NeuroImage **16,** 513–530 (2002)
doi:10.1006/nimg.2001.1044, available online at http://www.idealibrary.com on IDEAL®

Bayesian Estimation of Dynamical Systems: An Application to fMRI

K. J. Friston

The Wellcome Department of Cognitive Neurology, Institute of Neurology, Queen Square, London, United Kingdom WC1N 3BG

Received January 11, 2001

This paper presents a method for estimating the conditional or posterior distribution of the parameters of deterministic dynamical systems. The procedure conforms to an EM implementation of a Gauss–Newton search for the maximum of the conditional or posterior density. The inclusion of priors in the estimation procedure ensures robust and rapid convergence and the resulting conditional densities enable Bayesian inference about the model parameters. The method is demonstrated using an input–state–output model of the hemodynamic coupling between experimentally designed causes or factors in fMRI studies and the ensuing BOLD response. This example represents a generalization of current fMRI analysis models that accommodates nonlinearities and in which the parameters have an explicit physical interpretation. Second, the approach extends classical inference, based on the likelihood of the data given a null hypothesis about the parameters, to more plausible inferences about the parameters of the model given the data. This inference provides for confidence intervals based on the conditional density. © 2002 Elsevier Science (USA)

Key Words: **fMRI; Bayesian inference; nonlinear dynamics; model identification; hemodynamics; Volterra series; EM algorithm; Gauss–Newton method.**

1. INTRODUCTION

This paper is about the identification of deterministic nonlinear dynamical models. Deterministic here refers to models where the dynamics are completely determined by the state of the system. Random or stochastic effects enter only at the point that the system's outputs or responses are observed.[1] In this paper we focus on a particular model of how changes in neuronal activity translate into hemodynamic responses. By considering a voxel as an *input–state–output* system one can model the effects of an input (i.e., stimulus function) on some state variables (e.g., flow, volume, deoxyhemoglobin content) and the ensuing output (i.e., BOLD response). The scheme adopted here uses Bayesian estimation, where the aim is to identify the posterior or conditional distribution of the parameters, given the data. Knowing the posterior distribution allows one to characterize an observed system in terms of the parameters that maximize their posterior probability (i.e., those parameters that are most likely given the data) or, indeed, make inferences about whether the parameters are bigger or smaller than some specified value.

The primary aim of this paper is to present the methodology for Bayesian estimation of any deterministic nonlinear dynamical model. However, through demonstrating the approach using hemodynamic models pertinent to fMRI, we can also introduce the notion that biophysical and physiological models of evoked brain responses can be used to make Bayesian inferences about experimentally induced, regionally specific activations. This inference is enabled by including parameters that couple experimentally changing stimulus or task conditions (that are treated as inputs) to the system's dynamics. The posterior or conditional distribution of these parameters can then be used to make inferences about the efficacy of the inputs in eliciting a measured response. Because the parameters we want to make an inference about have an explicit physical interpretation, in the context of the hemodynamic model used, the face validity of the ensuing inference is more grounded in physiology. Furthermore, because the "activation" is parameterized in terms of processes that have natural biological constraints, these constraints can be used as priors in a Bayesian scheme.

The material presented here represents a convergence of work described in two previous papers. In the first paper (Friston *et al.*, 2002a) we discussed the utility of Bayesian inference in the context of hierarchical observation models commonly employed in fMRI. This paper focused on *empirical* Bayesian approaches in which the priors were derived from the data being analyzed. In this paper we use a *fully* Bayesian approach, where the priors are assumed to be known and apply it to the hemodynamic model described in the second paper (Friston *et al.*, 2000). In

[1] There is another important class of models where stochastic processes enter at the level of the state variables themselves (i.e., deterministic noise). These are referred to as stochastic dynamical models.

1053-8119/02 $35.00

Friston *et al.* (2000) we presented a hemodynamic model that embedded the Balloon/Windkessel (Buxton *et al.*, 1998; Mandeville *et al.*, 1999) model of flow to BOLD coupling to give a complete dynamical model of how neuronally mediated signals cause a BOLD response. In Friston *et al.* (2000) we restricted ourselves to single input–single output (SISO) systems by considering only one input. In this paper we demonstrate a general approach to nonlinear system identification using an extension of these SISO models to multiple input–single output (MISO) systems. This allows for a response to be caused by multiple experimental effects and we can assign a causal efficacy to any number of explanatory variables (i.e., stimulus functions). In subsequent papers we will generalize the approach taken in this paper to multiple input–multiple output systems (MIMO) such that interactions among brain regions, at a neuronal level can be addressed. What follows can be seen as a technical prelude to the latter generalization.

An important aspect of the proposed estimation scheme is that it can be reduced, exactly, to the scheme used in classical SPM-like analyses, where one uses the stimulus functions, convolved with a canonical hemodynamic response function, as explanatory variables in a general linear model. This classical analysis is a special case that obtains when the model parameters of interest (the efficacy of a stimulus) are treated as fixed effects with flat priors and the remaining biophysical parameters enter as known canonical values with infinitely small prior variance (i.e., high precision). In this sense the current approach can be viewed as a Bayesian generalization of that normally employed. The advantages of this generalization rest upon (i) the use of a nonlinear observation model and (ii) Bayesian estimation of that model's parameters. The fundamental advantage, of a nonlinear MISO model over linear models, is that only the parameters linking the various inputs to hemodynamics are input or trial specific. The remaining parameters, pertaining to the hemodynamics per se, are the same for each voxel. In conventional analyses the hemodynamic response function, for each input, is estimated in a linearly separable fashion (usually in terms of a small set of temporal basis functions) despite the fact that the (unknown) form of the impulse response function to each input is likely to be the same. In other words, a nonlinear model properly accommodates the fact that many of the parameters shaping input-specific hemodynamic responses are shared by all inputs. For example, the components of a compound trial (e.g., cue and target stimuli) might not interact at a neuronal level but may show subadditive effects in the measured response, due to nonlinear hemodynamic saturation. In contradistinction to conventional linear analyses the analysis proposed in this paper could, in principle, disambiguate between interactions at the neuronal and hemodynamic levels. The second advantage is that Bayesian inferences about input-specific parameters can be framed in terms of whether the efficacy for a particular cause exceeded some specified threshold or, indeed, the probability that it was less than some threshold (i.e., infer that a voxel did *not* respond). The latter is precluded in classical inference. These advantages should be weighed against the difficulties of establishing a valid model and the computational expense of identification.

This paper is divided into four sections. In the first we reprise briefly the hemodynamic model and motivate the four differential equations that it comprises. We will touch on the Volterra formulation of nonlinear systems to show the output can always be represented as a nonlinear function of the input and the model parameters. This nonlinear function is used as the basis of the observation model that is subject to Bayesian identification. This identification require priors which, in this paper, come from the distribution, over voxels, of parameters estimated in Friston *et al.* (2000). This estimation was in terms of the Volterra kernels associated with the model parameters. The second section describes these priors and how they were determined. Having specified the form of the nonlinear observation model and the prior densities on the model's parameters, the third section describes the estimation of their posterior densities. The derivation is sufficiently simple to be presented from basic principles. The ensuing scheme can be regarded as a Gauss–Newton search for the maximum posterior probability (as opposed to the maximum likelihood as in conventional applications). This section concludes with a note on integration, required to evaluate the local gradients of the objective function. The evaluation is greatly facilitated by the sparse nature of stimulus functions typically used in fMRI, enabling efficient integration using a bilinear approximation to the differential equations that constitute the model. The final section illustrates applications to empirical data. First, we revisit the same data used to construct the priors using a single input. We then apply the technique to a study of visual attention, first reported in Büchel and Friston (1998), to make inferences about the relative efficacy of multiple experimental effects in eliciting a BOLD response. These examples deal with single regions. We conclude by repeating the analysis at all voxels to form posterior probability maps (PPMs) for different experimental causes.

2. THE HEMODYNAMIC MODEL

The hemodynamic model considered here was presented in detail in Friston *et al.* (2000). Although relatively simple it is predicated on a substantial amount of previous careful theoretical work and empirical validation (e.g., Buxton *et al.*, 1998; Mandeville *et al.*,

1999; Hoge *et al.*, 1999; Mayhew *et al.*, 1998). The model is a SISO system with a stimulus function as input (that is supposed to elicit a neuronally mediated flow-inducing signal) and BOLD response as output. The model has six parameters and four state variables each with its corresponding differential equation. The differential or state equations express how each state variable changes over time as a function of the others. These state equations and the output nonlinearly (a static nonlinear function of the state variables that gives the output) specify the form of the model. The parameters determine any specific realisation of the model. In what follows we review the state equations, the output nonlinearity, extension to a MISO system, and the Volterra representation.

2.1. The State Equations

Assuming that the dynamical system linking synaptic activity and rCBF is linear (Miller *et al.*, 2000) we start with

$$\dot{f}_{in} = s, \tag{1}$$

where f_{in} is inflow and s is some flow-inducing signal. The signal is assumed to subsume many neurogenic and diffusive signal subcomponents and is generated by neuronal responses to the input (the stimulus function) $u(t)$

$$\dot{s} = \epsilon u(t) - \kappa_s s - \kappa_f(f_{in} - 1). \tag{2}$$

ϵ, κ_s, and κ_f are parameters that represent the efficacy with which input causes an increase in signal, the rate constant for signal decay or elimination, and the rate constant for autoregulatory feedback from blood flow. The existence of this feedback term can be inferred from: (i) poststimulus undershoots in rCBF (e.g., Irikura *et al.*, 1994) and (ii) the well-characterized vasomotor signal in optical imaging (Mayhew *et al.*, 1998). Both support the notion of local closed-loop feedback mechanisms as modeled in Eqs. (1) and (2). Inflow determines the rate of change of volume through

$$\begin{aligned} \tau\dot{v} &= f_{in} - f_{out}(v) \\ f_{out}(v) &= v^{1/\alpha}. \end{aligned} \tag{3}$$

Equation (3) says that normalized venous volume changes reflect the difference between inflow f_{in} and outflow f_{out} from the venous compartment with a time constant (transit time) τ. Outflow is a function of volume that models the balloon-like capacity of the venous compartment to expel blood at a greater rate when distended (Buxton *et al.*, 1998). It can be modeled with a single parameter (Grubb *et al.*, 1974) α based on the Windkessel model (Mandeville *et al.*, 1999). The change in normalized total deoxyhemoglobin voxel content $\dot{q}$ reflects the delivery of deoxyhemoglobin into the venous compartment minus that expelled (outflow times concentration)

$$\begin{aligned} \tau\dot{q} &= f_{in}\frac{E(f_{in}, E_0)}{E_0} - f_{out}(v)q/v \\ E(f_{in}, E_0) &= 1 - (1 - E_0)^{1/f_{in}}, \end{aligned} \tag{4}$$

where $E(f_{in}, E_0)$ is the fraction of oxygen extracted from inflowing blood. This is assumed to depend on oxygen delivery and is consequently flow dependent. This concludes the state equations, where there are six unknown parameters, namely efficacy ϵ, signal decay κ_s, autoregulation κ_f, transit time τ, Grubb's exponent α, and resting net oxygen extraction by the capillary bed E_0.

2.2. The Output Nonlinearity

The BOLD signal $y(t) = \lambda(v, q, E_0)$ is taken to be a static nonlinear function of volume (v) and deoxyhemoglobin content (q)

$$\begin{aligned} y(t) &= \lambda(v, q) = V_0(k_1(1 - q) \\ &\quad + k_2(1 - q/v) + k_3(1 - v)) \\ k_1 &= 7E_0 \qquad k_2 = 2 \qquad k_3 = 2E_0 - 0.2, \end{aligned} \tag{5}$$

where V_0 is resting blood volume fraction. This signal comprises a volume-weighted sum of extra- and intravascular signals that are functions of volume and deoxyhemoglobin content. A critical term in Eq. (5) is the concentration term $k_2(1 - q/v)$, which accounts for most of the nonlinear behaviour of the hemodynamic model. The architecture of this model is summarized in Fig. 1 and an illustration of the dynamics of the state variables in response to a transient input is provided in Fig. 2. The constants in Eq. (5) are taken from Buxton *et al.* (1998) and are valid for 1.5 T. Our data were acquired at 2 T; however, the use of higher field strengths may require different constants.

2.3. Extension to a MISO

The extension to a multiple input system is trivial and involves extending Eq. (2) to cover n inputs

$$\dot{s} = \epsilon_1 u(t)_1 + \ldots + \epsilon_n u(t)_n - \kappa_s s - \kappa_f(f_{in} - 1). \tag{6}$$

The model now has $5 + n$ parameters; five biophysical parameters κ_s, κ_f, τ, α, and E_0; and n efficacies $\epsilon_1, \ldots \epsilon_n$. Although all these parameters have to estimated we are only interested in making inferences

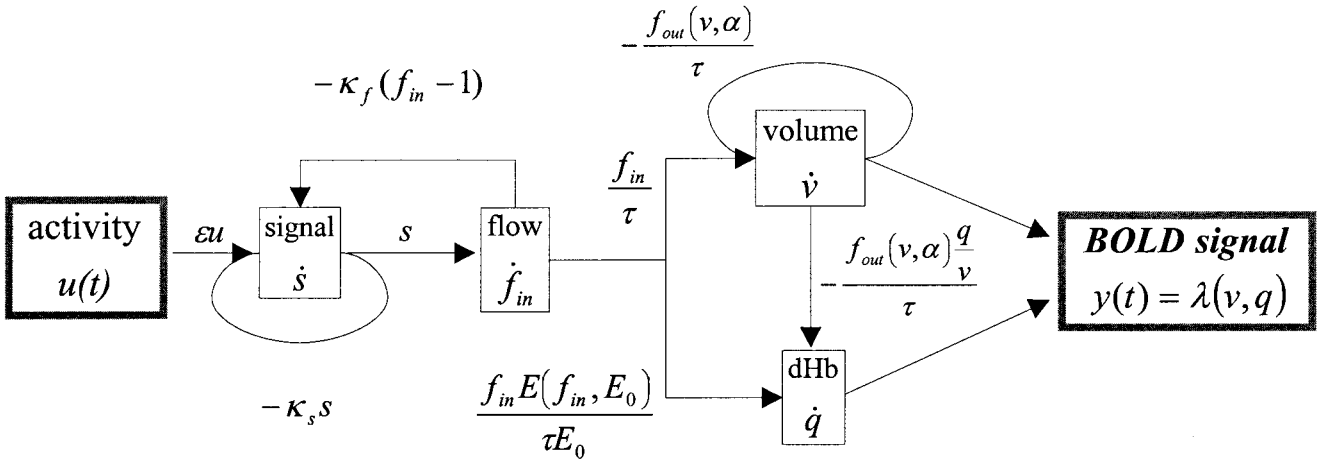

FIG. 1. Schematic illustrating the architecture of the hemodynamic model. This is a fully nonlinear single-input $u(t)$, single-output $y(t)$ state model with four state variables s, f_{in}, v, and q. The form and motivation for the changes in each state variable, as functions of the others, is described in the main text.

about the efficacies. Note that the biophysical parameters are the same for all inputs.

2.4. The Volterra Formulation

In our hemodynamic model the state variables are $X = \{x_1, \ldots, x_4\}^T = \{s, f_{in}, v, q\}^T$ and the parameters are $\theta = \{\theta_1, \ldots \theta_{5+n}\}^T = \{\kappa_s, \kappa_f, \tau, \alpha, E_0, \epsilon_1, \ldots, \epsilon_n\}^T$. The state equations and output nonlinearity specify a MISO model

$$
\begin{aligned}
\dot{X}(t) &= f(X, u(t)) \\
y(t) &= \lambda(X(t)) \\
\dot{x}_1 &= f_1(X, u(t)) \\
&= \epsilon_1 u(t)_1 + \ldots + \epsilon_n u(t)_n - \kappa_s x_1 - \kappa_f(x_2 - 1) \\
\dot{x}_2 &= f_2(X, u(t)) = x_1 \\
\dot{x}_3 &= f_3(X, u(t)) = \frac{1}{\tau}(x_2 - f_{\text{out}}(x_3, \alpha)) \\
\dot{x}_4 &= f_4(X, u(t)) = \frac{1}{\tau}\left(x_2 \frac{E(x_2, E_0)}{E_0} - f_{\text{out}}(x_3, \alpha)\frac{x_4}{x_3}\right) \\
y(t) &= \lambda(x_1, \ldots, x_4) \\
&= V_0(k_1(1 - x_4) + k_2(1 - x_4/x_3) + k_3(1 - x_3)).
\end{aligned}
\tag{7}
$$

This is the state–space representation. The alternative Volterra formulation represents the output $y(t)$ as a nonlinear convolution of the input $u(t)$, critically without reference to the state variables $X(t)$ (see Bendat, 1990). This series can be considered a nonlinear convolution that obtains from a functional Taylor expansion of $y(t)$ about $X(0)$ and $u(t) = 0$. For a single input this can be expressed as

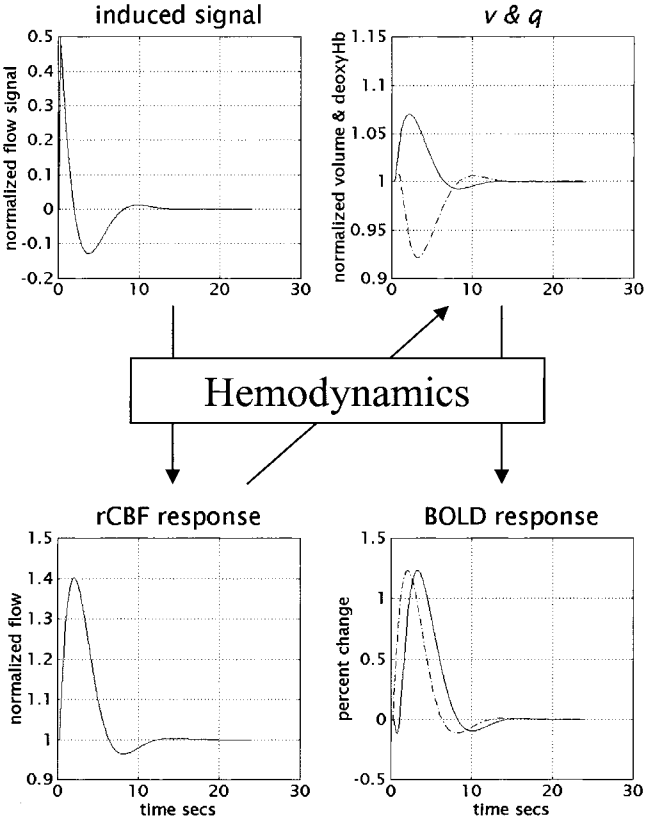

FIG. 2. Illustrative dynamics of the hemodynamic model. (Top left) The time-dependent changes in the neuronally induced perfusion signal that causes an increase in blood flow. (Bottom left) The resulting changes in normalized blood flow. (Top right) The concomitant changes in normalized venous volume (v) (solid line) and normalized deoxyhemoglobin content (q) (broken line). (Bottom right) The percentage of change in BOLD signal that is contingent on v and q. The broken line is inflow normalized to the same maximum as the BOLD signal. This highlights the fact that BOLD signal lags the rCBF signal by about 1 s.

$$y(t) = h(\theta, u) = \kappa_0 + \sum_{i=1}^{\infty} \int_0^{\infty} \cdots \int_0^{\infty} \kappa_i(\sigma_1, \ldots \sigma_i) \times u(t - \sigma_1) \ldots u(t - \sigma_i) d\sigma_1 \ldots d\sigma_i \quad (8)$$

$$\kappa_i(\sigma_1, \ldots \sigma_i) = \frac{\partial^i y(t)}{\partial u(t - \sigma_1) \ldots u(t - \sigma_i)},$$

where κ_i is the ith generalized convolution kernel (Fliess *et al.*, 1983). Equation (8) now expresses the output as a function of the input and the parameters whose posterior distribution we require. The Volterra kernels are a time-invariant characterization of the input–output behavior of the system and can be thought of as generalized high-order convolution kernels that are applied to a stimulus function to emulate the observed BOLD response. Integrating Eq. (7) and applying the output nonlinearity to the state variables is the same as convolving the inputs with the kernels. Both give the system's response in terms of the output. In what follows the response is evaluated by integrating Eq. (7). This means the kernels are not required. However, the Volterra formulation is introduced for several reasons. First, it demonstrates that the output is a nonlinear function of the inputs $y(t) = h(\theta, u)$. This is critical for the generality of the estimation scheme proposed below. (ii) Second, it provides an important connection with conventional analyses using the general linear model (see Section 3.5). (iii) Third, it was used in Friston *et al.* (2000) to estimate the parameters whose distribution, over voxels, constitutes the prior density in this paper and (iv) finally, we use the kernels to characterize evoked responses below (see Section 4).

3. THE PRIORS

Bayesian estimation requires informative priors on the parameters. Under Gaussian assumptions these prior densities can be specified in terms of their expectation and covariance. These moments are taken here to be the sample mean and covariance, over voxels, of the parameter estimates reported in Friston *et al.* (2000). Normally priors play a critical role in inference; indeed the traditional criticism leveled at Bayesian inference reduces to reservations about the validity of the priors employed. However, in the application considered here, this criticism can be discounted. This is because the priors, on those parameters about which inferences are made, are relatively flat. Only the five biophysical parameters have informative priors. In fact, in the limit of biophysical priors with zero variance, the procedure reduces to that used for conventional analyses (see Section 3.5) rendering the current scheme less dependent on the priors than conventional analyses. Given that only priors for the biophysical parameters are required these can be based on responses elicited by a single input.

In Friston *et al.* (2000) the parameters were identified as those that minimized the sum of squared differences between the Volterra kernels implied by the parameters and those derived directly from the data. This derivation used ordinary least square estimators, exploiting the fact that Volterra formulation Eq. (8) is linear in the unknowns, namely the kernel coefficients. The kernels can be thought of as a reparameterization of the model that does not refer to the underlying state representation. In other words, for every set of parameters there is a corresponding set of kernels (see Friston *et al.*, 2000, for the derivation of the kernels as a function of the parameters). The data and Volterra kernel estimation are described in detail in Friston *et al.* (1998). In brief, we obtained fMRI time series from a single subject at 2 T using a Magnetom Vision (Siemens, Erlangen, Germany) whole-body MRI system, equipped with a head volume coil. Multislice T_2^*-weighted fMRI images were obtained with a gradient echo-planar sequence using an axial slice orientation (TE = 40 ms, TR = 1.7 s, 64 × 64 × 16 voxels). After discarding initial scans (to allow for magnetic saturation effects) each time series comprised 1200 volume images with 3-mm isotropic voxels. The subject listened to monosyllabic or bisyllabic concrete nouns (i.e., "dog," "radio," "mountain," "gate") presented at five different rates (10, 15, 30, 60, and 90 words per minute) for epochs of 34 s, intercalated with periods of rest. The presentation rates were repeated according to a Latin Square design.

The distribution of the five biophysical parameters, over 128 voxels, was computed for the purposes of this paper to give their prior expectation η_θ and covariance C_θ. The expectations are reported in Friston *et al.* (2000): Signal decay κ_s had a mean of about 0.65 per second giving a half-life $t_{1/2} = \ln 2/\kappa_s \approx 1$ s consistent with spatial signaling with nitric oxide (Friston, 1995). Mean feedback rate κ_f was about 0.4 per second. The coupled differential equations (1) and (2) represent a damped oscillator with a resonance frequency of $\sqrt{\kappa_f + \kappa_s^2/4}/2\pi \approx 0.11$ per second. This is the frequency of the vasomotor signal that typically has a period of about 10 s (Mayhew *et al.*, 1998). Mean Transit time τ was 0.98 seconds. The transit time through the rat brain is roughly 1.4 s at rest and, according to the asymptotic projections for rCBF and volume, falls to 0.73 s during stimulation (Mandeville *et al.*, 1999). Under steady-state conditions Grubb's parameter α would be about 0.38. The mean over voxels was 0.326. This discrepancy, in relation to steady state levels, is anticipated by the Windkessel formulation and can be attributed to the fact that volume and flow are in a state of continuous flux during the evoked responses (Mandeville *et al.*, 1999). Mean resting oxygen extraction E_0 was about 34% and the range observed con-

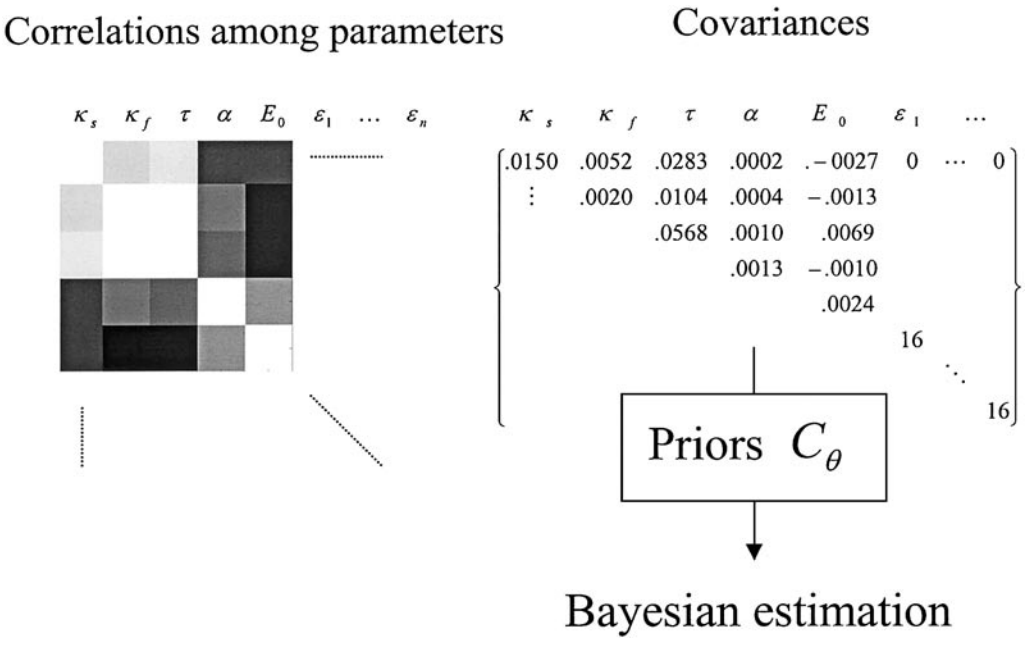

FIG. 3. Prior covariances for the five biophysical parameters of the hemodynamic model in Fig. 1. (Left) Correlation matrix showing the correlations among the parameters in image format (white = 1). (Right) Corresponding covariance matrix in tabular format. These priors represent the sample covariances of the parameters estimated by minimizing the difference between the Volterra kernels implied by the parameters and those estimated, empirically using ordinary least squares as described in Friston *et al.* (2000).

formed exactly with known values for resting oxygen extraction fraction (between 20 and 55%). Figure 3 shows the covariances among the biophysical parameters along with the correlation matrix (left-hand panel). The correlations suggest a high correlation between transit time and the rate constants for signal elimination and autoregulation.

The priors for the efficacies were taken to be relatively flat with an expectation of zero and a variance of 16 per second. The efficacies were assumed to be independent of the biophysical parameters with zero covariance. A variance of 16, or standard deviation of 4, corresponds to time constants in the range of 250 ms. In other words, inputs can elicit flow-inducing signal over wide range of time constants from infinitely slowly to very fast (250 ms) with about the same probability. A "strong" activation usually has an efficacy in the range of 0.5 to 0.6 per second. Notice that from a dynamical perspective "activation" depends upon the speed of the response not the percentage change. Equipped with these priors we can now pursue a fully Bayesian approach to estimating the parameters using new data sets and multiple input models.

3. SYSTEM IDENTIFICATION

3.1. Bayesian Estimation

This section describes Bayesian inference procedures for nonlinear observation models, with additive noise, of the form

$$y = h(\theta, u) + e \tag{9}$$

under Gaussian assumptions about the parameters θ and errors $e \sim N\{0, C_\epsilon\}$. These models can be adopted for any analytic dynamical system due to the existence of the equivalent Volterra series expansion in Eq. (8). Bayesian inference is based on the conditional probability of the parameters given the data $p(\theta|y)$. Assuming this *posterior* density is approximately Gaussian the problem reduces to finding its first two moments, the conditional mean $\eta_{\theta|y}$ and covariance $C_{\theta|y}$. We will denote the ith estimate of these moments by $\eta^{(i)}_{\theta|y}$ and $C^{(i)}_{\theta|y}$. Given the posterior density we can report its *mode*, i.e., the maximum *a posteriori* (MAP) estimate of the parameters (equivalent to $\eta_{\theta|y}$) or the probability that the parameters exceed some specified value. The posterior probability is proportional to the likelihood of obtaining the data, conditional on θ, times the prior probability of θ

$$p(\theta|y) \propto p(y|\theta)p(\theta), \tag{10}$$

where the Gaussian priors are specified in terms of their expectation η_θ and covariances C_θ as in the previous section. The likelihood can be approximated by expanding Eq. (9) about a working estimate of the conditional mean.

$$h(\theta, u) \approx h(\eta^{(i)}_{\theta|y}) + J(\theta - \eta^{(i)}_{\theta|y})$$
$$J = \frac{\partial h(\eta^{(i)}_{\theta|y})}{\partial \theta}. \tag{11}$$

Let $r = y - h(\eta^{(i)}_{\theta|y})$ such that $e \approx r - J(\theta - \eta^{(i)}_{\theta|y})$. Under Gaussian assumptions the likelihood and prior probabilities are given by

$$p(y|\theta) \propto \exp\{-\tfrac{1}{2}(r - J(\theta - \eta^{(i)}_{\theta|y}))^T \times C_\epsilon^{-1}(r - J(\theta - \eta^{(i)}_{\theta|y}))\}$$
$$p(\theta) \propto \exp\{-\tfrac{1}{2}(\theta - \eta_\theta)^T C_\theta^{-1}(\theta - \eta_\theta)\}. \tag{12}$$

Assuming the posterior density is approximately Gaussian, we can substitute 12 into 10 to give the posterior density

$$p(\theta|y) \propto \exp\{-\tfrac{1}{2}(\theta - \eta^{(i+1)}_{\theta|y})^T C^{-1}_{\theta|y}(\theta - \eta^{(i+1)}_{\theta|y})\}$$
$$C_{\theta|y} = (J^T C_\epsilon^{-1} J + C_\theta^{-1})^{-1}$$
$$\eta^{(i+1)}_{\theta|y} = \eta^{(i)}_{\theta|y} + C_{\theta|y}(J^T C_\epsilon^{-1} r + C_\theta^{-1}(\eta_\theta - \eta^{(i)}_{\theta|y})). \tag{13}$$

Equation (13) provides the basis for a recursive estimation of the conditional mean (and covariance) and corresponds to the **E**-step in an EM algorithm below. Equation (13) can be expressed in a more compact form by augmenting the residual data vector, design matrix, and covariance components

$$C_{\theta|y} = (\bar{J}^T \bar{C}_\epsilon^{-1} \bar{J})^{-1}$$
$$\eta^{(i+1)}_{\theta|y} = \eta^{(i)}_{\theta|y} + C_{\theta|y}(\bar{J}^T \bar{C}_\epsilon^{-1} \bar{y}) \tag{13a}$$

where

$$\bar{y} = \begin{bmatrix} y - h(\eta^{(i)}_{\theta|y}) \\ \eta_\theta - \eta^{(i)}_{\theta|y} \end{bmatrix}, \quad \bar{J} = \begin{bmatrix} J \\ I \end{bmatrix}, \quad \bar{C}_\epsilon = \begin{bmatrix} C_\epsilon & 0 \\ 0 & C_\theta \end{bmatrix}.$$

Equations (13) and (13a) are exactly the same but Eq. (13a) adopts the same formulation used in Friston *et al.* (2002a) to derive and explain the EM algorithm below. From the perspective of Friston *et al.* (2002a) the present problem represents a single-level hierarchical observation model with known priors.

The starting estimate of the conditional mean is generally taken to be the prior expectation. If Eq. (9) were linear, i.e., $h(\theta) = H\theta \Rightarrow J = H$, Eq. (13) would converge after a single iteration. However, when h is nonlinear J becomes a function of the conditional mean and several iterations are required. Note that in the absence of any priors, iterating Eq. (13) is formally identical to the Gauss–Newton method of parameter estimation. Furthermore, when h is linear, and the priors are flat, Eq. (13) reduces to the classical Gauss–Markov estimator, the minimum variance, linear maximum-likelihood estimator of the parameters.

The conditional covariance of the parameters is assumed to be Gaussian. The validity of this assumption depends on the rate of convergence of the Taylor expansion of h in Eq. (11). Because h is nonlinear the likelihood density will only be approximately Gaussian. However, the posterior or conditional density will be almost Gaussian, given a sufficiently long time series (Fahrmeir and Tutz, 1994, p. 58).

3.2. Covariance Component Estimation

So far we have assumed that the error covariance C_e is known. Clearly in many situations (e.g., serial correlations in fMRI) it is not. When the error covariance is unknown, it can be estimated through some hyperparameters λ_j, where $C_e = \sum \lambda_j Q_j$. $Q_j = \partial C_\epsilon / \partial \lambda_j$ represents a basis set that embodies the form of the variance components and could model different variances for different blocks of data or indeed different forms of serial correlations within blocks. Restricted Maximum likelihood (ReML) estimators of λ_j maximise the log likelihood $\log p(y|\lambda) = F(\lambda)$. This log likelihood obtains by integrating over the conditional distribution of the parameters as described in Neal and Hinton (1998). Under a Fisher-scoring scheme (see Friston *et al.*, 2002; Harville, 1977) this gives

$$\lambda^{(i+1)} = \lambda^{(i)} - \left\langle \frac{\partial^2 F}{\partial \lambda^2} \right\rangle^{-1} \frac{\partial F}{\partial \lambda}$$
$$\frac{\partial F}{\partial \lambda_j} = -\tfrac{1}{2} tr\{PQ_i\} + \tfrac{1}{2} \bar{y}^T P^T Q_i P \bar{y}$$
$$\left\langle \frac{\partial^2 F}{\partial \lambda^2_{jk}} \right\rangle = -\tfrac{1}{2} tr\{PQ_i PQ_j\}$$
$$P = \bar{C}_\epsilon^{-1} - \bar{C}_\epsilon^{-1} \bar{J} C_{\theta|y} \bar{J}^T \bar{C}_\epsilon^{-1}. \tag{14}$$

Although this expression may look complicated, in practice it is quick to implement due to the sparsity structure of the covariance basis set Q_j. If the basis set is the identity matrix, embodying *i.i.d.* assumptions about the errors, Eq. (14) is equivalent to the sum of squared residual estimator used in classical analysis of variance.

3.3. An EM Gauss–Newton Search

Recursive implementation of Eqs. (13) and (14) corresponds to an expectation maximisation or EM algorithm. The *E*-step (expectation step) computes the conditional expectations and covariances according to Eq. (13) using the error covariances specified by the hyper-

parameters from the previous **M**-step. Equation (14) corresponds to the **M**-step (maximum-likelihood step) that updates the ReML estimates of the hyperparameters by integrating the parameters out of the log-likelihood function using their conditional distribution from the **E**-step (Harville, 1977; Dempster *et al.*, 1977; Neal and Hinton, 1998). The ensuing EM algorithm is derived in full in Friston *et al.* (2002) and can be summarized in pseudo-code as

Initialize

$$\lambda^{(1)} = \lambda^{(0)}$$
$$\eta^{(1)}_{\theta|y} = \eta_\theta$$

Until convergence {

E-step

$$J = \partial h(\eta^{(i)}_{\theta|y})/\partial\theta$$

$$\bar{y} = \begin{bmatrix} y - h(\eta^{(i)}_{\theta|y}) \\ \eta_\theta - \eta^{(i)}_{\theta|y} \end{bmatrix},\ \bar{J} = \begin{bmatrix} J \\ I \end{bmatrix},\ \bar{C}_\epsilon = \begin{bmatrix} \sum \lambda^{(i)}_i Q_i & 0 \\ 0 & C_\theta \end{bmatrix}$$

$$C_{\theta|y} = (\bar{J}^T \bar{C}_\epsilon^{-1} \bar{J})^{-1}$$

$$\eta^{(i+1)}_{\theta|y} = \eta^{(i)}_{\theta|y} + C_{\theta|y}(\bar{J}^T \bar{C}_\epsilon^{-1} \bar{y})$$

M-step (15)

$$P = \bar{C}_\epsilon^{-1} - \bar{C}_\epsilon^{-1} \bar{J} C_{\theta|y} \bar{J}^T \bar{C}_\epsilon^{-1}$$

$$\frac{\partial F}{\partial \lambda_j} = -\tfrac{1}{2}\, tr\{PQ_i\} + \tfrac{1}{2}\, \bar{y}^T P^T Q_i P \bar{y}$$

$$\left\langle \frac{\partial^2 F}{\partial \lambda_{jk}^2} \right\rangle = -\tfrac{1}{2}\, tr\{PQ_i PQ_j\}$$

$$\lambda^{(i+1)} = \lambda^{(i)} - \left\langle \frac{\partial^2 F}{\partial \lambda^2} \right\rangle^{-1} \frac{\partial F}{\partial \lambda}$$

The convergence criterion, we used, is that the sum of squared change in conditional means falls below 10^{-6}.

This EM scheme is effectively a *Gauss–Newton* search for the posterior mode or MAP estimate of the parameters. A Gauss–Newton search can be regarded as a Newton–Raphson method, where the second derivative or curvature of the log-likelihood function is approximated by neglecting terms that involve the residuals r, which are assumed to be small. The relationship between the **E**-step and a conventional Gauss–Newton ascent can be seen easily in terms of the derivatives of their respective objective functions. For conventional Gauss–Newton this function is the *log likelihood*

$$l = \ln p(y|\theta)$$
$$= -\tfrac{1}{2}(y - h(\theta))^T C_\epsilon^{-1}(y - h(\theta)) + const.$$
$$\frac{\partial l}{\partial \theta}(\eta^{(i)}_{ML}) = J^T C_\epsilon^{-1} r \qquad (16)$$
$$\frac{\partial^2 l}{\partial \theta^2}(\eta^{(i)}_{ML}) \approx J^T C_\epsilon^{-1} J$$
$$\eta^{(i+1)}_{ML} = \eta^{(i)}_{ML} + (J^T C_\epsilon^{-1} J)^{-1} J^T C_\epsilon^{-1} r.$$

This is a conventional Gauss–Newton scheme. By simply augmenting the log likelihood with the log prior we get the *log posterior*

$$l = \ln p(\theta|y) = \ln p(y|\theta) + \ln p(\theta)$$
$$= -\tfrac{1}{2}(y - h(\theta))^T C_\epsilon^{-1} y - h(\theta)$$
$$-\tfrac{1}{2}(\eta_\theta - \theta)^T C_\theta^{-1}(\eta_\theta - \theta) + const.$$
$$\frac{\partial l}{\partial \theta}(\eta^{(i)}_{\theta|y}) = J^T C_\epsilon^{-1} r + C_\theta^{-1}(\eta_\theta - \eta^{(i)}_{\theta|y}) \qquad (17)$$
$$\frac{\partial^2 l}{\partial \theta^2}(\eta^{(i)}_{\theta|y}) \approx J^T C_\epsilon^{-1} J + C_\theta^{-1}$$
$$\eta^{(i+1)}_{\theta|y} = \eta^{(i)}_{\theta|y} + (J^T C_\epsilon^{-1} J + C_\theta^{-1})^{-1}$$
$$\times (J^T C_\epsilon^{-1} r + C_\theta^{-1}(\eta_\theta - \eta^{(i)}_{\theta|y})),$$

which is identical to the expression for the conditional expectation in the **E**-step. In fact Eq. (17) serves as an alternative derivation of the conditional mean in Eq. (13). Equation (17) is not sufficient for EM because the conditional covariance in Eq. (13) is required in the **E**-step, to provide the conditional density for the **M**-step. However, Eq. (17) does highlight the fact that the conditional covariance is approximately the inverse of the log posterior curvature.

An intuitive understanding of the **E**-step's update equation (formulated by one of the reviewers) is that the change in the conditional estimate is driven by two terms. The first $J^T C_\epsilon^{-1} r$ ensures a minimization of the residuals and the second $C_\theta^{-1}(\eta_\theta - \eta^{(i)}_{\theta|y})$ a minimization of the difference between the prior expectation and posterior estimate. The relative strength of these terms is moderated by the precisions with which the measurements are made and with which the priors are specified. If the error variance is small, relative to the prior variability, more weight is given to minimising the residuals and vice versa.

In summary, the only difference between the **E**-step and a conventional Gauss–Newton search is that priors are included in the objective log probability function converting it from a log likelihood into a log pos-

terior. The use of an EM algorithm rests upon the need to find not only the conditional density but also the hyperparameters of unknown variance components. The **E**-step finds (i) the current MAP estimate that provides the next expansion point for the Gauss–Newton search and (ii) the conditional covariance required by the **M**-step. The **M**-step then updates the ReML estimates of the covariance hyperparameters that are required to compute the conditional moments in the **E**-step. Technically Eq. (15) is a *generalized* EM (GEM) because the **M**-step increases the log likelihood of the hyperparameter estimates, as opposed to maximising it.

3.4. Relationship to Established Procedures

The procedure presented above represents a fairly obvious extension to conventional Gauss–Newton searches for the parameters of nonlinear observation models. The extension has two components: (i) First, maximization of the *posterior* density that embodies priors, as opposed to the likelihood. This allows for the incorporation of prior information into the solution and ensures uniqueness and convergence. (ii) Second, the estimation of unknown covariance components. This is important because it accommodates nonsphericity in the error terms. The overall approach engenders a relatively simple way of obtaining Bayes estimators for nonlinear systems with unknown additive observation error. Technically, the algorithm represents a *posterior mode estimation* for nonlinear observation models using EM. It can be regarded as approximating the posterior density of the parameters by replacing the conditional mean with the mode and the conditional precision with the curvature (at the current expansion point). Covariance hyperparameters are then estimated, which maximize the expectation of the log likelihood of the data over this approximate posterior density.

Posterior mode estimation is an alternative to full posterior density analysis, which avoids numerical integration (Fahrmeir and Tutz, 1994, p. 58) and has been discussed extensively in the context of *generalized linear models* (e.g., Leonard, 1972; Santner and Duffy, 1989). The departure from Gaussian assumptions in generalized linear models comes from non-Gaussian likelihoods, as opposed to nonlinearities in the observation model considered here, but the issues are similar. Posterior mode estimation usually assumes the error covariances and priors are known. If the priors are unknown constants then empirical Bayes can be employed to estimate the required hyperparameters. Fahrmeir and Tutz (1994, p. 59) discuss the use of an *EM-type algorithm* in which the posterior means and covariances appearing in the **E**-step are replaced by the posterior modes and curvatures (cf. Eq. (15)). Since C_θ^{-1} appears only in the log prior (see Eq. (17)) this leads to a simple EM scheme for generalized linear models, where the hyperparameters maximize the expected log prior. In this paper, we have dealt with the more general nonlinear problem in which the hyperparameters influence the likelihood, leading to the GEM scheme above.

It is important not to confuse this application of EM with Kalman filtering. Although Kalman filtering can be formulated in terms of EM and, indeed, posterior mode estimation, Kalman filtering is used with completely different observation models—*state–space models.* State space or dynamic models comprise a *transition* equation and an *observation* equation (cf. the state equation and output nonlinearity in Eq. (7)) and cover systems in which the underlying state is hidden and is treated as a stochastic variable. This is not the sort of model considered this paper, in which the inputs (experimental design) and the ensuing states are known. This means that the conditional densities can be computed for the entire time series simultaneously (Kalman filtering updates the conditional density recursively, by stepping through the time series). If we treated the inputs as unknown and random then the state equation of Eq. (7) could be rewritten as a stochastic differential equation (SDE) and a transition equation derived from it, using local linearity assumptions. This would form the basis of a state–space model. This approach may be useful for accommodating deterministic noise in the hemodynamic model but, in this treatment, we consider the inputs to be fixed. This means that the only random effects enter at the level of the observation or output nonlinearity. In other words, we are assuming that the measurement error in fMRI is the principal source of randomness in our measurements and that hemodynamic responses per se are determined by known inputs. This is the same assumption used in conventional analyses of fMRI data (see Section 3.5).

3.4. A Note on Integration

To iterate Eq. (15) the local gradients $J = \partial h(\eta_{\theta|y}^{(i)})/\partial\theta$ must be evaluated. This involves evaluating $h(\theta, u)$ around the current expansion point with the generalized convolution of the inputs for the current conditional parameter estimates according to Eq. (8) or, equivalently, the integration of Eq. (7). The latter can be accomplished efficiently by capitalizing on the fact that stimulus functions are usually sparse. In other words inputs arrive as infrequent events (e.g., event-related paradigms) or changes in input occur sporadically (e.g., boxcar designs). We can use this to evaluate $y(t) = h(\eta_{\theta|y}^{(i)}, u)$ at the times the data were sampled using a bilinear approximation to Eq. (7).

The Taylor expansion of $\dot{X}(t)$ about $X(0) = X_0 = [0, 1, 1, 1]^T$ and $u(t) = 0$

$$\dot{X}(t) \approx f(X_0, 0) + \frac{\partial f(X_0, 0)}{\partial X}(X - X_0) + \sum_i u(t)_i \left(\frac{\partial^2 f(X_0, 0)}{\partial X \partial u_i}(X - X_0) + \frac{\partial f(X_0, 0)}{\partial u_i} \right)$$

has a bilinear form, following a change of variables (equivalent to adding an extra state variable $x_0(t) = 1$)

$$\begin{aligned} \dot{\tilde{X}}(t) &\approx A\tilde{X} + \sum_i u(t)_i B_i \tilde{X} \\ \tilde{X} &= \begin{bmatrix} 1 \\ X \end{bmatrix} \\ A &= \begin{bmatrix} 0 & 0 \\ \left(f(X_0, 0) - \dfrac{\partial f(X_0, 0)}{\partial X} X_0 \right) & \dfrac{\partial f(X_0, 0)}{\partial X} \end{bmatrix} \\ B_i &= \begin{bmatrix} 0 & 0 \\ \left(\dfrac{\partial f(X_0, 0)}{\partial u_i} - \dfrac{\partial^2 f(X_0, 0)}{\partial X \partial u_i} X_0 \right) & \dfrac{\partial^2 f(X_0, 0)}{\partial X \partial u_i} \end{bmatrix}. \end{aligned} \tag{18}$$

This bilinear approximation is important because the Volterra kernels of bilinear systems have closed-form expressions. This means that the kernels can be derived analytically, and quickly, to provide a characterization of the impulse response properties of the system (see Section 4). The integration of Eq. (18) is predicated on its solution over periods $\Delta t_k = t_{k+1} - t_k$ within which the inputs are constant.

$$\begin{aligned} \tilde{X}(t_{k+1}) &\approx \exp\{\Delta t_k (A + \sum_i u(t_k)_i B_i)\} \tilde{X}(t_k) \\ y(t_{k+1}) &\approx \lambda(X(t_{k+1})). \end{aligned} \tag{19}$$

Equation (19) can be used to evaluate the state variables in a computationally expedient manner at every time t_{k+1} input changes or the response variable is measured, using the values from the previous time point t_k. After applying the output nonlinearity the resulting values enter into the numerical evaluation of the partial derivatives $J = \partial h(\eta_{\theta|y}^{(n)})/\partial\theta$ in the EM algorithm above. This quasi-analytical integration scheme can be 1 order of magnitude quicker than straightforward numerical integration, depending on the sparsity of inputs.

3.5. *Relation to Conventional fMRI Analyses*

Note that if we treated the five biophysical parameters as known canonical values and discounted all but the first-order terms in the Volterra expansion Eq. (8) the following linear model would result

$$\begin{aligned} h(u, \theta) &= \kappa_0 + \sum_{i=1}^{n} \int_0^t \kappa_1(\sigma) u(t - \sigma)_i d\sigma \\ &= \sum_{i=1}^{n} \kappa_1 * u(t)_i \\ &\approx \kappa_0 + \sum_{i=1}^{n} \left(\frac{\partial \kappa_1}{\partial \epsilon_i} * u(t)_i \right) \epsilon_i, \end{aligned} \tag{20}$$

where $*$ denotes convolution and the second expression is a first-order Taylor expansion around the expected values of the parameters.[2] Substituting this into Eq. (9) gives the general linear model adopted in conventional analysis of fMRI time series, if we elect to use just one (canonical) hemodynamic response function (*hrf*) to convolve our stimulus functions with. In this context the canonical *hrf* plays the role of $\partial\kappa_1/\partial\epsilon_i$ in Eq. (20). This partial derivative is shown in Fig. 4 (top) using the prior expectations of the parameters and conforms closely to the sort of *hrf* used in practice. Now by treating the efficacies as fixed effects (i.e., with flat priors) the MAP and ML estimators reduce to the same thing and the conditional expectation reduces to the Gauss–Markov estimator

$$\eta_{ML} = (J^T C_\epsilon^{-1} J)^{-1} J^T C_\epsilon^{-1} y,$$

where J is the design matrix. This is precisely the estimator used in conventional analyses when whitening strategies are employed.

Consider now the second-order Taylor approximation to Eq. (20) that obtains when we do not know the exact values of the biophysical parameters and they are treated as unknown

$$h(\theta, u) \approx \kappa_0 + \sum_{i=1}^{n} \left[\left(\frac{\partial \kappa_1}{\partial \epsilon_i} * u(t)_i \epsilon_i + \frac{1}{2} \sum_{j=1}^{5} \left(\frac{\partial^2 \kappa_1}{\partial \epsilon_i \partial \theta_j} * u(t)_i \right) \epsilon_i \theta_j \right) \right]. \tag{21}$$

This expression[3] is precisely the general linear model proposed in Friston *et al.* (1998) and implemented in SPM99: In this instance the explanatory variables comprise the stimulus functions, each convolved with a small temporal basis set corresponding to the canonical $hrf = \partial\kappa_1/\partial\epsilon_i$ and its partial derivatives with respect to

[2] Note that in this first-order Taylor approximation $\kappa_1 = 0$ when expanding around the prior expectations of the efficacies = 0. Furthermore, all the first-order partial derivatives $\partial\kappa_1/\partial\theta_i = 0$ unless they are with respect to an efficacy.

[3] Note that in this second-order Taylor approximation all the second-order partial derivatives $\partial^2\kappa_1/\partial\theta_i\partial\theta_j = 0$ unless they are with respect to an efficacy and one of the biophysical parameters.

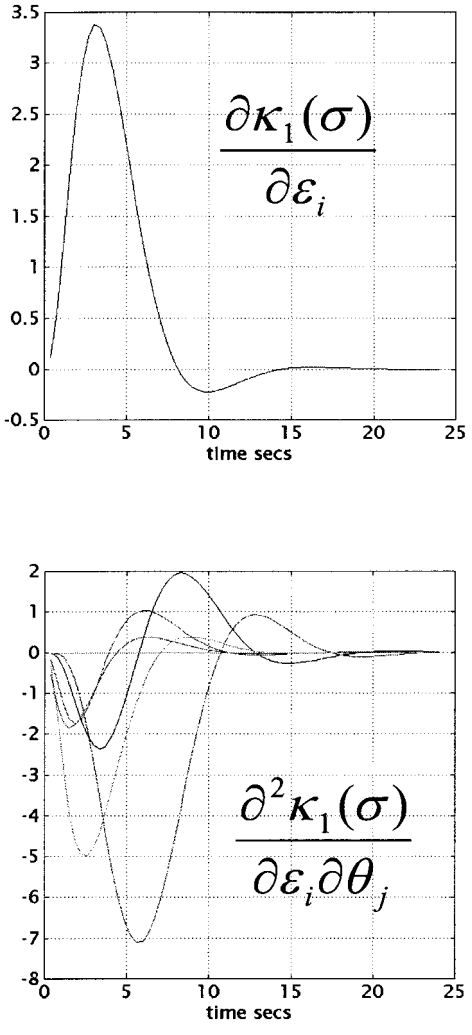

FIG. 4. Partial derivatives of the kernels with respect to parameters of the model evaluated at their prior expectation. (Top) First-order partial derivative with respect to efficacy. (Bottom) Second-order partial derivatives with respect to efficacy and the biophysical parameters. When expanding around the prior expectations of the efficacies = 0 the remaining first- and second-order partial derivatives with respect to the parameters are zero.

the biophysical parameters. Examples of these second-order partial derivatives are provided in the bottom panel of Fig. 4. The unknowns in this general linear model are the efficacies ϵ_i and the interaction between the efficacies and the biophysical parameters $\epsilon_i\theta_j$. Of course, the problem with this linearized approximation is that any generalized least squares estimates of the unknown coefficients $\beta = [\epsilon_1, \ldots, \epsilon_n, \epsilon_1\theta_1, \ldots, \epsilon_n\theta_1, \epsilon_1\theta_2, \ldots]^T$ are not constrained to factorize into stimulus-specific efficacies ϵ_i and biophysical parameters θ_j that are the same for all inputs. Only a nonlinear estimation procedure can do this.

In the usual case of using a temporal basis set (e.g., a canonical form and various derivatives) one obtains a ML or generalized least squares estimate of (functions of) the parameters in some subspace defined by the basis set. Operationally this is like specifying priors but of a very particular form. This form can be thought of as uniform priors over the support of the basis set and zero elsewhere. In this sense basis functions implement hard constraints that may not be very realistic but provide for efficient estimation. The soft constraints implied by the Gaussian priors in the EM approach are more plausible but are computationally more expensive to implement.

In summary this section has described an EM algorithm that can be viewed as a Gauss–Newton search for the conditional distribution of the parameters of deterministic dynamical system, with additive Gaussian error. It was shown that classical approaches to fMRI data analysis are special cases that ensue when considering only first-order kernels and adopting flat or uninformative priors. Put another way the proposed scheme can be regarded as a generalization of existing procedures that is extended in two important ways. (i) First the model encompasses nonlinearities and (ii) second it moves the estimation from a classical into a Bayesian frame.

4. AN EMPIRICAL ILLUSTRATION

4.1. Single-Input Example

In this, the first of the two examples, we revisit the original data set on which the priors were based. This constitutes a single-input study where the input corresponds to the aural presentation of single words, at different rates, over epochs. The data were subject to a conventional event-related analysis where the stimulus function comprised trains of spikes indexing the presentation of each word. The stimulus function was convolved with a canonical *hrf* and its temporal derivative. The data were high-pass filtered by removing low-frequency components modelled by a discrete cosine set. The resulting SPM{T}, testing for activations due to words, is shown in Fig. 5 (left) thresholded at $P = 0.05$ (corrected).

A single region in the left superior temporal gyrus was selected for analysis. The input comprised the same stimulus function used in the conventional analysis and the output was the first eigenvariate of high-pass filtered time series, of all voxels, within a 4-mm sphere, centered on the most significant voxel in the SPM{T} (marked by an arrow in Fig. 5). The error covariance basis set Q comprised two bases: an identity matrix modeling white or an *i.i.d.* component and a second with exponentially decaying off-diagonal elements modeling an AR(1) component (see Friston *et al.*, 2002b). This models serial correlations among the errors. The results of the estimation procedure are shown in the right-hand panel in terms of (i) the conditional distribution of the parameters and (ii) the conditional

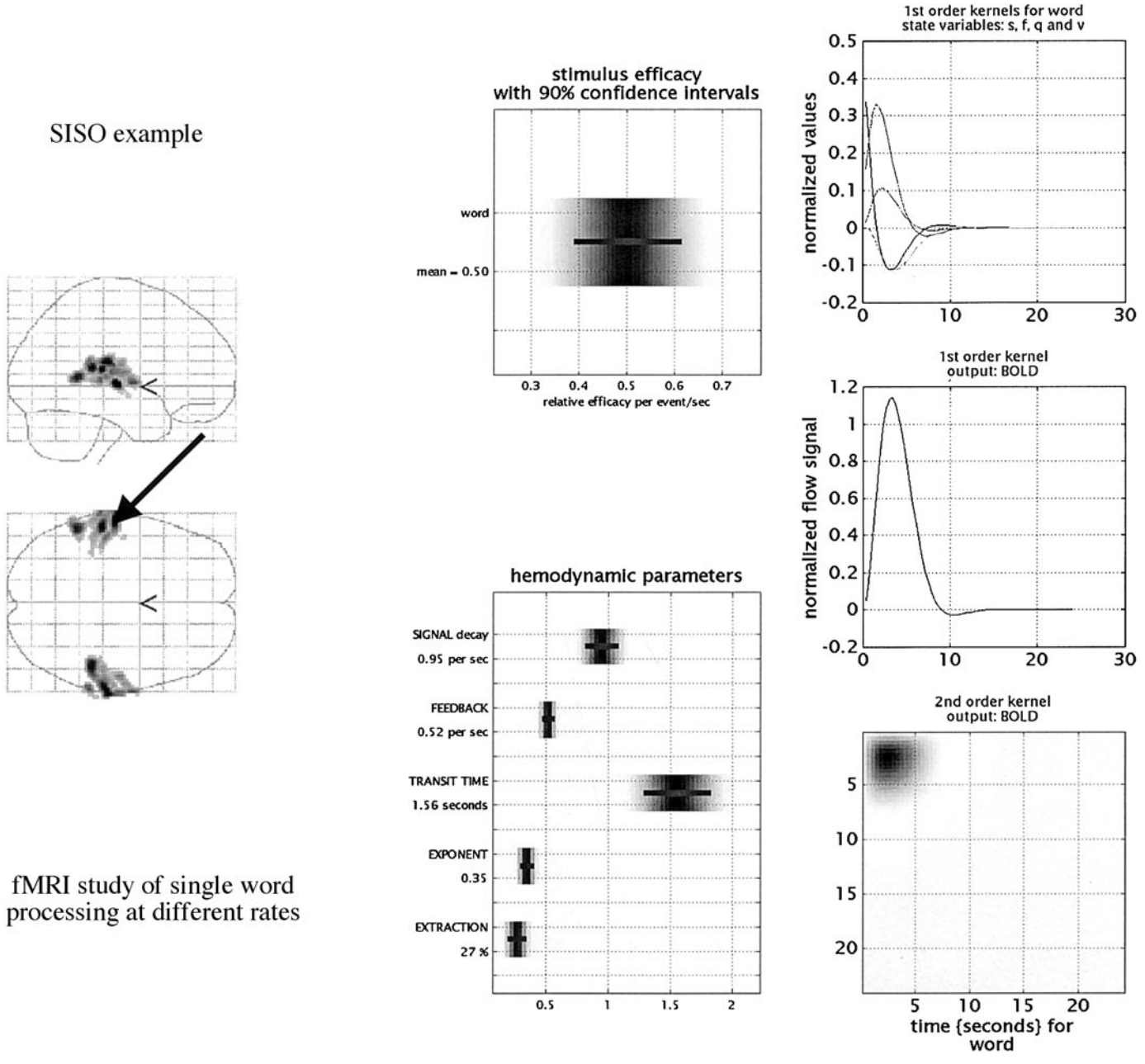

FIG. 5. A SISO example: (Left) Conventional SPM{T} testing for an activating effect of word presentation. The arrow shows the centre of the region (a sphere of 4-mm radius) whose response was entered into the Bayesian estimation procedure. The results for this region are shown in the right-hand panel in terms of (i) the conditional distribution of the parameters and (ii) the conditional expectation of the first- and second-order kernels. The top right panel shows the first-order kernels for the state variables (signal, inflow, deoxyhemoglobin content, and volume). The first- and second-order output kernels for the BOLD response are shown in the bottom right panels. The left-hand panels show the conditional or posterior distributions. That for efficacy is presented in the top panel, and those for the five biophysical parameters, in the bottom panel. The shading corresponds to the probability density and the bars to 90% confidence intervals.

expectation of the first- and second-order kernels. The kernels are a function of the parameters and their derivation using a bilinear approximation is described in Friston *et al.* (2000). The top right panel shows the first-order kernels for the state variables (signal, inflow, deoxyhemoglobin content, and volume). These can be regarded as impulse response functions detailing the response to a transient input. The first- and second-order output kernels for the BOLD response are shown in the bottom right panels. They concur with those derived empirically in Friston *et al.* (2000). Note the characteristic undershoot in the first-order kernel

and the pronounced negativity in the top left of the second-order kernel, flanked by two off-diagonal positivities at around 8 s. These lend the hemodynamics a degree of refractoriness when presenting paired stimuli less than a few seconds apart and a superadditive response with about 8 s separation. The left-hand panels show the conditional or posterior distributions. The density for the efficacy is presented in the top panel and those for the five biophysical parameters are shown in the bottom panel using the same format. The shading correspond to the probability density and the bars to 90% confidence intervals. The values of the biophysical parameters are all within a very acceptable range. In this example the signal elimination and decay appears to be slower than normally encountered, with the rate constants being significantly larger than their prior expectations. Grubb's exponent here is closer to the steady state value of 0.38 than the prior expectation of 0.32. Of greater interest is the efficacy. It can be seen that the efficacy lies between 0.4 and 0.6 and is clearly greater than 0. This would be expected given we chose the most significant voxel from the conventional analysis. Notice there is no null hypothesis here and we do not even need a *P* value to make the inference that words evoke a response in this region. The nature of Bayesian inference is much more straightforward and as discussed in Friston *et al.* (2002a) is relatively immune from the multiple comparison problem. An important facility, with inferences based on the conditional distribution and precluded in classical analyses, is that one can infer a cause did not elicit a response. This is demonstrated in the second example.

4.2. Multiple Input Example

In this example we turn to a new data set, previously reported in Büchel and Friston (1998) in which there are three experimental causes or inputs. This was a study of attention to visual motion. Subjects were studied with fMRI under identical stimulus conditions (visual motion subtended by radially moving dots) while manipulating the attentional component of the task (detection of velocity changes). The data were acquired from normal subjects at 2-T using a Magnetom Vision (Siemens) whole-body MRI system, equipped with a head volume coil. Here we analyze data from the first subject. Contiguous multislice T_2^*-weighted fMRI images were obtained with a gradient echo-planar sequence (TE = 40 ms, TR = 3.22 s, matrix size = 64 × 64 × 32, voxel size 3 × 3 × 3 mm). Each subject had four consecutive 100-scan sessions comprising a series of 10-scan blocks under five different conditions D F A F N F A F N S. The first condition (D) was a dummy condition to allow for magnetic saturation effects. F (Fixation) corresponds to a low-level baseline where the subjects viewed a fixation point at the center of a screen. In condition A (Attention) subjects viewed 250 dots moving radially from the center at 4.7° per second and were asked to detect changes in radial velocity. In condition N (No attention) the subjects were asked simply to view the moving dots. In condition S (Stationary) subjects viewed stationary dots. The order of A and N was swapped for the last two sessions. In all conditions subjects fixated the centre of the screen. In a prescanning session the subjects were given five trials with five speed changes (reducing to 1%). During scanning there were no speed changes. No overt response was required in any condition.

This design can be reformulated in terms of three potential causes, photic stimulation, visual motion, and directed attention. The F epochs have no associated cause and represent a baseline. The S epochs have just photic stimulation. The N epochs have both photic stimulation and motion whereas the A epochs encompass all three causes. We performed a conventional analysis using boxcar stimulus functions encoding the presence or absence of each of the three causes during each epoch. These functions were convolved with a canonical *hrf* and its temporal derivative to give two repressors for each cause. The corresponding design matrix is shown in the left panel of Fig. 6. We selected a region that showed a significant attentional effect in the lingual gyrus for Bayesian inference. The stimulus functions modeling the three inputs were the box functions used in the conventional analysis. The output corresponded to the first eigenvariate of high-pass filtered time series from all voxels in a 4-mm sphere centered on 0, −66, −3 mm (Talairach and Tournoux, 1998). The error covariance basis set was simply the identity matrix.[4] The results are shown in the right-hand panel of Fig. 6 using the same format as Fig. 5. The critical thing here is that there are three conditional densities, one for each of the input efficacies. Attention has a clear activating effect with more than a 90% probability of being greater than 0.25 per second. However, in this region neither photic stimulation per se or motion in the visual field evokes any real response. The efficacies of both are less than 0.1 and are centered on 0. This means that the time constants of the response to visual stimulation would range from about 10 s to never. Consequently these causes can be discounted from a dynamical perspective. In short this visually unresponsive area responds substantially to attentional manipulation *showing a true functional selectivity.* This is a crucial statement because classical inference does not allow one to infer any region does

[4] We could motivate this by noting the TR is considerably longer in these data than in the previous example. However, in reality, serial correlations were ignored because the loss of sparsity in the associated inverse covariance matrices considerably increases computation time and we wanted to repeat the analysis many times (see next subsection).

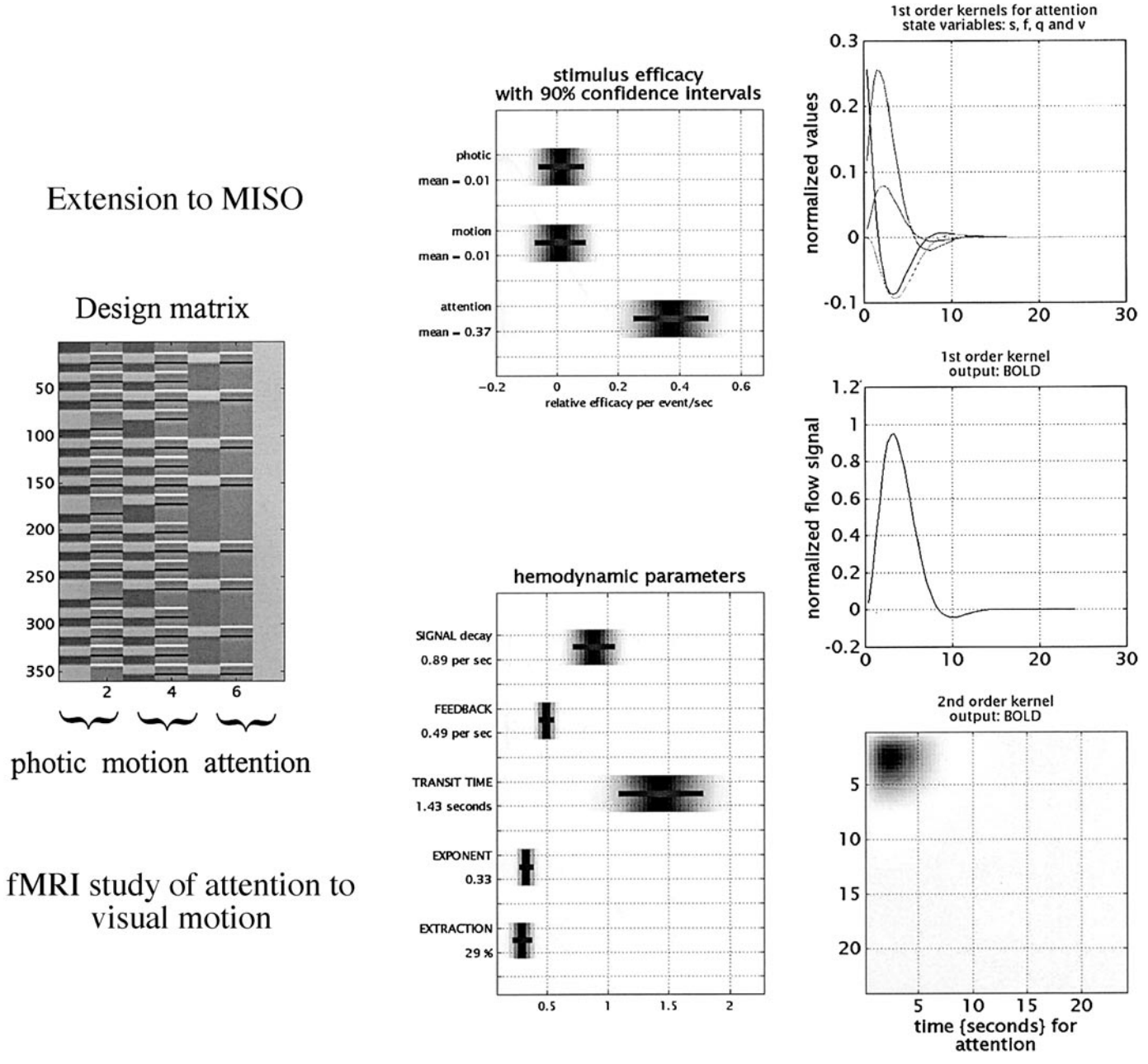

FIG. 6. A MISO example using visual attention to motion. (Left) The design matrix used in the conventional analysis; (right) the results of the Bayesian analysis of a lingual extrastriate region. This panel has the same format as Fig. 5.

not respond and therefore precludes a formal inference about the selectivity of regional responses. The only reason one can say "this region responds *selectively* to attention" is because Bayesian inference allows one to say "it does *not* respond to photic stimulation with random dots or motion."

4.3. Posterior Probabilities

Given the conditional densities we can compute the posterior probability that the efficacy for any input exceeds some specified threshold γ. This posterior probability is a function of the threshold chosen and the conditional moments

$$1 - \Phi\left(\frac{\gamma - c_i^T \eta_{\theta|y}}{\sqrt{c_i^T C_{\theta|y} c_i}}\right), \tag{22}$$

where Φ denotes the cumulative density function for the unit normal distribution and c is a vector of contrast weights specifying the linear compound of param-

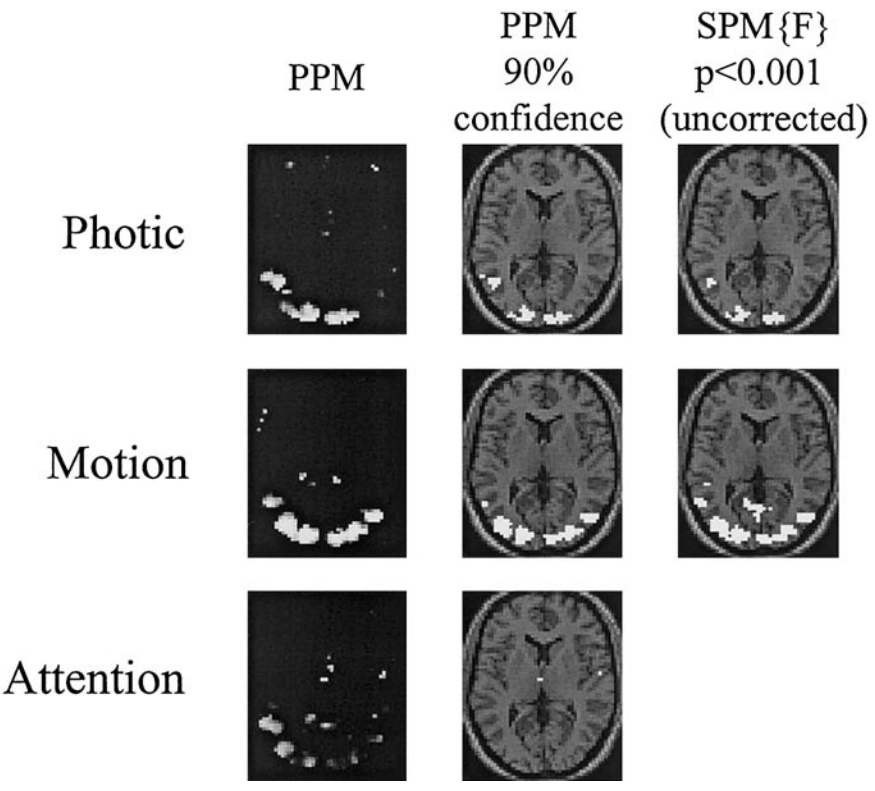

FIG. 7. Posterior probability maps (PPMs) for the study of attention to visual motion. The left-hand column shows the raw PPMs and the middle column after thresholding at 0.9. Voxels surviving this confidence threshold are those in which one can infer, with at least 90% confidence, that the parameters are greater than 0.1 per second. The right-hand column contains the equivalent SPM{F}s from a conventional analysis using the design matrix in Fig. 6 and thresholded at $P = 0.001$ (uncorrected).

eters one wants to make an inference about. For example c_{photic} would be a vector with zeros for all conditional estimators apart from the efficacy mediating photic input, where it would be one. Posterior probabilities were computed for all voxels in a slice through visually response areas ($z = 6$ mm) for each of the three efficacies. The resulting PPMs are shown in Fig. 7 for a threshold of 0.1 per second. The left-hand column shows the PPMs per se and the middle column shows them after thresholding at 0.9. Voxels surviving this confidence threshold are those in which one can infer, with at least 90% confidence, that the parameters are greater than 0.1 per second. In this slice attention has little effect with a few voxels in the mediodorsal thalamus. Conversely, photic stimulation and motion excite large areas of responses in striate and extrastriate areas. Interestingly motion appears to be more effective particularly around the V5 complex and pulvinar. For comparison the SPM{F}s from the equivalent classical analysis are shown on the right. There is a remarkable concordance between the PPMs and SPMs (note that the SPM{F} shows deactivations as well as activations). However, we have deliberately chosen a threshold that highlights the similarities. Had we chosen a corrected threshold, the PPMs would be (apparently) more sensitive. PPMs represent a potentially useful way of characterizing activation profiles of this sort.

5. CONCLUSION

In this paper we have presented a method that conforms to an EM implementation of the Gauss–Newton method, for estimating the conditional or posterior distribution of the parameters of a deterministic dynamical system. The inclusion of priors in the estimation procedure ensures robust and rapid convergence, and the resulting conditional densities enable Bayesian inference about the model's parameters. We have examined the coupling between experimentally designed causes or factors in fMRI studies and the ensuing BOLD response. This application represents a generalization of existing linear models to accommodate nonlinearities in the transduction of experimental causes to measured output in fMRI. Because the model is predicated on biophysical processes the parameters have a physical interpretation. Furthermore the approach extends classical inference about the likelihood

of the data, to more plausible inferences about the parameters of the model given the data. This inference provides confidence intervals based on the conditional density.

5.1. Limitations

The validity of any Bayesian inference rests upon the validity of (i) the priors and (ii) the model adopted. In this work the former concern is ameliorated by virtue of the fact that inferences are restricted to those parameters that have relatively uninformative priors. Indeed in the limit of flat priors, for the efficacies, one would revert to classical inference. The priors on the remaining biophysical parameters clearly play a role, similar to that played by temporal basis functions in conventional analyses. The more valid they are, the better the model fit and the smaller the error variance. It should be acknowledged that the priors used in this paper are based on the distribution over voxels in a single subject. Clearly is would be better to use the distribution over subjects in a single voxel. We chose this single-subject data set because the experimental paradigm and acquisition parameters were explicitly chosen to ensure robust estimation of physiological parameters (i.e., short TR, restricted field of view, covering regions known to be aurally responsive, and comprising long time series). This experiment was also repeated using the same subject and emulated fMRI stimulus-delivery conditions to ensure the linearity of the stimulus to rCBF coupling (Rees *et al.*, 1996). However, it is anticipated that the priors will be refined as more analyses of the sort proposed here are performed. One particular concern is that the correlations among the biophysical parameters may reflect artifacts due to things like slice timing in sequential acquisition; i.e., using the distribution over voxels means that the data sequences and stimulus functions show a variable temporal relationship from voxel to voxel, which may influence the parameter estimates used to construct the priors. Because this influence is correlated over voxels, spurious correlations may be induced in voxel to voxel estimates. One reason for using the data reported in Friston *et al.* (1998) was that the TR was relatively short and the voxels used were all roughly from the same slice. Similar coupling among the estimates may be caused by collinearity when projecting the effects of priors onto the observation space (note the similarity among some of the second partial derivatives in the bottom panel of Fig. 4). Despite these reservations, the fact that rather tight conditional densities for the efficacies are obtained using data from other subjects and paradigms suggests that they are sufficiently valid, or close to the veridical priors, for our purposes.

The validity of the hemodynamic model was established, to a certain level, in Friston *et al.* (2000). However, it is important to reiterate that biophysical models of this sort undergo continual refinement and elaboration. Indeed there are at least two components of the model that are already being improved (Mayhew *et al.*, personal communication). First we have assumed that the only cause of flow-inducing signal increase is the experimental input. Clearly oxygen tension itself is likely to be an important factor. Second we have followed Buxton *et al.* (1998) in assuming a fairly simply form for the coupling between oxygen delivery and flow that assumes oxygen tension is close to zero. Alternative formulations that embody the modulatory effect of oxygen tension on the extraction–flow coupling are currently being explored within the framework of the hemodynamic model using optical imaging (Zheng *et al.*, 2002). These considerations are important from the perspective of physiology and the interpretation of the model parameters. However, the main purpose of this paper was to present the methodology from a system identification perspective. In this sense the approach described above can easily accommodate any refinements or additions to the hemodynamic model. Furthermore, if one is only interested in the inference about stimulus efficacy, the interpretation of the biophysical parameters becomes subordinate. It is only necessary for the model to capture the transduction dynamics. Any model with 5 degrees of freedom is likely to be sufficient, provided that the system is only weakly nonlinear. This can be inferred anecdotally from the fact that the best results emerge when using two or three basis functions in classical analyses, using variants of Eq. (21). A more formal analysis can be envisaged using variational techniques and this will be the subject of future work.

The hemodynamic model is only weakly nonlinear. The nature of its nonlinearity is quite subtle and can be inferred from the analysis presented in Friston *et al.* (2000). In this analysis we showed that a bilinear approximation to the state equation, followed by an output nonlinearity, was sufficient to account for nonlinear responses observed empirically. Critically the bilinear approximation precludes interactions (i.e., nonlinear effects) among the state variables [see Eq. (18)]. Furthermore, the inputs enter linearly and do not interact with the state variables [see Eq. (7)]. This means that the dynamics of the state variables are essentially linear. The second-order kernel associated with the output is due solely to the output nonlinearity. These observations imply that the hemodynamic model could be formulated as two parallel first-order convolutions of the input to produce q and v followed by the static nonlinearity $y(t) = \lambda(q, v)$. Note that this is not quite the same as the first hemodynamic model proposed by Vazquez and Noll (1996) which comprised a single convolution followed by a static nonlinearity but a suitable transformation of the state variables might make the latter a good approximation. Perhaps the simplest defence of the model's validity is that, irre-

spective of its shortcomings, it is more plausible than the linear models in current use.

5.2. Extensions

Much of the discussion above has been preoccupied with nonlinearities in the hemodynamics per se. Interactions at the neuronal level are, of course, prevalent and motivate the extensive use of factorial designs in neuroimaging that look explicitly for interactions among the causes of neuronal responses. These interaction terms are simply accommodated in the current framework by forming additional inputs that enter linearly into the model. In practice this involves taking two mean-corrected stimulus functions and multiplying them together. This new stimulus function represents the interaction at the level of synaptic input or efficacy. Bayesian inferences about the interaction proceed in exactly the same way as the main effects.

On a more technical note, we have focused on the estimation of hyperparameters pertaining to the error covariance. Exactly the same algorithm can be used to estimate hyperparameters of unknown priors. The ensuing empirical determination of priors rests upon a hierarchical observation model in which variation over parameter estimates can be used as empirical priors on the estimates themselves. This hierarchical form for the model requires many estimates of the parameters (e.g., repeated measures in multiple sessions) or by linking the relative variances of different priors. The extension to hierarchical models is described in general terms in Friston *et al.* (2002a).

In this paper we have assumed that each stimulus function or cause elicits a flow-inducing signal and that this can be described by a single parameter. Clearly neuronal activity mediates between the stimulus and flow-inducing signal. Furthermore the neuronal dynamics may themselves differ in form over different causes or trials. For example, some stimuli may engage high level processing and evoke late or endogenous neuronal components whereas others may not, eliciting only early exogenous activity. The current model can be naturally extended to include neuronal activity and to embrace a distinction between early or transient and late or enduring responses. The simplest extension involves introducing two further state variables x_5 and x_6, representing transient and enduring neuronal activity in distinct cell assembles within the voxel, to give a new version of $\dot{X}(t) = f(X, u(t))$ in Eq. (7)

$$\begin{aligned}
\dot{x}_5 &= f_5(X, u(t)) = \epsilon_1 u(t)_1 + \ldots + \epsilon_n u(t)_n - x_5/\tau_e \\
\dot{x}_6 &= f_6(X, u(t)) = l_1 u(t)_1 + \ldots + l_n u(t)_n - x_6/\tau_l \\
\dot{x}_1 &= f_1(X, u(t)) = x_5 + x_6 - \kappa_s x_1 - \kappa_f(x_2 - 1) \\
\dot{x}_2 &= f_2(X, u(t)) = x_1 \\
\dot{x}_3 &= f_3(X, u(t)) = \frac{1}{\tau}(x_2 - f_{out}(x_3, \alpha)) \\
\dot{x}_4 &= f_4(X, u(t)) \\
&= \frac{1}{\tau}\left(x_2 \frac{E(x_2, E_0)}{E_0} - f_{out}(x_3, \alpha)\frac{x_4}{x_3}\right).
\end{aligned} \tag{23}$$

The model has n new input-specific parameters rendering the effect of any experimental cause bidimensional: ϵ_i represent the efficacy of the ith input in evoking an early response with time constant τ_e whereas l_i reflects the input's ability to invoke a sustained or enduring neuronal transient with time constant τ_l. The two new parameter τ_e and τ_l specify the dynamics of the early and late neuronal components and could be set at, say, 100 and 500 ms, respectively. Inputs with a larger efficacy for the late component will produce a slightly more protracted BOLD response with a peak latency shift, relative to trials evoking only early responses. This and related extensions will be developed in a subsequent paper. Perhaps the most important extension of the models described in this paper is to MIMO systems where we deal with multiple regions or voxels at the same time. The fundamental importance of this extension is that one can incorporate interactions among brain regions at the neuronal level. This provides a very promising framework for the dynamic causal modeling of functional integration in the brain and will be the subject of the next paper from our group pursuing the nonlinear modeling of fMRI time series.

ACKNOWLEDGMENT

The Wellcome Trust supported this work.

Software implementation note. The algorithm described in this paper has been implemented in the development version of the SPM software (SPMα), which will constitute the next release. Figures 5 and 6 represent the standard graphical output provided by SPMα. Currently, the analysis is restricted to a selected region or voxel and is invoked after conventional preprocessing and analysis. At the time of writing computation time prohibits the routine application to every voxel to produce PPMs. A standard 128-scan data set requires about 10–100 s to estimate the conditional densities for each voxel, taking many hours for the whole brain. However, because Bayesian inference does not incur a multiple comparison problem it is quite tenable to perform a conventional SPM analysis and then report the Bayesian inference at selected maxima. We are currently thinking about the application of this approach to the dynamics of spatial modes, which might provide a more computationally tractable way of making Bayesian inferences over the entire brain.

REFERENCES

Bendat, J. S. 1990. *Nonlinear System Analysis and Identification from Random Data.* Wiley, New York.

Büchel, C., and Friston, K. J. 1997. Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modelling and fMRI. *Cereb. Cortex* **7:** 768–778.

Buxton, R. B., Wong, E. C., and Frank, L. R. 1998. Dynamics of blood flow and oxygenation changes during brain activation: The Balloon model. *MRM* **39:** 855–864.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **39:** 1–38.

Fahrmeir, L., and Tutz, G. 1994. *Multivariate Statistical Modelling Based on Generalised Linear Models,* Springer-Verlag, New York. Pp. 355–356.

Fliess, M., Lamnabhi, M., and Lamnabhi-Lagarrigue, F. 1983. An algebraic approach to nonlinear functional expansions. *IEEE Trans. Circuits Syst.* **30:** 554–570.

Friston, K. J. 1995. Regulation of rCBF by diffusible signals: An analysis of constraints on diffusion and elimination. *Hum. Brain Map.* **3:** 56–65.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J-B., Frith, C. D., and Frackowiak, R. S. J. 1995. Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Map.* **2:** 189–210.

Friston, K. J., Josephs, O., Rees, G., and Turner, R. 1998. Nonlinear event-related responses in fMRI. *MRM* **39:** 41–52.

Friston, K. J., Mechelli, A., Turner, R., and Price, C. J. 2000. Nonlinear responses in fMRI: The Balloon model, Volterra kernels and other hemodynamics. *NeuroImage* **12:** 466–477.

Friston, K. J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., and Ashburner, J. 2002a. Classical and Bayesian inference in neuroimaging: Theory. *NeuroImage* **16:** 465–483.

Friston, K. J., Glaser, D. E., Henson, R. N. A., and Ashburner, J. 2002b. Classical and Bayesian inference in neuroimaging: Variance component estimation in fMRI. Submitted for publication.

Grubb, R. L., Rachael, M. E., Euchring, J. O., and Ter-Pogossian, M. M. 1974. The effects of changes in PCO_2 on cerebral blood volume, blood flow and vascular mean transit time. *Stroke* **5:** 630–639.

Harville, D. A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.* **72:** 320–338.

Hoge, R. D., Atkinson, J., Gill, B., Crelier, G. R., Marrett, S., and Pike, G. B. 1999. Linear coupling between cerebral blood flow and oxygen consumption in activated human cortex. *Proc. Natl. Acad. Sci. USA* **96:** 9403–9408.

Irikura, K., Maynard, K. I., and Moskowitz, M. A. 1994. Importance of nitric oxide synthase inhibition to the attenuated vascular responses induced by topical 1-nitro-arginine during vibrissal stimulation. *J. Cereb. Blood Flow Metab.* **14:** 45–48.

Leonard, T. 1972. Bayesian methods for Binomial data. *Biometrika* **59:** 581–589.

Mandeville, J. B., Marota, J. J., Ayata, C., Zararchuk, G., Moskowitz, M. A., Rosen, B., and Weisskoff, R. M. 1999. Evidence of a cerebrovascular postarteriole windkessel with delayed compliance. *J. Cereb. Blood Flow Metab.* **19:** 679–689.

Mayhew, J., Hu, D., Zheng, Y., Askew, S., Hou, Y., Berwick, J., Coffey, P. J., and Brown, N. 1998. An evaluation of linear models analysis techniques for processing images of microcirculation activity. *NeuroImage* **7:** 49–71.

Miller, K. L., Luh, W. M., Liu, T. T., Martinez, A., Obata, T., Wong, E. C., Frank, L. R., and Buxton, R. B. 2000. Characterizing the dynamic perfusion response to stimuli of short duration. *Proc. ISRM* **8:** 580.

Neal, R. M., and Hinton, G. E. 1998. A view of the EM algorithm that justifies incremental, sparse and other variants. In *Learning in Graphical Models* (M. I. Jordan, Ed.), pp. 355–368. Kluwer, Dordrecht.

Rees, G., Howseman, A., Josephs, O., Frith, C. D., Friston, K. J., Frackowiak, R. S. J., and Turner, R. 1997. Characterising the relationship between BOLD contrast and regional cerebral blood flow measurements by varying the stimulus presentation rate. *NeuroImage* **6:** 270–278.

Santner, T. J., and Duffy, D. E. 1989. *The Statistical Analysis of Discrete Data.* Springer Verlag, New York.

Talairach, J., and Tournoux, P. 1988. *A Co-planar Stereotaxic Atlas of a Human Brain.* Thieme, Stuttgart.

Vazquez, A. L., and Noll, D. C. 1996. Non-linear temporal aspects of the BOLD response in fMRI. *Proc. Int. Soc. Mag. Res. Med.* **3:** S1765.

Zheng, Y., Martindale, J., Johnston, D., and Mayhew, J. 2002. A model of the hemodynamic response and oxygen delivery to brain. Submitted for publication.