DTU Library

Modeling the hemodynamic response in fMRI using smooth FIR filters

Goutte, Cyril; Nielsen, Finn Årup; Hansen, Lars Kai

Modeling the Haemodynamic Response in fMRI Using Smooth FIR Filters

Cyril Goutte*, Finn Årup Nielsen, and Lars Kai Hansen, *Member, IEEE*

Abstract—**Modeling the haemodynamic response in functional magnetic resonance (fMRI) experiments is an important aspect of the analysis of functional neuroimages. This has been done in the past using parametric response function, from a limited family. In this contribution, we adopt a semi-parametric approach based on finite impulse response (FIR) filters. In order to cope with the increase in the number of degrees of freedom, we introduce a Gaussian process prior on the filter parameters. We show how to carry on the analysis by incorporating prior knowledge on the filters, optimizing hyper-parameters using the evidence framework, or sampling using a Markov Chain Monte Carlo (MCMC) approach. We present a comparison of our model with standard haemodynamic response kernels on simulated data, and perform a full analysis of data acquired during an experiment involving visual stimulation.**

Index Terms—**Evidence, FIR filters, fMRI, haemodynamic response, Markov Chain Monte Carlo, neuroimaging, smoothness prior, Tikhonov regularization.**

I. INTRODUCTION

MODELING the haemodynamic response is important for several reasons. First, an appropriate modeling leads to better statistical maps. With the increased temporal resolution of functional magnetic resonance (fMRI) images [as compared to positron emission tomography (PET)], a binary baseline-activation description of the data is insufficient and it is necessary to take into account the temporal pattern of activation due to the haemodynamic response to the activation. A second reason is the possibility of performing simulations with the model. The predicted behavior obtained from simulation can be used to formulate more explicit hypotheses about the fMRI signal, and possibly optimize the design and acquisition [1]. A last reason is the possibility, for some models, to give a physiological interpretation of the model parameters and, thus, better understand the neurophysiology [2], [3].

The haemodynamic response is usually, as a first approximation, modeled as a convolution of the experimental paradigm by a linear filter, and implemented as a linear time-invariant (LTI) system. This assumption is usually justified by the observation of additivity in the fMRI signal [4], which is consistent with the linear hypothesis. Although several groups have since reported small to strong departures from normality in a number of contexts [5]–[8], it is still believed that the linearity assumption holds in a wide range of experimental conditions [8]. The LTI approach has been pursued using several types of parametric models of the filter, for example using a Poisson filter [9], a Gamma filter [4], [10], a Gaussian filter [11], or a simple delay [12]. In addition, a number of investigators have used linear filters to model the haemodynamic response, but as they *set* and do not *fit* the parameters, they are somewhat out of the scope of this study (see, e.g., [6], [13]). In these models, the few parameters have a specific interpretation, measuring, e.g., delay, strength of activation, etc.

In this contribution, we use a different standpoint, where we do not impose a specific shape on the linear filter coefficients. The haemodynamic response is modeled as an FIR function, a particular case of autoregressive with exogenous input (ARX) model. This approach has been pioneered by [14]. Though it is undoubtedly parametric in the sense that it fits a number of parameters, these do not really have a physical or physiological meaning. We will, therefore, refer to it as a semi-parametric *modeling* approach. This approach is much more flexible than the use of a parametric filter shape. In particular, it can model reliably the early decrease in signal (initial dip, see, e.g., [15], [16]) or the post-activation undershoot [17], whereas, e.g., the Poisson, Gamma or Gaussian filter are intrinsically unable to do so.

As the number of parameters increases, there is a risk that the model will overfit or that parameters become ill determined. We deal with this problem by placing a Gaussian Process prior on the filter coefficients, forcing the filter to be smooth. The resulting model is determined by the data and three hyper-parameters which can be set beforehand or again fitted on the data using a probabilistic argument.

In the following sections, we describe the basic theory for smooth FIR filters. We discuss a number of topics like boundary conditions and link to traditional Tikhonov regularization. We then go on to describe how to use a Bayesian argument to estimate the hyper-parameters, either using the evidence argument or by integrating over nuisance parameters using Markov Chain Monte Carlo (MCMC) methods.

We illustrate the workings of this filter using several experiments. We first show how the model can implement some of the LTI models currently in use in the fMRI literature. We show

*C. Goutte was with the Department of Mathematical Modeling, Technical University of Denmark, DK-2800 Lyngby, Denmark. He is now with INRIA Rhone-Alpes, Zirst Montbonnot - 655 avenue de l'Europe F-38334 Saint Ismier Cedex France (e-mail: cyril.goutte@inrialpes.fr).

F. Å. Nielsen and L. K. Hansen are with the Department of Mathematical Modeling, Building 321, Technical University of Denmark, DK-2800 Lyngby, Denmark.

that the smooth FIR filter is able to implement additional features that these classical models cannot, for example a post-activation undershoot. We then perform a full analysis of fMRI data acquired during a visual stimulation experiment. In particular, we show how to derive from the resulting filter measures of support (p-values) for the null hypothesis of no activation, and meaningful physiological information like the strength or delay in activation.

II. DATA

The dataset was acquired at Hvidovre Hospital on a 1.5-T Magnetom Vision MR scanner by Egill Rostrup. The scanning sequence was a 2-D gradient echo EPI (T2* weighted) with 66 ms echo time and 50 degrees RF flip angle. The images were acquired with a matrix of 128×128 pixels, with FOV of 230-mm, and 10-mm slice thickness, in a para-axial orientation parallel to the calcarine sulcus. The region of interest (ROI) will be limited to a 68×82 two-dimensional (2-D) voxel map. The voxel dimension is $1.8 \times 1.8 \times 10$ mm.

The visual paradigm consists of a rest period of 20 s of darkness using a light fixation dot, followed by 10 s of full-field checker board reversing at 8 Hz, and ending with 20 s of rest (darkness). In total, 150 images were acquired in 50 s, corresponding to a period of approximately $\mathrm{TR} = 330$ ms/image. The experiment was repeated in ten separate runs containing 150 images each. In order to reduce saturation effects, the first 29 images were discarded, leaving 121 images for each run.

The datasets studied in this article were acquired on the same subject, but during two separate scanning sessions (d3711 and d3991), such that, e.g., the position and the shape of the slice are slightly different. In each case, the dataset was built by combining the ten runs into a single sequence of 1210 images. However, as the runs were acquired separately, it should be noted that there cannot be any causality between the activation in one run and the signal measured in the next. Note also that due to the haemodynamic delay, the signal measured in activated voxels will be roughly centered within the remaining 40 s of each run. In the dataset we use in this article, the brain has first been masked, and the data was preprocessed using the run-based detrending described by [18].

III. THEORY OF SMOOTH FIR FILTERS

Let us consider a fMRI signal $y(t)$ acquired in a given voxel using a stimulus $x(t)$. The image index t runs between one and T. The finite input response (FIR) filter of order d models the fMRI signal using linear coefficients $\boldsymbol{w} = (w_1, w_2, \cdots w_d)^\top$

$$\hat{y}(t) = \sum_{i=1}^{d} w_i x(t-i) = \boldsymbol{w}^\top \boldsymbol{x}(t) \tag{1}$$

where $\boldsymbol{x}(t) = (x(t-1), x(t-2), \cdots x(t-d))^\top$ is a vector of past values of the stimulus.

Assuming independent additive zero mean Gaussian noise, the likelihood of the model parameters becomes

$$p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w}, \sigma^2)$$
$$= \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{w}^\top \boldsymbol{x}(t) - y(t))^2\right] \tag{2}$$

$$= (2\pi\sigma^2)^{-T/2} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{Y})^2\right] \tag{3}$$

where $\boldsymbol{X} = [\boldsymbol{x}(d+1), \boldsymbol{x}(d+2), \ldots, \boldsymbol{x}(T)]^\top$ is a matrix containing the (transposed) input vectors $\boldsymbol{x}(t)$ for all values[1] of t, and $\boldsymbol{Y} = [y(d+1), y(d+2), \ldots, y(T)]^\top$ is the vector of measurements, the target values for our filter. Maximizing the likelihood with respect to $\boldsymbol{w}$ leads to the well-known maximum-likelihood (ML) solution

$$\boldsymbol{w}_{\mathrm{ML}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} \tag{4}$$

When the ratio of the number of independent data to the filter order is small, the matrix $(\boldsymbol{X}^\top \boldsymbol{X})$ tends to be badly conditioned and the ML solution becomes unstable. It is necessary to regularize the solution. Alternatively, in a Bayesian context we will impose constraints on $\boldsymbol{w}$ by specifying a prior $p(w)$. We will focus on Gaussian priors, of the general form

$$p(\boldsymbol{w}|\boldsymbol{R}) = (2\pi)^{-d/2} \sqrt{|\boldsymbol{R}|} \exp\left[-\tfrac{1}{2}\boldsymbol{w}^\top \boldsymbol{R}\boldsymbol{w}\right] \tag{5}$$

where $|\cdot|$ indicates the determinant of a matrix. The posterior distribution of $\boldsymbol{w}$, conditioned on the data and the hyper-parameters becomes

$$p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{R}, \sigma^2)$$
$$\propto \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{w} - \boldsymbol{w}_{\mathrm{MAP}})^\top (\boldsymbol{X}^\top \boldsymbol{X} + \sigma^2 \boldsymbol{R})(\boldsymbol{w} - \boldsymbol{w}_{\mathrm{MAP}})\right] \tag{6}$$

which is largest for the maximum *a posteriori* parameters

$$\boldsymbol{w}_{\mathrm{MAP}} = (\boldsymbol{X}^\top \boldsymbol{X} + \sigma^2 \boldsymbol{R})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} \tag{7}$$

Note that this is also the ridge regression solution when $\boldsymbol{R}$ is a diagonal matrix with identical elements on the diagonal.

The matrix $\boldsymbol{R}$ implements the constraints that we impose on the model. Here, we want to obtain *smooth* filters, i.e. filters such as neighboring parameters (e.g., w_4 and w_5) have similar value. This corresponds to saying that neighboring filter parameters should be somehow correlated. Accordingly, $\boldsymbol{R}$ will be the inverse of a covariance matrix where the covariance is a decreasing function of the distance between two parameters

$$\boldsymbol{R} = \boldsymbol{\Sigma}^{-1} \qquad \text{with } \Sigma_{ij} = \nu \exp\left(-\frac{h}{2}(i-j)^2\right) \tag{8}$$

In (8), the covariance decreases as a Gaussian parameterized by ν and h, but any nonnegative decreasing function of the distance $|i - j|$ could be used. This corresponds to putting a Gaussian process prior on the filter parameters themselves, rather than on the predictions [19], [20].

With this expression, the MAP estimate of $\boldsymbol{w}$ becomes $\boldsymbol{w}_{\mathrm{MAP}} = (\boldsymbol{X}^\top \boldsymbol{X} + \sigma^2 \boldsymbol{\Sigma}^{-1})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$ which can be efficiently calculated as $\boldsymbol{w}_{\mathrm{MAP}} = \boldsymbol{\Sigma}(\boldsymbol{\Sigma}\boldsymbol{X}^\top \boldsymbol{X} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$, avoiding the additional inversion of $\boldsymbol{\Sigma}$. The resulting estimate $\boldsymbol{w}_{\mathrm{MAP}}$ depends on three hyper-parameters: the noise level σ^2, the strength of the prior ν and the smoothness factor h. We will see below how it is possible to estimate the values of these parameters using a Bayesian argument.

---

[1] Values of $x(t)$, $t < 1$ can be treated in several ways. For a block design involving baseline-activation-baseline patterns, they will naturally take the value of the baseline. Alternatively, all $x(t)$, $t < 1$ can be treated as nuisance parameters and integrated out of the model.
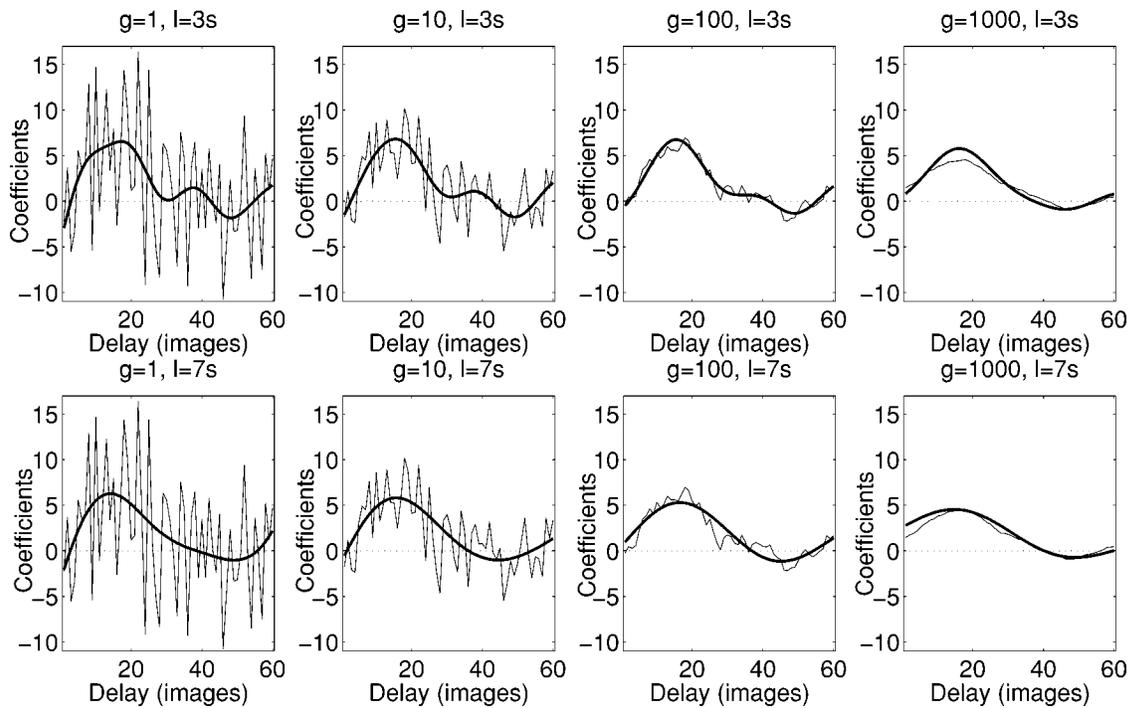
Fig. 1.   Resulting filter using ridge regression (thin line) and the smooth FIR filter approach (thick line). Top: characteristic length $\ell = 5$ s; bottom: $\ell = 7.5$ s. From left to right, increasing levels of regularization $g = \sigma^2/\nu$. Data: one voxel from the visual cortex (voxel 429) displaying a large activation.

The hyper-parameter h controls the smoothness of the resulting filter. For large values of h, Σ_{ij} will go to zero very fast for increasing values of $|i - j|$, such that there will be very little correlation between parameters: the filter will be very unsmooth. For $h \to +\infty$, we recover the ridge regression solution. For small values of h, Σ_{ij} will stay close to one for all (i, j), indicating perfect correlation between the filter parameters. The filter will be over-smooth. In the limit $h \to 0$ all parameters are identical and the filter performs a local averaging of the stimulus $x(t)$. It is useful to think of h as corresponding to a "characteristic length" of the filter, i.e., the typical length in which the filter varies. The characteristic length ℓ can be defined here as $\ell = 1/\sqrt{h}$. This is quite useful in fMRI modeling because it is widely believed on the basis of empirical studies [21], [3] that the haemodynamic response has a characteristic length on the scale of seconds, typically between 5 and 10 s. In first approximation it is then possible to use this prior information such that ℓ corresponds to, e.g., 7 s. For an fMRI experiment where $TR = 1/3$ s, this corresponds to $\ell = 21$ filter parameters.

Fig. 1 displays a comparison of smooth filters with the result of ridge regression. It is quite clear that ridge regression does not yield smooth filters, and the fluctuation in the parameters is important. It is possible to reduce this fluctuation by increasing the amount of regularization (from left to right on Fig. 1), but this also reduces the amplitude of the filter. This fluctuation means that the filter contains high frequency components, which potentially have a strong influence on summary statistics like the maximum parameter or the delay, even though they seem to contribute little to the modeling on average. The influence of the smoothness factor is clear from the comparison between the top and bottom row in the figure. The smooth filters in the bottom
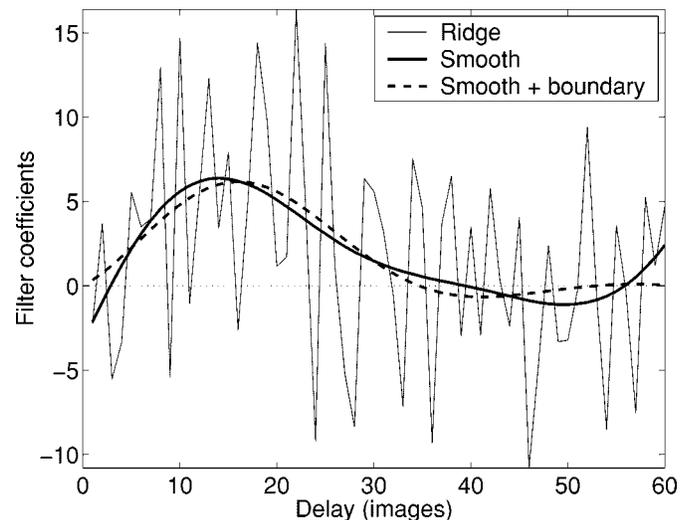
Fig. 2.   Comparison of the filters obtained by ridge regression, the smooth FIR filter without boundary condition (thick solid) and the FIR filter with boundary conditions (thick dashed). Notice how the endpoint goes (smoothly) to zero. Data: one voxel from the visual cortex (voxel 429) displaying a large activation.

row are clearly smoother, as expected from the larger characteristic length (the filters obtained from ridge regression are obviously identical).

A. Boundary Conditions

It is apparent from Fig. 1 that the first and (especially) the last filter parameters can take clearly positive or negative values. This cannot be avoided using the above equations. However, by causality, all w_i for $i \leq 0$ should be zero, as the influence of the stimulus at time t, $x(t)$ should be felt only for $t > 0$. This

corresponds to saying that a hypothetical filter parameter w_0 should be equal to zero. According to our prior, this will have a decreasing influence on w_1, w_2, etc. which will be forced (by smoothness) to be close to zero. Similarly, it is sensible that the influence of an activation should vanish in the past, hence, vanishing filter parameters for large delays. This can be again implemented by forcing an additional parameter w_{d+1} to be zero.

In practice, we are still interested only in estimating the values of filter parameters w_1, w_2, ..., w_d. This is done by defining a $(d+2) \times (d+2)$ matrix $\tilde{\Sigma}$, such that $\tilde{\Sigma}_{ij} = \nu \exp t(-h(i - j)^2)$, for $(i, j) \in \{0, 1, \ldots, d+1\}^2$. The matrix $\boldsymbol{R}$ is then defined as the central $d \times d$ part of $\tilde{\Sigma}^{-1}$, i.e., taking away the first and last rows and columns. This operation can be easily defined mathematically by introducing the $(d+2) \times d$ matrix $\boldsymbol{J}_{d+2,d}$, constructed as the superposition of a row of zeros on top, a unit $d \times d$ matrix in the central d rows and a row of zeros at the bottom. We then have $\boldsymbol{R} = \boldsymbol{J}_{d+2,d}^\top \tilde{\Sigma}^{-1} \boldsymbol{J}_{d+2,d}$.

Note that we cannot use the same trick as above to avoid inverting the parameter covariance matrix. However, $\tilde{\Sigma}$ is a band-diagonal (Toeplitz) matrix, such as efficient methods exist to perform an inversion in quadratic time instead of cubic for general matrices. Furthermore, d is usually quite small such that inversion of a $d \times d$ matrix is quite fast.

Fig. 2 shows the effect of the boundary conditions. The hyper-parameters are set to the same values as the left-most bottom plot on Fig. 1. The smooth FIR filter obtained above had clearly negative values for the first delays as well as for the longer delays (around 75). This effect disappears when boundary conditions are used. In particular, for large delays, the coefficients go smoothly to zero, as expected.

B. Link to Tikhonov Regularization

Regularization is often performed using Tikhonov regularization, which imposes a constraint on derivatives of the target function. In the context of this work, this would correspond to imposing smoothness by constraining the derivatives of the filter. The regularized solution is then obtained by minimizing the penalized cost

$$C(\boldsymbol{w}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w})^2 + \gamma \sum_i \left(\frac{\partial^n w_i}{\partial i^n} \right)^2 \qquad (9)$$

where n is the order of the derivatives used for smoothing.

Of course the true derivatives are unknown, such that we typically use instead the central differences approximation, where, e.g., the first derivative (gradient) is approximated by the difference between neighboring filter coefficients: $\partial \boldsymbol{w} / \partial i \approx (w_{i+1} - w_i)$. The regularized cost can then be formulated as

$$C(\boldsymbol{w}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w})^2 + \boldsymbol{w}^\top \boldsymbol{R}\boldsymbol{w} \qquad (10)$$

Note that we have kept the notation $\boldsymbol{R}$ because equation (10) can actually be obtained (up to an additive and a multiplicative constants) as the negative logarithm of the product of the likelihood (2) and the prior (5), i.e., as the log-posterior, and $\boldsymbol{R}$ plays the same role as the prior covariance. The expression of $\boldsymbol{R}$ depends
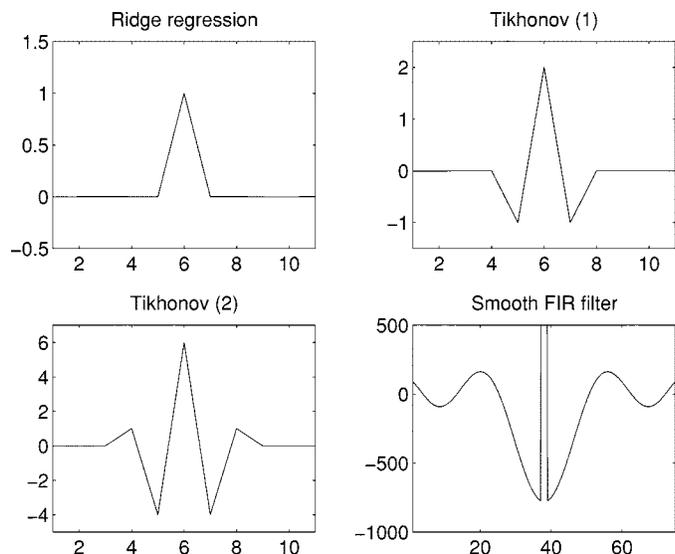

Fig. 3. Comparison of the neighboring influences for Ridge regression, Tikhonov regularization on the gradient ["Tikhonov (1)"] and the curvature ["Tikhonov (2)"] and the smooth filter approach.

on the derivative used. For $n = 1$ (gradient) and $n = 2$ (curvature), we have $\boldsymbol{R} = \gamma \boldsymbol{R}_n$, with

$$\boldsymbol{R}_1 = \begin{bmatrix} 1 & -1 & 0 & \cdots & & 0 \\ -1 & 2 & -1 & & & 0 \\ 0 & -1 & 2 & & & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix}$$

and

$$\boldsymbol{R}_2 = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & & 0 \\ -2 & 5 & -4 & 2 & & & 0 \\ 1 & -4 & 6 & -4 & & & 0 \\ \vdots & & & & \ddots & & \vdots \\ 0 & & 2 & -4 & 5 & -2 \\ 0 & \cdots & 0 & 1 & -2 & 1 \end{bmatrix} \qquad (11)$$

Tikhonov regularization is, thus, implemented by using a band-diagonal regularization matrix $\boldsymbol{R}$. However, whereas for smooth FIR filters $\boldsymbol{R} = \Sigma^{-1}$ has nonzero elements on (almost) all diagonals, the number of diagonals used by Tikhonov regularization depends on the order of the derivative and the approximation used. This means that whereas for smooth FIR filters the effect of one given parameter is far reaching, it is really limited for Tikhonov regularization (2 neighbors for $n = 2$ here).

By plotting one row (or one column) of the regularization matrix, one can picture the influence of the values of neighboring parameters on the solution. This is done on Fig. 3 for ridge regression, Tikhonov regularization on the gradient and curvatures, and the smooth FIR filter. The influence of the smooth filter can be seen here as a generalization of the Tikhonov approach of regularizing on the (approximate) amplitude of higher order derivatives.

C. Error Bars

The posterior distribution of $\boldsymbol{w}$ makes it possible to estimate the uncertainty of $\boldsymbol{w}_{\mathrm{MAP}}$. Note however that (6) is a multivariate
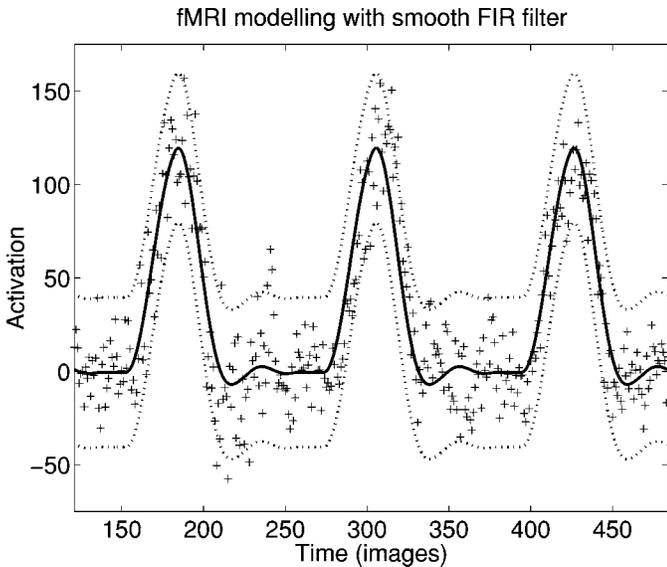
Fig. 4. fMRI time series $(+)$ measured in one strongly activated voxel (voxel 429), with the activation estimated by the smooth FIR filter (solid) and the prediction error bars on this activation, obtained from (12) (dotted). The error bars appear slightly over estimated because the first run has a larger amplitude than the last nine runs, thus inflating the apparent noise level.

Gaussian distribution with a general covariance matrix, such as the individual components of $\boldsymbol{w}$ are correlated. In that context it is not easy to represent graphically the uncertainty. Conditional error bars obtained from $p(w_i|\boldsymbol{w}_{-i}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{R}, \sigma^2)$, where $\boldsymbol{w}_{-i}$ contains all filter parameters except w_i, give a good idea of how close the filter parameters should be from each other, but greatly underestimate the possible range of variation of $\boldsymbol{w}$. This range is well estimated by the marginally error bars obtained from

$$p(w_i|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{R}, \sigma^2) = \int_{\boldsymbol{w}_{-i}} p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{R}, \sigma^2)\,d\boldsymbol{w}_{-i},$$

but these error bars overlook the fact that filter parameters are very correlated to each other, such that it is impossible for example that w_4 lies at the top of its marginal error bar while w_5 lies at the bottom.

The conditional error bars are easily obtained from the covariance matrix of the posterior: $\sigma_C^2 = \sigma^2/a_{kk}$ where a_{kk} is the kth diagonal element of $(\boldsymbol{X}^\top \boldsymbol{X} + \sigma^2 \boldsymbol{R})^{-1}$.

It is especially interesting to find error bars on the resulting predictions of the MAP filter, i.e., the estimated (de-noised) response pattern. Using the Gaussian noise assumption, the posterior for the prediction $\hat{y}$ associated with an input $\boldsymbol{x}_0$ is also Gaussian

$$p(\hat{y}|\boldsymbol{X}, \boldsymbol{x}_0, \boldsymbol{Y}, \boldsymbol{R}, \sigma^2)$$
$$= \mathcal{N}\left(\boldsymbol{x}_0^\top \boldsymbol{w}_{\text{MAP}}, \sigma^2 + \boldsymbol{x}_0^\top \left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} + \boldsymbol{R}\right)^{-1} \boldsymbol{x}_0\right) \quad (12)$$

This is illustrated for a particular time series (measured in the visual cortex) in Fig. 4. We have plotted three runs out of 10, and we see clearly that the data fits well inside the error bars. These actually seem slightly overestimated, for two main reasons. First the Gaussian assumption might be violated, though there is little evidence on this data of outliers. Second, the impression is actu-

ally due to the fact that we represent only three of ten runs on the figure. Over the 10 runs, 79 measurements exceed the interval given by the estimate plus or minus 1.96 standard deviations. This should be compared to an expected 5% of 1210, or 61.

The time series modeled by the smooth FIR filter shows a clear post-activation undershoot, followed by an overshoot of similar amplitude. It should be noted that this might be a pre-processing artefact. Note also that (on this data at least) it is not possible to observe an "initial dip."

D. Significance of Activation

In the context of functional neuroimaging, it is not sufficient to estimate the haemodynamic response in each location of the brain. One has to use the estimated model in the purpose of finding regions that are activated by a given stimulus sequence. This is traditionally done by testing the null hypothesis of no activation using various statistical tests. In the context of this study, the null hypothesis takes the form of $\{H_0: \boldsymbol{w} = 0\}$, i.e., the filter parameters are identically equal to zero. The alternative hypothesis is $\{H_1: \boldsymbol{w} \neq 0\}$. In a Bayesian context, this problem is fundamentally ill posed. The posterior probability for each hypothesis $\Pr(H_0|\mathcal{D})$ and $\Pr(H_1|\mathcal{D})$ can, of course, be estimated, but as H_0 corresponds to a single point in parameter $(\boldsymbol{w})$ space, the associated volume is zero, yielding $\Pr(H_0|\mathcal{D}) = 0$ and rejection of the null-hypothesis in favor of the alternative H_1.

In a Bayesian context, the comparison of a point hypothesis with an interval hypothesis will, thus, usually lead to the adoption of the latter. In order to derive a measure of support for our point-null hypothesis in a Bayesian context, we will use the concept of highest posterior density (HPD), described, e.g., by [22] and used in a functional neuroimaging context by [12]. Given a posterior density function $p(\boldsymbol{w}|\mathcal{D})$, the HPD region of content $(1-\alpha)$ is the region R of parameter space such that [22, section 2.8]

1) $\Pr(\boldsymbol{w} \in R|\mathcal{D}) = 1 - \alpha$;
2) $\forall \boldsymbol{w}_1 \in R, \boldsymbol{w}_2 \notin R, P(\boldsymbol{w}_1|\mathcal{D}) > P(\boldsymbol{w}_2|\mathcal{D})$.

For a given significance level α, we can test whether H_0 lies within the HPD region of content R. If so, the null hypothesis would be accepted at level α. Otherwise H_0 would be rejected and the voxel declared activated. Alternatively, we can use the HPD as a measure of support by calculating the volume of the region $R_0 = \{\boldsymbol{w}, p(\boldsymbol{w}|\mathcal{D}) < p(\boldsymbol{w} = 0|\mathcal{D})\}$. This is the region in parameter space that lies outside the equiprobability curve going through $\boldsymbol{w} = 0$. Clearly, if H_0 is to be accepted, this volume will be larger than α, as the HPD contains $\boldsymbol{w} = 0$. This region is large when zero is close to the MAP (hence, should not be rejected), and small when zero is far of the MAP (and should be rejected). It, thus, seems justified to use the volume of R_0 as a measure of support for the null hypothesis.

It is important to note that the use of the HPD to construct a measure of support for point hypotheses is not exempt of some of the logical flaws of other traditional measures of support like p-values or Bayes factors [23], [24]. In particular, it potentially suffers from inconsistency in some pathological cases. However, it has been noted that in a number of standard situations, it yields results that are similar to classical statistical tests [22].

In our case, the posterior density of the filter parameter is Gaussian, such that it is possible to get an efficient closed-form solution for the measure of support of H_0 in each voxel. With the notation $r^2 = \boldsymbol{w}_{\mathrm{MAP}}(\boldsymbol{X}^\top \boldsymbol{X}/\sigma^2 + \boldsymbol{R})\boldsymbol{w}_{\mathrm{MAP}}$, we have

$$
\begin{aligned}
p_{\mathrm{HPD}} &= \frac{1}{\Gamma(d/2)} \int_0^{r^2/2} u^{(d/2)-1} e^{-u}\, du \\
&= \mathrm{gammainc}\left(\frac{\boldsymbol{w}_{\mathrm{MAP}}(\boldsymbol{X}^\top \boldsymbol{X}/\sigma^2 + \boldsymbol{R})\boldsymbol{w}_{\mathrm{MAP}}}{2}, d/2 \right)
\end{aligned}
\tag{13}
$$

where gammainc is the two-parameter incomplete Gamma function, and we have adopted the notation p_{HPD} for the support for H_0 by similarity with traditional p-values.

E. Estimation of the Delay

Whereas standard parametric models of the haemodynamic response have one or several parameters representing the delay (parameter of the Poisson, mean of the Gaussian, ratio of the Gamma parameters), the FIR filter does not model this directly. It is necessary to estimate the delay from the many filter parameters. One approach is to use the group delay described, e.g., by Oppenheim and Schafer [25]

$$
\tau_{\mathrm{FIR}} = \frac{\sum_i w_i \times i}{\sum_i w_i}
\tag{14}
$$

i.e., the average of the delay in each filter parameter, weighted by the parameter values. In some situations, this estimate will be unreliable. This is the case for example when the denominator of (14) is small or when the filter has high frequency components. Note that by construction, the smooth FIR filter contain only low frequency components, such that the latter does not occur. Furthermore, the denominator will take small values when the mean filter coefficient is close to zero, indicating a nonactivated voxel. Overall, the estimation of the group delay in activated regions will give a reliable idea of the delay implemented by the FIR filter.

A second interesting measure is the delay necessary to reach 90% percent of activation after onset of the stimulus, or to return within 10% of maximum activation after offset of the stimulus. This delay has been reported to be between 5 and 8 s [21]. Note that linear filters implement symmetrical responses, such that the shape of the activation (and, thus, delay) after stimulus onset is identical to the deactivation after stimulus offset. For block design involving binary baseline-activation stimulus, the delay is easily calculated from the cumulative sums of the filter parameters.

F. Tuning of Hyper-Parameters

For given values of σ^2, ν and h, we have been able to give the expression of the posterior (6) and derive the MAP estimate of the filter parameters and some error bars. We will now see how we can find proper values for these parameters, using again a probabilistic approach. In a fully Bayesian approach, we would integrate over "nuisance" parameters to obtain the posterior distribution of interest. If we are interested in $\boldsymbol{w}$, for example, we

would integrate over the three hyper-parameters, after endowing them with suitable priors (i.e., reflecting prior knowledge or lack thereof). In the context of this study, it is impractical to carry out the marginalization analytically. Classical MCMC techniques [26] are able to perform numerical integration, but these techniques are computationally intensive and not practical for a full brain (or even a full slice) analysis. We give a quick guideline and an example of application in the appendix to this article.

In the following, we will use an intermediate approach, and select the hyper-parameters according to their likelihood $p(\boldsymbol{Y}|\boldsymbol{X}, \sigma^2, \nu, h)$. Note that using uniform priors on the hyper-parameters,[2] the posterior distribution of the hyper-parameters is proportional to the likelihood, $p(\sigma^2, \nu, h|\boldsymbol{X}, \boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{X}, \sigma^2, \nu, h)$. The hyper-parameters that we wish to optimize will, thus, be chosen so as to maximize the likelihood, also known as the *evidence*. This is obtained by integrating over the distribution of the weights

$$
p(\boldsymbol{Y}|\boldsymbol{X}, \sigma^2, \nu, h) = \int_{\boldsymbol{w}} p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w}, \sigma^2) p(\boldsymbol{w}|\nu, h)\, d\boldsymbol{w}
\tag{15}
$$

As the product of the two terms inside the integral has a Gaussian form, integration can be performed analytically, leading to

$$
\begin{aligned}
p(\boldsymbol{Y}|\boldsymbol{X}, \sigma^2, \nu, h) = (2\pi\sigma^2)^{-(T/2)} \sqrt{\frac{\sigma^{2d}|\boldsymbol{R}|}{|\boldsymbol{X}^\top \boldsymbol{X} + \sigma^2 \boldsymbol{R}|}} \\
\cdot \exp\left[-\frac{\boldsymbol{Y}^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w}_{\mathrm{MAP}})}{2\sigma^2} \right].
\end{aligned}
\tag{16}
$$

The evidence (16) can be optimized over several hyper-parameters using standard nonlinear optimization techniques [27]. As this can still be computationally intensive, we can use an approximation from the so-called "evidence framework" [28], [29, sec. 10.4], which provides a re-estimation formula for the noise level σ^2 and the prior strength ν. In that framework, these hyper-parameters can be estimated iteratively. For example, given a noise level σ^2_{old}, we estimate the filter $\boldsymbol{w}_{\mathrm{MAP}}$, and the noise level σ^2_{new} is updated according to the resulting filter fit [29, sec. 10.4].

The model size d can also be thought of as playing the role of a hyper-parameter. A sensible choice is to take d sufficiently large such that the corresponding filter contains the entire hemodynamic response. In this paper we have chosen to take $d = 60$, corresponding to 20 s. Additional experiments using $d = 75$ (corresponding to 25 s) showed little difference in the results. As a comparison, the length of the filter in SPM99 (from the file SPM_HRF.M) is 32 s.

IV. EXPERIMENTS

A. Can Smooth FIR Filter Estimate Standard Kernel Shapes?

In a first experiment, we look at the ability of the smooth filter to recover the shape of traditional linear filters: the Poisson filter proposed by [9], the Gamma filter of [10] and the Gaussian filter [11].

We use the same sequence of 1210 images with ten runs of 31 baseline, 30 activation and 60 baseline. Hence, the paradigm

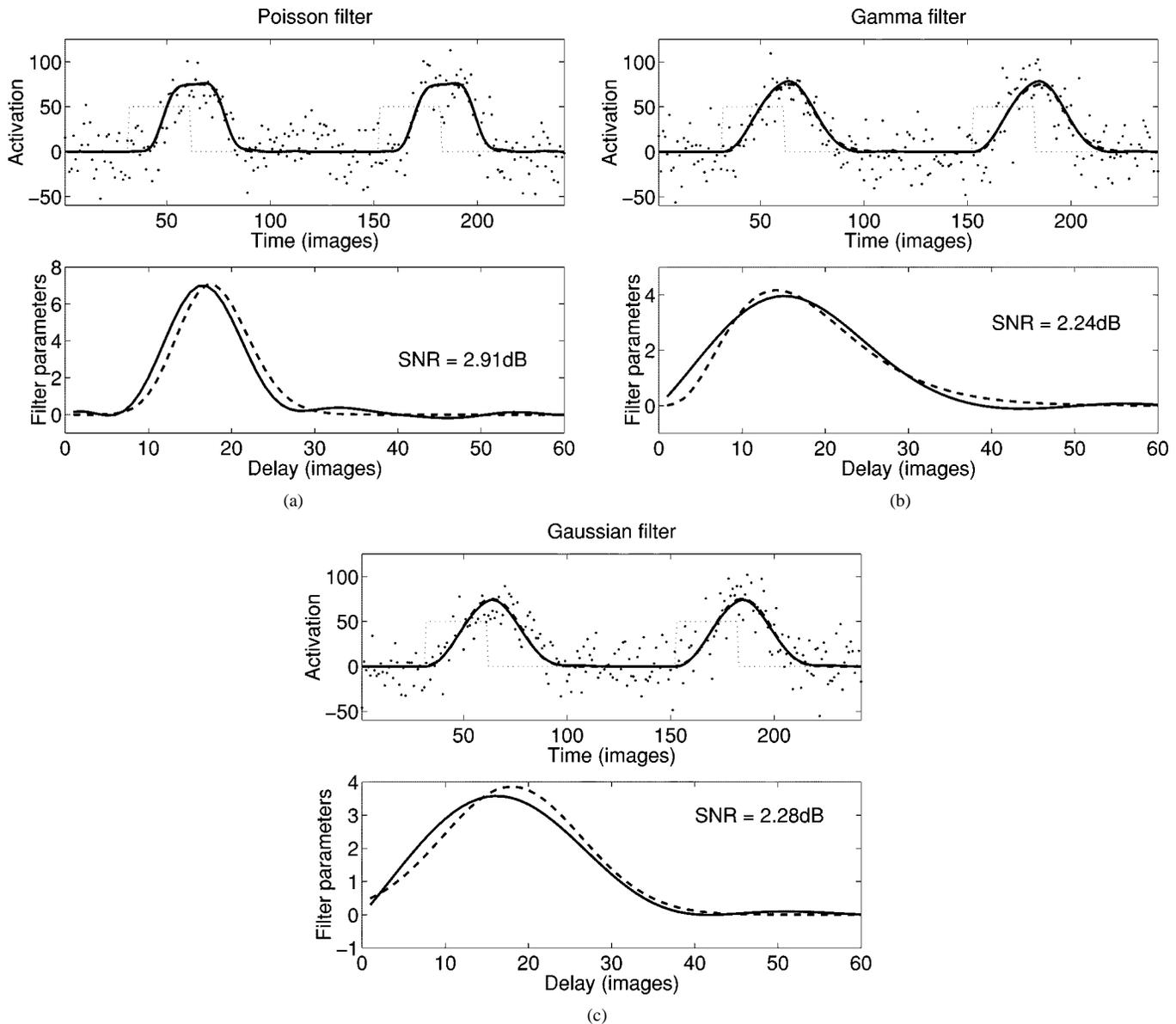[2]Technically, the priors are uniform on the log-domain

Fig. 5. Smooth FIR filter obtained on data generated by convolution of a square wave by a fixed-shape kernel. (a) Poisson, (b) Gamma, and (c) Gaussian filters. Top: paradigm (dotted), signal (dashed), noisy data (dots) and modeled signal (solid); Bottom: generating filter (dashed) and estimated smooth FIR filter (solid). Notice that the target and modeled signals (top row) are almost indistinguishable.

is a vector with 1210 elements and consists of a series of square waves. For all three filters, the mean is taken to be 18 images or 6 s, while the variance of the Gamma and Gaussian filters are set equal to 70. The variance of the Poisson filter is by construction equal to the mean, i.e., 18. All filter parameters where scaled such that the amplitude of the signal was roughly the same as what we observed in activated voxels in the actual experiment. Additive white noise of variance $\sigma^2 = 400$ was added to the convolved signal, giving a signal-to-noise ratio between 2.24 and 2.91 dB (i.e., the variance of the signal is only around 40% larger than that of the noise), cf. Fig. 5.

Fig. 5 shows the results obtained for the three filters. The resulting smooth FIR filter was better estimated in the case of the Gamma and Gaussian generating filter. The Poisson filter is more difficult to estimate because the "length scale" of the filter varies widely depending on the delay. There are steep changes in the filter coefficients around the maximum, while

the second half of the filter is virtually flat. So in a way this is the least smooth of the three filters. Note that the estimation of the Gamma filter is also slightly impaired for the same reason: the variation in filter coefficients is faster before the mode, and smoother afterwards (middle plot). Note however that when taking into account the rather large amount of noise in the data, the fit is quite satisfactory in all cases. Furthermore, the misfit in the filter parameters does not prevent the smooth FIR filter to model the data almost exactly (cf. top row in Fig. 5).

This simulation indicates that for the three basic filter shapes, the smooth FIR filter is able to recover the filter shape efficiently. In particular, in all three cases studied here, the recovered FIR filter showed little or no post-activation undershoot, in accordance with the strictly positive target filters. This suggests that any difference between the smooth FIR filter and these classical filters observed on real data is not due to the inability of the FIR filter to reconstruct the true fMRI response, but rather
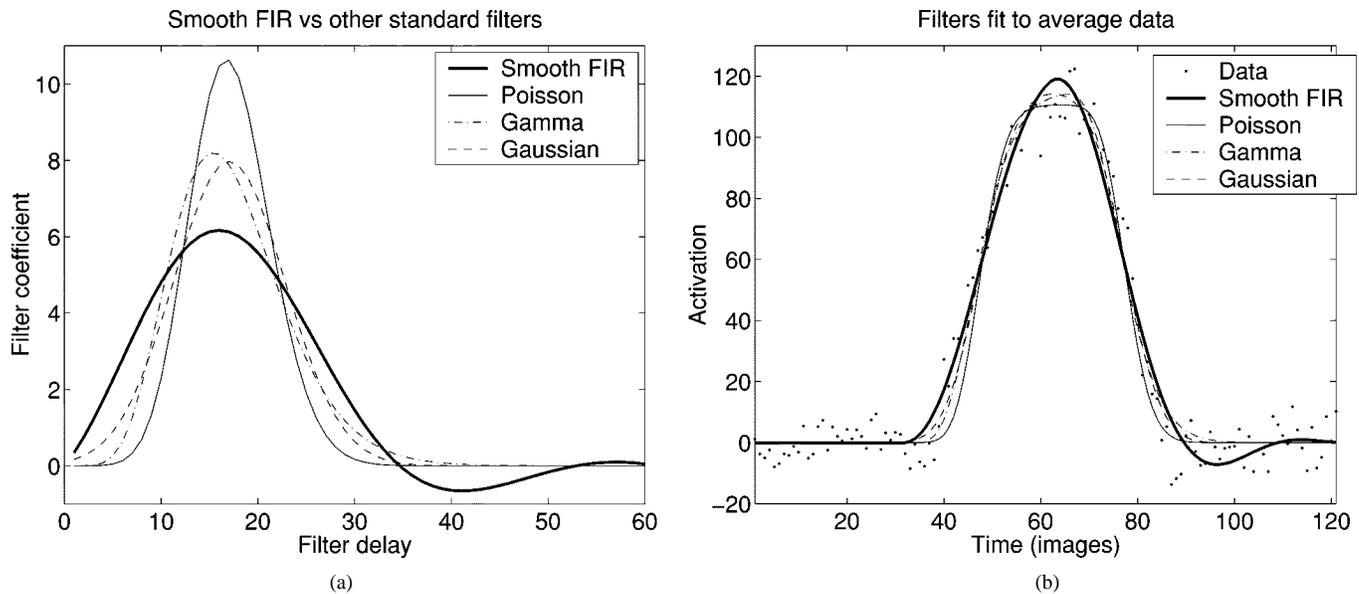
Fig. 6. (a) Smooth FIR filter obtained for a particular voxel (bold solid), together with the best fit obtained using a Poisson shape (solid), a Gamma shape (dash dotted) and a Gaussian shape (dashed). (b) Fitted signal, for one run, of the four filters from (a), plotted together with the data, averaged across the ten runs.
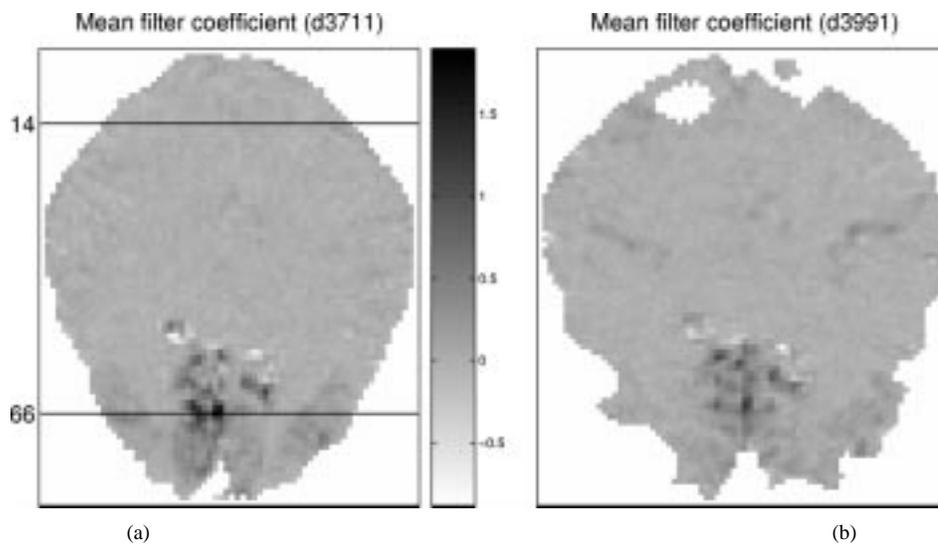

Fig. 7. Brain maps indicating the mean filter coefficient in each voxel. The colorbar is common to both images. The two maps are from the same subject, but two different scanning sessions. The differing shapes are due to different alignments in the two sessions. Notice the good agreement between the activation patterns, indicating good reproducibility. Voxels from the primary visual cortex (and to the lesser extent the supplementary visual cortex) display a strong (and asymmetric) response to the stimulus. Notice that locally the haemodynamic response yields a "negative activation" (white dots above the activated area). (a) The horizontal lines at row 14 and 66 indicate the voxels that we will study in more detail further down (cf. Fig. 8).

to the built-in limitations of these classical filters. In particular, the modeling of the post-activation undershoot observed on real data is not due to the Gaussian process prior used to constrain the FIR coefficients, but reflects a feature that standard filter shapes are unable to model.

B. What is the Shape of the Haemodynamic Response?

Let us now take the opposite standpoint and compare the smooth filter obtained on real data to the best fit using the other three standard kernel shapes. On the same data, we estimate the maximum *a posteriori* filter parameters w_{MAP}, as well as the Poisson, Gamma and Gaussian filters with the best fit to the data.

The results are presented on Fig. 6, where we have plotted the result of the smooth FIR filter together with the best fit obtained using the three standard kernel shapes introduced above.

One obvious result is that the Poisson filter seems to be quite inappropriate for estimating the haemodynamic filter. This is due to the fact that the one-parameter filter has identical mean and variance. In some particular cases, notably when TR is large and the filter only covers a few images, this might not be too limiting. However, this clearly introduces a strong constraint on the shape of the filter, which leads to an inappropriate filter on our data. The Poisson parameter is here 17.8, corresponding to a mean activation delay of 5.9 s, which is reasonable. We would on the other hand appreciate getting a wider filter as the three other filters are wider, but due to the restriction of the Poisson
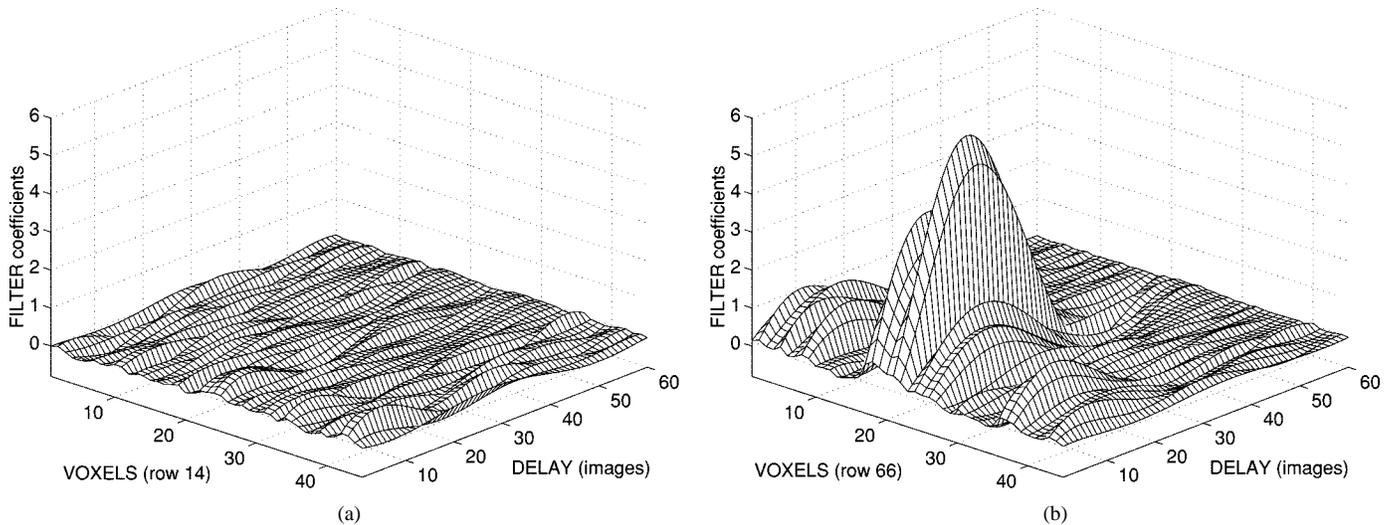
Fig. 8.   The smooth FIR filters obtained on the rows indicated in black on Fig. 7 (row 14 and 66). The X-axis (VOXELS) runs along the "cut" indicated in Fig. 7, the Y-axis (DELAY) runs along the delays in the FIR filter (like the X-axis in, e.g., Figs. 1 and 2). Notice the strong activation in row 66, in the middle of the range, which corresponds to voxels from the primary visual cortex (V1). There is also a more limited response in the lateral areas. By contrast, the filters in row 14 are almost flat, indicating no activation.

filter, this would increase the activation delay beyond reasonable. For comparison, the activation delays are 5.3 s and 5.9 s for the Gamma and Gaussian filters, respectively, and 5.8 s for the smooth FIR filter. But the width, measured by the standard deviation, is 1.4 s for the Poisson filter, versus 1.8 s and 1.9 s for Gamma and Gaussian.

A second salient feature is that, by construction, none of the three basic filter shapes is able to model the post activation undershoot evidenced by the smooth filter. The ability of the Gaussian filter to model the first activation "bump" nicely, and go to zero fast afterwards gives a slightly better fit to the data. By construction, the Gamma filter is skewed and has significant mass in the tail (i.e., for large delays). This proves to be harmful as it introduces additional misfit around the post-activation undershoot (as the filter can not go to zero fast enough). This is also the reason why the maximum of the filter seems to be reached slightly ahead of what is expected. Because of the skewness, the maximum of the filter is attained sensibly earlier (at 4.7 s) than the mean (5.9 s). Furthermore, in order to minimize the misfit around the post-activation undershoot, the mode has to be shifted toward zero.

This result indicates that the smooth FIR filter will be able to model additional features in the data, when traditional filter shapes fail. This is important because we know from previous studies (e.g., [17]) that there is a post-activation undershoot in fMRI data. It might also be possible to model the initial negative response if it is present in the data (cf. Section V-A).

C. Full Analysis

The slice that we study in this dataset contains 3891 voxels. We estimate the smooth FIR filter in each voxel, using $d = 60$ delays. This leads to 3891 distinct filters. Note however that in the MAP estimation procedure (7), the fMRI signal comes into the pictures through the vector Y, while X is identical for all voxels. An important methodological question is whether the hyper-parameters σ^2, ν and h should be kept constant across

the brain or locally estimated. The local estimation requires a significant increase in computation which makes the estimation process impractical on current workstations for several thousand voxels. Accordingly, we will here adopt a hybrid approach, which is computationally easier. Parameter h is fixed from *a priori* knowledge to a value corresponding to a characteristic length of 7 s; σ is set by iterative re-estimation, which converges very fast; and ν is fixed for the whole brain to a value optimized using the evidence on a given activated voxel.

Note that using a global set of hyper-parameters does not imply that the filter itself should be constant. As argued by [10], the characteristics of the filter should vary spatially. However, it is desirable to impose some constraint on the filter such that the filters would not vary beyond reasonable from one voxel to the next. The use of global hyper-parameters, beyond its computational justification, forces such a high level constraint between the filters.

The results are summarized in Fig. 7. As we have 60 filter parameters/voxel, it is necessary to design a summary statistic in each voxel for presentation purposes. In Fig. 7 we present the mean filter coefficient. The rationale behind this choice is that positive responses will display a positive mean coefficient, even when the post-activation undershoot is taken into account. Alternatively, we could use the most extreme coefficient, or the mean absolute coefficient, but the latter loses the sign of the activation, and we know from previous studies [30], [18] that some areas display a negative BOLD response to the stimulus.

In order to represent the filter themselves, additional dimensions are required to accommodate the filter delay and the coefficient values. Accordingly, we will illustrate the difference between the filters on two rows of voxels, one from a nonactivated area (row 14), and one taken from a cut through the visual cortex corresponding to row 66 in the summary images (Fig. 7). The results are presented in Fig. 8. In row 14, the filters are almost flat, reflecting the fact that there is no activation. Small fluctuations around zero reflect the presence of noise. On the other hand, the
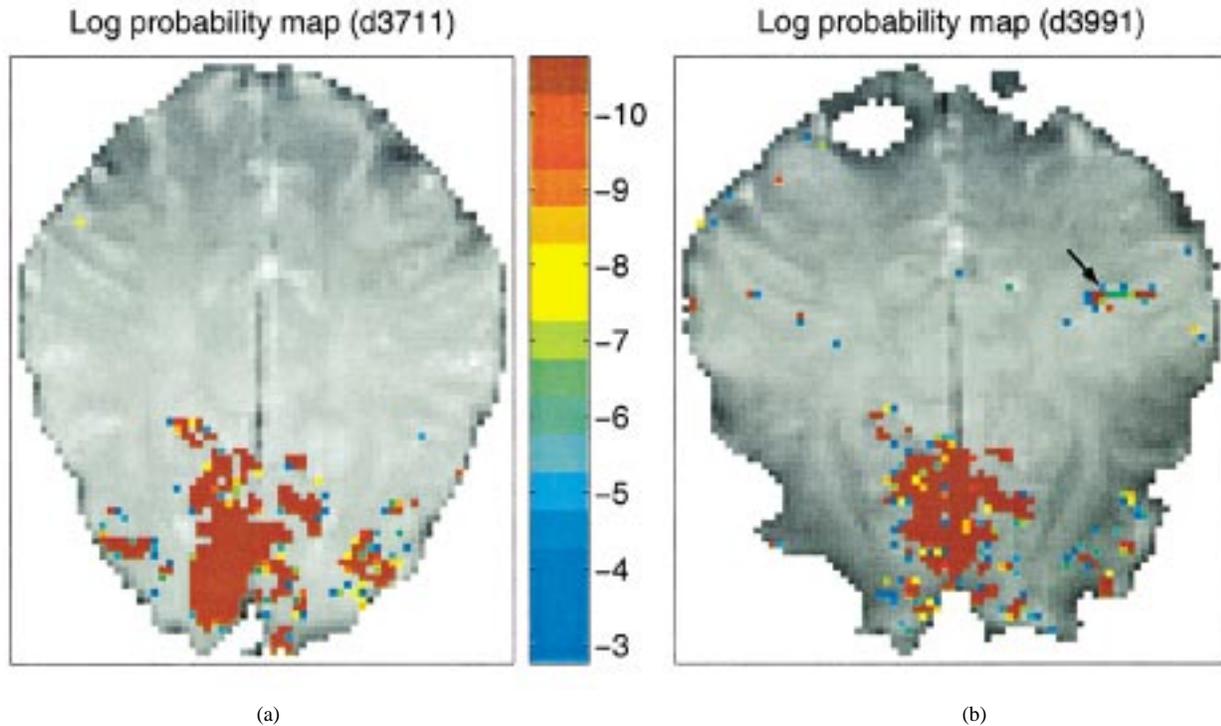
(a)   (b)

Fig. 9. Brain maps indicating the log support p_{HPD} in each voxel, calculated using (13). Values have been thresholded at 10^{-3} and superimposed over a reference background. (a) Displays concentrated activation in the primary visual cortex, as well as in lateral areas. In (b), the lateral activation is more diffuse, on the other hand artefactual false positives appear (see, e.g., arrow), probably due to movement effects. The colorbar indicates the (base 10) log of p_{HPD}.

filters identified in the voxels corresponding to the primary visual cortex (V1), in the middle of the range on Fig. 8(b), display a strong positive activation, followed by a post-activation undershoot, modeled by a series of negative filter coefficients around 40 to 60 images delay. Voxels located in the lateral visual cortex display a moderate positive activation. In some cases, the estimated filter displays a corresponding under-shoot, but the amplitudes are so limited that the relevance of this feature is clearly debatable.

Let us investigate the significance of activation using the highest probability density approach outlined above. Fig. 9 presents the location of the voxels for which $p_{\mathrm{HPD}} < 0.001$, superimposed on a background reference. This allows for a quantitative characterization of the activation pattern outlined above (Fig. 7). In both experiments there is a clear activation in the primary visual cortex. Notice that the asymmetric nature of the activation reproduces well in both experiments. Another finding is that the negative activation that were apparent above the main (positive) activation area turn out as highly significant in Fig. 9. We also note that whereas some significant activation is present on both sides of the lateral visual cortex in the first experiment [d3711, Fig. 9(a)], only traces of significant activation are observed in the second experiment. On the other hand, experiment d3991 displays a higher number of scattered artefactual activation, including a very consistent area [arrow on Fig. 9(b)] which could be due to movement artefacts. Finally, the activated area may seem larger than the cortical area (visual cortex) especially for d3711. One factor explaining this is the spatial blurring in the hemodynamic signal, which to our knowledge is still imperfectly understood.

We will now characterize the delay in activation modeled by the FIR filters, using the group delay [25] described earlier. Fig. 10 plots the resulting delays on brain maps where only the voxels that exceeded the threshold used in Fig. 9 ($p < 0.001$) are retained, and superimposed on the background reference. There is a similarity in the spread of the delays, as well as in the fact that the group delay seems longer in the posterior region of activation. There is however a striking difference in the actual delay values. In the first experiments, the delays range roughly between nine images (3 s) and 18 images (6 s), while in the second, the range is between 15 and 24 images (5–8 s). This difference could be the sign of an inconsistency in the time registration during the experiments. The activation periods might have actually occurred earlier than registered in experiment d3711, or later than registered in d3991, or a combination of both.

Apart from this possible inconsistency, Fig. 10 reveals that the estimation of the group delay yields values that seem biologically plausible, though noticeably smaller than the on-line/off-line delay. The estimates seem to be locally similar, but display noticeable differences at larger scale, some regions reacting with shorter delays. This difference in activation delay was previously spotted using clustering [18].

V. DISCUSSION

A. Initial Dip

In addition to modeling the post-activation undershoot, as shown, e.g., in Fig. 6, the smooth FIR filter is potentially able to model the initial negative response or initial dip [15], [16],
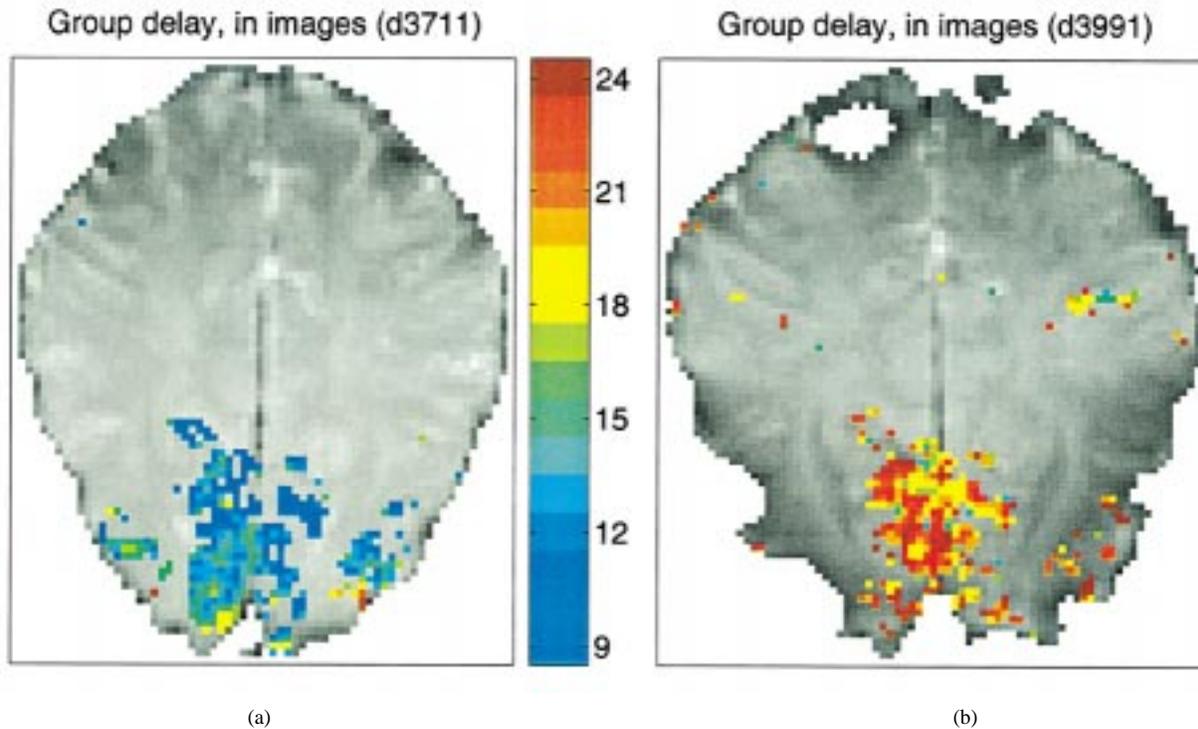
Fig. 10. Brain maps indicating the group delay measured in each activated voxel ($p < 0.001$), cf. Fig. 9. The delays in the activated areas range from nine to 18 images for d3711 (3–6 s) and from 15 to 24 images for d3991 (5–8 s), indicating a possible inconsistency in time registration during the experiments. The colorbar indicates delay (in images) between nine images (3 s) and 24 images (8 s).

[31]–[33], . Despite the relatively low field intensity, the initial dip was observed here in a number of voxels, mainly in experiment d3991. Due to the temporal alignment problem that we have uncovered above, it was difficult to check the reproducibility of the observed negative response across the two experiment.

However, we can report that this feature has been observed in 20%–30% of the activated voxels, in both experiments, after correction of the temporal alignment. A complete analysis of the initial dip is out of the scope of the current article and will be reported elsewhere.

B. Dependency on Neuronal Activation

The relationship between the experimental paradigm and the observed images involves both a neuronal activation induced by the paradigm, and the hemodynamic response to this activation, which leads to the actual measurements. By using the paradigm as input to the linear filter, we assume that the pattern of activation follows the paradigm closely. Although this is a reasonable assumption is the context of a strong visual stimulation, in other cases the stimulus might not be the same as the actual neuronal activation.

This limitation is common to all the parametric models mentioned in this study and has, therefore, little relevance for the comparison between these methods. It should, however, be kept in mind when more complicated experiments (perhaps involving cognitive tasks) are involved. To determine both the neuronal activation and the hemodynamic response would require blind deconvolution using a latent variable model. The model of [34]

is an attempt to do this. It is, however, outside of the scope of this article.

C. Computational Issues

One of the biggest challenge of this approach lies in the practical implementation for whole brain analysis, or for the analysis of a reasonable subset of voxel, e.g., after sieving with an omnibus F-test. While the calculation of the MAP estimate for a given set of hyper-parameters σ^2, ν and h is straightforward and not more computationally intensive than traditional approaches based on ridge regression or singular value decomposition, the tuning of the hyper-parameter is usually time consuming. While we have argued for example that the parameter h controlling the typical length scale could be set *a priori* to between 5 and 10 s (7 s here), there is no guarantee that this is optimal in any sense.

A full nonlinear optimization over two or three hyper-parameters is feasible for a limited number of voxels, but too computationally demanding for a whole volume or even a slice with current computing facilities. Similarly, sampling from the posterior using an MCMC technique is only practical for a limited subset of voxels.

One simplification would be the use of fixed hyper-parameters for the whole volumes. In that case only one nonlinear optimization or Markov Chain would be needed to yield a set of hyper-parameters which are applied to all voxels. Though some researchers (e.g., [10]) have argued that the characteristics of the haemodynamic response vary spatially, note that having fixed hyper-parameters would allow the filters themselves to be spatially different, while tying them at a higher level in a hierarchical manner. A drawback of this approach is that it might lead

to averaging some of the characteristics like the noise level or the length scale. Typically, nonactivated voxels could be modeled using large length scales, corresponding to flat filters, while activated voxels would benefit from the flexibility introduced by smaller length scales. Note that this is not necessarily a problem as far as the *predictions* themselves are concerned, as suggested by Fig. 5.

In the experiments described above, we have adopted an intermediate approach, where h is set *a priori* to give a length scale of 7 s, the regularization strength ν is optimized once and for all based on some activated voxels, while the noise level σ^2 is estimated locally using an iterative procedure similar to the re-estimation formulas in the so-called "evidence framework" (e.g., [28], [29, sec. 10.4]). There is an obvious benefit in terms of computationally time. In our Matlab implementation running on a 450 Mhz Pentium II, the full estimation of the filters and associated measures of support p_{HPD} takes around 90 s for 4000 voxels. An added benefit of the local estimation of σ^2 is that we do not need to make the assumption that the noise is spatially stationary.

VI. CONCLUSION

In this paper, the use of smooth FIR filters for analyzing functional magnetic resonance imaging data was described. Smoothness is implemented using a correlated Gaussian prior, and analysis is carried on using Bayesian inference. The smooth FIR filter has a number of advantages over standard (Poisson, Gamma, Gaussian) parametric families for modeling the haemodynamic response. In particular, it can model a long post-activation undershoot or the initial negative response. The generality and flexibility of the smooth FIR approach was illustrated on simulated data. A full analysis of data acquired during a visual stimulation experiment with high temporal resolution was performed. The ability of the smooth FIR filter to find activated regions was demonstrated using a measure of support derived from the highest posterior density approach.

APPENDIX
SAMPLING VIA MCMC

As noted in Section III-F, an ideal Bayesian analysis would not optimize parameters, but obtain distributions of the relevant quantities by integrating over nuisance parameters. This can be useful here is at least three contexts.

1) Marginalize the hyper-parameters to obtain the posterior of the filter parameters $p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{Y})$, in order to obtain maximum posterior parameters or the covariance of these parameters;
2) Marginalize the hyper-parameters in the distribution of the prediction (12), in order to obtain the distribution of y_0 conditioned only on the actual data;
3) Obtain the posterior distribution of the hyper-parameters conditioned on the data $p(\sigma^2, \nu, h|\boldsymbol{Y}, \boldsymbol{X})$ in order to check for example whether the hyper-parameters are well-determined by the data.

Numerical integration methods will be necessary for all three problems, and can in principle be easily performed using
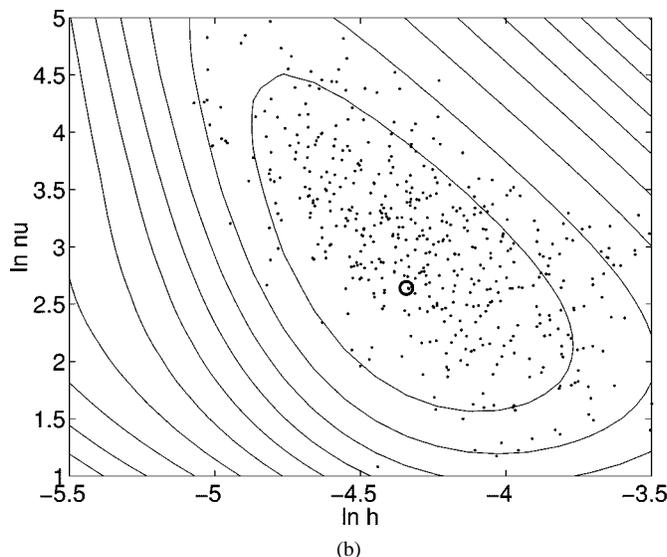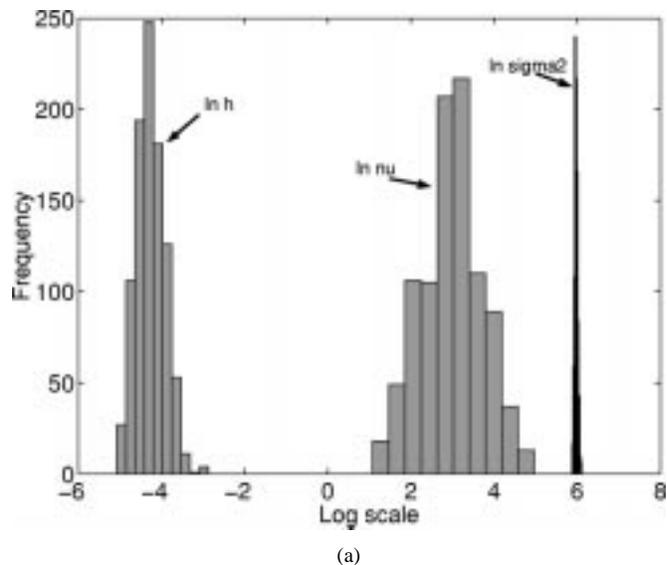

(a)

(b)

Fig. 11. (a) Histogram for the three hyper-parameters obtained from sampling the posterior. (b) Bivariate samples of ν and h (in the log domain) as dots, superimposed on a contour plot of the joint density of ν and h conditioned on $\sigma^2 = 400$.

MCMC [26], in particular Metropolis–Hastings, or hybrid Monte Carlo [35] if derivatives are available.

A. Priors

The first step is to setup some priors for the three hyper-parameters that we use here. We will put a Gamma prior on the variance of the noise σ^2 and prior strength ν. For normalized data, we would take a mean of $\mu = 1$ and a shape factor of $\alpha = 0.5$, because it is unlikely that the noise level far exceeds the variance of the data, and there should be significant mass toward zero in order to allow small noise levels (or little regularization). Accordingly, we will scale the prior with the empirical variance calculated on the actual data σ_y

$$\sigma^2, \nu \sim \mathcal{G}(\sigma_y^2, 0.5)$$
$$p(\sigma^2) \propto \sigma^{-3/2} \exp(-\sigma^2/4\sigma_y^2)$$
$$p(\nu) \propto \nu^{-3/4} \exp(-\nu/4\sigma_y^2).$$

For the length scale, we have strong *a priori* information suggesting that typical length scales for the haemodynamic response should be between 5 and 10 s. However, we want to allow larger length scales, which might be useful for nonactivated voxels, where the underlying filter should be uniformly zero. Accordingly, we will model this prior information using a log-normal distribution, such that the log of the characteristic length covers $[\log 5, \log 10]$ over two standard deviations. This leads to

$$\ln h \sim \mathcal{N}(2\ln \mathrm{TR} - 3.92, 0.52)$$

$$p(h) = \frac{1}{h\sqrt{1.04\pi}} \exp\left[-\frac{(\ln h - 2\ln\mathrm{TR} + 3.92)^2}{1.04}\right].$$

B. Sampling from the Posterior

We will now illustrate how to sample from the posterior of the hyper-parameters $p(\sigma^2, \nu, h | \boldsymbol{X}, \boldsymbol{Y})$, in order to check how well determined these hyper-parameters are. The posterior is easily obtained (from Bayes' rule) as it is proportional to the product of the evidence (16) by the priors described previously. As the derivative of the evidence with respect to h is nontrivial, we will simply use a Metropolis–Hastings algorithm [36], [26] with a Gaussian proposal (in the log domain), which has the advantage of being symmetric.

After setting the proposal such that we get a acceptance rate between 50% and 60%, we run the chain for 1000 iterations and discard the 50 first samples as "burn-in." The histograms of the sample distribution for the three hyper-parameters are presented in Fig. 11 (above). The log scale gives a good indication of the relative spread of the hyper-parameters around their mean. Clearly, the noise level σ^2 is very well determined by the data. The prior strength ν is badly determined, meaning that a wide range of values have large probability. The situation for the length scale h is somewhat intermediate. These results show that it is sensible to optimize the noise level σ^2 to a fixed value, as its posterior distribution is close to a delta function. On the other hand, it would be interesting to integrate over hyper-parameters ν and h, which have broader marginal distributions.

In Fig. 11(b), we investigate the joint distribution of ν and h. The background contour plot has been obtained from the expression of the posterior $p(\sigma^2, \nu, h | \boldsymbol{Y}, \boldsymbol{X})$ by setting $\sigma^2 = 400$. As our previous investigation showed that the noise level is well determined in the neighborhood of this value, this gives a probably accurate description of the marginal joint distribution $p(\sigma^2, \nu | \boldsymbol{Y}, \boldsymbol{X})$. The sample (dots on Fig. 11, right) seems to support this approximation, and indicates that, due to correlation between ν and h, the hyper-parameters are slightly better determined by the data than it is suggested by the marginal histograms.

This result shows that obtaining a sample from the hyper-parameters posterior is potentially useful. Unfortunately, as indicated earlier, it is not computationally possible to perform this sampling on a large scale.

ACKNOWLEDGMENT

REFERENCES

[1] P. A. Bandettini and R. W. Cox, "Functional contrast in event-related fMRI: Interstimulus dependency and blocked design comparison," in *Proceedings of the Fourth International Conference on Functional Mapping of the Human Brain*. ser. NeuroImage, T. Paus, A. Gjedde, and A. Evans, Eds.   New York: Academic, May 1998, pt. 2 of 3, p. 522.

[2] C. G. Thomas and R. S. Menon, "Amplitude response and stimulus presentation frequency response of human primary visual cortex using BOLD EPI at 4 T," *Magn. Reson. Medicine*, vol. 40, no. 2, pp. 203–209, 1998.

[3] R. B. Buxton, E. C. Wong, and L. R. Frank, "Dynamics of blood flow and oxygenation changes during brain activation: The ballon model," *Magn. Reson. Med*, vol. 39, no. 6, pp. 855–864, 1998.

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

[5] J. R. Binder, S. M. Rao, T. A. Hammeke, J. A. Frost, P. A. Bandettini, and J. S. Hyde, "Effects of stimulus on signal response during functional magnetic-resonance-imaging of auditory-cortex," *Cogn. Brain Res.*, vol. 2, no. 1, pp. 31–38, 1994.

[6] A. M. Dale and R. L. Buckner, "Selective averaging of rapidly presented individual trials using fMRI," *Human Brain Mapping*, vol. 5, no. 5, pp. 329–340, 1997.

[7] M. D. Robson, J. L. Dorosz, and J. C. Gore, "Measurements of the temporal fMRI response of the human auditory cortex to trains of tones," *NeuroImage*, vol. 7, no. 3, pp. 185–198, 1998.

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

[9] K. J. Friston, P. Jezzard, and R. Turner, "The analysis of functional MRI time-series," *Human Brain Mapping*, vol. 1, pp. 153–174, 1994.

[10] N. Lange and S. L. Zeger, "Non-linear Fourier time series analysis for human brain mapping by functional magnetic resonance imaging," *J. Roy. Statistical Soc.*, ser. C, Appl. Stat., vol. 46, no. 1, pp. 1–30, 1997.

[11] J. C. Rajapakse, F. Kruggel, J. M. Maisog, and D. Y. von Cramon, "Modeling hemodynamic response for analysis of functional MRI time-series," *Human Brain Mapping*, vol. 6, pp. 283–300, 1998.

[12] J. Kershaw, B. A. Ardekani, and I. Kanno, "Application of Bayesian inference to fMRI data analysis," *IEEE Trans. Med. Imag.*, vol. 18, pp. 1138–1153, Dec. 1999.

[13] M. S. Cohen, "Parametric analysis of fMRI data using linear systems methods," *NeuroImage*, vol. 6, no. 2, pp. 93–103, Aug. 1997.

[14] F. Å. Nielsen, L. K. Hansen, P. Toft, C. Goutte, N. Lange, S. C. Strother, N. Mørch, C. Svarer, R. Savoy, B. Rosen, E. Rostrup, and B. Peter, "Comparison of two convolution models for fMRI time series," in *Friberg et al. [37]*, L. Friberg, A. Gjedde, S. Holm, N. A. Lassen, and M. Nowak, Eds.   New York: Academic, May 1997, pt. 2 of 4 in NeuroImage, vol. 5, p. S473.

[15] R. S. Menon, S. Ogawa, X. Hu, J. P. Strupp, P. Anderson, and K. Ugurbil, "BOLD based functional MRI at 4 Tesla includes a capillary bed contribution: Echo-planar imaging correlates with previous optical imaging using intrinsic signals," *Magn. Reson. Med*, vol. 33, no. 3, pp. 453–459, 1995.

[16] E. Yacoub and X. Hu, "Detection of the early negative response in fMRI at 1.5 tesla," *Magn. Reson. Med*, vol. 41, no. 6, pp. 1088–1092, 1999.

[17] G. Krüger, A. Kleinschmidt, and J. Frahm, "Dynamic MRI sensitized to cerebral blood oxygenation and flow during sustained activation of human visual cortex," *Magn. Reson. Med*, vol. 35, no. 6, pp. 797–800, 1996.

[18] C. Goutte, L. K. Hansen, M. G. Liptrot, and E. Rostrup, "Feature space clustering for fMRI meta-analysis," IMM, Tech. Rep. IMM-REP-1999-13, 1999.

[19] C. E. Rasmussen, "Evaluation of Gaussian processes and other methods for nonlinear regression," Ph.D. dissertation, Dept. Comput. Sci., Univ. Toronto, Toronto, Canada, 1996.

[20] C. K. I. Williams, "Prediction with Gaussian processes: From linear regression to linear prediction and beyond," in *Learning and Inference in Graphical Models*, M. I. Jordan, Ed.   Norwell, MA: Kluwer, 1998.

[21] 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, no. 2, pp. 161–173, August 1993.

[22] G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*. New York: Wiley, 1992.

[23] M. J. Schervish, "*P*-values: What they are and what they are not," *Amer. Statistician*, vol. 50, no. 3, pp. 203–206, Aug. 1996.

[24] M. Lavine and M. J. Schervish, "Bayes factors: What they are and what they are not," *Amer. Statistician*, vol. 53, no. 2, pp. 119–122, May 1999.

[25] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[26] D. J. C. MacKay, "Introduction to Monte Carlo methods," in *Learning in Graphical Models*. ser. NATO SCIENCE: D Behavioral and Social Sciences, M. I. Jordan, Ed. Dordrecht, The Netherlands: Kluwer Academic, 1998, vol. 89.

[27] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 1992.

[28] D. MacKay, "A practical Bayesian framework for backprop networks," *Neural Computation*, vol. 4, pp. 448–472, 1992.

[29] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Clarendon, 1995.

[30] C. Goutte, P. Toft, E. Rostrup, F. Å. Nielsen, and L. K. Hansen, "On clustering fMRI time series," *NeuroImage*, vol. 9, no. 3, pp. 298–310, 1999.

[31] X. Hu, T. H. Le, and K. Ugurbil, "Evaluation of the early response in fMRI in individual subjects using short stimulus duration," *Magn. Reson. Med*, vol. 37, no. 6, pp. 877–884, 1997.

[32] E. Yacoub, T. H. Le, K. Ugurbil, and X. Hu, "Further evaluation of the initial negative response in functional magnetic resonance imaging," *Magn. Reson. Med*, vol. 41, no. 3, pp. 436–441, 1999.

[33] G. M. Hathout, B. Varjavand, and R. K. Gopi, "The early response in fMRI: A modeling approach," *Magn. Reson. Med*, vol. 41, no. 3, pp. 550–554, 1999.

[34] P. A. d. F. R. Højen-Sørensen, L. K. Hansen, and C. E. Rasmussen, "Bayesian modeling of fMRI time series," in *Advances in Neural Information Processing Systems 12: Proceedings of the 1999 Conference*, S. A. Solla, T. K. Leen, and K.-R. Müller, Eds. Cambridge, MA: MIT Press, 2000, pp. 754–760.

[35] R. M. Neal, *Bayesian Learning for Neural Networks*. New York: Springer, 1996, vol. 118 of Lecture Notes in Statistics.

[36] S. Chib and E. Greenberg, "Understanding the Metropolis–Hastings algorithm," *The American Statistician*, vol. 49, no. 4, pp. 327–335, 1995.

[37] L. Friberg, A. Gjedde, S. Holm, N. A. Lassen, and M. Nowak, Eds., *Proceedings of the Third International Conference on Functional Mapping of the Human Brain*. New York: Academic, May 1997, part 2 of 4 in NeuroImage.