Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2010, Article ID 621670, 34 pages
doi:10.1155/2010/621670

Research Article

Generalised Filtering

Karl Friston,[1] Klaas Stephan,[1,2] Baojuan Li,[1,3] and Jean Daunizeau[1,2]

[1] *Wellcome Trust Centre for Neuroimaging, University College London, Queen Square, London WC1N 3BG, UK*
[2] *Laboratory for Social and Neural Systems Research, Institute of Empirical Research in Economics, University of Zurich, 8006 Zurich, Switzerland*
[3] *College of Mechatronic Engineering and Automation, National University of Defense Technology, Changsha, Hunan 410073, China*

Correspondence should be addressed to Karl Friston, k.friston@fil.ion.ucl.ac.uk

Received 29 January 2010; Accepted 17 March 2010

Academic Editor: Massimo Scalia

We describe a Bayesian filtering scheme for nonlinear state-space models in continuous time. This scheme is called Generalised Filtering and furnishes posterior (conditional) densities on hidden states and unknown parameters generating observed data. Crucially, the scheme operates online, assimilating data to optimize the conditional density on time-varying states and time-invariant parameters. In contrast to Kalman and Particle smoothing, Generalised Filtering does not require a backwards pass. In contrast to variational schemes, it does not assume conditional independence between the states and parameters. Generalised Filtering optimises the conditional density with respect to a free-energy bound on the model's log-evidence. This optimisation uses the generalised motion of hidden states and parameters, under the prior assumption that the motion of the parameters is small. We describe the scheme, present comparative evaluations with a fixed-form variational version, and conclude with an illustrative application to a nonlinear state-space model of brain imaging time-series.

1. Introduction

This paper is about the inversion of dynamic causal models based on nonlinear state-space models in continuous time. These models are formulated in terms of ordinary or stochastic differential equations and are ubiquitous in the biological and physical sciences. The problem we address is how to make inferences about the hidden states and unknown parameters generating data, given only observed responses and prior beliefs about the form of the underlying generative model, and its parameters. The parameters here include quantities that parameterise the model's equations of motion and control the amplitude (variance or inverse precision) of random fluctuations. If we consider the parameters and precisions as separable quantities, model inversion represents a triple estimation problem. There are relatively few schemes in the literature that can deal with problems of this sort. Classical filtering and smoothing schemes such as those based on Kalman and Particle filtering (e.g., [1, 2]) deal only with estimation of hidden states. Recently, we introduced several schemes that solve the triple estimation problem, using variational or ensemble learning [3–5]. Variational schemes of this sort simplify the problem by assuming conditional independence among sets of unknown quantities, usually states, parameters and precisions. This is a natural partition, in terms of time-varying hidden states and time-invariant parameters and precisions. The implicit mean-field approximation leads to schemes that optimise the posterior or conditional density on time-varying hidden states, while accumulating the sufficient statistics necessary to optimise the conditional density of parameters and precisions, after all the data have been observed.

In this paper, we dispense with the mean-field approximation and treat all unknown quantities as conditionally dependent variables, under the prior constraint that the changes in parameters and precisions are very small. This constraint is implemented by representing all unknown variables in generalised coordinates of motion, which allows one to optimise the moments of the joint posterior as data arrive. The resulting scheme enables an efficient assimilation of data and the possibility of online and real-time deconvolution. We refer to this Bayesian filtering in generalised coordinates as Generalised Filtering (GF). Furthermore, by assuming a fixed form for the conditional density (the Laplace assumption) one can reduce the triple estimation problem to integrating or solving a set of relatively simple ordinary differential equations. In this paper, we focus on GF under the Laplace assumption.

We have previously described Variational filtering in [3] and smoothing in [5] for probabilistic inference on the hidden states of generative models based upon stochastic differential equations (see also [6–8], for recent and related advances). Variational filtering and its fixed form variant (Dynamic Expectation Maximization) outperform extended Kalman filtering and Particle filtering (provided the true posterior density is unimodal) in terms of the accuracy of estimating hidden states [3]. This improvement rests upon using generalised coordinates of motion. In other words, instead of just trying to estimate the conditional density of hidden states, one optimises the conditional density on their generalised motion to arbitrarily high order. The implicit modelling of generalised states has several fundamental advantages. First, it can accommodate temporal correlations in random fluctuations on the hidden states (e.g., fractal time and $1/f$ spectra in biological systems [9]). In other words, random terms in the model's stochastic differential equations can have analytic autocovariance functions, whose smoothness can be estimated. This allows one to eschew standard Weiner assumptions, which is important in many realistic settings, particularly in the analysis of biological time-series. Second, generalised states afford a very simple filtering scheme that is formally distinct from Kalman and Particle filtering. In brief, Variational filtering uses a gradient descent on a free-energy bound on the model's (negative) log-evidence. This means that filtering can be reduced to the solution of differential equations that necessarily entail continuity constraints on the conditional estimates. Clearly, the free-energy is a functional of time, which means that the gradient descent has to "hit a moving target." Generalised coordinates finesse this problem by placing the gradient descent in a frame of reference that moves with the conditional expectation or mean (see [4] for details). This is heuristically related to the separation of temporal scales in centre manifold theory [10], where the motion of free-energy minima (the centre manifold) enslaves the dynamics of the gradient descent.

Variational filtering of this sort is fundamentally different in its mathematical construction from conventional schemes like Kalman filtering because of its dynamical formulation. It can be implemented without any assumptions on the form of the conditional density by using an ensemble of “particles” that are subject to unit (standard) Wiener perturbations. The ensuing ensemble density tracks the conditional mean of the hidden states and its dispersion encodes conditional uncertainty. Variational filtering can be further simplified by assuming the ensemble density (conditional density) is Gaussian, using the Laplace assumption. Crucially, under this approximation, the conditional covariance (second moment of the conditional density) becomes an analytic function of the conditional mean. In other words, only the mean *per se* has to be optimized. This allows one to replace an ensemble of particles, whose motion is governed by stochastic differential equations, with a single ordinary differential equation describing the motion of the conditional mean. The solution of this ordinary differential equation corresponds to the *D*-step in Dynamic Expectation Maximisation (DEM). DEM comprises additional *E* (expectation) and *M* (maximization) steps that optimise the conditional density on parameters and precisions after the *D* (deconvolution) step has been solved. Iteration of these steps proceeds until convergence, in a way that is formally related to conventional variational schemes (cf. [5, 11]).

In this work, we retain the Laplace approximation to the conditional density but dispense with the mean-field approximation; in other words, we do not assume conditional independence between the states, parameters, and precisions. We implement this by absorbing parameters and precisions into the hidden states. This means that we can formulate a set of ordinary differential equations that describe the motion of time-dependent conditional means and implicitly the conditional precisions (inverse covariances) of all unknown variables. This furnishes (marginal) conditional densities on the parameters and precisions that are functionals of time. The associated conditional density of the average parameters and precisions over time can then be accessed using Bayesian parameter averaging. Treating time-invariant parameters (and precisions) as states rests on modelling their motion. Crucially, we impose prior knowledge that this motion is zero, leading to a gradient descent on free-energy, which is very smooth (cf. the use of filtering as a “second-order” technique for learning parameters [12]). In brief, the resulting scheme assimilates evidence in the generalised motion of data to provide a time-varying density on all the model’s unknown variables, where the marginal density on the parameters and precisions converges slowly to a steady-state solution. The scheme can be iterated until the time or path-integral of free-energy (free-action) converges and may represent a useful and pragmatic (online) alternative to variational schemes.

This paper comprises four sections. In the first, we describe the technical details of Generalised Filtering from the first principles. This section starts with the objective (to maximize the path-integral of a free-energy bound on a model’s log-evidence). It ends with set of ordinary differential equations, whose solution provides the conditional moments of a Gaussian approximation to the conditional density we seek. The second section reviews a generic model that embodies both dynamic and structural (hierarchical) constraints. We then look at Generalised Filtering from the first section, under this model. The third section presents comparative evaluations of GF using a simple linear convolution model, which is a special case of the model in Section 2. These evaluations are restricted to a comparison with DEM because DEM is formally the closest scheme to GF and has been compared with Extended Kalman filtering and Particle filtering previously [4]. In the final section, we apply Generalised Filtering to a neurobiological problem, namely, inferring the parameters and hidden physiological states generating a time-series of functional magnetic resonance imaging (fMRI) data from the human brain. We use this to illustrate the effect of the mean-field assumption implicit in DEM and establish the face validity of Generalised Filtering in terms of known neurobiology. We conclude with a brief discussion.

2. Generalised Filtering

In this section, we present the conceptual background and technical details behind Generalised Filtering, which (in principle) can be applied to any nonlinear state-space or dynamic causal model formulated with stochastic differential equations. Given the simplicity of the ensuing scheme, we also take the opportunity to consider state-dependant changes in the precision of random fluctuations. This represents a generalisation of our previous work on dynamic causal models and will be exploited in a neurobiological context, as a metaphor for attention (Feldman et al.; in preparation). However, we retain a focus on cascades of state-space models, which we have referred to previously as hierarchical dynamic models [13].

2.1. Filtering from Basic Principles

Given a model m and generalised sensor data $\tilde{s} = [s, s', s'', \ldots]^T \in \Re^p$ comprising real values, their velocity, acceleration, jerk, and so forth, we want to evaluate the log-evidence integrated over the time $t \in [0, T]$ that data are observed (cf. [14, 15]):

$$\varepsilon = \int_0^T dt \ \ln p(\tilde{s}(t) \mid m). \tag{2.1}$$

Generally, this path-integral cannot be evaluated directly; however, one can induce an upper bound $\mathcal{S} \geq -\varepsilon$ that can be evaluated with a recognition density $q(t) := q(\vartheta(t))$ on the causes (i.e., states and parameters) of the data. We will see later that these causes comprise time-varying states $u(t) \subset \vartheta$ and slowly varying parameters $\varphi(t) \subset \vartheta$. This bound is based on free-energy, which is constructed by simply adding a nonnegative term $D_{KL}(t)$ to the (negative) log-evidence [11, 16, 17]. The resulting integral is called free-action because it is a path-integral of free-energy (see also [18])

$$\begin{gathered}
\mathcal{S} = \int dt \mathcal{F}(t) \geq -\varepsilon, \\
\mathcal{F}(t) = -\ln p(\tilde{s}(t) \mid m) + D_{KL}(t), \\
D_{KL}(t) = \langle \ln q(\vartheta(t)) - \ln p(\vartheta(t) \mid \tilde{s}(t), m) \rangle_q.
\end{gathered} \tag{2.2}$$

By Gibb's inequality the Kullback-Leibler divergence $D_{KL}(t) \geq 0$ is greater than zero, with equality when $q(t) = p(\vartheta(t) \mid \tilde{s}(t), m) : \forall t \in [0, T]$ is the true conditional density. In this case, (negative) free-action becomes accumulated log-evidence $\mathcal{S} = -\varepsilon$. Minimising this bound, by optimising the recognition density at each time point, makes it an approximate conditional density on the causes and makes free-action a bound approximation to the accumulated log-evidence. The approximate conditional density can then be used for inference on the states or parameters of a particular model, and the accumulated log-evidence can be used to compare different models (e.g., [19]).

Crucially, the free-energy can be evaluated easily because it is a function of $q(\vartheta(t))$ and a generative model $p(\widetilde{s}(t), \vartheta(t) \mid m)$ entailed by m

$$\begin{gathered} \mathcal{F}(t) = \langle \mathcal{L}(t) \rangle_q - \mathcal{H}(t), \\ \mathcal{L}(t) = -\ln p(\widetilde{s}(t), \vartheta(t) \mid m), \\ \mathcal{H}(t) = -\langle \ln q(t) \rangle_q. \end{gathered} \tag{2.3}$$

The free-energy has been expressed here in terms of $\mathcal{H}(t)$, the negentropy of $q(t)$, and an energy $\mathcal{L}(t)$ expected under $q(t)$. In physics, $\mathcal{L}(t)$ is called Gibb's energy and is a log-probability that reports the joint surprise about data and their causes. If we assume that $q(\vartheta(t)) = \mathcal{N}(\mu(t), C(t))$ is Gaussian (the Laplace assumption), then we can express free-energy in terms of its sufficient statistics, the mean and covariance of the recognition density

$$\mathcal{F} = \mathcal{L}(\mu) + \frac{1}{2}\operatorname{tr}(C\mathcal{L}_{\mu\mu}) - \frac{1}{2}\ln|C| - \frac{n}{2}\ln 2\pi e, \tag{2.4}$$

where $n = \dim(\mu)$. Here and throughout, subscripts denote derivatives. We can now minimise free-action with respect to the conditional precisions $P(t) = C(t)^{-1}$ (inverse covariances) by solving $\delta_C \mathcal{S} = 0 \Rightarrow \partial_C \mathcal{F} = 0$ to give

$$\mathcal{F}_\Sigma = \frac{1}{2}\mathcal{L}_{\mu\mu} - \frac{1}{2}P = 0 \Longrightarrow P = \mathcal{L}_{\mu\mu}. \tag{2.5}$$

Equation (2.5) shows that the precision is an analytic function of the mean, which means all we have worry about is the (approximate) conditional mean. One can see this clearly by eliminating the conditional covariance to express the free-energy as a function of (and only of) the conditional means

$$\mathcal{F} = \mathcal{L}(\mu) + \frac{1}{2}\ln\left|\mathcal{L}_{\mu\mu}\right| - \frac{n}{2}\ln 2\pi. \tag{2.6}$$

The conditional means, which minimise free-energy, are the solutions to the following ordinary differential equations. For the generalised states $\widetilde{u}(t) \subset \vartheta$ these equations are

$$\begin{gathered} \dot{\widetilde{\mu}}^{(u)} = \mathcal{D}\widetilde{\mu}^{(u)} - \mathcal{F}_{\widetilde{u}} \\ \Longleftrightarrow \\ \dot{\mu}^{(u)} = \mu'^{(u)} - \mathcal{F}_u \\ \dot{\mu}'^{(u)} = \mu''^{(u)} - \mathcal{F}_{u'} \\ \dot{\mu}''^{(u)} = \mu'''^{(u)} - \mathcal{F}_{u''} \\ \vdots \end{gathered} \tag{2.7}$$

where $\mathfrak{D}$ is a derivative matrix operator with identity matrices above the leading diagonal, such that $\mathfrak{D}\tilde{u} = [u', u'', \ldots]^T$. Here and throughout, we assume that all gradients are evaluated at the mean; here $\tilde{u}(t) = \tilde{\mu}^{(u)}(t)$. The stationary solution of (2.7), in a frame of reference that moves with the generalised motion of the mean, minimises free-energy and action. This can be seen by noting that the variation of free-action with respect to the solution is zero:

$$\dot{\tilde{\mu}}^{(u)} - \mathfrak{D}\tilde{\mu}^{(u)} = 0 \Longrightarrow \mathcal{F}_{\tilde{u}} = 0 \Longrightarrow \delta_{\tilde{u}}\mathcal{S} = 0. \tag{2.8}$$

This ensures that when Gibb's energy is minimized, the mean of the motion is the motion of the mean, that is, $\mathcal{F}_{\tilde{u}} = 0 \Rightarrow \dot{\tilde{\mu}}^{(u)} = \mathfrak{D}\tilde{\mu}^{(u)}$. For slowly varying parameters $\varphi(t) \subset \vartheta$ this motion disappears and we can use the following scheme:

$$\begin{aligned} \dot{\mu}^{(\varphi)} &= \mu'^{(\varphi)}, \\ \dot{\mu}'^{(\varphi)} &= -\mathcal{F}_{\varphi} - \kappa\mu'^{(\varphi)}. \end{aligned} \tag{2.9}$$

Here, the solution $\dot{\tilde{\mu}}^{(\varphi)} = 0$ minimises free-energy, under constraint that the motion of the mean is small; $\mu'^{(\varphi)} \to 0$. This can be seen by noting

$$\dot{\mu}^{(\varphi)} = \dot{\mu}'^{(\varphi)} = 0 \Longrightarrow \mathcal{F}_{\varphi} = 0 \Longrightarrow \delta_{\varphi}\mathcal{S} = 0. \tag{2.10}$$

Equations (2.7) and (2.9) prescribe recognition (filtering) dynamics for expected states and parameters respectively. The dynamics for states can be regarded as a gradient descent in a frame of reference that moves with the expected motion (cf. a moving target). Conversely, the dynamics for the parameters can be thought of as a gradient descent that resists transient fluctuations in free-energy with a damping term $-\kappa\mu'^{(\varphi)}$. This instantiates our prior belief that fluctuations in the parameters are small, where κ can be regarded as a prior precision. Figure 1 shows the implicit kernel (solid line) that is applied to fluctuating free-energy gradients to produce changes in the conditional mean. It can be seen that the height of the kernel is $1/\kappa$. We find that using $\kappa = 8T$ ensures stable convergence in most situations. This renders the integral of the kernel over the time-series about one eighth (see Appendix A for a discussion of this assimilation scheme in relation to classical methods).

The solutions of (2.7) and (2.9) produce the time-dependent moments of an approximate Gaussian conditional density on the unknown states and parameters. These can be used for inference on states (or parameters) directly or for model comparison using the (bound on) accumulated log-evidence (see (2.3) and (2.6)). Because the conditional means of the parameters change slowly over time (unlike their conditional precisions), we can summarise the conditional estimates with the conditional density on the average over time $\overline{\varphi}$ using Bayesian parameter averaging

$$\begin{aligned} q(\overline{\varphi}) &= \mathcal{N}\left(\overline{\mu}^{(\varphi)}, \overline{C}^{(\varphi)}\right), \\ \overline{\mu}^{(\varphi)} &= \overline{C}^{(\varphi)} \int \mathcal{P}^{(\varphi)}(t)\mu^{(\varphi)}(t)dt, \\ \overline{C}^{(\varphi)-1} &= \int \mathcal{P}^{(\varphi)}(t)dt. \end{aligned} \tag{2.11}$$

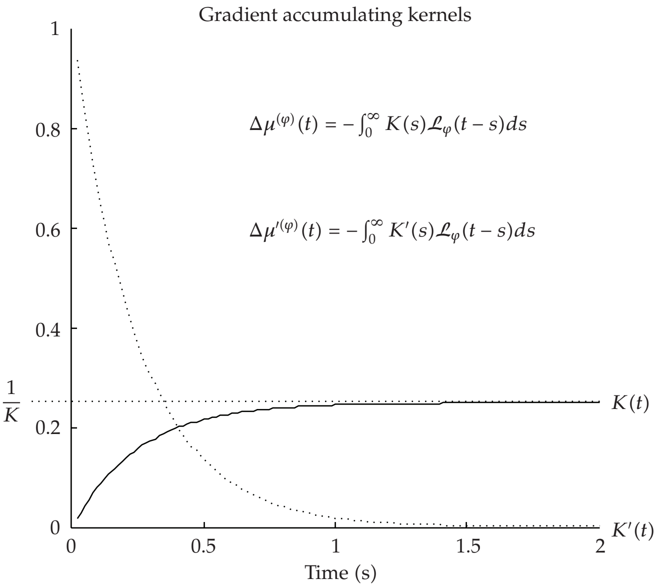


Figure 1: This graph shows the kernels implied by the recognition dynamics in (2.9) that accumulate evidence for changes in the conditional estimates of model parameters. These kernels are applied to the history of free-energy gradients to produce changes in the conditional mean (solid line) and its motion (broken line). The kernels are derived from a standard Volterra-series expansion about the true conditional mean, where, in this example, $\kappa = 4$.

Here, $\overline{\mu}^{(\varphi)}$ and $\overline{\mathcal{C}}^{(\varphi)}$ are the conditional moments of the average while $\mu^{(\varphi)}(t) \subset \mu(t)$ and $\mathcal{P}^{(\varphi)}(t)^{-1} = \mathcal{C}^{(\varphi)}(t) \subset \mathcal{C}(t)$ are the mean and precision of the marginal conditional density on the parameters at each point in time. This summary will be overconfident because it neglects conditional dependencies due to temporally correlated random terms (and the use of generalised data). However, it usefully connects the conditional density with posterior distributions estimated in conventional schemes that treat the parameters as static.

In Generalised Filtering, changes in the conditional uncertainty about the parameters are modelled explicitly as part of a time-varying conditional density on states and parameters. In contrast, variational schemes optimise a conditional density on static parameters, that is, $q(\overline{\varphi})$ that is not a functional of time (although fluctuations in the conditional dependence between states and parameters are retained as mean-field terms that couple their respective marginals). This difference is related to a formal difference between the free-energy bound on log-evidence in variational schemes and free-action. Variational schemes return the free-energy of the accumulated data (e.g., see [3, equation (9)]). Conversely, in Generalised Filtering, the free-energy is accumulated over time. This means that the variational free-energy $\mathcal{F}^{\mathcal{V}}$ bounds the log-evidence of accumulated data, while free-action $\mathcal{S} = \int dt \mathcal{F}(t)$ bounds the accumulated log-evidence of data

$$\begin{aligned}
\mathcal{S} &\leq \int dt \ln p(\tilde{s}(t) \mid m), \\
\mathcal{F}^{\mathcal{V}} &\leq \ln p\left(\bigcup_t \tilde{s}(t) \mid m\right).
\end{aligned} \tag{2.12}$$

These are not the same because evidence is a nonlinear function of the data (see, Appendix B for an example). This distinction does not mean that one form of evidence accumulation is better than another; although it raises interesting issues for model comparison that will be pursued elsewhere (Li et al; in preparation). For the moment, we will approximate the free-action associated with the conditional densities from variational schemes by assuming that $q(u(t),\varphi) = q(u(t))q(\varphi)$, where $q(\varphi) = \mathcal{N}(\overline{\mu}^{(\varphi)}, \mathcal{C}^{(\varphi)})$ and $\mathcal{C}^{(\varphi)} = T\overline{\mathcal{C}}^{(\varphi)}$ (Appendix B). This allows us to compare fixed-form schemes with (DEM) and without (GF) the mean field assumption, in terms of free-action.

2.2. *Summary*

In summary, we have derived recognition or filtering dynamics for expected states and parameters (in generalised coordinates of motion), which cause data. The solutions to these equations minimise free-action (at least locally) and therefore minimise a bound on the accumulated evidence for a model of how we think the data were caused. This minimisation furnishes the conditional density on the unknown variables in terms of conditional means and precisions. The precise form of the filtering depends on the energy $\mathcal{L} = -\ln p(\tilde{s}, u \mid m)$ associated with a particular generative model. In what follows, we examine the forms entailed by hierarchical dynamic models.

3. Hierarchical Dynamic Models

In this section, we review the form of models that will be used in subsequent sections to evaluate Generalised Filtering. Consider the state-space model

$$\begin{aligned} s &= f^{(v)}(x,v,\theta) + z^{(v)} : z^{(v)} \sim \mathcal{N}\Big(0, \Sigma^{(v)}(x,v,\gamma)\Big), \\ \dot{x} &= f^{(x)}(x,v,\theta) + z^{(x)} : z^{(x)} \sim \mathcal{N}\Big(0, \Sigma^{(x)}(x,v,\gamma)\Big). \end{aligned} \tag{3.1}$$

Using $\sigma^{(u)}\sigma^{(u)T} = \Sigma^{(u)} : u \in v, x$ and unit noise $w^{(u)} \sim \mathcal{N}(0, I)$, this model can also be written as

$$\begin{aligned} s &= f^{(v)}(x,v,\theta) + \sigma^{(v)}(x,v,\gamma)w^{(v)}, \\ \dot{x} &= f^{(x)}(x,v,\theta) + \sigma^{(x)}(x,v,\gamma)w^{(x)}. \end{aligned} \tag{3.2}$$

The nonlinear functions $f^{(u)} : u \in v, x$ represent a mapping from hidden states to observed data and the equations of motion of the hidden states. These equations are parameterised by $\theta \subset \varphi$. The states $v \subset u$ are variously referred to as sources or causes, while the hidden states $x \subset u$ meditate the influence of causes on data and endow the system with memory. We assume that the random fluctuations $z^{(u)}$ are analytic, such that the covariance of $\tilde{z}^{(u)}$ is well defined. Unlike our previous treatment of these models, we allow for state-dependent changes in the amplitude of random fluctuations. These effects are meditated by the vector and matrix functions $f^{(u)} \in \Re^{\dim(u)}$ and $\Sigma^{(u)} \in \Re^{\dim(u)\times\dim(u)}$ respectively, which are parameterised by first and second-order parameters $\{\theta, \gamma\} \subset \varphi$.

Under local linearity assumptions, the generalised motion of the data and hidden states can be expressed compactly as

$$\begin{aligned}\tilde{s} &= \tilde{f}^{(v)} + \tilde{z}^{(v)},\\ \mathfrak{D}\tilde{x} &= \tilde{f}^{(x)} + \tilde{z}^{(x)},\end{aligned} \tag{3.3}$$

where the generalised predictions are (with $u \in v, x$)

$$\tilde{f}^{(u)} = \begin{bmatrix} f^{(u)} = f^{(u)}(x, v, \theta) \\ f'^{(u)} = f_x^{(u)} x' + f_v^{(u)} v' \\ f''^{(u)} = f_x^{(u)} x'' + f_v^{(u)} v'' \\ \vdots \end{bmatrix}. \tag{3.4}$$

Gaussian assumptions about the random fluctuations prescribe a generative model in terms of a likelihood and empirical priors on the motion of hidden states

$$\begin{aligned} p(\tilde{s} \mid \tilde{x}, \tilde{v}, \theta, m) &= \mathcal{N}\left(\tilde{f}^{(v)}, \tilde{\Sigma}^{(v)}\right),\\ p(\mathfrak{D}\tilde{x} \mid x, \tilde{v}, \theta, m) &= \mathcal{N}\left(\tilde{f}^{(x)}, \tilde{\Sigma}^{(x)}\right).\end{aligned} \tag{3.5}$$

These probability densities are encoded by their covariances $\tilde{\Sigma}^{(u)}$ or precisions $\tilde{\Pi}^{(u)} := \tilde{\Pi}^{(u)}(x, v, \gamma)$ with precision parameters $\gamma \subset \varphi$ that control the amplitude and smoothness of the random fluctuations. Generally, the covariances $\tilde{\Sigma}^{(u)} = V^{(u)} \otimes \Sigma^{(u)}$ factorise into a covariance proper and a matrix of correlations $V^{(u)}$ among generalised fluctuations that encode their autocorrelation [4, 20]. In this paper, we will deal with simple covariance functions of the form $\Sigma^{(u)} = \exp(-\gamma^{(u)}) I^{(u)}$. This renders the precision parameters log-precisions.

Given this generative model, we can now write down the energy as a function of the conditional expectations, in terms of a log-likelihood $\mathcal{L}^{(v)}$ and log-priors on the motion of hidden states $\mathcal{L}^{(x)}$ and the parameters $\mathcal{L}^{(\varphi)}$ (ignoring constants):

$$\begin{aligned} \mathcal{L} &= \mathcal{L}^{(v)} + \mathcal{L}^{(x)} + \mathcal{L}^{(\varphi)},\\ \mathcal{L}^{(v)} &= \frac{1}{2}\tilde{\varepsilon}^{(v)T}\tilde{\Pi}^{(v)}\tilde{\varepsilon}^{(v)} - \frac{1}{2}\ln\left|\tilde{\Pi}^{(v)}\right|,\\ \mathcal{L}^{(x)} &= \frac{1}{2}\tilde{\varepsilon}^{(x)T}\tilde{\Pi}^{(x)}\tilde{\varepsilon}^{(x)} - \frac{1}{2}\ln\left|\tilde{\Pi}^{(x)}\right|,\\ \mathcal{L}^{(\varphi)} &= \frac{1}{2}\tilde{\varepsilon}^{(\varphi)T}\tilde{\Pi}^{(\varphi)}\tilde{\varepsilon}^{(\varphi)} - \frac{1}{2}\ln\left|\tilde{\Pi}^{(\varphi)}\right|,\\ \tilde{\varepsilon}^{(v)} &= \tilde{s} - \tilde{f}^{(v)}(\mu),\\ \tilde{\varepsilon}^{(x)} &= \mathfrak{D}\tilde{\mu}^{(x)} - \tilde{f}^{(x)}(\mu),\\ \tilde{\varepsilon}^{(\varphi)} &= \tilde{\mu}^{(\varphi)} - \tilde{\eta}^{(\varphi)}.\end{aligned} \tag{3.6}$$

This energy has a simple quadratic form, where the auxiliary variables $\tilde{\varepsilon}^{(j)} : j \in v, x, \varphi$ are prediction errors for data, the motion of hidden states and parameters respectively. The predictions of the states are $\tilde{f}^{(u)}(\mu) : u \in v, x$ and the predictions of the parameters are their prior expectations. Equation (3.6) assumes flat priors on the states and that priors on the parameters are Gaussian $p(\varphi \mid m) = \mathcal{N}(\tilde{\eta}^{(\varphi)}, \tilde{\Sigma}^{(\varphi)})$, where

$$\tilde{\varphi} = \begin{bmatrix} \varphi \\ \varphi' \end{bmatrix}, \qquad \tilde{\Sigma}^{(\varphi)} = \begin{bmatrix} \Sigma^{(\varphi)} & 0 \\ 0 & \kappa \end{bmatrix}. \tag{3.7}$$

This assumption allows us to express the learning scheme in (2.9) succinctly as $\ddot{\mu}^{(\varphi)} = -\mathcal{F}_{\varphi} - \mathcal{F}_{\varphi'}$, where $\mathcal{F}_{\varphi'} = \mathcal{L}_{\varphi'} = -\kappa\varphi'$. Equation (3.6) provides the form of the free-energy and its gradients required for filtering, where (with a slight abuse of notation),

$$\begin{aligned}
\mathcal{F} &= \mathcal{L} + \frac{1}{2}\ln|\mathcal{P}| - \frac{p}{2}\ln 2\pi, \\
\dot{\tilde{\mu}}^{(v)} &= -\mathcal{F}_{\tilde{v}} + \mathcal{D}\tilde{\mu}^{(v)}, \\
\dot{\tilde{\mu}}^{(x)} &= -\mathcal{F}_{\tilde{x}} + \mathcal{D}\tilde{\mu}^{(x)}, \\
\ddot{\mu}^{(\theta)} &= -\mathcal{F}_{\theta} - \mathcal{F}_{\theta'}, \\
\ddot{\mu}^{(\gamma)} &= -\mathcal{F}_{\gamma} - \mathcal{F}_{\gamma'}, \\
\mathcal{F}_{\tilde{v}} &= \mathcal{L}_{\tilde{v}} + \frac{1}{2}\operatorname{tr}(\mathcal{P}_{\tilde{v}}\mathcal{C}), \\
\mathcal{F}_{\tilde{x}} &= \mathcal{L}_{\tilde{x}} + \frac{1}{2}\operatorname{tr}(\mathcal{P}_{\tilde{x}}\mathcal{C}), \\
\mathcal{F}_{\theta} &= \mathcal{L}_{\theta} + \frac{1}{2}\operatorname{tr}(\mathcal{P}_{\theta}\mathcal{C}), \\
\mathcal{F}_{\gamma} &= \mathcal{L}_{\gamma} + \frac{1}{2}\operatorname{tr}(\mathcal{P}_{\gamma}\mathcal{C}).
\end{aligned} \tag{3.8}$$

Note that the constant in the free-energy just includes $p = \dim(\tilde{s})$ because we have used Gaussian priors. The derivatives are provided in Appendix C. These have simple forms that comprise only quadratic terms and trace operators.

3.1. Hierarchical Forms

We next consider hierarchical forms of this model. These are just special cases of Equation (3.1), in which we make certain conditional independences explicit. Although they may look more complicated, they are simpler than the general form above. They are useful because they provide an empirical Bayesian perspective on inference [21, 22]. Hierarchical dynamic models have the following form

$$\begin{aligned}
s &= f^{(v)}\left(x^{(1)}, v^{(1)}, \theta\right) + z^{(1,v)} \\
\dot{x}^{(1)} &= f^{(x)}\left(x^{(1)}, v^{(1)}, \theta\right) + z^{(1,x)} \\
&\vdots \\
v^{(i-1)} &= f^{(v)}\left(x^{(i)}, v^{(i)}, \theta\right) + z^{(i,v)} \\
\dot{x}^{(i)} &= f^{(x)}\left(x^{(i)}, v^{(i)}, \theta\right) + z^{(i,x)} \\
&\vdots \\
v^{(h-1)} &= \eta^{(v)} + z^{(h,v)}.
\end{aligned} \tag{3.9}$$

Again, $f^{(i,u)} := f^{(u)}(x^{(i)}, v^{(i)}, \theta) : u \in v, x$ are continuous nonlinear functions and $\eta^{(v)}(t)$ is a prior mean on the causes at the highest level. The random terms $z^{(i,u)} \sim \mathcal{N}(0, \Sigma(x^{(i)}, v^{(i)}, \gamma^{(i,u)})) : u \in v, x$ are conditionally independent and enter each level of the hierarchy. They play the role of observation noise at the first level and induce random fluctuations in the states at higher levels. The causes $v = v^{(1)} \oplus v^{(2)} \oplus \cdots$ link levels, whereas the hidden states $x = x^{(1)} \oplus x^{(2)} \oplus \cdots$ link dynamics over time. In hierarchical form, the output of one level acts as an input to the next. This input can enter nonlinearly to produce quite complicated generalised convolutions with “deep” (i.e., hierarchical) structure [22]. The energy for hierarchical models is (ignoring constants)

$$\begin{aligned}
\mathcal{L} &= \sum_i \mathcal{L}^{(i,v)} + \sum_i \mathcal{L}^{(i,x)} + \mathcal{L}^{(\varphi)}, \\
\mathcal{L}^{(i,v)} &= \frac{1}{2}\tilde{\varepsilon}^{(i,v)T}\tilde{\Pi}^{(i,v)}\tilde{\varepsilon}^{(i,v)} - \frac{1}{2}\ln\left|\tilde{\Pi}^{(i,v)}\right|, \\
\mathcal{L}^{(i,x)} &= \frac{1}{2}\tilde{\varepsilon}^{(i,x)T}\tilde{\Pi}^{(i,x)}\tilde{\varepsilon}^{(i,x)} - \frac{1}{2}\ln\left|\tilde{\Pi}^{(i,x)}\right|, \\
\tilde{\varepsilon}^{(i,v)} &= \tilde{v}^{(i-1)} - \tilde{f}^{(i,v)}, \\
\tilde{\varepsilon}^{(i,x)} &= \mathfrak{D}\tilde{x}^{(i)} - \tilde{f}^{(i,x)}.
\end{aligned} \tag{3.10}$$

This is exactly the same as (3.6) but now includes extra terms that mediate empirical (structural) priors on the causes $\mathcal{L}^{(i,v)} = -\ln p(\tilde{v}^{(i-1)} \mid \tilde{x}^{(i)}, \tilde{v}^{(i)}, m)$. These are induced by the conditional independence assumptions in hierarchical models. Note that the data enter the prediction errors at the lowest level such that $\tilde{\varepsilon}^{(1,v)} = \tilde{s} - \tilde{f}^{(1,v)}$.

3.2. *Summary*

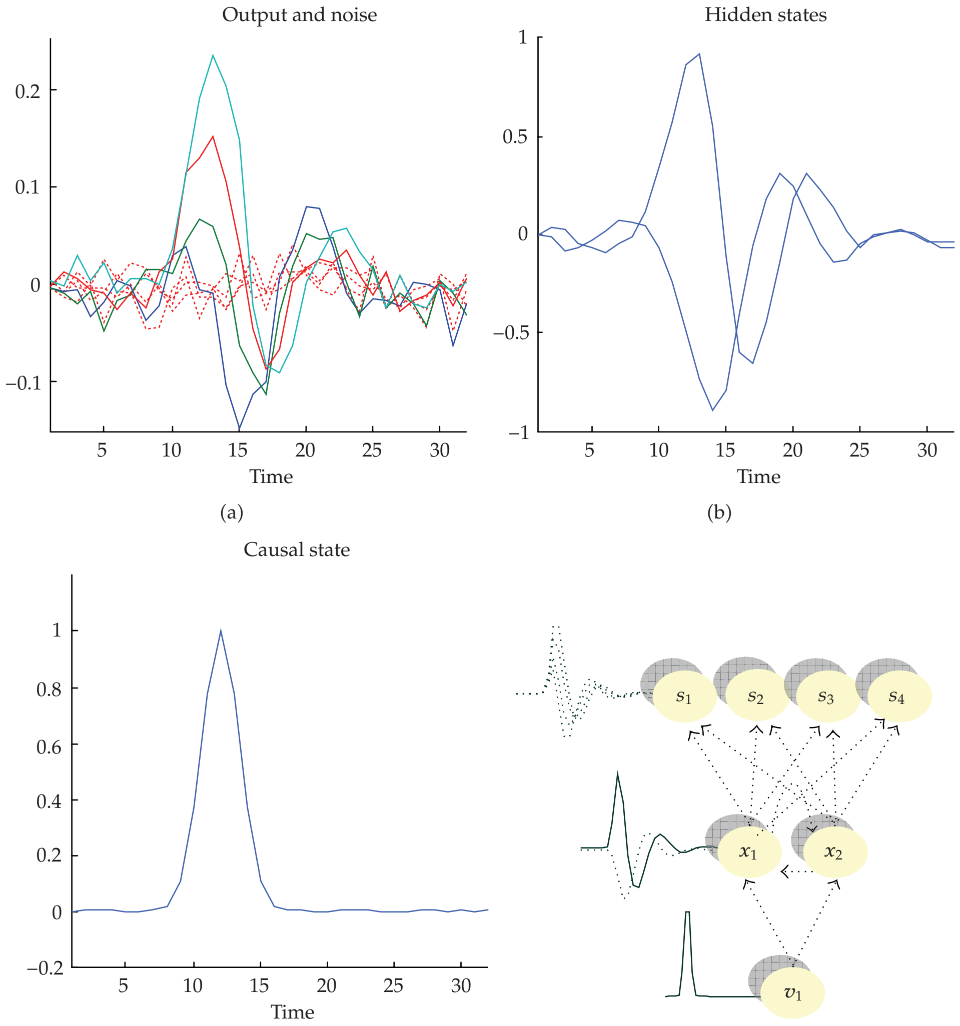
In summary, hierarchical dynamic models are nearly as complicated as one could imagine; they comprise causal and hidden states, whose dynamics can be coupled with arbitrary (analytic) nonlinear functions. Furthermore, the states can be subject to random fluctuations with state-dependent changes in amplitude and arbitrary (analytic) autocorrelation functions. A key aspect is their hierarchical form that induces empirical priors on the causes that link successive levels and complement the dynamic priors afforded by the model's equations of motion (see [13] for more details). These models provide a form for the free-energy and its gradients that are needed for filtering, according to (3.8) (see Appendix C for details). We now evaluate this scheme using simulated and empirical data.

4. Comparative Evaluations

In this section, we generate synthetic data using a simple linear convolution model used previously to cross-validate Kalman filtering, Particle filtering, Variational filtering and DEM [3, 4]. Here we restrict the comparative evaluations to DEM because it is among the few schemes that can solve the triple estimation problem in the context of hierarchical models, and provides a useful benchmark in relation to other schemes. Our hope was to show that the conditional expectations of states, parameters and precisions were consistent between GF and DEM but that the GF provided more realistic conditional precisions that are not confounded by the mean-field assumption in implicit in DEM. To compare DEM and GF, we used the a model based on (3.9) that is specified with the functions

$$\begin{gathered}
f^{(1,x)} = Ax^{(1)} + Bv^{(1)}, \\
f^{(1,v)} = Cx^{(1)} + Dv^{(1)}, \\
f^{(2,v)} = \eta^{(v)}, \\
A = \begin{bmatrix} -0.25 & 1.00 \\ -0.50 & -0.25 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad C = \begin{bmatrix} 0.1250 & 0.1633 \\ 0.1250 & 0.0676 \\ 0.1250 & -0.0676 \\ 0.1250 & -0.1633 \end{bmatrix}, \qquad D = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.
\end{gathered} \tag{4.1}$$

Here, the parameters $\theta \supseteq \{A, B, C, D\}$ comprise the state and other matrices. In this standard state-space model, noisy input $v^{(1)} = \eta^{(v)} + z^{(2,v)}$ perturbs hidden states, which decay exponentially to produce an output that is a linear mixture of hidden states. Our example used a single input, two hidden states and four outputs. This model was used to generate data for the examples below. This entails the integration of stochastic differential equations in generalised coordinates, which is relatively straightforward (see [4, Appendix 2]). We generated data over 32 time bins, using innovations sampled from Gaussian densities with the log-precisions of eight and six for observation noise $z^{(1,v)}$ and state noise $z^{(1,x)}$ respectively. We used $d = 4$ orders of generalised motion in all simulations and all random fluctuations were smoothed with a Gaussian kernel whose standard deviation was one quarter of a time bin.



(a)

(b)

(c)

(d)

Figure 2: The linear state-space model and an example of the data it generates: the upper left panel shows simulated data in terms of the output due to hidden states (coloured lines) and observation noise (red lines). The (noisy) dynamics of the hidden states are shown in the upper right panels (blue lines), which are the response to the cause or input on the lower left. The generative model is shown as a Bayesian dependency graph on the lower right.

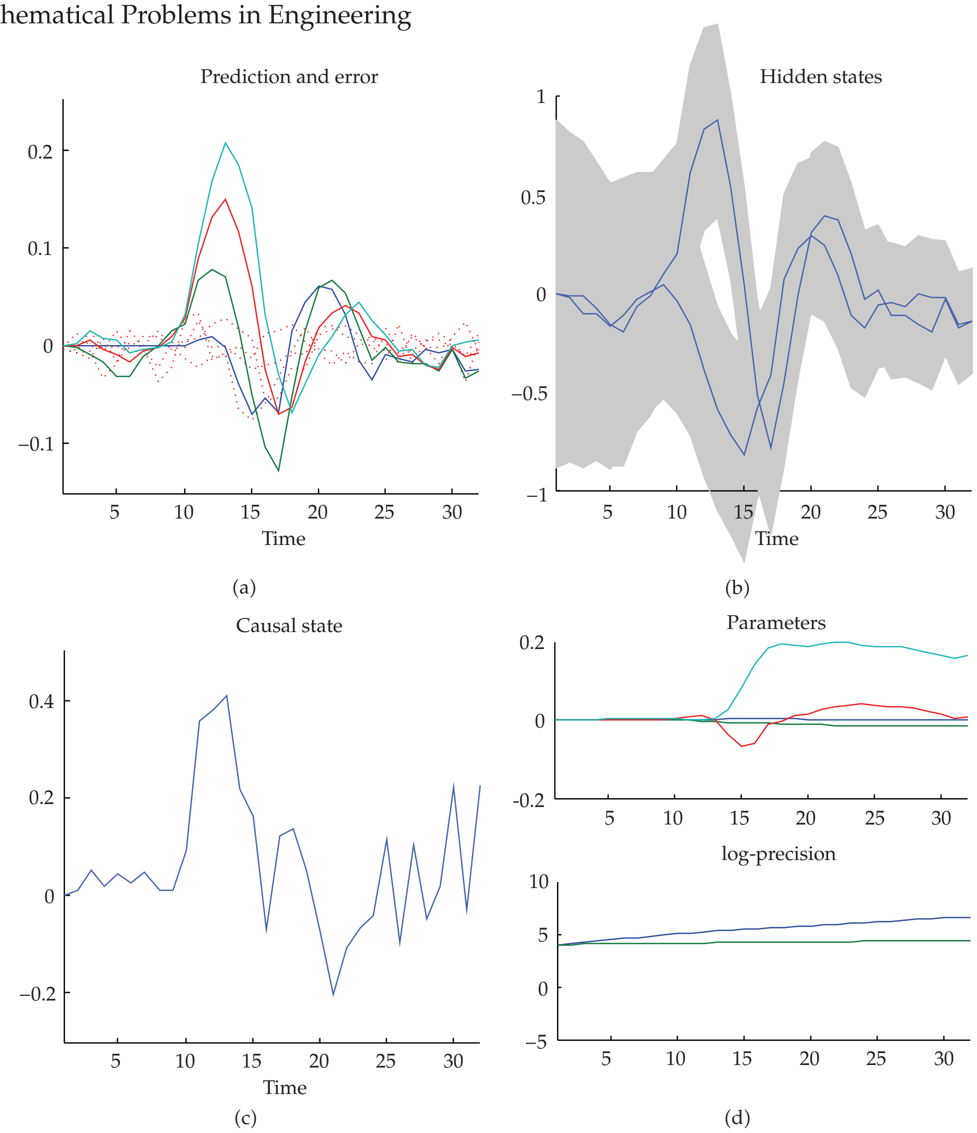
When generating data, we used a deterministic Gaussian bump function $v^{(1)}(t) = \exp((1/4)(t - 12)^2)$ centred on $t = 12$. However, when inverting the model, this cause was unknown and was subject to mildly informative shrinkage priors with zero mean and unit precision; $p(v^{(1)} \mid m) = \mathcal{N}(0, 1)$. These were implemented by fixing the log-precision $\gamma^{(2,v)} = 0$. This model and an example of the data it generates are shown in Figure 2. The format of Figure 2 will be used in subsequent figures and shows the data and hidden states in the top panels and the causes below. The upper left panel shows the simulated data in terms of the output due to hidden states (coloured lines) and observation noise (red lines). The (noisy) dynamics of the hidden states are shown in the upper right panels, which are the response to the cause or input on the lower left. The model is shown as a Bayesian dependency graph on the lower right.

These data were then subject to GF and DEM to recover the conditional densities on the hidden states, unknown cause, parameters and log-precisions. For both schemes, we used uninformative priors on four parameters; $p(\theta_i \mid m) = \mathcal{N}(0, 32)$, and set the remainder to their true value, with infinitely precise priors. The four parameters included two from the equations of motion A_{11}, A_{21} and two from the observer function C_{11}, C_{12}. The log-precision priors were $p(\gamma^{(1,u)} \mid m) = \mathcal{N}(4, 1)$, where $\Pi^{(1,u)} = \exp(\gamma^{(1,u)})I^{(1,u)} : u \in x, v$. In effect, model inversion or filtering entails estimating the hidden states and unknown inputs perturbing these states while, at the same time, identifying the parameters underlying the influence of the hidden states on their motion (and the system's response) and the precision of both observation and state noise. In addition, we estimated the smoothness of the random fluctuations (see Appendix D) but placed very tight priors on the smoothness parameters (whose prior expectations were the true values) to ensure a fair comparison with DEM. Note that we are trying to infer the inputs to the system, which makes this a particularly difficult problem. This inference is precluded in conventional filtering, which usually treats inputs as noise (or as known).

Figure 3 shows the conditional estimates of the states during the first iteration (pass through the time series) of the GF scheme. Here the predicted response and prediction error are shown on the upper left while the conditional expectations of the hidden states and cause are shown on the upper right and lower left respectively. For the hidden states (upper right) the conditional means are depicted by blue lines and the 90% conditional confidence regions by grey areas. These are sometimes referred to as "tubes". Here, the confidence tubes are based upon the marginal conditional density of the states. The key thing to observe here is that the conditional confidence increases with time. This reflects smooth increases in the conditional log-precisions (lower right panel) as they assimilate gradients and are drawn from their initial value (prior expectation or four) toward the true values (of eight and six). Note further how the parameter estimates show similar drifts; however, there is a marked movement towards the true values (from the prior expectation of zero) when their role is disclosed by the perturbation at around fifteen time bins.

The dynamics of the conditional states are prescribed by (2.7) and the slower assimilation dynamics of the conditional parameters and log-precisions by (2.9). By passing repeatedly through the time-series, the parameter and log-precision estimates converge to their stationary solution, at which point free-action stops increasing. As they become better estimates, the estimates of the states become more veridical and confident. By about sixteen iterations we get the estimates shown in Figure 4 (each iteration takes a second or so on a modern PC). Figure 4 uses the same format as Figure 3 but replaces the time-dependent evolution of the parameters and log-precisions with the conditional estimates of their Bayesian average (lower right; see (2.11)). Here, we see that the confidence tubes on the hidden states have shrunk to the extent we can be very confident about the conditional means. In this figure, the confidence tube around the cause is shown and suggests there is much less confidence here; despite the fact that the estimates are remarkably close to the true values (shown as broken grey lines) during the onset and offset of the true cause.

The conditional estimates of the parameters show good agreement with the true values, with the exception of the first parameter of the equation of motion A_{11}. Crucially, this parameter has the highest conditional uncertainty, shown in terms of 90% confidence intervals (red bars). Happily the true values lie within or near these intervals, with a slight overconfidence for the parameters of the observer function (second pair). These estimates of



(a)

(b)

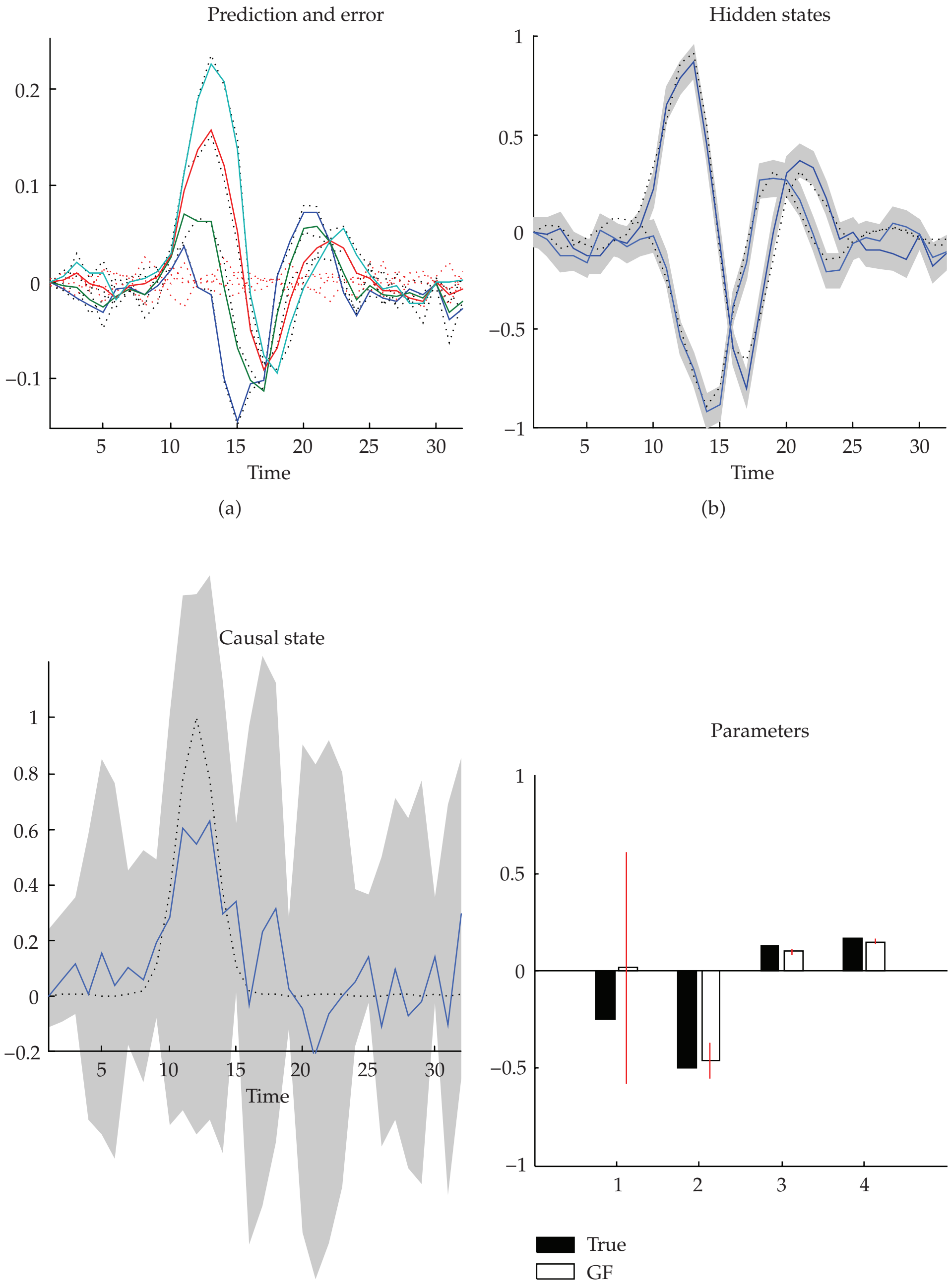
(c)

(d)

Figure 3: Conditional estimates during the first iteration of Generalised Filtering. This format will be used in subsequent figures and summarizes the predictions and conditional densities on the states of a hierarchical dynamic model. The first (upper left) panel shows the predicted response (coloured lines) and the error (red lines) on this response (their sum corresponds to observed data). For the hidden states (upper right) and causes (lower left) the conditional mode is depicted by a blue line and the 90% conditional confidence intervals (regions) by the grey area. The lower right panels show the optimisation of the conditional means of the free parameters (above) and log-precisions (below) as a function of time.

conditional uncertainty should now be compared with the equivalent results from the DEM scheme.

Figure 5 shows the results of DEM using exactly the same format as Figure 4. The first thing to note is the remarkable similarity between the conditional expectations, both for the states and parameters, which are virtually indistinguishable. The most poignant difference between the two schemes lies in the confidence intervals: Although the results of the DEM scheme look much “tighter” in terms of the confidence tubes on states, they fail to contain



(a)

(b)

(c)

(d)

Figure 4: Conditional estimates after convergence of the GF scheme. This figure uses the same format as the previous figure but includes the true values of states used to generate the response (broken grey lines). Furthermore, time-dependent estimates of the parameters and precisions have been replaced with the conditional moments of the Bayesian average of the parameters over time. The black bars are the true values used to generate the data and the white bars are the conditional means. 90% confidence intervals are shown in red.

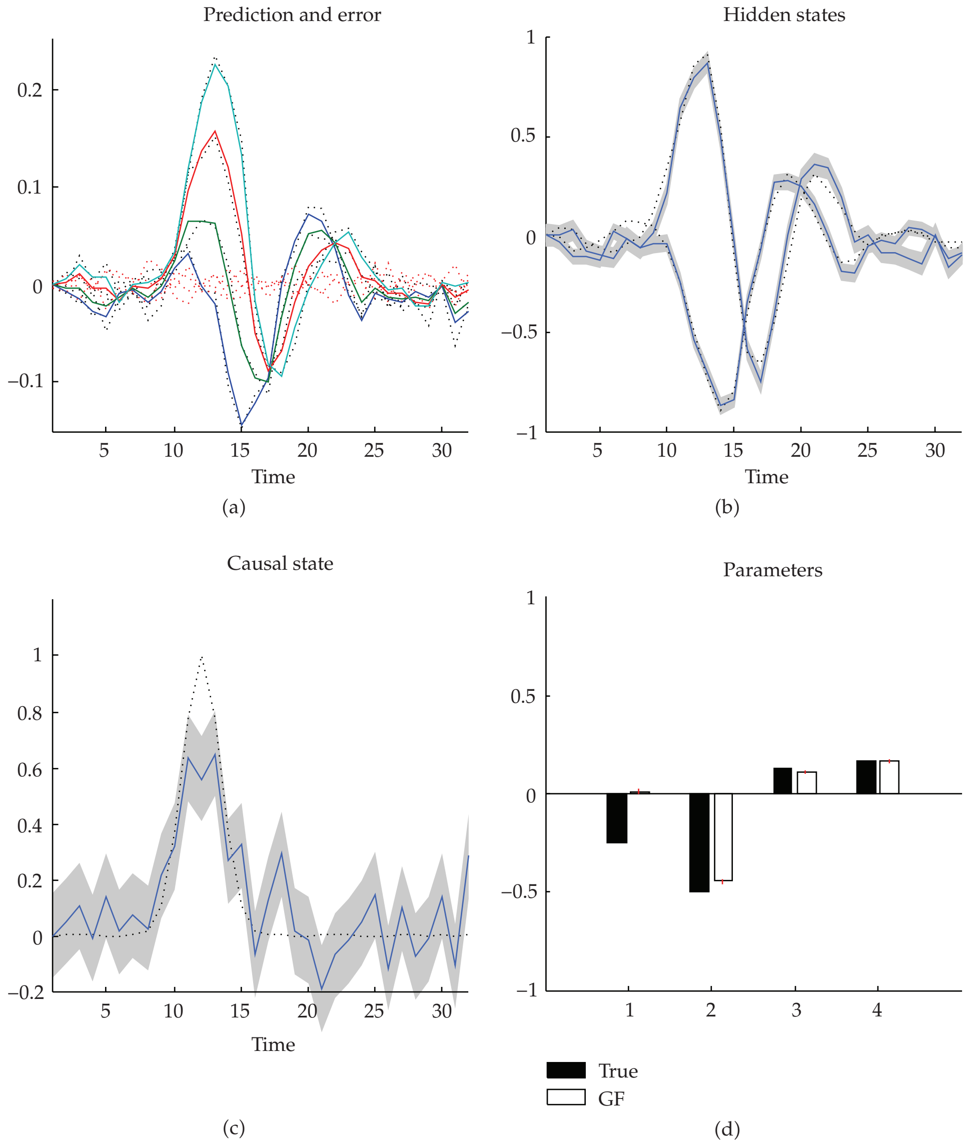


Figure 5: Conditional estimates after convergence of the DEM scheme. This is exactly the same as the previous figure but shows the conditional estimates following Dynamic Expectation Maximisation of the data in Figure 2. The key difference is manifested in terms of smaller confidence tubes and intervals on the states and parameters, respectively, which are overconfident in relation to GF (cf. Figure 4).

the true values at all times. This implicit overconfidence problem is even more acute for the parameter estimates that have very tight posterior precisions, most notably on the first parameter that was identified with low confidence by GF. This example illustrates how the mean-field assumption implicit in DEM can manifest in terms of overconfidence; and how relaxing this assumption with GF partly resolves this problem. We are not claiming that this is a ubiquitous behaviour but are just demonstrating that such differences are easy to find.

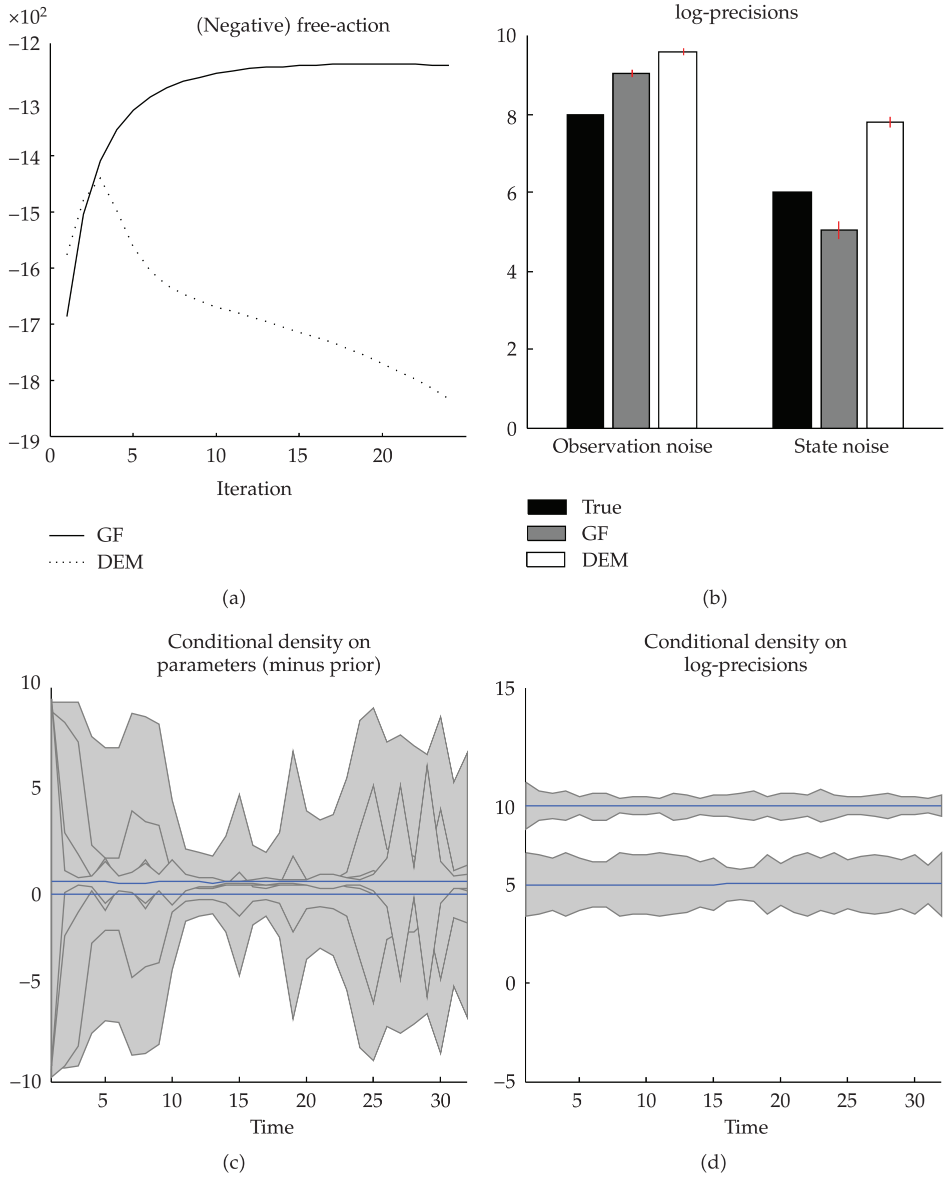


Figure 6: Comparison of Generalised Filtering and DEM. Upper left: negative free-action for GF (solid line) and DEM (dotted line) as a function of iteration number. Upper right: conditional moments of the log-precisions, shown for GF (grey bars) and DEM (white bars), in relation to the true values used to generate the data (black bars). 90% confidence intervals are shown in red. The lower panels show the conditional moments of the parameters (left) and log-precisions (right) as a function of time, after convergence of the GF scheme. The conductional means (minus the prior expectation of the parameters) are shown as blue lines within their 90% confidence tubes (grey regions).

The free-action bound on accumulated log-evidence is shown in Figure 6 as a function of (twenty four) iterations of the GF and DEM schemes. The first thing to note is that GF furnishes a much tighter bound than DEM. This is not surprising because DEM optimises the free-energy of the accumulated data under mean-field assumptions, not the accumulated free-energy. However, it does speak to a profound difference in the bounds on evidence that might have important implications for Bayesian model comparison (this will be pursued in the work of Li et al.—in preparation). GF extremises free-action until the last few iterations, when there is a paradoxical (if small) reversal. We suppose that this is due to the first-order approximations employed during integration of the scheme (see Appendix C). This effect can be mitigated by using smaller step-sizes during solution of the recognition dynamics and by increasing the precision of the random fluctuation in parameters: in practice, we increase κ with each iteration, in proportion to the length of the time-series). The negative free-action increases for the first few iterations of DEM and then decreases to well below its starting point. A quantitative examination of the contributions to free-action suggests that this decrease is largely due to the high conditional log-precisions estimated by DEM (upper right panel).

It can be seen that DEM overestimates the precision of both observation and state noise while GF overestimates observation noise but underestimates state noise. Both schemes are overconfident about their estimate, in that the true values lie outside the 90% confidence intervals (red bars). These confidence intervals are based on accumulating the conditional precisions at each time step. For DEM, this accumulation is an integral part of optimisation whereas for GF it rests on the Bayesian parameter averaging of time-dependent precisions. These are shown on the lower right in terms of the corresponding confidence regions (grey areas). This panel shows a mild contraction of the confidence tube for the precision of state noise, when the hidden states are changing the most (shortly after the cause arrives). This is sensible because state noise is on the motion of hidden sates. A similar but more pronounced effect is seen in the equivalent confidence tubes for the parameters (lower left). Here, all the parameters estimates enjoy a transient decrease in conditional uncertainty during the perturbation because there is more information in the data about their putative role at these times.

4.1. Summary

In this section, we have tried to illustrate some of the basic features of Generalised Filtering and provide some comparative evaluations using an established and formally similar variational scheme (DEM). In this example, the estimates of the conditional means were very similar. The main difference emerged in the estimation of posterior confidence intervals and the behaviour of the free-action bound on accumulated log-evidence. These differences are largely attributable to the mean-field approximation inherent in DEM and related variational schemes. In the next section, we turn to a more complicated (and nonlinear) model to show that GF can recover causal structure from data, which DEM fails to disclose.

5. An Empirical Illustration

In this section, we turn to a model that is more representative of real-world applications and involves a larger number of states, whose motion is coupled in a nonlinear fashion. This model and the data used for its inversion have been presented previously in a comparative evaluation of variational filtering and DEM. Here, we use it to illustrate that the GF scheme operates with nonlinear models and to provide a face validation in this context. This validation rests upon analysing data from a part of the brain known to be functionally selective for visual motion processing [23]. We hoped to show that GF could

Table 1: (a) Biophysical parameters (state-equation). (b) Biophysical parameters (observer).

(a)

	Description	Value (and prior mean
k	rate of signal decay	$1.2\,\mathrm{s}^{-1}$
χ	rate of flow-dependent elimination	$0.31\,\mathrm{s}^{-1}$
τ	transit time	$2.14\,\mathrm{s}$
α	Grubb's exponent	0.36
ϕ	resting oxygen extraction fraction	0.36

(b)

	Description	Value
V_0	Blood volume fraction	0.04
ε	Intra/extra-vascular ratio	1

establish a significant response to visual motion using evoked brain imaging responses. This example was chosen because inference about brain states from noninvasive neurophysiologic observations is an important issue in cognitive neuroscience and functional imaging (e.g., [24–26]), and GF may play a useful role in stochastic dynamic causal modelling.

5.1. The Biophysical Model

We used a hemodynamic model of brain responses to explain evoked neuronal activity that has been described extensively in previous communications (e.g., [27, 28]). In brief, neuronal activity causes an increase in a vasodilatory signal h_1 that is subject to autoregulatory feedback. Blood flow h_2 responds in proportion to this signal and causes change in blood volume h_3 and deoxyhaemoglobin content h_4. The observed signal is a nonlinear function of volume and deoxyhaemoglobin. These dynamics are modelled by the differential equations

$$\begin{aligned}
\dot{h}_1 &= Av - k(h_1 - 1) - \chi(h_2 - 1),\\
\dot{h}_2 &= h_1 - 1,\\
\dot{h}_3 &= \tau(h_2 - F(h_3)),\\
\dot{h}_4 &= \tau\left(h_2 E(h_2) - \frac{F(h_3)h_4}{h_3}\right).
\end{aligned} \tag{5.1}$$

In this model, changes in vasodilatory signal h_1 are elicited by neuronal input v. Relative oxygen extraction $E(h_2) = (1/\phi)(1 - (1-\phi)^{1/h_2})$ is a function of flow, where ϕ is resting oxygen extraction fraction and outflow is a function of volume $F(h_3) = h_3^{1/\alpha}$ with Grubb's exponent α. A description of the parameters of this model and their assumed values (prior expectations) are provided in Table 1. Blood flow, volume, and deoxyhaemoglobin concentration are all nonnegative quantities. One can implement this formal constraint with the transformation $x_i = \ln h_i \Leftrightarrow h_i = \exp(x_i) : i \in 2, 3, 4$. Under this transformation the differential equations above can be written as

$$\dot{h}_i = \frac{\partial h_i}{\partial x_i}\frac{\partial x_i}{\partial t} = h_i \dot{x}_i = f_i(h, v). \tag{5.2}$$

This allows us to formulate the model in terms of hidden states $x_i = \ln h_i$ with unbounded support (i.e., the conditional means can be positive or negative) to give the following functions (see (3.9)):

$$\begin{gathered}
f^{(1,x)} = \begin{bmatrix} \dot{x}_1^{(1)} \\ \dot{x}_2^{(1)} \\ \dot{x}_3^{(1)} \\ \dot{x}_4^{(1)} \end{bmatrix} = \begin{bmatrix} Av^{(2)} - kx_1^{(1)} - \chi(h_2 - 1) \\ \dfrac{x_1^{(1)}}{h_2} \\ \dfrac{\tau(h_2 - F(h_3))}{h_3} \\ \dfrac{\tau(h_2 E(h_2) - F(h_3)h_4/h_3)}{h_4} \end{bmatrix}, \\
f^{(1,v)} = V_0\left(6.93\phi(1 - h_2) + \varepsilon\phi\left(1 - \frac{h_4}{h_3}\right) + (1 - \varepsilon)(1 - h_3)\right), \\
h_i = \exp\left(x_i^{(1)}\right), \\
f^{(2,v)} = \eta^{(v)}.
\end{gathered} \tag{5.3}$$

This model represents a multiple-input, single-output model with four hidden states. The parameters $\theta \supseteq \{k, \chi, \tau, \alpha, \phi, A\}$ of interest here were $A_i \subset \theta : i \in 1, 2, 3$ that couple three distinct neuronal responses $v_i^{(1)} : i \in 1, 2, 3$ to the vasodilatory signal $x_1^{(1)}$. These evoked responses correspond to neuronal activity elicited experimentally, as described next:

5.2. Data and Preprocessing

Data were acquired from a normal subject at 2-Tesla using a Magnetom VISION (Siemens, Erlangen) whole body MRI system, during a visual attention study. Contiguous multislice images were obtained with a gradient echo-planar sequence (TE = 40 ms; TR = 3.22 seconds; matrix size = 64 × 64 × 32, voxel size 3 × 3 × 3 mm). Four consecutive hundred-scan sessions were acquired, comprising a sequence of 10-scan blocks under five conditions. The first was a dummy condition to allow for magnetic saturation effects. In the second, *Fixation*, subjects viewed a fixation point at the centre of a screen. In an *Attention* condition, subjects viewed 250 dots moving radially from the centre at 4.7 degrees per second and were asked to detect changes in radial velocity. In *no attention,* the subjects were asked simply to view the moving dots. In another condition, subjects viewed stationary dots. The order of the conditions alternated between *Fixation* and visual stimulation. In all conditions subjects fixated the centre of the screen. No overt response was required in any condition and there were no actual speed changes. The data were analysed using a conventional SPM analysis (http://www.fil.ion.ucl.ac.uk/spm). The activity from extrastriate cortex (motion-sensitive area V5 [23]) was summarised using the principal local eigenvariate of a region centred on the maximum of a contrast testing for the effect of visual motion. The first 256 samples from this regional response were used for Generalised Filtering and DEM.

The three potential causes of neuronal activity were encoded as box-car functions corresponding to the presence of a visual stimulus, motion in the visual field, and attention. These stimulus functions constitute the priors $\eta_i^{(v)} : i \in 1, 2, 3$ on the three causes in the model. The associated parameters, A_i encode the degree to which these experimental effects induce hemodynamic responses. Given we selected a motion-selective part of the brain; one would anticipate that the conditional probability that A_2 exceeds zero would be large. The regional data were subject to GF using the model in (5.3) and informative shrinkage priors $p(\theta_i \mid m) = \mathcal{N}(\eta^{(\theta)}, 1/32)$ on all but the neuronal coupling parameters A_i (see Table 1 for the prior means). We used mildly informative priors $p(A_i \mid m) = \mathcal{N}(0, 1)$ for the coupling parameters and similarly for the log-precisions $p(\gamma^{(1,u)} \mid m) = \mathcal{N}(2, 1) : u \in v, x$. Otherwise, the analysis was identical to the analysis of the simulated data of the previous section (including unit prior precisions on the causes).

The ensuing conditional means and 90% confidence regions for the causal and hidden states are shown in Figure 7. The dynamics of inferred activity, flow and other biophysical states are physiologically plausible. For example, activity-dependent changes in flow are around 50% $\approx$ exp(0.4), producing about a 3% change in fMRI signal. The conditional estimates of the neuronal coupling parameters, $\theta_i : i = 1, 2, 3$ are shown on the lower right. As anticipated, we can be almost certain that neuronal activity encoding visual motion induces a response. Interestingly, both visual stimulation *per se* and motion appear to elicit responses in this area. This was somewhat unexpected because this area is meant to be selective for motion processing. The interpretation of these results is confounded by profound negative correlations between the visual and motion coupling parameters as evidenced by the conditional covariance among the parameters in Figure 8 (these correlations mean that one can infer confidently the sum of the two parameters but not their differences). This figure shows the conditional means of the eight (Bayesian average) parameters (upper left) and their conditional covariance (upper right). The first five (biophysical) parameters have been estimated with a high conditional confidence while the neuronal parameters are less precise and are interdependent. These conditional dependencies arise from the experiential design, which is highly inefficient for disambiguating between visually evoked responses and motion-specific responses. This is because there were only three blocks of static stimuli, compared to twelve blocks of visual motion. We confirmed this by simulating data using the conditional estimates of the parameters and precision from the current analysis but enforcing motion-specific responses by making $A_1 = 0$, $A_2 = 0.2$ and $A_3 = 0$. The conditional estimates were virtually identical to those from the empirical analysis in Figure 8 (results not shown). This highlights the importance of interpreting not just the marginal conditional density of a single parameter but the conditional density over all parameters estimated.

Figure 8 also shows the time-dependent changes in conditional confidence during the experimental session for the parameters (lower left) and log-precisions (lower right). We have focused on the conditional precision of the visual coupling parameter (grey area) to highlight the increase in precision during the two periods of stationary visual stimulation. These are the only times that the data inform the strength of this parameter in a way that is not confounded by the simultaneous presence of motion in the visual stimulus. Similar transient increases in the conditional precision of the log-precision estimates are seen at the onset and offset of visual stimulation on the lower right.

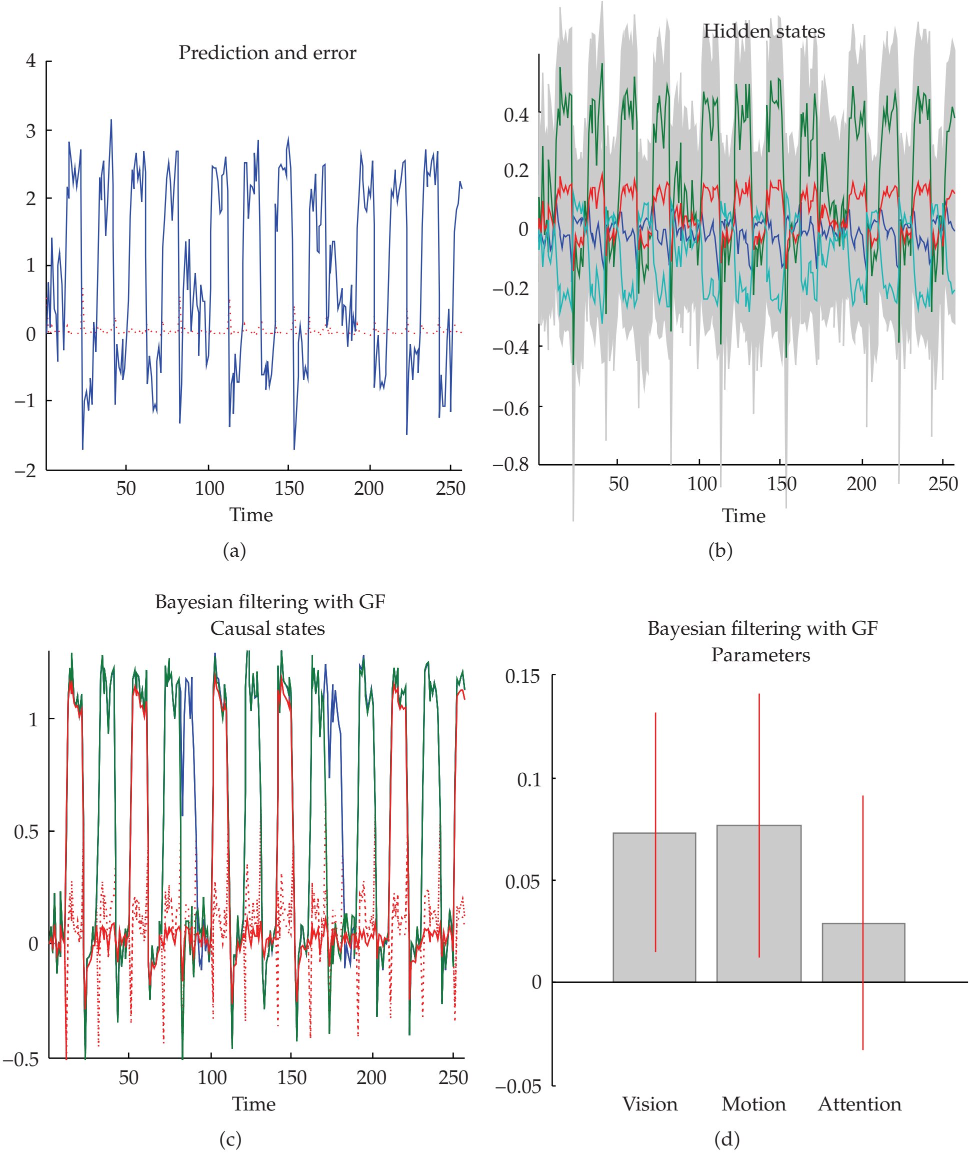
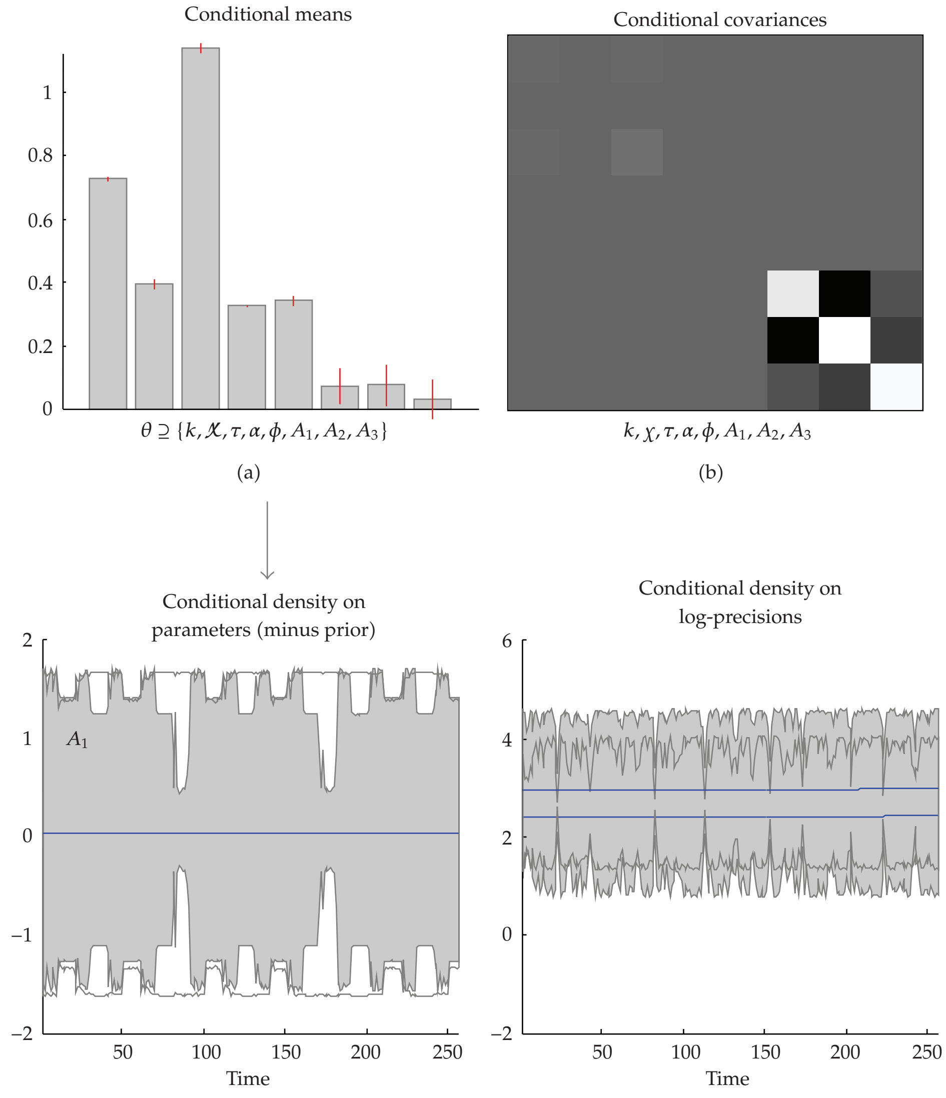


Figure 7: Conditional estimates following Generalised Filtering of the empirical brain imaging time-series. This figure adopts the same format as Figure 4; however, only three (of the eight) parameter estimates are shown (lower right). These are the conditional means (grey bars) and 90% confidence intervals (red bars) of coupling parameters that mediate the effects of vision, motion, and attention on neuronal activity.

A detailed summary of the hemodynamics is shown in Figure 9. This figure focuses on the first 120 time bins and plots the hemodynamic states in terms of the conditional mean of $h_i = \exp(x_1^{(1)})$. Each time bin corresponds to 3.22 seconds. The hidden states are shown in the upper panel, overlaid on periods (grey) of visual motion. It can be seen that a mixture of neuronal activity (the conditional estimate of $Av^{(1)}$ shown in the lower panel), induces a transient burst of vasodilatory signal (blue), which is suppressed rapidly by the resulting increase in flow (green). The increase in flow dilates the venous capillary bed to increase



(a)

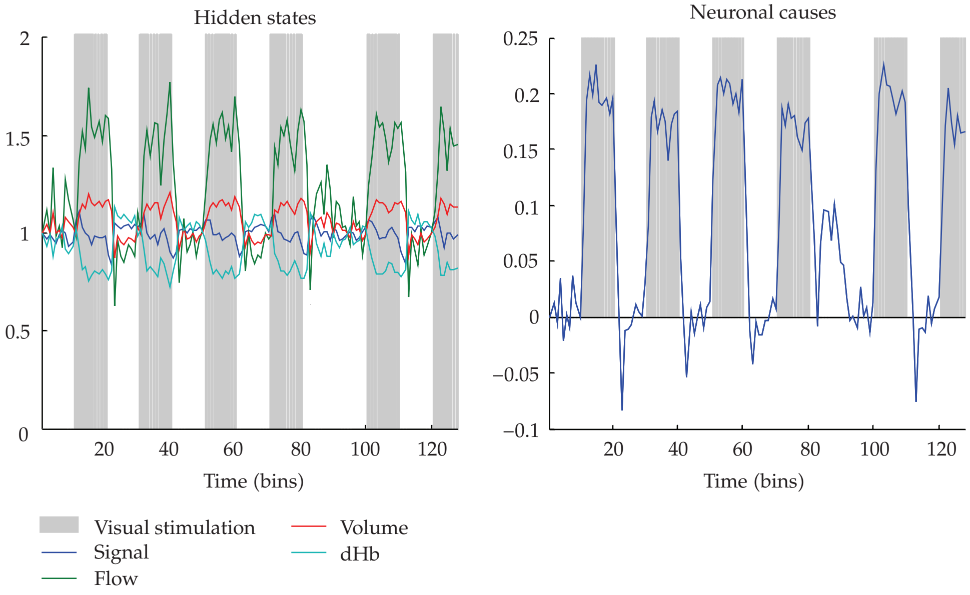
(b)

(c)

(d)

Figure 8: Parameter and precision estimates from the analysis of the empirical data presented in the previous figure. Upper left: conditional means (grey bars) and 90% confidence intervals (red bars) of all (eight) free parameters in this nonlinear model. Upper right: the corresponding conditional covariances are shown in image format (with arbitrary scaling). The lower panels show the time-dependent changes in conditional moments as a function of scan number for the parameters (minus their prior expectation; left) and the log-precisions (right). We have focused on the precision of the first (vision) coupling parameter (grey area) in this figure.

volume (red) and dilute deoxyhaemoglobin (cyan). The concentration of deoxyhaemoglobin determines the measured activity. Interestingly, the underlying neuronal activity appears to show adaptation during visual stimulation and a marked offset transient in nearly all the epochs shown. Note that the conditional densities of the hemodynamic states are non-Gaussian (i.e., lognormal) despite the Laplace assumption entailed by the filtering scheme



(a)

(b)

Figure 9: These are the same results shown in Figure 7 but focussing on the conditional expectations of the hidden states and neuronal causes over the first 128 (3.22 second) time bins. Left panel: the hidden states are overlaid on periods (grey bars) of visual motion. These hidden states correspond to flow-inducing signal, flow, volume, and deoxyhaemoglobin (dHb). It can be seen that neuronal activity, shown in the right panel, induces a transient burst of signal (blue), which is rapidly suppressed by the resulting increase in flow (green). The increase in flow dilates the venous capillary bed to increase volume (red) and dilute deoxyhaemoglobin (cyan). The concentration of deoxyhaemoglobin (involving volume and dHb) determines the measured response.

This is an example of how nonlinear models, under Gaussian assumptions, can be used to model non-Gaussian densities.

5.3. Comparison with DEM

Finally, we analysed the same data using DEM. The results are shown in Figure 10 using the same format as Figure 8. It is immediately obvious that DEM has failed to detect visual motion-dependent responses in this brain area. The coupling parameter estimates are small, both quantitatively and statistically (their confidence intervals include zero). This failure is also evident in the (negative) free-action bound on accumulated log-evidence in Figure 11. This shows that GF provides a much tighter bound relative to DEM. Again, we do not mean to suggest that this is a generic behaviour of variational schemes, just that one can find examples where the mean-field assumption can have a profound effect on inference. Having said this, it took some time to find the length of the time-series and priors on the log-precisions that revealed this difference. In most of our analyses, the conditional estimates from GF and DEM were very similar. The behaviour of the GF free-action over iterations is interesting and speaks to the important point that online evidence accumulation leads to an efficient inversion scheme for long time-series. This is because the conditional moments of the parameters and precisions are near optimal at the end of long sequences (as in conventional

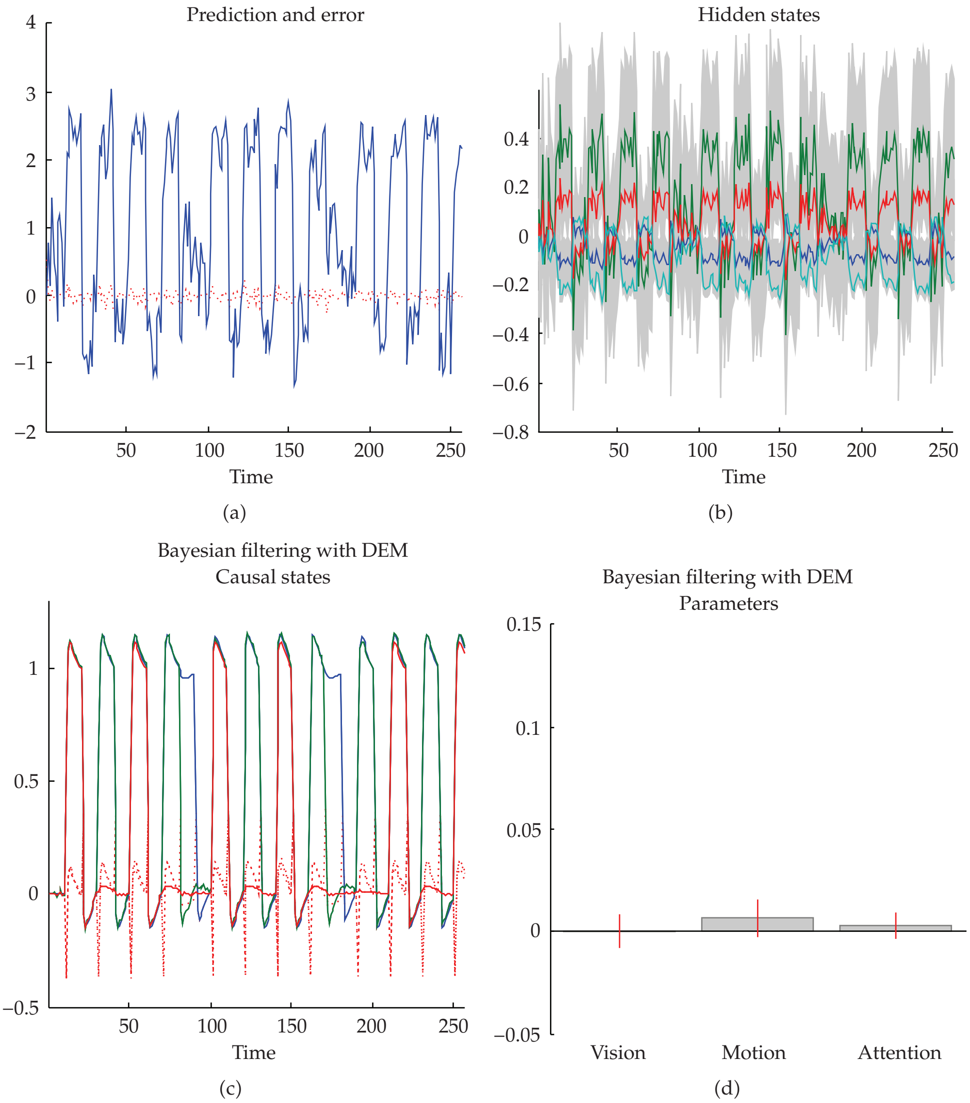


Figure 10: The equivalent results for the hemodynamic deconvolution using *DEM*. These densities should be compared with those in **Figure 7** that were obtained using Generalised Filtering.

assimilation schemes). This is manifest by the convergence of free-action within a handful of iterations (four in the example here, where each iteration took about ten seconds). This computationally efficient form of data assimilation was one of the pragmatic motivations for developing Generalised Filtering.

5.4. Summary

As noted in [4], it is perhaps remarkable that so much conditional information about the underlying neuronal and hemodynamics can be extracted from a single time-series, given only the functional form of its generation. This speaks to the power of generative

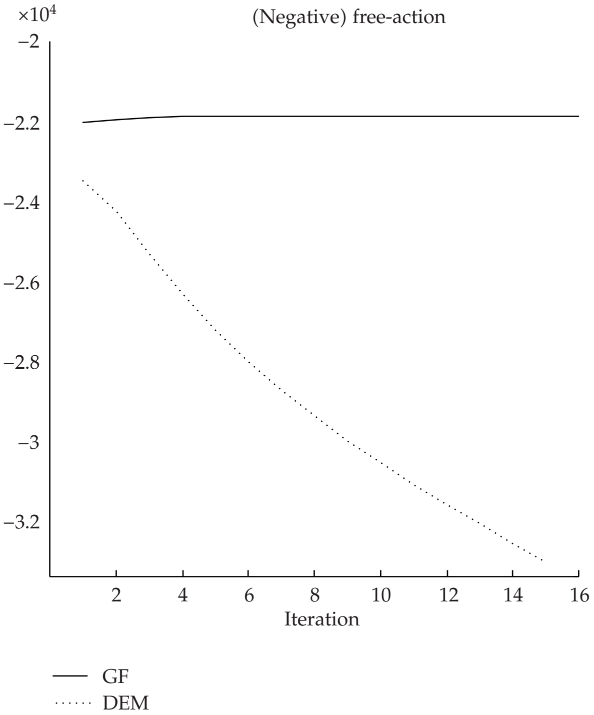


Figure 11: Comparison of Generalised Filtering and DEM for hemodynamic deconvolution. Upper left: negative free-action for GF (solid line) and DEM (dotted line) as a function of iteration number.

modelling, in which constraints on the form of the model allow one to access hidden quantities. To date, dynamic causal models of neuronal systems, measured using fMRI or electroencephalography (EEG), have used known, deterministic causes and have ignored state-noise (see, [25, 26] for important exceptions). One of the motivations for Generalised Filtering was to develop a computational efficient inference scheme that can deal with correlated state-noise, of the sort seen in biological time-series.

6. Conclusion

In this paper, we have introduced Generalised Filtering, an online Bayesian scheme for inverting generative models cast as stochastic differential equations in generalised coordinates of motion. This scheme is based upon a path-integral optimisation of free-energy, where free-energy bounds the log-evidence for a model. Under a Laplace approximation to the true posterior density on the model's unknown variables, one can formulate deconvolution or model inversion as a set of ordinary differential equations, whose solution provides their conditional mean (which implicitly prescribes their conditional precision). Crucially, this density covers not only time-varying hidden states but also parameters, and precisions that change slowly. We have seen that its performance is consistent with equivalent fixed-form variational schemes (Dynamic Expectation Maximisation) that entail the extra assumption that the states, parameters and precisions are conditionally independent.

Although not emphasised in this paper, the basic approach on which Generalised Filtering is based was developed with neurobiological implementation in mind. In other words, we have tried to construct a scheme that could be implemented by the brain in a neurobiologically plausible fashion. This was one of the primary motivations for a dynamical optimisation of the parameter and precision estimates. In future communications, we will focus on the neurobiological interpretation of Generalised Filtering and how it might relate to the optimisation of synaptic activity, efficacy, and gain during perceptual inference in the brain. Our particular focus here will be on state-dependent changes in precision as a model of visual attention (Feldman et al; in preparation). In this context, the recognition dynamics entailed by optimisation can be regarded as simulations of neuronal responses to sensory inputs.

In a more practical setting, this sort of filtering may find a useful role, not only in data analysis but also in online applications, such as speech recognition or active noise cancellation. Indeed, we have already used DEM to infer the hidden states of chaotic systems (hierarchically coupled Lorentz attractors) that were used to simulate bird songs [29]. Applications of Generalised Filtering to similar time-series from chaotic systems may finesse model optimisation in a variety of systems. This is clearly speculative but highlights the potential importance of assimilating data to make inference dynamically, even when the unknown quantities do not change with time.

Appendices

A. Parameter Optimisation, Newton's Method, and Stability

There is a close connection between the updates implied by (2.9) and Newton's method for optimisation. Consider the update under a local linearisation [30], assuming that $\mathcal{L}_\varphi \approx \mathcal{F}_\varphi$

$$\begin{aligned}
\Delta\tilde{\mu}^{(\varphi)} &= \left(\exp\left(t\mathfrak{I}^{(\varphi)}\right) - I\right)\mathfrak{I}^{(\varphi)^{-1}}\dot{\tilde{\mu}}^{(\varphi)},\\
\dot{\tilde{\mu}}^{(\varphi)} &= \begin{bmatrix} \mu'^{(\varphi)} \\ -\mathcal{L}_\varphi - \kappa\mu'^{(\varphi)} \end{bmatrix},\\
\mathfrak{I}_{(\varphi)} &= \frac{\partial\dot{\tilde{\mu}}^{(\varphi)}}{\partial\tilde{\mu}^{(\varphi)}} = \begin{bmatrix} 0 & I \\ -\mathcal{L}_{\varphi\varphi} & -\kappa \end{bmatrix}.
\end{aligned} \tag{A.1}$$

As time proceeds, the change in generalised mean becomes

$$\begin{aligned}
\lim_{t\to\infty}\Delta\tilde{\mu}^{(\varphi)} &= -\mathfrak{I}^{(\varphi)-1}\dot{\tilde{\mu}}^{(\varphi)} = \begin{bmatrix} \Delta\mu^{(\varphi)} \\ \Delta\mu'^{(\varphi)} \end{bmatrix} = -\begin{bmatrix} \mathcal{L}_{\varphi\varphi}^{-1}\mathcal{L}_\varphi \\ \mu'^{(\varphi)} \end{bmatrix},\\
\mathfrak{I}^{(\varphi)-1} &= \begin{bmatrix} \kappa\mathcal{L}_{\varphi\varphi}^{-1} & -\mathcal{L}_{\varphi\varphi}^{-1} \\ 1 & 0 \end{bmatrix}.
\end{aligned} \tag{A.2}$$

The first line means the motion cancels itself and becomes zero, while the change in the conditional mean $\Delta\mu^{(\varphi)} = -\mathcal{L}_{\varphi\varphi}^{-1}\mathcal{L}_\varphi$ becomes a classical Newton update.

An intuition about the need for a high prior precision on the fluctuations of the model parameters can be motivated by a linear stability analysis of the associated recognition dynamics (see (2.9)), with Jacobian in (A.1)

$$\operatorname{eig}\left(\mathfrak{J}^{(\varphi)}\right) = \lambda^{(\varphi)} = -\frac{\kappa}{2} \pm \sqrt{\frac{-4\mathcal{L}_{\varphi\varphi} + \kappa^2}{2}}. \tag{A.3}$$

For the ensuing dynamics to converge exponentially to the conditional mean (when $\mathcal{L}_{\varphi} = 0$) we require $\operatorname{imag}(\lambda^{(\varphi)}) = 0 \Rightarrow \kappa^2 \geq 4\mathcal{L}_{\varphi\varphi}$. In other words, the prior precision κ on the motion of parameters should be greater than the (twice the root) conditional precision of the parameters *per se*. Otherwise, the conditional estimate will exhibit (damped) oscillations.

B. Free-Energy and Action under Mean-Field Approximations

In this appendix, we compare the free-energy $\mathcal{F}^{\mathcal{V}}$ and action $\mathcal{S}^{\mathcal{V}}$ under the mean-field approximation assumed in variational schemes, where

$$\begin{aligned} \mathcal{F}^{\mathcal{V}} &\leq \ln p\left(\bigcup_t \tilde{s}(t) \mid m\right), \\ \mathcal{S}^{\mathcal{V}} &\leq \int dt \ln p(\tilde{s}(t) \mid m). \end{aligned} \tag{B.1}$$

The mean-field approximation finesses evaluation of the free-energy of the accumulated data by making the (implausible) assumption that $q(\bigcup_t u(t), \varphi) = \prod_t q(u(t))q(\varphi)$ for discrete samples at $t \in 1, \ldots, T$. This means that, under the Laplace assumption,

$$\begin{aligned} \mathcal{F}^{\mathcal{V}} &= \sum_t \left(\mathcal{L}^{(u,t)} + \frac{1}{2}\ln\left|\mathcal{P}^{(u,t)}\right| - \frac{p}{2}\ln 2\pi\right) + \mathcal{L}^{(\varphi)} + \frac{1}{2}\ln\left|\sum_t \mathcal{P}^{(\varphi,t)}\right|, \\ \mathcal{S}^{\mathcal{V}} &= \sum_t \left(\mathcal{L}^{(u,t)} + \frac{1}{2}\ln\left|\mathcal{P}^{(u,t)}\right| - \frac{p}{2}\ln 2\pi\right) + \sum_t \left(\mathcal{L}^{(\varphi)} + \frac{1}{2}\ln\left|\mathcal{P}^{(\varphi,t)}\right|\right), \\ \mathcal{L}^{(u,t)} &= \frac{1}{2}\tilde{\varepsilon}_t^{(v)T}\tilde{\Pi}^{(v)}\tilde{\varepsilon}_t^{(v)} - \frac{1}{2}\ln\left|\tilde{\Pi}^{(v)}\right| + \frac{1}{2}\tilde{\varepsilon}_t^{(x)T}\tilde{\Pi}^{(x)}\tilde{\varepsilon}_t^{(x)} - \frac{1}{2}\ln\left|\tilde{\Pi}^{(x)}\right|, \\ \mathcal{L}^{(\varphi)} &= \frac{1}{2}\tilde{\varepsilon}^{(\varphi)T}\tilde{\Pi}^{(\varphi)}\tilde{\varepsilon}^{(\varphi)} - \frac{1}{2}\ln\left|\tilde{\Pi}^{(\varphi)}\right|, \\ \mathcal{P}^{(u,t)} &= \mathcal{L}_{uu}^{(u,t)} + \mathcal{L}_{uu}^{(\varphi)}, \\ \mathcal{P}^{(\varphi,t)} &= \mathcal{L}_{\varphi\varphi}^{(u,t)} + \mathcal{L}_{\varphi\varphi}^{(\varphi)}. \end{aligned} \tag{B.2}$$

Note that (under flat priors on the parameters) the key difference between free-energy and action lies in the contribution from conditional uncertainty (precision) about the parameters. The free-energy contains the log-determinant of the sum of precisions (cf. the precision of the Bayesian parameter average in (2.11)), while the free-action contains the sum of the log-determinant of precisions.

C. Integrating Recognition Dynamics (Filtering)

Filtering involves integrating the ordinary differential equations (2.7) and (2.9) to optimise the conditional means. We can simplify the numerics for hierarchical dynamic models by first collapsing the hierarchy and then collapsing over causal and hidden states

$$\mu^{(v)} = \begin{bmatrix} \mu^{(1,v)} \\ \mu^{(2,v)} \\ \vdots \end{bmatrix}, \quad f^{(v)} = \begin{bmatrix} f^{(1,v)} \\ f^{(2,v)} \\ \vdots \end{bmatrix}, \quad \Pi^{(v)} = \begin{bmatrix} \Pi^{(1,v)} & & \\ & \Pi^{(2,v)} & \\ \vdots & & \ddots \end{bmatrix}, \quad \tilde{\varepsilon}^{(v)} = \begin{bmatrix} \tilde{s} \\ \tilde{\mu}^{(v)} \end{bmatrix} - \begin{bmatrix} \tilde{f}^{(v)} \\ \tilde{\eta}^{(v)} \end{bmatrix},$$

$$\mu^{(x)} = \begin{bmatrix} \mu^{(1,x)} \\ \mu^{(2,v)} \\ \vdots \end{bmatrix}, \quad f^{(x)} = \begin{bmatrix} f^{(1,x)} \\ f^{(2,x)} \\ \vdots \end{bmatrix}, \quad \Pi^{(x)} = \begin{bmatrix} \Pi^{(1,x)} & & \\ & \Pi^{(2,x)} & \\ \vdots & & \ddots \end{bmatrix}, \quad \tilde{\varepsilon}^{(x)} = \mathfrak{D}\tilde{\mu}^{(x)} - \tilde{f}^{(u)},$$

$$\tilde{u} = \begin{bmatrix} \tilde{v} \\ \tilde{x} \end{bmatrix}, \quad \tilde{\varepsilon}^{(u)} = \begin{bmatrix} \tilde{\varepsilon}^{(v)} \\ \tilde{\varepsilon}^{(x)} \end{bmatrix}, \quad \tilde{\Pi}^{(u)} = \begin{bmatrix} R^{(v)} \otimes \Pi^{(v)} & \\ & R^{(x)} \otimes \Pi^{(x)} \end{bmatrix}. \tag{C.1}$$

This gives a simple form for the Gibbs energy that comprises a log-likelihood and log-prior

$$\begin{aligned} \mathcal{L} &= \mathcal{L}^{(u)} + \mathcal{L}^{(\varphi)}, \\ \mathcal{L}^{(u)} &= \frac{1}{2}\tilde{\varepsilon}^{(u)T}\tilde{\Pi}^{(u)}\tilde{\varepsilon}^{(u)} - \frac{1}{2}\ln\left|\tilde{\Pi}^{(u)}\right|, \\ \mathcal{L}^{(\varphi)} &= \frac{1}{2}\tilde{\varepsilon}^{(\varphi)T}\tilde{\Pi}^{(\varphi)}\tilde{\varepsilon}^{(\varphi)} - \frac{1}{2}\ln\left|\tilde{\Pi}^{(\varphi)}\right| \end{aligned} \tag{C.2}$$

with the following integration (Generalised Filtering) scheme:

$$\dot{y} = \begin{bmatrix} \dot{\tilde{s}} \\ \dot{\tilde{\mu}} \end{bmatrix} = \begin{bmatrix} \dot{\tilde{s}} \\ \dot{\tilde{\mu}}^{(u)} \\ \dot{\mu}^{(\theta)} \\ \dot{\mu}^{(\gamma)} \\ \dot{\mu}'^{(\theta)} \\ \dot{\mu}'^{(\gamma)} \end{bmatrix} = \begin{bmatrix} \mathfrak{D}\tilde{s} \\ \mathfrak{D}\tilde{\mu}^{(u)} - \mathcal{F}_{\tilde{u}} \\ \mu'^{(\theta)} \\ \mu'^{(\gamma)} \\ -\mathcal{F}_{\theta} - \kappa\mu'^{(\theta)} \\ -\mathcal{F}_{\gamma} - \kappa\mu'^{(\gamma)} \end{bmatrix}, \qquad \mathfrak{I} = \frac{\partial \dot{y}}{\partial y} = \begin{bmatrix} \mathfrak{D} & & & & & \\ -\mathcal{F}_{\tilde{u}\tilde{s}} & \mathfrak{D} - \mathcal{F}_{\tilde{u}\tilde{u}} & & & & \\ & & & & I & \\ & & & & & I \\ -\mathcal{F}_{\theta\tilde{s}} & -\mathcal{F}_{\theta\tilde{u}} & -\mathcal{F}_{\theta\theta} & -\mathcal{F}_{\theta\gamma} & -\kappa & \\ -\mathcal{F}_{\gamma\tilde{s}} & -\mathcal{F}_{\gamma\tilde{u}} & -\mathcal{F}_{\gamma\theta} & -\mathcal{F}_{\gamma\gamma} & & -\kappa \end{bmatrix}. \tag{C.3}$$

This system can be solved (integrated) using a local linearisation [30] with updates $\Delta y = (\exp(\Delta t \Im) - I)\Im(t)^{-1}\dot{y}$ over time steps Δt, where $\Im(t)$ the filter's Jacobian. Note that we have omitted terms that mediate changes in the motion of state estimates due to changes in parameter estimates. This is because changes in parameter estimates are negligible at the time scale of changes in state estimates. The requisite gradients (evaluated at the conditional expectation) are, with a slight abuse of notion when dealing with derivatives w.r.t. vectors

$$\begin{gathered}
\mathcal{F}_{\tilde{u}} = \frac{1}{2}\tilde{\varepsilon}^{(u)T}\tilde{\Pi}_{\tilde{u}}^{(u)}\tilde{\varepsilon}^{(u)} + \tilde{\varepsilon}_{\tilde{u}}^{(u)T}\tilde{\Pi}^{(u)}\tilde{\varepsilon}^{(u)} - \frac{1}{2}\operatorname{tr}\left(\tilde{\Pi}_{\tilde{u}}^{(u)}\tilde{\Sigma}^{(u)}\right) + \frac{1}{2}\operatorname{tr}(\mathcal{P}_{\tilde{u}}\mathcal{C}), \\
\mathcal{F}_{\gamma} = \frac{1}{2}\tilde{\varepsilon}^{(u)T}\tilde{\Pi}_{\gamma}^{(u)}\tilde{\varepsilon}^{(u)} + \Pi^{(\gamma)}\mu^{(\gamma)} - \frac{1}{2}\operatorname{tr}\left(\tilde{\Pi}_{\gamma}^{(u)}\tilde{\Sigma}^{(u)}\right) + \frac{1}{2}\operatorname{tr}(\mathcal{P}_{\gamma}\mathcal{C}), \qquad \text{(C.4)} \\
\mathcal{F}_{\theta} = \tilde{\varepsilon}_{\theta}^{(u)T}\tilde{\Pi}^{(u)}\tilde{\varepsilon}^{(u)} + \Pi^{(\theta)}\mu^{(\theta)} + \frac{1}{2}\operatorname{tr}(\mathcal{P}_{\theta}\mathcal{C}).
\end{gathered}$$

The corresponding curvatures are (neglecting second-order terms involving states and parameters and second-order derivatives of the conditional entropy)

$$\begin{gathered}
\mathcal{F}_{\tilde{u}\tilde{s}} \approx \mathcal{L}_{\tilde{u}\tilde{s}} = \tilde{\varepsilon}_{\tilde{u}}^{(u)T}\tilde{\Pi}^{(u)}\tilde{\varepsilon}_{\tilde{s}}^{(u)}, \\
\mathcal{F}_{\gamma\tilde{s}} \approx \mathcal{L}_{\gamma\tilde{s}} = \tilde{\varepsilon}^{(u)T}\tilde{\Pi}_{\gamma}^{(u)}\tilde{\varepsilon}_{\tilde{s}}^{(u)}, \\
\mathcal{F}_{\theta\tilde{s}} \approx \mathcal{L}_{\theta\tilde{s}} = \tilde{\varepsilon}_{\theta}^{(u)T}\tilde{\Pi}^{(u)}\tilde{\varepsilon}_{\tilde{s}}^{(u)}, \\
\mathcal{F}_{\tilde{u}\tilde{u}} \approx \mathcal{L}_{\tilde{u}\tilde{u}} = \tilde{\varepsilon}_{\tilde{u}}^{(u)T}\tilde{\Pi}^{(u)}\tilde{\varepsilon}_{\tilde{u}}^{(u)} + \tilde{\varepsilon}_{\tilde{u}}^{(u)T}\tilde{\Pi}_{\tilde{u}}^{(u)}\tilde{\varepsilon}^{(u)} + \tilde{\varepsilon}^{(u)}\tilde{\Pi}_{\tilde{u}}^{(u)}\tilde{\varepsilon}_{\tilde{u}}^{(u)}, \\
\mathcal{F}_{\tilde{u}\theta} \approx \mathcal{L}_{\tilde{u}\theta} = \tilde{\varepsilon}_{\tilde{u}}^{(u)T}\tilde{\Pi}^{(u)}\tilde{\varepsilon}_{\theta}^{(u)} + \tilde{\varepsilon}^{(u)T}\tilde{\Pi}_{\tilde{u}}^{(u)}\tilde{\varepsilon}_{\theta}^{(u)}, \qquad \text{(C.5)} \\
\mathcal{F}_{\theta\theta} \approx \mathcal{L}_{\theta\theta} = \tilde{\varepsilon}_{\theta}^{(u)T}\tilde{\Pi}^{(u)}\tilde{\varepsilon}_{\theta}^{(u)} + \Pi^{(\theta)}, \\
\mathcal{F}_{\gamma\theta} \approx \mathcal{L}_{\gamma\theta} = \tilde{\varepsilon}^{(u)T}\tilde{\Pi}_{\gamma}^{(u)}\tilde{\varepsilon}_{\theta}^{(u)} \approx \mathcal{F}_{\theta\gamma}^{T}, \\
\mathcal{F}_{\tilde{u}\gamma} \approx \mathcal{L}_{\tilde{u}\gamma} = \tilde{\varepsilon}_{\tilde{u}}^{(u)T}\tilde{\Pi}_{\gamma}^{(u)}\tilde{\varepsilon}^{(u)}, \\
\mathcal{F}_{\gamma\gamma} \approx \mathcal{L}_{\gamma\gamma} = \frac{1}{2}\tilde{\varepsilon}^{(u)T}\tilde{\Pi}_{\gamma\gamma}^{(u)}\tilde{\varepsilon}^{(u)} + \Pi^{(\gamma)}.
\end{gathered}$$

Finally, the conditional precision and its derivatives are given by the curvature of Gibb's energy

$$\begin{gathered}
\mathcal{C}^{-1} = \mathcal{P} = \mathcal{L}_{\mu\mu} \approx \begin{bmatrix} \mathcal{L}_{\tilde{u}\tilde{u}} & \mathcal{L}_{\tilde{u}\theta} & 0 \\ \mathcal{L}_{\theta\tilde{u}} & \mathcal{L}_{\theta\theta} & 0 \\ 0 & 0 & \mathcal{L}_{\gamma\gamma} \end{bmatrix}, \\
\mathcal{P}_{\tilde{u}} \approx \begin{bmatrix} & \tilde{\varepsilon}_{\tilde{u}}^{(u)T} \tilde{\Pi}^{(u)} \tilde{\varepsilon}_{\theta\tilde{u}}^{(u)} & \\ \tilde{\varepsilon}_{\theta\tilde{u}}^{(u)T} \tilde{\Pi}^{(u)} \tilde{\varepsilon}_{\tilde{u}}^{(u)} & 2\tilde{\varepsilon}_{\theta}^{(u)T} \tilde{\Pi}^{(u)} \tilde{\varepsilon}_{\theta\tilde{u}}^{(u)} & \\ & & \tilde{\varepsilon}^{(u)T} \tilde{\Pi}_{\gamma\gamma}^{(u)} \tilde{\varepsilon}_{\tilde{u}}^{(u)} \end{bmatrix}, \\
\mathcal{P}_{\theta} \approx \begin{bmatrix} 2\tilde{\varepsilon}_{\tilde{u}}^{(u)T} \tilde{\Pi}^{(u)} \tilde{\varepsilon}_{\tilde{u}\theta}^{(u)} & \tilde{\varepsilon}_{\theta}^{(u)T} \tilde{\Pi}^{(u)} \tilde{\varepsilon}_{\tilde{u}\theta}^{(u)} & \\ \tilde{\varepsilon}_{\tilde{u}\theta}^{(u)T} \tilde{\Pi}^{(u)} \tilde{\varepsilon}_{\theta}^{(u)} & & \\ & & \tilde{\varepsilon}^{(u)T} \tilde{\Pi}_{\gamma\gamma}^{(u)} \tilde{\varepsilon}_{\theta}^{(u)} \end{bmatrix}, \\
\mathcal{P}_{\gamma} \approx \begin{bmatrix} \tilde{\varepsilon}_{\tilde{u}}^{(u)T} \tilde{\Pi}_{\gamma}^{(u)} \tilde{\varepsilon}_{\tilde{u}}^{(u)} & \tilde{\varepsilon}_{\tilde{u}}^{(u)T} \tilde{\Pi}_{\gamma}^{(u)} \tilde{\varepsilon}_{\theta}^{(u)} & \\ \tilde{\varepsilon}_{\theta}^{(u)T} \tilde{\Pi}_{\gamma}^{(u)} \tilde{\varepsilon}_{\tilde{u}}^{(u)} & \tilde{\varepsilon}_{\theta}^{(u)T} \tilde{\Pi}_{\gamma}^{(u)} \tilde{\varepsilon}_{\theta}^{(u)} & \\ & & \tilde{\varepsilon}^{(u)T} \tilde{\Pi}_{\gamma\gamma\gamma}^{(u)} \tilde{\varepsilon}^{(u)} \end{bmatrix}.
\end{gathered} \tag{C.6}$$

Note that we have simplified the numerics here by neglecting conditional dependencies between the precisions and the states or parameters. This is easy to motivate because one is not interested in the conditional precision of the precisions but in the (conditional expectation of the) precisions *per se*; cf. the mean-field approximation implicit in variational approximations. We have also ignored terms due to state-dependent noise, which are not called on in this paper. Finally, we find that the integration of (C.3) is much more stable (in the first few iterations) if we make $\mathcal{P}_{\tilde{u}} = 0$. Again, this rests on a mean-field like approximation, where we ignore the effects of rapid fluctuations in the states on the entropy of the parameters (but not *vice versa*).

These equations may look complicated but can be evaluated automatically using numerical derivatives. All the simulations in this paper used just one routine—*spm_LAP.m.* All the user has to supply are equations defining the generative model. Demonstrations of this scheme are available as part of the SPM software (http://www.fil.ion.ion.ucl.ac.uk.com/spm; *DEM_demo.m*) and reproduce the examples in the main text.

D. Optimising Smoothness Hyperparameters

When including a hyperparameterisation of the smoothness of the random fluctuations, encoded by the precision matrix on generalised motion $R^{(u)}(\gamma) : u \in v, x$, one needs the following derivatives:

$$\begin{gathered}
\partial_{\gamma} \ln\left|\tilde{\Pi}^{(u)}\right| = n^{(u)} \operatorname{tr}\left(R_{\gamma}^{(u)} V^{(u)}\right) + d^{(u)} \operatorname{tr}\left(\Pi_{\gamma}^{(u)} \Sigma^{(u)}\right), \\
\partial_{\gamma\gamma'} \ln\left|\tilde{\Pi}^{(u)}\right| = -n^{(u)} \operatorname{tr}\left(R_{\gamma}^{(u)} V^{(u)} R_{\gamma'}^{(u)} V^{(u)}\right) - d^{(u)} \operatorname{tr}\left(\Pi_{\gamma}^{(u)} \Sigma^{(u)} \Pi_{\gamma'}^{(u)} \Sigma^{(u)}\right), \\
\tilde{\Pi}_{\gamma}^{(u)} = R_{\gamma}^{(u)} \otimes \Pi^{(u)} + R^{(u)} \otimes \Pi_{\gamma}^{(u)}, \\
\tilde{\Pi}_{\gamma\gamma'}^{(u)} = R_{\gamma\gamma'}^{(u)} \otimes \Pi^{(u)} + R^{(u)} \otimes \Pi_{\gamma\gamma'}^{(u)} + R_{\gamma}^{(u)} \otimes \Pi_{\gamma'}^{(u)} + R_{\gamma'}^{(u)} \otimes \Pi_{\gamma}^{(u)},
\end{gathered} \tag{D.1}$$

where $d^{(u)} = \operatorname{rank}(R^{(u)})$ is the order of generalised motion. We have not included these derivatives above, because we used assumed values for smoothness in this paper. However, our software implementation of this scheme estimates smoothness by default.

Software Note

The schemes described in this paper are implemented in Matlab code and are available freely http://www.fil.ion.ucl.ac.uk.com/spm. A DEM toolbox provides several demonstrations from a graphical user interface. These demonstrations reproduce the figures of this paper (see *spm_LAP.m* and ancillary routines).

Acknowledgments

This work was funded by the Wellcome Trust and supported by a grant from the China Scholarship Council (CSC). The author would like to thank Marcia Bennett for help preparing this paper.

References

[1] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[2] R. var der Merwe, A. Doucet, N. de Freitas, and E. Wan, "The unscented particle filter," Tech. Rep. CUED/F-INFENG/TR 380, 2000.

[3] K. J. Friston, "Variational filtering," *NeuroImage*, vol. 41, no. 3, pp. 747–766, 2008.

[4] K. J. Friston, N. Trujillo-Barreto, and J. Daunizeau, "DEM: a variational treatment of dynamic systems," *NeuroImage*, vol. 41, no. 3, pp. 849–885, 2008.

[5] J. Daunizeau, K. J. Friston, and S. J. Kiebel, "Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models," *Physica D*, vol. 238, no. 21, pp. 2089–2118, 2009.

[6] G. L. Eyink, J. M. Restrepo, and F. J. Alexander, "A mean field approximation in data assimilation for nonlinear dynamics," *Physica D*, vol. 195, no. 3-4, pp. 347–368, 2004.

[7] C. Archambeau, D. Cornford, M. Opper, and J. Shawe-Taylor, "Gaussian process approximations of stochastic differential equations," in *Proceedings of the Journal of Machine Learning Research Workshop and Conference*, vol. 1, pp. 1–16, October 2007.

[8] B. Balaji, "Continuous-discrete path integral filtering," *Entropy*, vol. 11, no. 3, pp. 402–430, 2009.

[9] V. A. Billock, G. C. de Guzman, and J. A. Scott Kelso, "Fractal time and 1/f spectra in dynamic images and human vision," *Physica D*, vol. 148, no. 1-2, pp. 136–146, 2001.

[10] J. Carr, *Applications of Centre Manifold Theory*, vol. 35 of *Applied Mathematical Sciences*, Springer, New York, NY, USA, 1981.

[11] M. J. Beal and Z. Ghahramani, "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures," in *Bayesian Statistics*, J. M. Bernardo, M. J. Bayarri, J. O. Berger, et al., Eds., chapter 7, Oxford University Press, Oxford, UK, 2003.

[12] G. V. Puskorius and L. A. Feldkamp, "Decoupled extended Kalman filter training of feedforward layered networks," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '91)*, vol. 1, pp. 771–777, 1991.

[13] K. Friston, "Hierarchical models in the brain," *PLoS Computational Biology*, vol. 4, no. 11, Article ID e1000211, 24 pages, 2008.

[14] P. Whittle, "Likelihood and cost as path integrals," *Journal of the Royal Statistical Society. Series B*, vol. 53, no. 3, pp. 505–538, 1991.

[15] G. L. Eyink, "Action principle in nonequilibrium statistical dynamics," *Physical Review E*, vol. 54, no. 4, part A, pp. 3419–3435, 1996.

[16] G. E. Hinton and D. van Camp, "Keeping neural networks simple by minimising the description length of weights," in *Proceedings of the 6th ACM Conference on Computational Learning Theory (COLT '93)*, pp. 5–13, Santa Cruz, Calif, USA, July 1993.

[17] D. J. C. MacKay, "Free energy minimisation algorithm for decoding and cryptanalysis," *Electronics Letters*, vol. 31, no. 6, pp. 446–447, 1995.

[18] G. L. Eyink, "A variational formulation of optimal nonlinear estimation," Tech. Rep. LA-UR00-5264, University of Arizona, 2001, http://arxiv.org/abs/physics/0011049.

[19] W. D. Penny, K. E. Stephan, A. Mechelli, and K. J. Friston, "Comparing dynamic causal models," *NeuroImage*, vol. 22, no. 3, pp. 1157–1172, 2004.

[20] D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes*, Methuen, London, UK, 1965.

[21] B. Efron and C. Morris, "Stein's estimation rule and its competitors—an empirical Bayes approach," *Journal of the American Statistical Association*, vol. 68, pp. 117–130, 1973.

[22] R. E. Kass and D. Steffey, "Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models)," *Journal of the American Statistical Association*, vol. 84, no. 407, pp. 717–726, 1989.

[23] S. Zeki, J. D. G. Watson, C. J. Lueck, K. J. Friston, C. Kennard, and R. S. J. Frackowiak, "A direct demonstration of functional specialization in human visual cortex," *Journal of Neuroscience*, vol. 11, no. 3, pp. 641–649, 1991.

[24] R. B. Buxton, K. Uludaỹ, D. J. Dubowitz, and T. T. Liu, "Modeling the hemodynamic response to brain activation," *NeuroImage*, vol. 23, supplement 1, pp. S220–S233, 2004.

[25] J. J. Riera, J. Watanabe, I. Kazuki, et al., "A state-space model of the hemodynamic approach: nonlinear filtering of BOLD signals," *NeuroImage*, vol. 21, no. 2, pp. 547–567, 2004.

[26] R. C. Sotero and N. J. Trujillo-Barreto, "Biophysical model for integrating neuronal activity, EEG, fMRI and metabolism," *NeuroImage*, vol. 39, no. 1, pp. 290–309, 2008.

[27] R. B. Buxton, E. C. Wong, and L. R. Frank, "Dynamics of blood flow and oxygenation changes during brain activation: the balloon model," *Magnetic Resonance in Medicine*, vol. 39, no. 6, pp. 855–864, 1998.

[28] K. J. Friston, L. Harrison, and W. Penny, "Dynamic causal modelling," *NeuroImage*, vol. 19, no. 4, pp. 1273–1302, 2003.

[29] K. J. Friston and S. J. Kiebel, "Attractors in song," *New Mathematics and Natural Computation*, vol. 5, no. 1, pp. 83–114, 2009.

[30] T. Ozaki, "A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: a local linearization approach," *Statistica Sinica*, vol. 2, no. 1, pp. 113–135, 1992.