. 7 shows the Maximum-likelihood HRF estimates in three different voxels from the left superior temporal gyrus of one participant: (Talairach coordinates in millimeters: $V_1\,(X = -60,\ Y = -24,\ Z = 4)$; $V_2\,(X = -68,\ Y = -28,\ Z = 8)$; $V_3\,(X = -64,\ Y = -40,\ Z = 16)$. These time courses have been computed without oversampling *i.e.,* just by estimating the amplitude of each HRF at every TR. The bottom row of Fig. 7 shows the MAP HRF estimates in the same voxels. These results have been obtained from model ($\mathcal{M}$), that is discarding the influence of the trend and the baseline and taking the six datasets into account. We have also considered the same prior variance for the three modeled hemodynamic responses.

Not surprisingly, Fig. 7(a_1)-(a_2) proves that the stimuli elicited very similar responses in Heschel gyrus (primary auditory cortex, V_1). The two other voxels were located in the planum temporale. Fig. 7(b_1)-(b_2) shows for V_2 that there is differential treatment when the stimuli differed, regardless of the type of difference (phonological or acoustic). By contrast, Fig. 7(c_1)-(c_2) shows a specific increment for phonological contrasts (speech processing).

The main error that appears on the ML solutions in Fig. 7(a_1)-(c_1) concerns the *hemodynamic* delay *i.e.,* the time-to-peak since the instant-matching error is about $TR/2$. To circumvent this problem, the HRFs should also oversampled but this leads to an increased number of parameters and thus to unstable ML solutions. This situation therefore requires regularization.

[Figure 7 about here.]

VII. Discussion

In this paper, we have described and tested a general method for estimating the hemodynamic response function in fMRI data. The method is general enough to deal with all specific features of fMRI data, including the ability to work on several sessions and several experimental conditions in the context of an asynchronous sampling in event-related paradigm. In addition, our optimization scheme is sufficiently efficient to allow large fMRI time series to be processed. Finally, physiological artifacts can be correctly taken into account, provided that they are modeled with a small number of parameters (typically four per session). This work has been implemented in a MATLAB toolbox[13] and interfaced with the SPM99 software.

To our knowledge, this work presents the only comprehensive robust non-parametric estimation of the fMRI brain response to a task or a stimulus. Applications of the technique are manifold. This approach should improve the observation of significant differences between the HRFs estimated for different stimuli within the same brain region. Using the HRF estimate to specify a *subject-dependent* general linear model, we could address the problem of the validity of the choice of the regressors[14] of such a model and may improve the localization of the signal (see [35] for a multivariate approach of this problem). In the domain of fusion of information (in particular with EEG signals) it is also important to have such a robust estimate of the HRF, to investigate the relationship between metabolic and electrical measurements.

Physiologically, the HRF should be zero-valued at $k = 0$. However, in practice, one sometimes detects voxels where this assertion is not true. The most common cause

[13]the HRF toolbox can be downloaded at the following URL site: `http://www.madic.org/download/HRFTBx`

[14]These regressors has been selected empirically so far.

of this effect seems to be uncorrected stimulus-correlated motion. Since motion correction in fMRI data is hardly perfect, it is often important to remove the zero constraint at $k = 0$. In the proposed software, this constraint can be either changed to a less stringent prior (for instance the first-order derivative at the extremal time samples of the HRF) or shifted in time such that the start of the HRF estimate can be chosen arbitrarily before the actual occurrence of the stimuli (for instance a few seconds before). This allows the user to check the influence of this constraint on the HRF estimation.

The method can be extended in several ways:

First, to reduce the bias on the error bars we could resort to a second order Taylor expansion on the drift estimate as proposed for hyperparameters in [33]. This scheme needs an important computational effort. Alternatively, the usual Bayesian methodology leads to integrate the joint posterior distribution $p(\boldsymbol{h}, \mathbb{l} \,|\, \mathrm{y}, \mathcal{M} \,;\, \boldsymbol{\theta})$ over $\mathbb{l}$. Further developments should be done to develop the MMAP estimate and check its ability to provide error bars that allow for the dispersion of the nuisance variables, in particular if a large number of nuisance variables is necessary to model drift terms as well as other additive effects of no interest. In this framework, a particular attention should be paid to hyperparameter estimation. If an EM algorithm was considered to perform this step, the *complete* data should be $(\mathrm{y}, \boldsymbol{h}, \mathbb{l})$ and function $\mathcal{Q}$ should be defined by

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^k; \mathrm{y}, \mathcal{M}) &= \cdots \\ &= \int \log p(\mathrm{y}, \boldsymbol{h}, \mathbb{l} | \mathcal{M}; \boldsymbol{\theta})\, p(\boldsymbol{h}, \mathbb{l} | \mathrm{y}, \mathcal{M}; \boldsymbol{\theta}^k)\, d\boldsymbol{h}\, d\mathbb{l} \\ &= \mathrm{E}_{\boldsymbol{H},\mathbb{L}}\big[\log p(\mathrm{y}, \boldsymbol{h}, \mathbb{l} | \mathcal{M} \,;\, \boldsymbol{\theta}) \,|\, \mathrm{y}, \mathcal{M}; \boldsymbol{\theta}^k\big].\end{aligned}$$

Maximizing $\mathcal{Q}$ would require the computation of terms similar to (24), which could depend on the joint covariance matrix $\Sigma_{\boldsymbol{H},\mathbb{L}}$ of size $M(K-1)Q \times M(K-1)Q$. This matrix is obtained after a block matrix inversion that will be more expensive than the computation of the covariance matrix $\boldsymbol{\Sigma}$ (see (5)) since the number of nuisance variables $Q = \sum_i Q_i$ would be larger than the number of HRF coefficients $M(K-1)$. On the other hand, since vector $\boldsymbol{\theta}$ would have a smaller size, the EM algorithm should converge in fewer iterations with a higher cost per iteration.

Second, another interesting extension would consist in considering a spatial model. Since the BOLD signal is known to have some spatial structure [36], estimation of the HRF over a region of interest should also provide a more robust estimation, as demonstrated by [37] who use a general prior on the spatial extension of the signal.

Third, the model presented here assumes that the response is constant in time. While this assumption is reasonable in a first instance (as long as the ISIs do not decrease below about two seconds), it is likely that there exists some variations in time due to physiological or neural adaptation to the stimulus or task. This is the subject of ongoing research.

Fourth, it is not yet clear if the responses coming from different subjects can easily be averaged in a single response. It may be that several subjects have too different brain responses such that the averaging of those signals would be difficult to interpret. The extension of the method to deal with multisubject data should be developed, although data can be analysed at the same time using a different HRF per subject, with a generalization of model ($\mathcal{M}$) that takes different noise variances across subjects into account. Such a generalization requires to overcome the actual difficulties encountered in group analysis: the most salient ones consist in removing subject-dependent movement artifacts and normalizing all subjects in the same space reference.

Lastly, we hope that this method can be the basis for some work that would take advantage of the recent advance in joint recording of electrical activity at the surface of the scalp in the MR scanner. In particular, using such (joint) recordings, it may be possible to partially retrieve the local field potential information from the BOLD signal, which would provide a better understanding of the neural computation well resolved in time and space.

Appendix I
Decoupling the complete likelihoold optimization

Vector $\widetilde{\boldsymbol{\theta}}$ can be partitionned into two subvectors $\boldsymbol{\theta}_{\mathbb{Y}\,|\,\boldsymbol{H}}$ and $\boldsymbol{\theta}_{\boldsymbol{H}}$ which respectively control the conditional pdfs $p(\mathrm{y} \,|\, \boldsymbol{h}, \mathcal{M} \,;\, \boldsymbol{\theta}_{\mathbb{Y}\,|\,\boldsymbol{H}})$ and $p(\boldsymbol{h} \,;\, \boldsymbol{R}, \boldsymbol{\theta}_{\boldsymbol{H}})$. The M-step of the EM algorithm can be divided into two simpler independent maximization problems. The complete likelihood which enters in the definition of $\mathcal{Q}$ in (12) can be expressed as

$$p(\mathrm{y}, \boldsymbol{h} \,|\, \mathcal{M} \,;\, \widetilde{\boldsymbol{\theta}}) = p(\mathrm{y} \,|\, \boldsymbol{h}, \mathcal{M} \,;\, \boldsymbol{\theta}_{\mathbb{Y}\,|\,\boldsymbol{H}})\, p(\boldsymbol{h} \,;\, \boldsymbol{R}, \boldsymbol{\theta}_{\boldsymbol{H}}). \tag{18}$$

For any set value of parameters vector $\widetilde{\boldsymbol{\theta}}^0$, define functions $\mathcal{Q}_{\mathbb{Y}\,|\,\boldsymbol{H}}$ and $\mathcal{Q}_{\boldsymbol{H}}$ as

$$\mathcal{Q}_{\mathbb{Y}|\boldsymbol{H}}(\boldsymbol{\theta}_{\mathbb{Y}|\boldsymbol{H}}, \widetilde{\boldsymbol{\theta}}^0; \mathrm{y}) = \mathrm{E}\big[\log p(\mathrm{y} \,|\, \boldsymbol{h}, \mathcal{M}; \boldsymbol{\theta}_{\boldsymbol{Y}|\boldsymbol{H}}) | \mathrm{y}; \widetilde{\boldsymbol{\theta}}^0\big] \tag{19}$$

$$\mathcal{Q}_{\boldsymbol{H}}(\boldsymbol{\theta}_{\boldsymbol{H}}, \widetilde{\boldsymbol{\theta}}^0; \mathrm{y}) = \mathrm{E}\big[\log p(\boldsymbol{h}; \boldsymbol{R}, \boldsymbol{\theta}_{\boldsymbol{H}}) | \mathrm{y}; \widetilde{\boldsymbol{\theta}}^0\big]. \tag{20}$$

It can be immediately deduced from (12) and (18) that function $\mathcal{Q}$ can be expressed as

$$\mathcal{Q}(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}^0; \mathbb{y}) = \mathcal{Q}_{\mathbb{Y}|\boldsymbol{H}}(\boldsymbol{\theta}_{\mathbb{Y}|\boldsymbol{H}}, \widetilde{\boldsymbol{\theta}}^0; \mathbb{y}) + \mathcal{Q}_{\boldsymbol{H}}(\boldsymbol{\theta}_{\boldsymbol{H}}, \widetilde{\boldsymbol{\theta}}^0; \mathbb{y}), \tag{21}$$

which shows that the M step of the EM algorithm can be decoupled into two operations: maximization of $\mathcal{Q}_{\mathbb{Y}\,|\,\boldsymbol{H}}$ w.r.t. $\boldsymbol{\theta}_{\mathbb{Y}\,|\,\boldsymbol{H}}$ and maximization of $\mathcal{Q}_{\boldsymbol{H}}$ w.r.t. $\boldsymbol{\theta}_{\boldsymbol{H}}$.

APPENDIX II
UPDATING $\boldsymbol{\theta}_{\mathbb{Y}\,|\,\boldsymbol{H}}$

In the present case, $\mathbb{y}$ is made up of independent realizations $(\boldsymbol{y}_i)_{1 \leqslant i \leqslant I}$. As a consequence, the expression of $\mathcal{Q}_{\mathbb{Y}\,|\,\boldsymbol{H}}$ can be rewritten as

$$\mathcal{Q}_{\mathbb{Y}|\boldsymbol{H}}(\boldsymbol{\theta}'_{\mathbb{Y}|\boldsymbol{H}}, \widetilde{\boldsymbol{\theta}}; \mathbb{y}) = \sum_{i=1}^{I} \mathcal{Q}_{\boldsymbol{Y}_i|\boldsymbol{H}}(\boldsymbol{\theta}'_{\boldsymbol{Y}_i|\boldsymbol{H}}, \widetilde{\boldsymbol{\theta}}; \mathbb{y}). \tag{22}$$

Handling successively the likelihood (3), the Gaussianity of $\boldsymbol{H}$, definition (19) and (22), we are able to derive a closed form expression for $\mathcal{Q}_{\boldsymbol{Y}_i\,|\,\boldsymbol{H}}$:

$$\mathcal{Q}_{\boldsymbol{Y}_i\,|\,\boldsymbol{H}}(\boldsymbol{\theta}'_{\boldsymbol{Y}_i|\boldsymbol{H}}, \widetilde{\boldsymbol{\theta}}; \mathbb{y}) = -\frac{1}{2}\left[N_i \log r'_b + \frac{1}{r'_b}\mathcal{E}_i(\boldsymbol{l}'_i)\right], \tag{23}$$

where function $\mathcal{E}_i(\cdot)$ is defined by the following expectation

$$\mathcal{E}_i(\boldsymbol{l}'_i) = \mathrm{E}_{\boldsymbol{H}}\left[\left\|\boldsymbol{y}_i - \boldsymbol{P}_i\boldsymbol{l}'_i - \boldsymbol{X}_i\boldsymbol{h}\right\|^2 \mid \mathbb{y}, \widetilde{\boldsymbol{\theta}}\right]. \tag{24}$$

Since the posterior pdf $p(\boldsymbol{h} \mid \mathbb{y}, \mathcal{M}\ ;\ \widetilde{\boldsymbol{\theta}})$ is $\mathcal{N}(\widehat{\boldsymbol{h}}^{\mathrm{MAP}}, \boldsymbol{\Sigma})$-distributed, analytic calculations give for $\mathcal{E}_i$:

$$\mathcal{E}_i(\boldsymbol{l}'_i) = ||\boldsymbol{y}_i - \boldsymbol{P}_i\boldsymbol{l}'_i - \boldsymbol{X}_i\widehat{\boldsymbol{h}}^{\mathrm{MAP}}||^2 + \mathrm{tr}\,\left(\boldsymbol{X}_i\boldsymbol{\Sigma}\boldsymbol{X}_i^{\mathrm{t}}\right).$$

Here, we would like to stress the separability property, expressed through (22), that allows to think about a parallel update of vectors $\boldsymbol{\theta}_{\boldsymbol{Y}_i\,|\,\boldsymbol{H}}$, for $i \in \mathbb{N}_I^*$, except for r_b. Such a property results from model $(\mathcal{M})$ and implies that the maximization of $\mathcal{Q}_{\mathbb{Y}\,|\,\boldsymbol{H}}$ is block-wise decoupled for vectors $\boldsymbol{l}_i$. Finally, replacing $\mathcal{E}_i$ by its value in (23), taking the first derivative of $\mathcal{Q}_{\boldsymbol{Y}_i\,|\,\boldsymbol{H}}$ w.r.t. $\boldsymbol{\theta}'_{\boldsymbol{Y}_i\,|\,\boldsymbol{H}}$ and equating to zero yields (14) and (15) for $\boldsymbol{l}_i$ and r_b, respectively.

APPENDIX III
UPDATING $\boldsymbol{\theta}_{\boldsymbol{H}}$

From the prior pdf (4) and the definition (20), function $\mathcal{Q}_{\boldsymbol{H}}$ is defined by

$$\begin{aligned}\mathcal{Q}_{\boldsymbol{H}}(\boldsymbol{\theta}'_{\boldsymbol{H}}, \widetilde{\boldsymbol{\theta}}; \mathbb{y}) &= -\frac{(K-1)}{2}\sum_{m=1}^{M}\log\tau'_m - \frac{\mathcal{E}(\boldsymbol{\theta}'_{\boldsymbol{H}})}{2} \\ &+ \frac{M}{2}\log\det(\boldsymbol{R})\end{aligned} \tag{25}$$

with

$$\begin{aligned}\mathcal{E}(\boldsymbol{\theta}'_{\boldsymbol{H}}) &= \mathrm{E}_{\boldsymbol{H}}\left[\boldsymbol{h}^{\mathrm{t}}\boldsymbol{R}'_{\boldsymbol{H}}\boldsymbol{h} \mid \mathbb{y}, \mathcal{M}, \widetilde{\boldsymbol{\theta}}\right] \\ &= (\widehat{\boldsymbol{h}}^{\mathrm{MAP}})^{\mathrm{t}}\boldsymbol{R}'_{\boldsymbol{H}}\widehat{\boldsymbol{h}}^{\mathrm{MAP}} + \mathrm{tr}\,\left(\boldsymbol{R}_{\boldsymbol{H}|\mathbb{Y}}\boldsymbol{R}'_{\boldsymbol{H}}\right),\end{aligned} \tag{26}$$

and $\boldsymbol{R}'_{\boldsymbol{H}} = [\tau'_1\boldsymbol{R}, \tau'_2\boldsymbol{R}, \ldots, \tau'_M\boldsymbol{R}]$. Here again, replacing $\mathcal{E}$ by (26) in (25) and taking the first derivative of $\mathcal{Q}_{\boldsymbol{H}}$ w.r.t. $\boldsymbol{\theta}'_{\boldsymbol{H}}$ allow us to find the updating equations (16).

ACKNOWLEDGMENT

The authors would like to thank the anonymous referees for helpful comments and remarks

REFERENCES

[1] B. R. Rosen, R. L. Buckner, and A. M. Dale, "Event-related functional MRI: past, present and future." *Proc. Natl. Acad. Sci. USA*, vol. 95, pp. 773–780, 1998.

[2] 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.

[3] P. A. Bandettini, A. Jesmanowicz, E. C. Wong, and J. S. Hyde, "Processing strategies for time-course data sets in functional mri of the human brain," *Magn. Reson. Med.*, vol. 30, pp. 161–173, 1993.

[4] R. L. Buckner, J. Goodman, M. Burock, M. Rotte, W. Koutstaal, D. L. Schachter, B. R. Rosen, and A. M. Dale, "Functional-anatomic correlates of object priming in humans revealed by rapid presentation event-related fMRI," *Neuron*, vol. 20, pp. 285–296, February 1998.

[5] R. 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, 1997.

[6] R. B. Buxton, E. C. Wong, and F. L. R., "Dynamics of blood flow and oxygenation changes during brain activation: the balloon model," *Magn. Reson. Med.*, vol. 39, pp. 855–864, June 1998.

[7] T.-Q. Li, T. N. Haefelin, B. Chan, A. Kastrup, T. Jonsson, G. Glover, and M. E. Moseley, "Assessment of hemodynamic response during focal neural activity in human using bolus tracking, arterial spin labeling, and BOLD techniques," *Neuroimage*, vol. 12, no. 4, pp. 442–451, October 2000.

[8] N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann, "Neurophysiological investigation of the basis of the fMRI signal," *Nature*, vol. 412, no. 6843, pp. 150–157, July, 12 2001.

[9] K. J. Friston, "Statistical parametric mapping," in *Functional Neuroimaging : Technical Foundations*, R. Thatcher, M. Hallet, T. Zeffiro, E. John, and M. Huerta, Eds., 1994, pp. 79–93.

[10] 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, pp. 4207–4221, 1996.

[11] N. Lange, "Empirical and substantive models, the Bayesian paradigm, and meta-analysis in functional brain imaging," *Hum. Brain Mapp.*, vol. 5, pp. 259–263, 1997.

[12] M. S. Cohen, "Parametric analysis of MRI data using linear systems methods," *Neuroimage*, vol. 6, pp. 93–103, 1997.

[13] J. C. Rajapakse, F. Kruggel, J. M. Maisog, and D. Von Cramon, "Modeling hemodynamic response for analysis of functional MRI time-series," *Hum. Brain Mapp.*, vol. 6, pp. 283–300, 1998.

[14] G. H. Glover, "Deconvolution of impulse response in event-related BOLD fMRI," *Neuroimage*, vol. 9, pp. 416–429, 1999.

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

[16] C. Gössl, L. Fahrmeir, and D. P. Auer, "Bayesian modeling of the hemodynamic response function in BOLD fMRI," *Neuroimage*, vol. 14, pp. 140–148, 2001.

[17] F. A. Nielsen, L. K. Hansen, P. Toft, C. Goutte, N. Lange, S. C. Stroher, N. Morch, C. Svarer, R. Savoy, B. Rosen, E. Rostrup, and P. Born, "Comparison of two convolution models for fMRI time series," *Neuroimage*, vol. 5, p. S473, 1997.

[18] G. Marrelec, H. Benali, P. Ciuciu, and J.-B. Poline, "Bayesian estimation of the hemodynamic response function in functional MRI," in *Bayesian Inference and Maximum Entropy Methods*, R. Fry, Ed. Baltimore, MD: MaxEnt Workshops, August 2001.

[19] E. T. Bullmore, M. Brammer, S. C. Williams, S. Rabe-Hesketh, N. Janot, A. David, J. Mellers, R. Howard, and P. Sham, "Statistical methods of estimation and inference for functional MR image analysis," *Magn. Reson. Med.*, vol. 35, pp. 261–277, 1996.

[20] R. Birn, R. Cox, and P. A. Bandettini, "Detection versus estimation in event-related fMRI: choosing the optimal stimulus timing," *Neuroimage*, vol. 15, no. 1, pp. 252–264, January 2002, rapid communication.

[21] K. Worsley, C. Liao, J. Aston, V. Petre, G. Duncan, F. Morales, and A. Evans, "A general statistical analysis for fMRI data," *Neuroimage*, vol. 15, no. 1, pp. 1–15, January 2002.

[22] E. Zarahn, G. K. Aguirre, and M. D'Esposito, "Empirical analysis of BOLD fMRI statistics. I. Spatially unsmoothed data collected under null-hypothesis conditions," *Neuroimage*, vol. 5, pp. 179–197, 1997.

[23] G. H. Golub and C. F. Van Loan, *Matrix computations*, Third ed. Baltimore, Maryland: The Johns Hopkins University Press, 1996.

[24] F. Champagnat, Y. Goussard, and J. Idier, "Unsupervised deconvolution of sparse spike trains using stochastic approximation," *IEEE Trans. Signal Processing*, vol. 44, no. 12, pp. 2988–2998, December 1996.

[25] S. S. Saquib, C. A. Bouman, and K. D. Sauer, "ML parameter estimation for Markov random fields with applications to Bayesian tomography," *IEEE Trans. Image Processing*, vol. 7, no. 7, pp. 1029–1044, July 1998.

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc. B*, vol. 39, pp. 1–38, 1977.

[27] J. Idier, Y. Goussard, and A. Ridolfi, "Unsupervised image segmentation using a telegraph parameterization of Pickard random fields," in *Spatial statistics. Methodological aspects and some applications*, ser. Lecture notes in Statistics, M. Moore, Ed. New York, NY: Springer Verlag, 2001, vol. 159, pp. 115–140.

[28] X.-L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm: A general framework," *Biometrika*, vol. 80, pp. 267–278, 1993.

[29] C. F. J. Wu, "On the convergence of the EM algorithm," *The Annals of Statistics*, vol. 11, no. 1, pp. 95–103, 1983.

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

[31] A. M. Dale, "Optimal experimental design for event-related fMRI," *Hum. Brain Mapp.*, vol. 8, pp. 109–114, 1999.

[32] T. D. Wager and T. E. Nichols, "Optimization of experimental design in fMRI: a general framework using a genetic algorithm," *Neuroimage*, vol. 18, pp. 293–309, 2003.

[33] R. E. Kass and D. Steffey, "Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models)," *J. Amer. Statist. Soc.*, vol. 84, no. 407, pp. 717–726, September 1989.

[34] G. Marrelec, H. Benali, P. Ciuciu, M. Pélégrini-Issac, and J.-B. Poline, "Robust Bayesian estimation of the hemodynamic response function in event-related BOLD MRI using basic physiological information," *Hum. Brain Mapp.*, vol. 19, no. 1, pp. 1–17, May 2003.

[35] F. Kherif, J.-B. Poline, G. Flandin, H. Benali, O. Simon, S. Dehaene, and K. Worsley, "Multivariate model specification for fMRI data," *Neuroimage*, vol. 16, no. 4, pp. 1068–1083, August 2002.

[36] X. Descombes, F. Kruggel, and D. Y. von Cramon, "Spatio-temporal fMRI analysis using Markov Random Fields," *IEEE Trans. Medical Imaging*, vol. 17, no. 6, pp. 1028–1039, December 1998.

[37] C. Gössl, D. P. Auer, and L. Fahrmeir, "Bayesian spatio-temporal inference in functional magnetic resonance imaging," *Biometrics*, vol. 57, 2001.

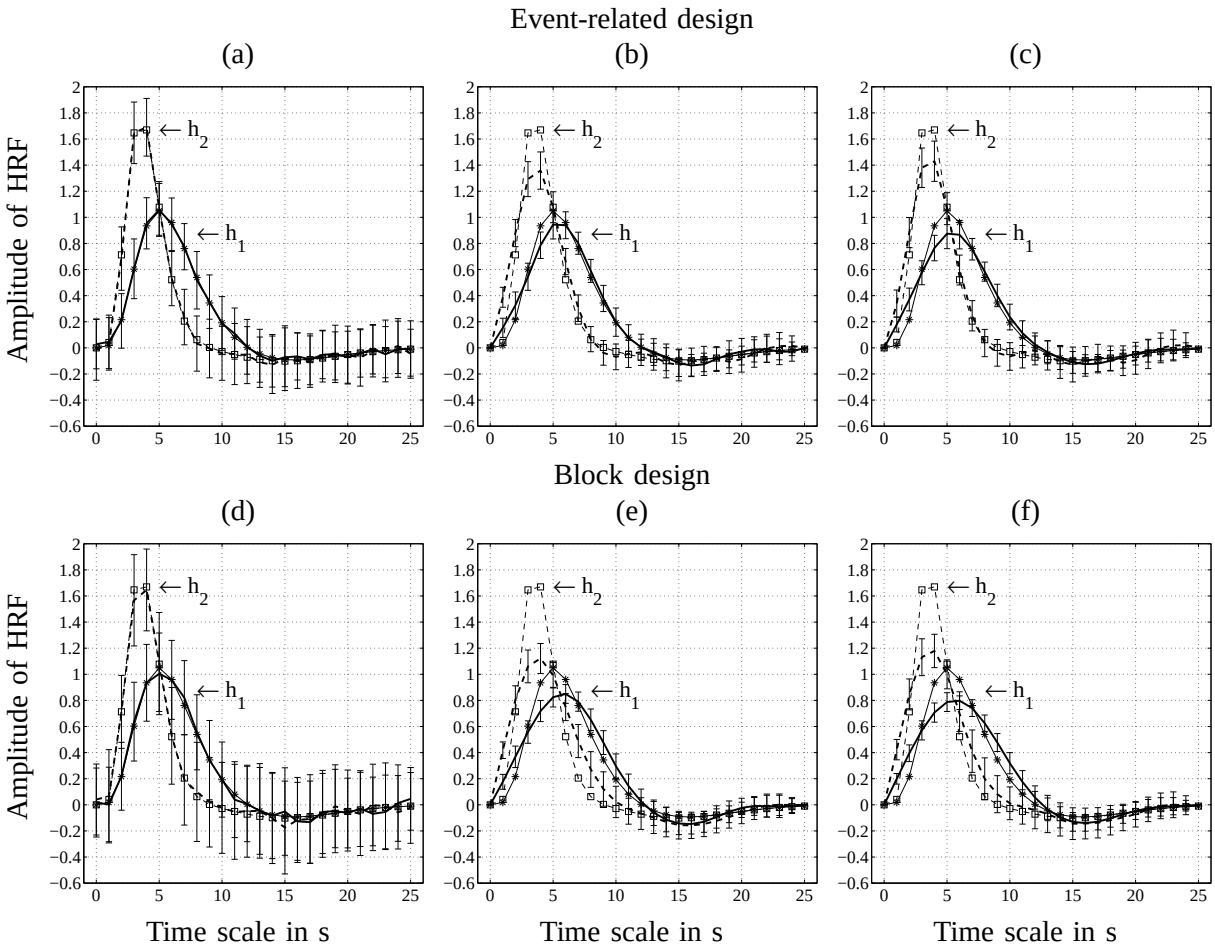

Fig. 1. Average HRF estimates over $J = 100$ drawings of the noise distribution in the case of synchronous paradigm ($\Delta t = \mathrm{TR} = 1$ s) and for a low CNR value. (a) & (d): Maximum-likelihood HRF estimates. (b)-(c) & (e)-(f): MAP HRF estimates computed with *constant* (b)-(e) and *adaptative* prior models (c)-(f). Solid and dashed lines represent $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$, respectively. Fine and thick lines code for true and HRF estimates, respectively. In addition, the time samples of $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ are marked with $\star$ and $\square$, respectively.

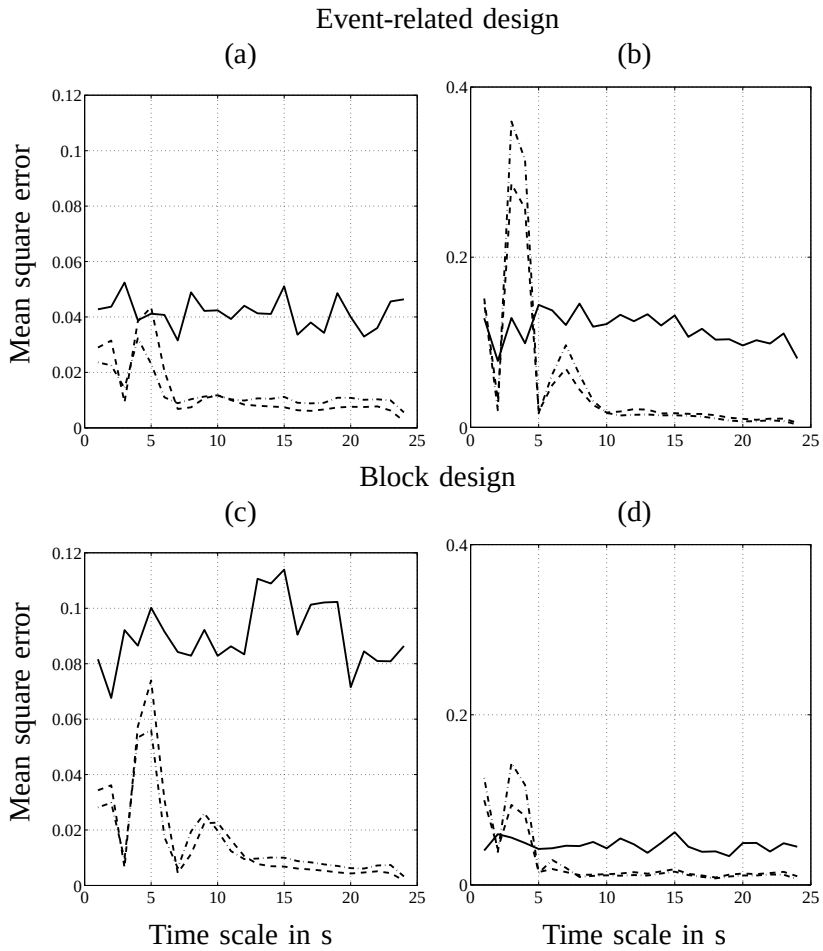

Fig. 2. Mean square error computed on $\boldsymbol{h}_1$ in (a) & (c) and on $\boldsymbol{h}_2$ in (b) & (d). Solid lines depict the MSE value for the ML estimates. Dash-dotted and dashed lines give the MSE values for the MAP estimates computed with constant and adaptative prior models, respectively.

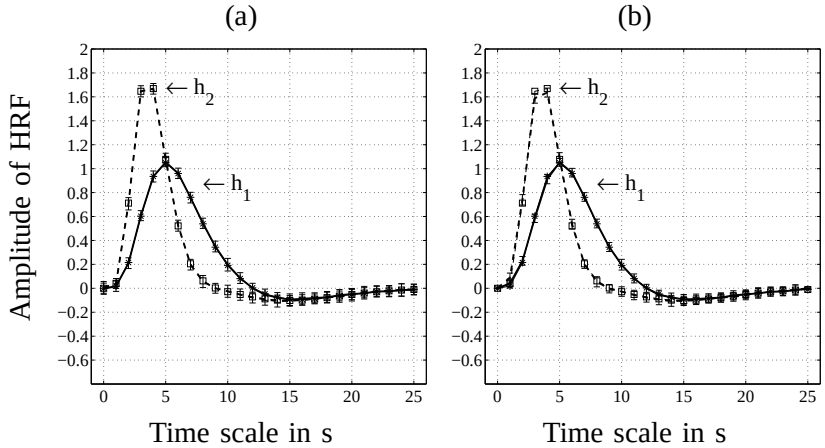

Fig. 3. Average HRF estimates over $J = 100$ drawings of the noise distribution in the case of synchronous paradigm (Δt = TR = 1 s) and for a high CNR case. (a): Maximum-likelihood HRF estimates. (b): MAP HRF estimates computed with a constant prior model. Solid and dashed lines represent $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$, respectively. Fine and thick lines code for true and HRF estimates, respectively but cannot be distinguished at this level of CNR. The time samples of $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ are marked with $\star$ and $\square$, respectively.

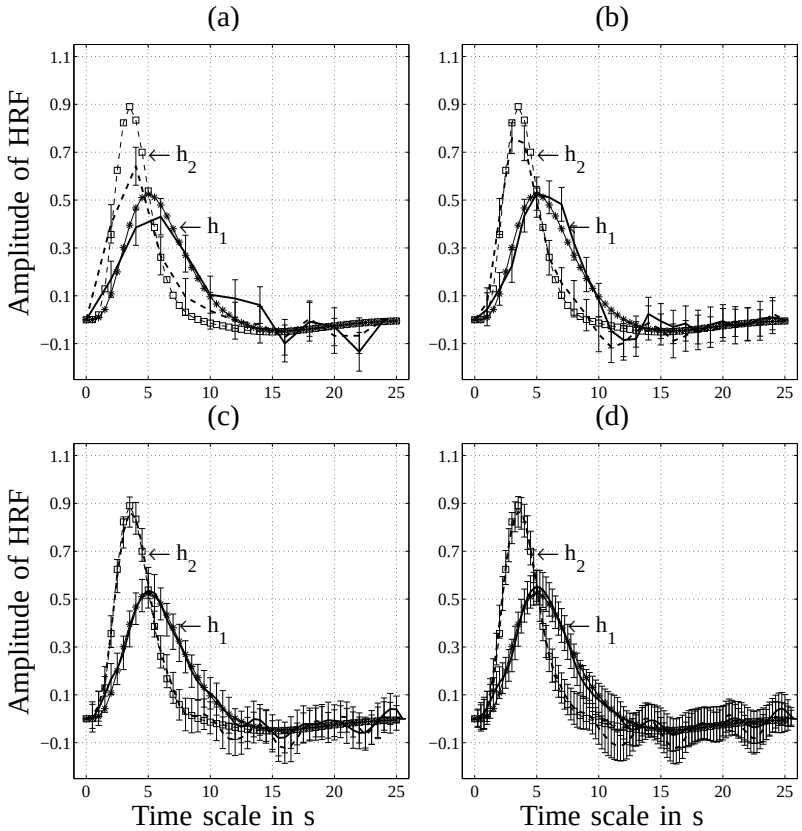

Fig. 4. HRF estimates for synchronous *vs* asynchronous estimation technique. Top row: HRFs computed using $\Delta t = \mathrm{TR}$ (a) and $\Delta t = \mathrm{TR}/2$ (b). Bottom row: HRFs estimated using $\Delta t = \mathrm{TR}/4$ (c) and $\Delta t = \mathrm{TR}/8$ (d). Solid and dashed lines represent $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$, respectively. Fine and thick lines code for true and HRF estimates, respectively. The time samples of $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ are marked with $\star$ and $\square$, respectively.

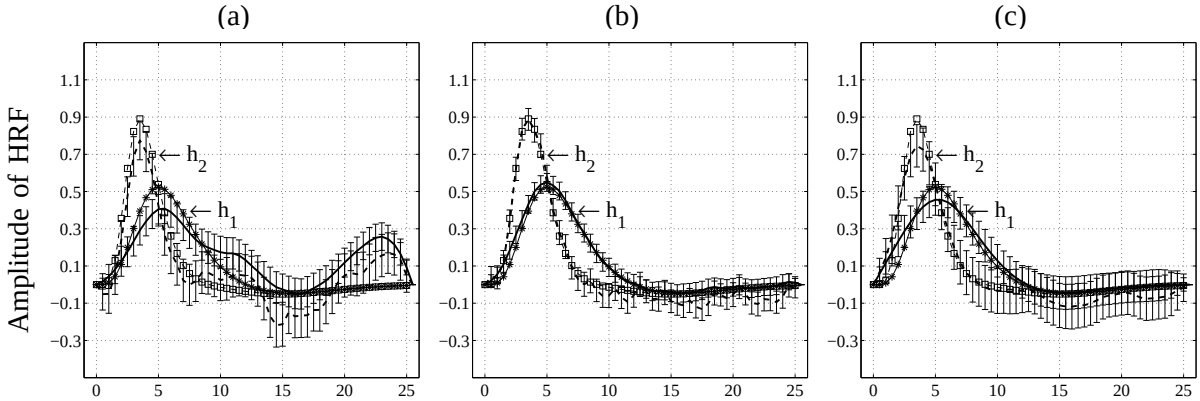

Fig. 5. (a) and (b): comparison of the HRF estimates computed from *one vs four* sessions, respectively when the low frequency drift included in the data is modeled with $Q_i = 4$ nuisance variables for each session. (b) and (c): comparison of drift modeled (b) and drift not modeled (c) for HRF estimates computed from four sessions. In all cases, we have considered the adaptative prior model ($\tau_1 \neq \tau_2$).

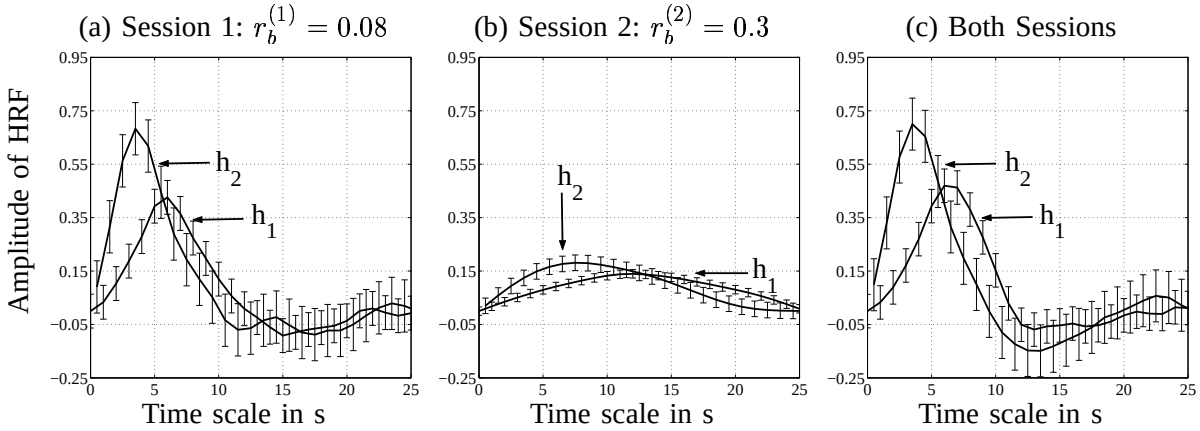

Fig. 6. Testing for the robustness to the equal noise variance hypothesis. (a)-(b): HRFs estimated from a single session where the Gaussian noise has for variance $r_b^{(1)} = 0.08$ and $r_b^{(2)} = 0.3$, respectively. In (c), the HRFs have been estimated from both sessions with model ($\mathcal{M}$), that is assuming $r_b^{(1)} = r_b^{(2)}$.

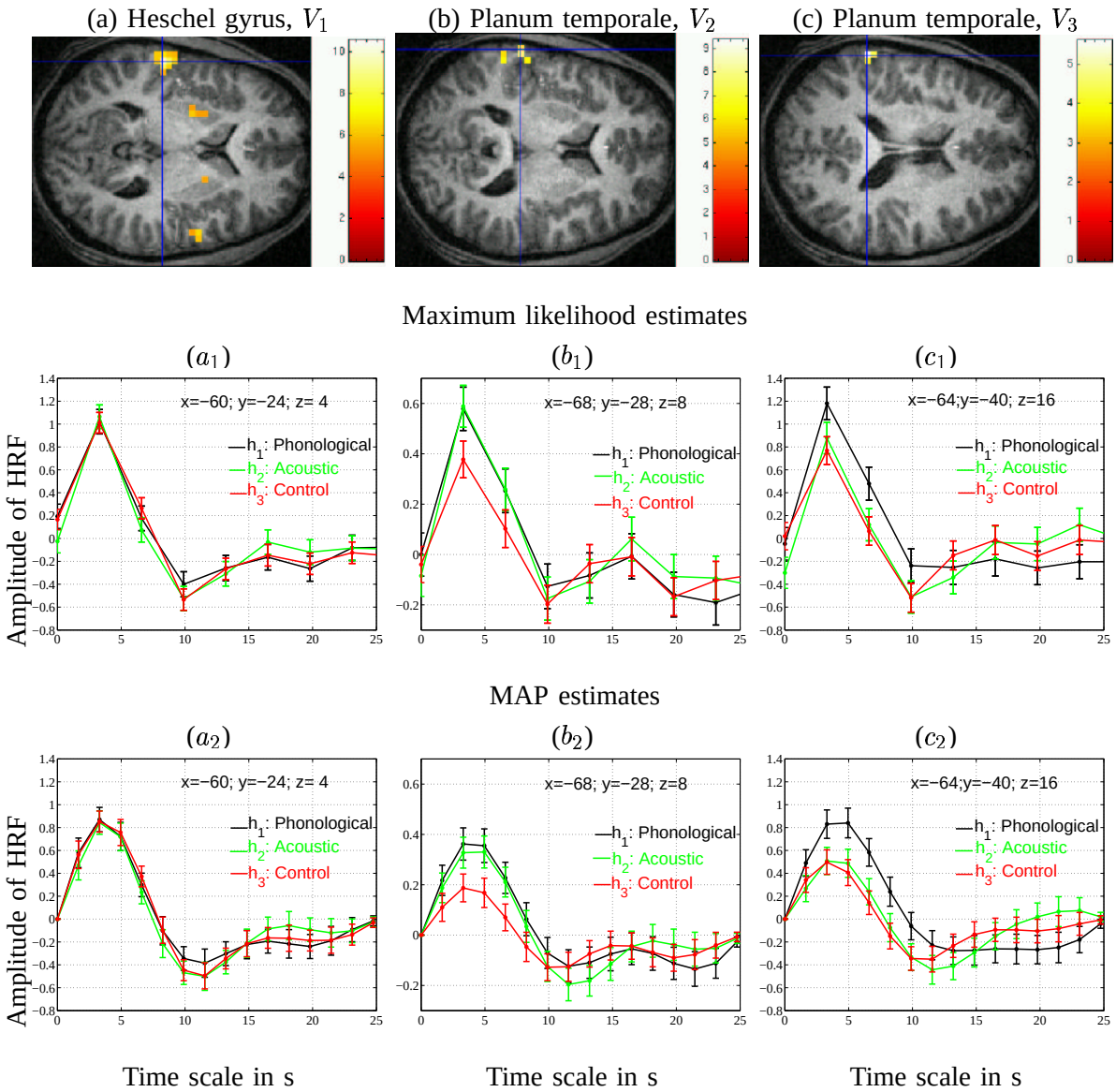

Fig. 7. Real data originating from a speech perception experiment. Top row: statistical t maps yielded by SPM 99 (thresholded at $P = 0.001$), superimposed on axial slices of averaged T1-weighted images. Middle row (a_1)-(c_1): Maximum likelihood HRF estimates computed from six sessions in voxels marked by blue crosses (V_1-V_3). Bottom row (a_2)-(c_2): MAP HRF estimates computed from six sessions in the same voxels. Voxel coordinates are indicated near the time courses. 'Phonological', 'Acoustic' and 'Control' conditions are coupled with $\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3$, respectively.

TABLE I

LIST OF NOTATIONS

$\boldsymbol{y}_i \in \mathbb{R}^{N_i}$	Data of session i for a given voxel
$\boldsymbol{b}_i \in \mathbb{R}^{N_i}$	Gaussian noise vector of session i
$K+1$	Number of HRF coefficients for each condition
$\boldsymbol{X}_i^{(m)} \in \mathbb{R}^{N_i} \times \mathbb{R}^{K+1}$	Binary "onsets matrix" of stimulus m for session i
$\boldsymbol{h}_m \in \mathbb{R}^{K+1}$	HRF for the mth condition
M	Number of different conditions in the experiment
$\boldsymbol{h} = [\boldsymbol{h}_1^{\mathrm{t}}, \ldots, \boldsymbol{h}_M^{\mathrm{t}}]^{\mathrm{t}} \in \mathbb{R}^{M(K+1)}$	Set of HRFs
$\boldsymbol{X}_i = [\boldsymbol{X}_i^{(1)} \vert \ldots \vert \boldsymbol{X}_i^{(M)}] \in \mathbb{R}^{N_i} \times \mathbb{R}^{M(K+1)}$	Binary "onsets matrix" of all conditions for session i
Q_i	Number of drift parameters for session i
$\boldsymbol{P}_i = [P_{i,1}, \ldots, P_{i,Q_i}] \in \mathbb{R}^{N_i} \times \mathbb{R}^{Q_i}$	Low frequency orthogonal matrix for session i
$\boldsymbol{l}_i \in \mathbb{R}^{Q_i}$	nuisance parameters for session i
I	Sessions number
$N = \sum_{i=1}^{I} N_i$	Global number of data for all sessions
$\mathbb{y} = \left[\boldsymbol{y}_1^{\mathrm{t}}, \ldots, \boldsymbol{y}_I^{\mathrm{t}}\right]^{\mathrm{t}} \in \mathbb{R}^N$	Complete set of data for all sessions
$\mathbb{X} = \left[\boldsymbol{X}_1^{\mathrm{t}} \mid \ldots \mid \boldsymbol{X}_I^{\mathrm{t}}\right]^{\mathrm{t}}$	Binary "onsets matrix" for all sessions
$\mathbb{P} = \mathrm{diag}\left[\boldsymbol{P}_1, \ldots, \boldsymbol{P}_I\right]$	Low frequency orthogonal matrix for all sessions
$\mathbb{l} = \left[\boldsymbol{l}_1^{\mathrm{t}}, \ldots, \boldsymbol{l}_I^{\mathrm{t}}\right]^{\mathrm{t}} \in \mathbb{R}^{\sum_i Q_i}$	nuisance parameters for all sessions
$\boldsymbol{\theta}_{\boldsymbol{H}} = [\tau_1, \ldots, \tau_M] \in \mathbb{R}^M$	Hyperparamters of the prior pdf $p(\boldsymbol{h}\ ;\ \boldsymbol{R}, \boldsymbol{\theta}_{\boldsymbol{H}})$
$\boldsymbol{\theta} = [r_b, \boldsymbol{\theta}_{\boldsymbol{H}}^{\mathrm{t}}]^{\mathrm{t}}$	Complete set of hyperparameters
$\boldsymbol{R}$	Prior covariance matrix of $\boldsymbol{h}_m$
$\boldsymbol{R}_{\boldsymbol{H}} = \mathrm{diag}\left[\tau_1 \boldsymbol{R}, \tau_2 \boldsymbol{R}, \ldots, \tau_M \boldsymbol{R}\right]$	Prior covariance matrix of $\boldsymbol{h}$
$\widehat{\boldsymbol{h}}^{\mathrm{MAP}}$	MAP HRFs estimate
$\boldsymbol{\Sigma}$	Posterior covariance matrix of $\widehat{\boldsymbol{h}}^{\mathrm{MAP}}$
$\boldsymbol{\Upsilon}$	Covariance matrix of the data $\mathbb{y}$
$\mathcal{E}(\widehat{\boldsymbol{h}}_m\ ;\ \boldsymbol{h}_m)$	Quadratic error between $\boldsymbol{h}_m$ and its estimate $\widehat{\boldsymbol{h}}_m$
$\mathcal{S}(\widehat{\boldsymbol{h}}_m)$	Mean standard deviation error measure for $\widehat{\boldsymbol{h}}_m$

TABLE II

ECM ALGORITHM FOR HYPERPARAMETERS AND NUISANCE VARIABLE ESTIMATION.

1) Initialize $\widetilde{\boldsymbol{\theta}}^0$.
2) Iteration k ($k \geqslant 1$)
   - Estimate $\boldsymbol{\theta}_{\mathbb{Y} \mid \boldsymbol{H}}^k = [\boldsymbol{l}_1^k, \ldots, \boldsymbol{l}_I^k, r_b^k]^{\mathrm{t}}$:
     - Compute $\boldsymbol{l}_i^k$ using (14) for $i \in \mathbb{N}_I^*$;
     - Compute r_b^k using (15);
   - Estimate $\boldsymbol{\theta}_{\boldsymbol{H}}^k = [\tau_1^k, \ldots, \tau_M^k]^{\mathrm{t}}$ using (16);
3) Compute $\mathcal{Q}(\widetilde{\boldsymbol{\theta}}^k, \widetilde{\boldsymbol{\theta}}^{k-1}\ ;\ \mathbb{y}, \mathcal{M})$ using (21)–(25);
4) Iterate $k = k+1$ and $\widetilde{\boldsymbol{\theta}}^k = [\boldsymbol{\theta}_{\mathbb{Y} \mid \boldsymbol{H}}^k, \boldsymbol{\theta}_{\boldsymbol{H}}^k]^{\mathrm{t}}$ until stop rules are satisfied;

TABLE III

SIMULATION PARAMETERS FOR FOUR DIFFERENT DATASETS.

Subsect.	Drift		r_b	CNR	# Sess.
	Flag	Cut-Off-Period (s)			
§ V-B	No		[0.7 0.028]	[0.3 1.53]	1
§ V-C	No		0.008	1.46	1
§ V-D	Yes	[180 170 160 180]	0.03	0.73	4
§ V-E	Yes	[180 170 160 180]	[0.03 0.2]	[0.73 0.29]	4
		[40 50 50 40]			
§ V-F	No		[0.08 0.3]	[0.46 0.24]	2
§ V-G	Yes	160	0.02	0.92	1

TABLE IV

GLOBAL MSE OF THE HRF ESTIMATES COMPUTED BY ML AND MAP METHODOLOGIES.

		$gMSE(\cdot\ ;\ \boldsymbol{h}_m) \times 100$			
		ER design		Block design	
		$\boldsymbol{h}_1$	$\boldsymbol{h}_2$	$\boldsymbol{h}_1$	$\boldsymbol{h}_2$
CNR=0.3	ML estimates	5.45	5.37	8.83	11.26
	MAP estimates ($\tau_1 = \tau_2$)	1.47	3.39	1.58	5.31
	MAP estimates ($\tau_1 \neq \tau_2$)	1.46	2.97	1.6	4.78
CNR=1.53	ML estimates	0.21	0.22		
	MAP estimates ($\tau_1 = \tau_2$)	0.15	0.19		
	MAP estimates ($\tau_1 \neq \tau_2$)	0.12	0.18		

TABLE V

QUADRATIC ERROR $\mathcal{E}$ AND MEAN STANDARD DEVIATION $\mathcal{S}$ OF THE MAP ESTIMATES COMPUTED FOR DIFFERENT VALUES OF Δt.

	$\mathcal{E}(\cdot\ ;\ \boldsymbol{h}_m)$		$\mathcal{S}(\widehat{\boldsymbol{h}}_m)$	
	$\boldsymbol{h}_1$	$\boldsymbol{h}_2$	$\boldsymbol{h}_1$	$\boldsymbol{h}_2$
$\Delta t = \mathrm{TR}$	0.02	0.021	0.078	0.077
$\Delta t = \mathrm{TR}/2$	$9.3\ 10^{-3}$	$9.2\ 10^{-3}$	0.07	0.067
$\Delta t = \mathrm{TR}/4$	$3.5\ 10^{-3}$	$\boxed{4.2\ 10^{-3}}$	0.069	$\boxed{0.065}$
$\Delta t = \mathrm{TR}/8$	$\boxed{3.1\ 10^{-3}}$	$5.3\ 10^{-3}$	$\boxed{0.068}$	0.065

TABLE VI

QUANTITATIVE ASSESSMENT OF THE HRFS ESTIMATED BY THE SUCCESSIVE MODELS OF SECTION II.

		Quadratic error $\mathcal{E}(\cdot\ ;\ \boldsymbol{h}_m)$		Dispersion $\mathcal{S}(\widehat{\boldsymbol{h}}_m)$	
		$\boldsymbol{h}_1$	$\boldsymbol{h}_2$	$\boldsymbol{h}_1$	$\boldsymbol{h}_2$
	Single Session & No drift & $\tau_1 = \tau_2$	0.017	0.018	0.13	0.13
	Single Session & No drift & $\tau_1 \neq \tau_2$	0.016	0.017	0.12	0.13
	Single Session & drift & $\tau_1 = \tau_2$	0.015	0.015	0.11	0.1
Fig.5(a)	Single Session & drift & $\tau_1 \neq \tau_2$	0.015	0.014	0.08	0.11
	multisession & No drift & $\tau_1 = \tau_2$	$5\ 10^{-3}$	0.01	0.1	0.1
Fig.5(c)	multisession & No drift & $\tau_1 \neq \tau_2$	$6\ 10^{-3}$	$8\ 10^{-3}$	0.08	0.1
	multisession & drift & $\tau_1 = \tau_2$	$5\ 10^{-3}$	$7\ 10^{-3}$	0.06	0.06
Fig.5(b)	multisession & drift & $\tau_1 \neq \tau_2$	$\boxed{4\ 10^{-3}}$	$\boxed{6,5\ 10^{-3}}$	$\boxed{0.04}$	$\boxed{0.05}$