ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 41 (2008) 747–766

Technical Note

Variational filtering

K.J. Friston*

The Wellcome Deptartment of Imaging Neuroscience, University College London, United Kingdom

Received 4 January 2008; revised 14 February 2008; accepted 12 March 2008
Available online 20 March 2008

This note presents a simple Bayesian filtering scheme, using variational calculus, for inference on the hidden states of dynamic systems. Variational filtering is a stochastic scheme that propagates particles over a changing variational energy landscape, such that their sample density approximates the conditional density of hidden and states and inputs. The key innovation, on which variational filtering rests, is a formulation in generalised coordinates of motion. This renders the scheme much simpler and more versatile than existing approaches, such as those based on particle filtering. We demonstrate variational filtering using simulated and real data from hemodynamic systems studied in neuroimaging and provide comparative evaluations using particle filtering and the fixed-form homologue of variational filtering, namely dynamic expectation maximisation.

Keywords: Variational Bayes; Free-energy; Action; Dynamic expectation maximisation; Dynamical systems; Nonlinear; Bayesian filtering; Variational filtering; Generalised coordinates

Introduction

Recently, we introduced a generic scheme for inverting dynamic causal models of systems with random fluctuations on exogenous inputs and hidden states (Friston et al., 2008). This scheme was called dynamic expectation maximisation (DEM) and assumed that the conditional densities on the system's states and parameters were Gaussian. This assumption is know as the Laplace approximation and imposes a fixed form on the conditional density. In this note, we present the corresponding free-form scheme, which allows the conditional density to take any form. This scheme is stochastic and propagates particles over a free-energy landscape to approximate the conditional density with their sample density. Both the ensuing variational filtering and DEM are formulated in generalised coordinates of motion, which finesses many issues that attend the inversion of dynamic models and furnishes a novel approach to Bayesian filtering.

The novel contribution of this work is to formulate the Bayesian inversion of dynamic causal or state-space models in generalised coordinates of motion. Furthermore, we show how the resulting inversion scheme can be applied to hierarchical dynamical models to disclose both the hidden states and unknown inputs, driving a cascade of nonlinear dynamical processes.

This paper comprises four sections. The first reviews variational approaches to ensemble learning, starting with static models and generalising to dynamic systems. We introduce the notion of generalised coordinates and the ensemble dynamics they entail. The ensuing time-varying ensemble density corresponds to a conditional density on the paths or trajectory of hidden states. In the second section, we look at a generic hierarchical dynamic model and its inversion with variational filtering. In the third section, we demonstrate inversion of linear and nonlinear dynamic systems to compare their performance with fixed-form approximations and standard (particle) filtering techniques. In the final section, we provide an illustrative application, in an empirical setting, by deconvolving hemodynamic states and neuronal activity from fMRI responses observed in the brain.

Notation

To simplify notation we will use $f_x = \partial_x f = \partial f/\partial x$ to denote the partial derivative of the function f, with respect to the variable x. We also use $\dot{x} = \partial_t x$ for temporal derivatives. Furthermore, we will be dealing with variables in generalised coordinates of motion, which will be denoted by a tilde; $\tilde{x} = [x, x', x'', \ldots]$. This specifies the position, velocity and higher-order motion of a variable. A point in generalised coordinates can be regarded as encoding the instantaneous trajectory of a variable. However, the motion of this point does not have to be consistent with the trajectory encoded; in other words, the rate of change of position $\dot{x}$ is not necessarily the motion encoded by x' (although it will be under Hamilton's principle of stationary action, as we will see later). Much of what follows recapitulates the material in Friston et al. (2008) so that interested readers can see how the Laplace assumption builds on the basics used in this paper.

* The Wellcome Trust Centre for Neuroimaging, Institute of Neurology, UCL, 12 Queen Square, London, WC1N 3BG, UK. Fax: +44 207 813 1445.
E-mail address: k.friston@fil.ion.ucl.ac.uk.
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2008.03.017

Variational Bayes and ensemble learning

This section reprises Friston et al. (2008), with a special focus on ensemble dynamics that form the basis of variational filtering. Variational Bayes or ensemble learning (Feynman, 1972; Hinton and von Cramp, 1993; MacKay, 1995; Attias, 2000) is a generic approach to model inversion that approximates the conditional density $p(\vartheta|y,m)$ on some model parameters, ϑ, given a model m and data y. We will call the approximating conditional density, $q(\vartheta)$ a variational or ensemble density. Variational Bayes also provides a lower-bound on the evidence (marginal or integrated likelihood) $p(y|m)$ of the model itself. These two quantities are used for inference on parameter and model-space respectively. In what follows, we review variational approaches to inference on static models and their connection to the dynamics of an ensemble of solutions for the model parameters. We then generalise the approach for dynamic systems that are formulated in generalised coordinates of motion. In generalised coordinates, a solution encodes a trajectory; this means inference is on the paths or trajectories of a system's hidden states.

Archambeau et al. (2007) motivate the importance of inference on paths for models based on stochastic differential equations and present a clever approach based on Gaussian process approximations. In the current work, the use of generalised motion makes inference on paths relatively straightforward, because they are represented explicitly (Friston et al., 2008). From the point of view of dynamical systems, inference is on the temporal derivatives of a system's hidden states, which are the bases of the functionals of the free-flow manifold (Gary Green — personal communication).

Other recent developments in this area include extensions of conventional Kalman filtering; for example, Särkkä (2007) considers the application of the unscented Kalman filter to continuous-time filtering problems, where both the state and measurement processes are modelled as stochastic differential equations. In this instance a continuous-discrete filter is derived as a special case of the continuous-time filter. Eyink et al. (2004) consider the problem of data assimilation into nonlinear stochastic dynamic equations using a variational formulation that reduces the approximate calculation of conditional statistics to the minimization of 'effective action'. In what follows, we will show that effective action is a special case of a variational action that can be treated in generalised coordinates.

Variational Bayes

The log-evidence for any parametric model can be expressed in terms of a free-energy and divergence term

$$\begin{aligned} \ln p(y|m) &= F + D(q(\vartheta)||p(\vartheta|y,m)) \\ F &= G + H \\ G(y) &= \langle \ln p(y,\vartheta)\rangle_q \\ H(\vartheta) &= -\langle \ln q(\vartheta)\rangle_q \end{aligned} \tag{1}$$

The free-energy comprises, $G(y)$, which is the internal energy, $U(y,\vartheta)=\ln p(y,\vartheta)$ expected under the ensemble density and the entropy, $H(\vartheta)_q$ which is a measure on that density. In this paper, energies are the negative of the corresponding quantities in physics; this ensures the free-energy increases with log-evidence. Eq. (1) indicates that $F(y,q)$ is a lower-bound on the log-evidence because the Kullback-Leibler cross-entropy or divergence term, $D(q(\vartheta)||p(\vartheta|y,m))$ is always positive. In other words, if the approximating density equals the true posterior density, the divergence is zero and the free-energy is exactly the log-evidence.

The objective is to compute $q(\vartheta)$ for each model by maximising the free-energy and then use $F\approx \ln p(y|m)$ as a lower-bound approximation to the log-evidence for model comparison (e.g., Penny et al., 2004) or averaging (e.g., Trujillo-Barreto et al., 2004). Maximising the free-energy minimises the divergence, rendering the variational density $q(\vartheta)\approx p(\vartheta|y,m)$ an approximate posterior, which is exact for simple (e.g., linear) systems. This can then be used for inference on the parameters of the model selected.

Invoking $q(\vartheta)$ effectively converts a difficult integration problem, inherent in marginalising $p(y,\vartheta|m)$ over the unknown parameters to compute the evidence, into an easier optimisation problem. This rests on inducing a bound that can be optimised with respect to $q(\vartheta)$. To finesse optimisation (e.g., to obtain a tractable solution or suppress computational load), one usually assumes $q(\vartheta)$ factorises over a partition[1] of the parameters

$$q(\vartheta) = \prod_i q(\vartheta^i) \tag{2}$$

Generally, this factorisation appeals to separation of temporal scales or some other heuristic that ensures strong correlations are retained within each subset and discounts weak correlations between them. Usually, one tries to use the most parsimonious partition (and if possible, no factorisation at all). We will not concern ourselves with this partitioning here because our focus on one set of variables, namely time-dependent states.

In statistical physics this is called a mean-field approximation. Under this approximation, it is relatively simply to show that the ensemble density on one parameter set, ϑ^i is a functional of the energy, $U=\ln p(y,\vartheta)$ averaged over the others. When there is only one set, this density reduces to a simple Boltzmann distribution.

Lemma 1. (Free-form variational density; see Corduneanu and Bishop, 2001). *The free-energy is maximised with respect to* $q(\vartheta^i)$ *when*

$$\begin{aligned} \ln q(\vartheta^i) &= V(\vartheta^i) - \ln Z^i \Leftrightarrow \\ q(\vartheta^i) &= \frac{1}{Z^i}\exp\left(V\left(\vartheta^i\right)\right) \\ V\left(\vartheta^i\right) &= \langle U(\vartheta)\rangle_{q\left(\vartheta^{\setminus i}\right)} \end{aligned} \tag{3}$$

where Z^i *is a normalisation constant* (*i.e., partition function*). *We will call* $V(\vartheta^i)$ *the variational energy.* $\vartheta^{\setminus i}$ *denotes parameters not in the i-th set or, more exactly, its Markov blanket. Note that the mode of the ensemble density maximises variational energy.*

Proof. The Fundamental Lemma of variational calculus states that $F(y,q)$ is maximised with respect to $q(\vartheta^i)$ when, and only when

$$\begin{aligned} &\delta_{q(\vartheta^i)}F = 0 \Leftrightarrow \partial_{q(\vartheta^i)}f^i = 0 \\ &\int d\vartheta^i f^i = F \end{aligned} \tag{4}$$

[1] A set of subsets in which each parameter belongs to one, and only one, subset.

$\delta_{q(\vartheta^i)}F$ is the variation of the free-energy with respect to $q(\vartheta^i)$. From Eq. (1)

$$\begin{aligned} f^i &= \int q(\vartheta^i)q(\vartheta^{\setminus i})U(\vartheta)d\vartheta^{\setminus i} - \int q(\vartheta^{\setminus i})q(\vartheta^{\setminus i})\ln q(\vartheta)d\vartheta^{\setminus i} \\ &= q(\vartheta^i)V(\vartheta^i) - q(\vartheta^i)\ln q(\vartheta^i) + q(\vartheta^i)H(\vartheta^{\setminus i}) \Rightarrow \\ \partial_{q(\vartheta^i)}f^i &= V(\vartheta^i) - \ln q(\vartheta^i) - \ln Z^i \end{aligned} \tag{5}$$

We have lumped terms that do not depend on ϑ^i into $\ln Z^i$. The extremal condition is met when $\partial_{q(\vartheta^i)}f^i = 0$, giving Eq. (3). □

If the analytic form in Eq. (3) was tractable (e.g., through the use of conjugate priors) it could be used directly (Attias, 2000). See Beal and Ghahramani (2003) for an excellent treatment of conjugate-exponential models. An alternative approach to optimising $q(\vartheta^i)$ is to consider the density over an ensemble of time-evolving solutions $q(\vartheta^i,t)$ and use its equilibrium solution. This rests on a formulating the ensemble density in terms of ensemble dynamics:

Ensemble densities and the Fokker-Planck formulation

This formulation considers an ensemble of solutions or particles for each parameter set. Each ensemble populates the *i*-th parameter space and is subject to two forces; a deterministic force that causes the particles to drift up the gradients established by the variational energy, $V(\vartheta^i)$ and a random fluctuation $\Gamma(t)$ (i.e., a Langevin force)[2] that disperses the particles. This enforces a local diffusion and exploration of the energy field. The effect of particles in other ensembles is mediated only through their average effect on the internal energy, $V(\vartheta^i)=\langle U(\vartheta)\rangle_{q(\vartheta^{\setminus i})}$, hence mean-field. The equations of motion for each particle are

$$\dot{\vartheta}^i = \nabla V(\vartheta^i) + \Gamma(t) \tag{6}$$

where, $\nabla V(\vartheta^i) = V(\vartheta^i)_{\vartheta^i}$ is the variational energy gradient. Because particles are conserved, the density of particles over parameter space is governed by the free-energy Fokker-Plank equation (also known as the *Kolmogorov* forward equation)

$$\dot{q}(\vartheta^i) = \nabla\cdot\left[\nabla q(\vartheta^i) - q(\vartheta^i)\nabla V(\vartheta^i)\right] \tag{7}$$

This describes the change in local density due to dispersion and drift of the particles. It is trivial to show that the stationary solution for $q(\vartheta^i,t)$ is the ensemble density above by substituting

$$\begin{aligned} q(\vartheta^i) &= \frac{1}{Z^i}\exp(V(\vartheta^i)) \Rightarrow \\ \nabla q(\vartheta^i) &= q(\vartheta^i)\nabla V(\vartheta^i) \Rightarrow \\ \dot{q}(\vartheta^i) &= 0 \end{aligned} \tag{8}$$

at which point the ensemble density is at equilibrium. The Fokker-Planck formulation affords a useful perspective on the variational results above and shows why the variational density is also referred to as the ensemble density; it is the stationary solution to a density on an ensemble of solutions.

[2] *I.e.*, a random fluctuation, whose variance scales linearly with time; in statistical thermodynamics and simulated annealing, this corresponds to a temperature of one, where, $\Omega=\langle\Gamma(t)\Gamma(t)^T\rangle=2I$.

Ensemble learning for dynamic systems

In dynamic systems some parameters change with time. We will call these states and denote them by $u(t)$. The remaining parameters are time-invariant, creating states and parameters; $\vartheta\rightarrow u,\vartheta$. This means the ensemble or variational density $q=q(u,t)q(\vartheta)$ and associated energies become functionals of time. To keep thing as simple, we will focus on optimising the approximate conditional density on the states, $q(u,t)$. Once $q(u,t)$ has been optimised it can be used to optimise $q(\vartheta)$ as described in Friston et al. (2008), to give a variational expectation maximisation (VEM) scheme; this is implemented in our software by summarising $q(u,t)$ in terms of its mean and covariance and optimising the remaining sets of parameters under the Laplace assumption of a Gaussian form. However, from now on, we will assume that ϑ are known, which means the states are the only set of unknowns. In this case, their variational and internal energy become the same thing; i.e., $V(u)=U(u)$ (see Eq. (3)).

By analogy with Lagrangian mechanics, time-varying states have action; the time-integral (or more exactly, anti-derivative) of energy. We will denote action with a bar over the corresponding energy; i.e., $\overline{F}$ $\overline{U}$ and $\overline{V}$ for the free, internal and variational action respectively. The free-action can be expressed as

$$\overline{F} = \int dt\langle U(u,t|\vartheta)\rangle_{q(u,t)} - \int dt\langle \ln q(u,t)\rangle_{q(u,t)} \tag{9}$$

Where $\partial_t\overline{F}=F$ and $U(u,t|\vartheta)=\ln p(y(t),u(t)|\vartheta)$ is the instantaneous energy given the parameters. The free-action, or henceforth action, is simply the path-integral of free-energy. Path-integral is used here in the sense of Whittle (1991), who considers path-integrals of likelihood functions, in the context of optimal estimators in time-series analysis. When $q(u,t)$ shrinks to a point estimator, action reduces to the 'effective action' in variational formulations of optimal estimators for nonlinear state-space models (Eyink, 1996). Under linear dynamics, the effective action coincides with the Onsager–Machlup action in statistical physics (Onsager and Machlup, 1953; Graham, 1978).

The action represents a lower-bound on the integral of log-evidence over time, which, in the context of uncorrelated noise, is simply the log-evidence of the time-series. We now seek $q(u,t)$ which maximises action[3]. By the fundamental Lemma, action is maximised with respect to the ensemble density when, and only when

$$\begin{aligned} \delta_{q(u,t)}\overline{F} = 0 &\Leftrightarrow \partial_{q(u,t)}f = 0 \\ \int duf &= \partial_t\overline{F} = F \end{aligned} \tag{10}$$

It can be seen that the solution is the same as in the static case (Eq. (4)); implying that the ensemble density of the states remains a functional of their variational energy $V(u,t)$

$$q(u,t) = \frac{1}{Z}\exp(V(u,t)) \tag{11}$$

Consider the density of an ensemble that flows on the variational energy manifold. Because this manifold evolves with time,

[3] Subject to the constraint, $\int q(u,t)du=1$.

the ensemble will deploy itself in a time-varying fashion that maximises free-energy and action. Unlike the static case, it will not attain a stationary solution because the manifold is changing. However, the ensemble density will be stationary in a frame of reference that moves with the manifold's topology (assuming its topology does not change too quickly). The equations of motion subtending this stationarity rest on formulating ensemble dynamics in generalised coordinates of motion (*c.f.*, position and momentum in statistical physics):

Ensemble dynamics in generalised coordinates of motion

In a dynamic setting, the ensemble density $q(u,t)$ evolves in a changing variational energy field, $V(u,t)$, which is generally a function of the states and their motion[4]; for example, $V(u,t):=V(v,v',t)$. This induces a variational density in generalised coordinates, where $q(u,t):=q(v,v',t)$ covers position, v and velocity, v'. The use of generalised coordinates is important and lends the ensuing generative models and their inversion useful properties that elude conventional schemes. Critically, generalised coordinates support a conditional density on trajectories or paths, as opposed to the position or state of the generative process. To construct a scheme based on ensemble dynamics we require the equations of motion for an ensemble whose variational density is stationary in a frame of reference that moves with its mode. This can be achieved by coupling different orders of motion through mean-field effects:

Lemma 2. (Ensemble dynamics in generalised coordinates). *The variational density* $q(u,t) = \frac{1}{Z}\exp(V(u,t))$ *is the stationary solution, in a moving frame of reference, for an ensemble whose equations of motion and ensemble dynamics are*

$$\begin{aligned} \dot{v} &= V(u,t)_v + \mu' + \Gamma(t) \\ \dot{v}' &= V(u,t)_{v'} + \Gamma(t) \\ \dot{q}(u,t) &= \nabla_v \cdot q(u)\mu' + \nabla_u \cdot [\nabla_u q(u) - q(u)\nabla_u V(u,t)] \end{aligned} \tag{12}$$

Where μ' *is the mean velocity over the ensemble* (*i.e., a mean-field effect*) *and* $\nabla_v V(u,t) = V(u,t)_v$ *is the variational energy gradient.*

Proof. Substituting $q(u,t) = \frac{1}{Z}\exp(V(u,t))$ and its derivatives into Eq. (12) gives

$$\dot{q}(u,t) = \nabla_v \cdot q(u)\mu' \tag{13}$$

This describes a stationary density in a moving frame of reference, with velocity, μ', as seen using the coordinate transform

$$\begin{aligned} \upsilon &= v - \mu' t \\ q(\upsilon, v', t) &= q(v - \mu' t, v', t) \\ \dot{q}(\upsilon, v', t) &= \dot{q}(v, v', t) - \nabla_v \cdot q(u)\mu' = 0 \end{aligned} \tag{14}$$

Under this coordinate transform, the change in the ensemble density is zero. □

Heuristically, the motion of the particles is coupled through the mean of the ensemble's velocity. In this moving frame of reference, the only forces acting on particles are the deterministic effects exerted by the gradients of the field, which drive particle towards its peak and the random forces, which disperse the particles. Critically, the gradients and peak move with the same velocity and are stationary in the moving frame of reference. This enables particles to 'hit a moving target' because, from the point of view of particles driven by mean-field effects, the target (i.e., peak) is not moving.

The conditional mode and the principle of stationary action

In static systems, the peak or mode of the conditional density maximises variational energy (Lemma 1). Similarly, in dynamic systems, the trajectory of the conditional mode $\tilde{\mu} = \{\mu, \mu'\}$ maximises variational action. This can be seen easily by noting the gradient of the variational energy at the mode is zero

$$\begin{aligned} \partial_u V(\tilde{\mu}, t) = 0 &\Leftrightarrow \delta_u \overline{V}(\tilde{\mu}) = 0 \\ \partial_t \overline{V}(u) &= V(u,t) \end{aligned} \tag{15}$$

This means the mode maximises variational action (by the Fundamental lemma). In other words, changes in variational action, $\overline{V}(u)$, with respect to variations of the mode's path are zero (c.f., Hamilton's principle of stationary action). Intuitively, it means the evolution of the mode follows the peak of the variational energy as it evolves over time, such that tiny perturbations to its path do not change the variational energy. This path has the greatest variational action (i.e., path-integral of variational energy) of all possible paths.

Recall that the position of motion in generalised coordinates is not the same as the motion of the position. This is the counter-intuitive power of generalised coordinates; they allow the state of any particle to move freely along variational energy gradients, irrespective of their generalised motion. Generalised motion only influences movement through the mean-field terms above; such that the motion x' and movement $\dot{x}$ are consistent when, and only when, there are no variational forces (i.e., at the mode of the variational density, where there are no gradients). At this point the motion and movement are consistent; i.e., $\dot{\mu} = \mu'$ and Hamilton's principle of stationary action prevails. In summary, coupling the generalised motion of states and their movement with the mean-field term μ' creates a moving cloud of particles that enshroud the peak, tracking the mode and encoding conditional uncertainty with its dispersion.

See Fig. 1 for a schematic summary and Kerr and Graham (2000) for a related example in statistical physics. Kerr and Graham use ensemble dynamics in generalised coordinates to provide a generalised phase-space version of Langevin and associated Fokker–Planck equations. See also Weissbach et al. (2002) for an example of variational perturbation theory for the free-energy.

Variational filtering

Above, we assumed that the variational energy was a function of position and velocity. We will see later that for most dynamical systems, the variational density and its energy depend on generalised motion to much higher orders. In this instance, the formalism above can be extended to give ensemble dynamics in generalised coordinates, $u = \tilde{v} = (v, v', v'', \ldots)$

$$\begin{aligned} \dot{v} &= V(u,t)_v + \mu' + \Gamma(t) \\ \dot{v}' &= V(u,t)_{v'} + \mu'' + \Gamma(t) \\ \dot{v}'' &= \ldots \end{aligned} \quad \Rightarrow \quad \begin{aligned} \dot{\mu} &= \mu' \\ \dot{\mu}' &= \mu'' = \ddot{\mu} \\ \dot{\mu}'' &= \ldots \end{aligned} \tag{16a}$$

[4] We will just state this to be the case here; it will become obvious why the energy of dynamical systems depends on motion in the next section.

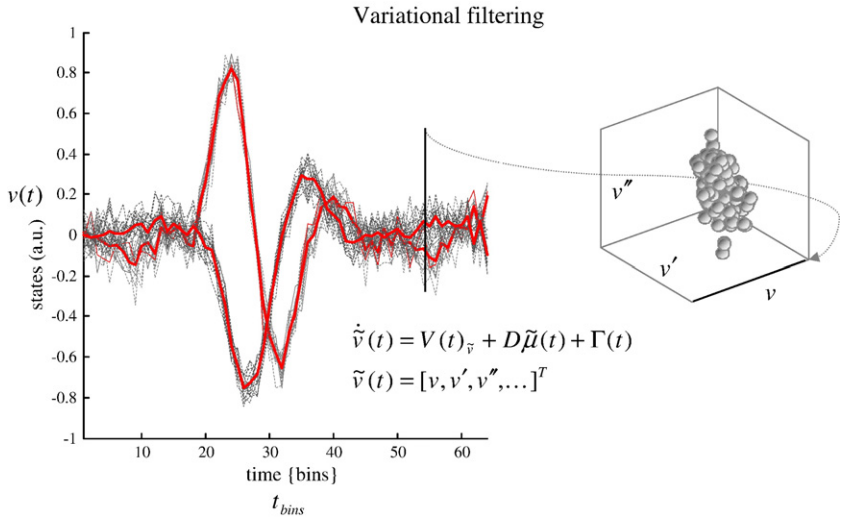

Fig. 1. Schematic illustrating the nature of variational filtering. The left panel shows the evolution of 32 particles over time as they negotiate a changing variational energy landscape. The peak or mode of this landscape is depicted by the red line. Particles flow deterministically towards this mode but are dispelled by random fluctuations to form a cloud that is centred on the mode (insert on the right). The dispersion of this cloud reflects the curvature of the landscape and, through this, the conditional precision of the states. The sample density of the particles in the insert approximates the ensemble or variational density we require. This example comes from a system that will be analyzed in detail in the next section (see Fig. 3). Here we focus on one state in six generalised coordinates of motion, three of which are shown in the insert.

This can be expressed more compactly in terms of a derivative operator D whose first leading-diagonal contains identity matrices

$$\dot{u} = V(u,t)_u + D\tilde{\mu} + \Gamma(t)$$
$$u = \begin{bmatrix} v \\ v' \\ v'' \\ \vdots \end{bmatrix} \quad D = \begin{bmatrix} 0 & I & & \\ & 0 & \ddots & \\ & & \ddots & I \\ & & & 0 \end{bmatrix} \tag{16b}$$

Here, the mode $\tilde{\mu}=(\mu,\mu',\mu'')$ satisfies $V(\tilde{\mu},t)_u=0$ such that the motion of the mode is the mode of the motion; i.e., $\dot{\mu}=\mu'$; this is only true for the mode. Eq. (16a) is the basis for a stochastic, free-form approximation to non-stationary ensemble densities. This entails integrating the path of multiple particles according to the stochastic differential equations in Eq. (16b) and using their sample distribution to approximate $q(u,t)$. We refer to this as variational filtering.

Summary

In this section, we have seen that inference on model parameters can proceed by optimising a free-energy bound on the log-evidence of data, given a model. This bound is a functional of an ensemble density on a mean-field partition of parameters. Using variational calculus, the ensemble or variational density can be expressed in terms of its variational energy. This is simply the internal energy $\ln(p(y,\vartheta|m)$ expected under the Markov Blanket of each set in the partition. When there is only one set, the variational energy reduces to the internal energy *per se*. For dynamic systems, we introduced time-varying states and replaced energies with actions to create a bound that is a functional of time. In the absence of closed-form solutions for the variational densities, they can be approximated using ensemble dynamics that flow on a variational energy manifold, in generalised coordinates of motion. These particles are subject to forces exerted by the variational energy field and mean-field terms from their generalised motion. Free-form approximations obtain by integrating the paths of an ensemble of such particles.

To implement this scheme we need the gradients of the variational energy, which, in the absence of unknown parameters, is simply the internal energy, $V(u,t)=U(u,t|\vartheta)$. This is defined by a generative model. Next, we consider generative models for dynamic systems and the variational filtering entailed.

Nonlinear dynamic models

In this section, we apply the theory of the previous section to an input-state-output model with additive noise. This model has many conventional models as special cases. Critically, it is formulated in generalised coordinates, such that the evolution of the states is subject to empirical priors (Efron and Morris, 1973). This makes the states accountable to their conditional velocity through empirical priors on the dynamics (similarly for high-order motion). Special cases of this generalised model include state-space models used by Bayesian filtering that ignore high-order motion.

Dynamic causal models

To simplify exposition we will deal with a non-hierarchical model and generalise to hierarchical models *post hoc*. A dynamic causal input-state-output model (DCM) can be written as

$$\begin{aligned} y &= g(x,v) + z \\ \dot{x} &= f(x,v) + w \end{aligned} \tag{17}$$

The continuous nonlinear functions f and g of the states are parameterised by $\theta\subset\vartheta$. The states $v(t)$ can be deterministic, stochastic, or both. They are variously referred to as inputs, sources or causes. The states $x(t)$ meditate the influence of the input on the output and endow the system with memory. They are often referred

to as hidden states because they are not observed directly. We assume the stochastic innovations (i.e., observation noise) $z(t)$ are analytic such that the covariance of $\tilde{z}=[z,\dot{z},\ddot{z},\ldots]^T$ is well defined; similarly for the system or state noise, $w(t)$, which represents random fluctuations on the motion of the hidden states. Note that we eschew Ito calculus because we are working in generalised coordinates. This allows us to model innovations that are not limited to Weiner processes (e.g., Brownian motion and other diffusions, whose innovations do not have well-defined derivatives).

Under local linearity assumptions, the motion of the response $\tilde{y}$ is given by

$$\begin{array}{ll} y = g(x,v) + z & x' = f(x,v) + w \\ \dot{y} = g_x x' + g_v v' + \dot{z} & x'' = f_x x' + f_v v' + \dot{w} \\ \dot{y} = g_x x'' + g_v v'' + \ddot{z} & x''' = f_x x'' + f_v v'' + \ddot{w} \\ \vdots & \vdots \end{array} \tag{18}$$

The first (observer) equation shows that the generalised states $u=\{\tilde{v},\tilde{x}\}=\{v,v',\ldots,x,x',\ldots\}$ are needed to generate a response trajectory. This induces a variational density, $q(u,t):=q(\tilde{v},\tilde{x},t)$ on the generalised states. The second (state) equations enforce a coupling between the motions of the hidden states, which confers memory on the dynamics.

The energy functions

The energy function associated with this system; $U(u,t|\vartheta) = \ln p(\tilde{y}|u,\vartheta) + \ln p(u|\vartheta)$ comprises a log-likelihood and prior. Gaussian assumptions about the random fluctuations $p(\tilde{z})=N(0,\tilde{\Sigma}^z)$ and $p(\tilde{w})=N(0,\tilde{\Sigma}^w)$ furnish a likelihood and empirical prior respectively

$$\begin{array}{l} U(u,t|\vartheta) = \ln p(\tilde{y}|u,\vartheta) + \ln p(\tilde{x}|\tilde{v},\vartheta) + \ln p(\tilde{v}) \\ p(\tilde{y}|u,\vartheta) = N(\tilde{y} : \tilde{g}, \tilde{\Sigma}^z) \\ p(\tilde{x}|\tilde{v},\vartheta) = N(D\tilde{x} - \tilde{f} : 0, \tilde{\Sigma}^w) \end{array} \tag{19}$$

This is because these random terms affect the mapping from prediction to response and the evolution of hidden states respectively

$$\begin{array}{ll} \tilde{y} = \tilde{g} + \tilde{z} & D\tilde{x} = \tilde{f} + \tilde{w} \\ g = g(x,v) & f = f(x,v) \\ g' = g_x x' + g_v v' & f' = f_x x' + f_v v' \\ g'' = g_x x'' + g_v v'' & f'' = f_x x'' + f_v v'' \\ \vdots & \vdots \end{array} \tag{20}$$

Here, $\tilde{g}$ and $\tilde{f}$ are the predicted response and motion of the hidden states. To simplify things, we will assume priors on the generalised causes $p(\tilde{v})$ are flat and re-instate informative empirical priors with hierarchical models below. The covariances of the fluctuations $\tilde{\Sigma}(\lambda)^z$ and $\tilde{\Sigma}(\lambda)^w$ depend on known hyperparameters, $\lambda\subset\vartheta$. We will denote the inverse of these covariances as the precisions $\tilde{\Pi}^z$ and $\tilde{\Pi}^w$.

Fig. 2 (left panel) shows the directed graph depicting the conditional dependencies implied by this model. Note that in generalised coordinates there is no explicit temporal dependency and the only constraints on the hidden states are their empirical

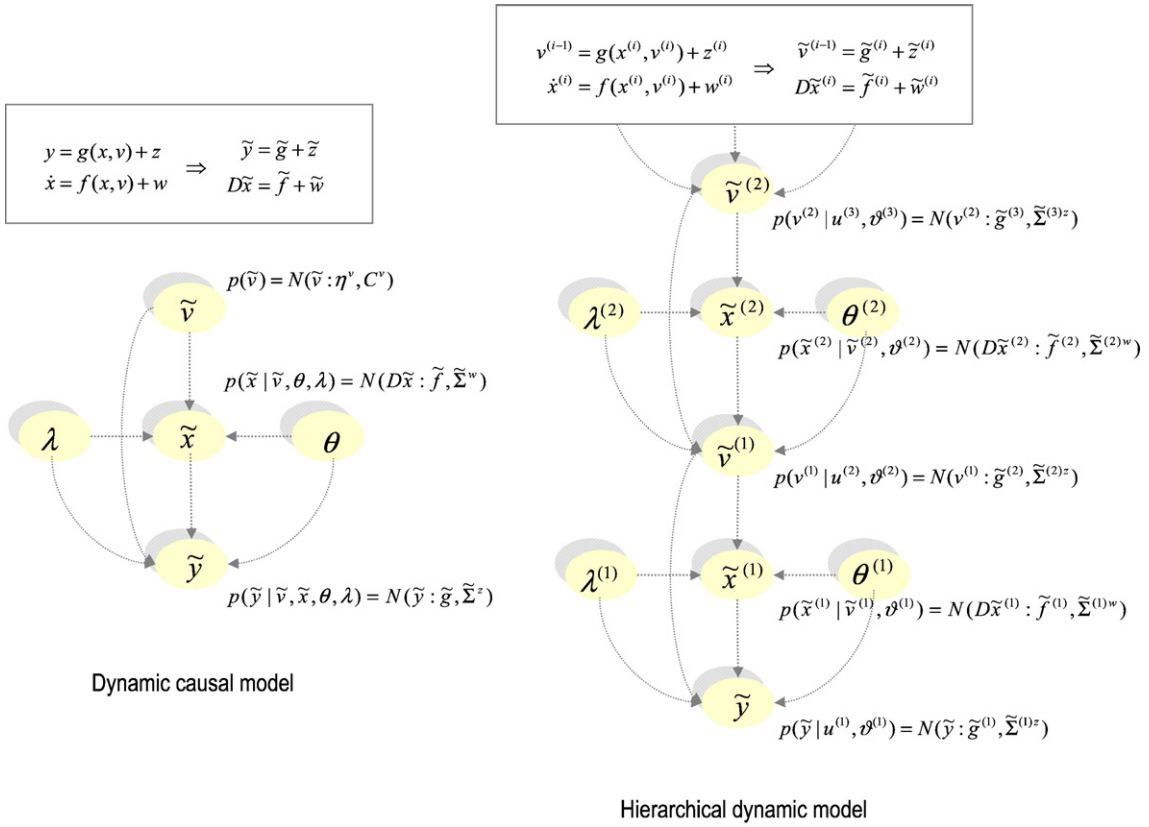

Fig. 2. Conditional dependencies of dynamic (left) and hierarchical (right) models, shown as directed Bayesian graphs. The nodes of these graphs correspond to quantities in the model and the responses they generate. The arrows or edges indicate conditional dependencies between these quantities. The form of the models is provided, both in terms of their state-space formulation (above) and in terms of the prior and conditional probabilities (below). The hierarchal structure of these models induces empirical priors, which depend on states in the level above and provide constraints on the level below.

priors. Readers who are familiar with conventional treatments of state-space models may wonder where all these generalised terms have come from. In fact, they are always present but can be ignored if the precision of the generalised motion of random fluctuations is zero. This is the case for Weiner processes, under which Eq. (18) can be reduced to Eq. (17) with impunity. However, in biophysical systems this is inappropriate because the fluctuations are themselves the product of dynamical systems and are differentiable to high order (this is because the output of a dynamical system is a generalised convolution or smoothing of its input). In short, approximating random effects with a Weiner process is a convenient but specious approximation that precludes an important source of constraints on the dynamics prescribed by state-space models.

For these generative models, the internal energy and gradients are simply (omitting constants)

$$U(t) = -\frac{1}{2}\,\tilde{\varepsilon}^T\,\tilde{\Pi}\,\tilde{\varepsilon} \qquad (21)$$
$$\tilde{\Pi} = \begin{bmatrix} \tilde{\Pi}^z & \\ & \tilde{\Pi}^w \end{bmatrix} \quad \tilde{\varepsilon}(t) = \begin{bmatrix} \tilde{\varepsilon}^v = \tilde{y} - \tilde{g} \\ \tilde{\varepsilon}^x = D\tilde{x} - \tilde{f} \end{bmatrix}$$

The auxiliary variables $\tilde{\varepsilon}(t)$ are prediction errors for the response and the generalised motion of hidden states. The precision of the predictions is encoded by $\tilde{\Pi}$, which depends on the magnitude of the random effects. The gradient of the variational and internal energy is simply[5]

$$V(u,t)_u = U_u = -\tilde{\varepsilon}_u^T\,\tilde{\Pi}\,\tilde{\varepsilon} \qquad (22)$$

Where

$$u = \begin{bmatrix} \tilde{v} \\ \tilde{x} \end{bmatrix} \quad \tilde{\varepsilon}_u = \begin{bmatrix} \tilde{\varepsilon}_v^v & \tilde{\varepsilon}_x^v \\ \tilde{\varepsilon}_v^x & \tilde{\varepsilon}_x^x \end{bmatrix} = -\begin{bmatrix} I \otimes g_v & I \otimes g_x \\ I \otimes f_v & I \otimes f_x - D \end{bmatrix}$$

The form of the generative model (Eq. (17)) means that the partial derivatives of the generalised errors, with respect to the generalised states, comprise diagonal block matrices formed with the Kronecker Tensor product, ⊗. Note the derivative matrix operator in the block encoding $\tilde{\varepsilon}_x^x$. This comes from the prediction error of generalised motion $D\tilde{x} - \tilde{f}$ and ensures the generalised motion of the hidden states conforms to the dynamics entailed by the state equation.

Before describing how these gradients are used to integrate the path of the particles, we consider an important generalisation that endows variational filtering with empirical priors on the causes.

Hierarchical nonlinear dynamic models

Hierarchical dynamic models are important because they subsume many other models. In fact (with the exception of mixture models), they cover most parametric models one could conceive of; from independent component analysis to generalised convolution models. The relationship among these special cases is itself a large area (see Choudrey and Roberts, 2001), to which we will devote a subsequent paper. Here, we simply describe the general form of these models and their inversion. Hierarchical models have the following form, which generalises the ($m=1$) DCM above

$$\begin{aligned} y &= g\left(x^{(1)}, v^{(1)}\right) + z^{(1)} \\ \dot{x}^{(1)} &= f\left(x^{(1)}, v^{(1)}\right) + w^{(1)} \\ &\vdots \\ v^{(i-1)} &= g\left(x^{(i)}, v^{(i)}\right) + z^{(i)} \\ \dot{x}^{(i)} &= f\left(x^{(i)}, v^{(i)}\right) + w^{(i)} \\ &\vdots \\ v^{(m)} &= \eta^v + z^{(m+1)} \end{aligned} \qquad (23)$$

Again, $f^{(i)}$ and $g^{(i)}$ are continuous nonlinear functions of the states. The innovations $z^{(i)}$ and $w^{(i)}$ are conditionally independent fluctuations at each level of the hierarchy. These play the role of observation error or noise at the first level and induce random fluctuations in the states at higher levels. The causes $v^{(i)}$ link levels, whereas the hidden states $x^{(i)}$ are intrinsic to each level. The corresponding directed graphical model, summarising these conditional dependencies, is shown in Fig. 2 (right panel).

The conditional independence of the fluctuations induces a Markov property over levels, which simplifies the architecture of attending inference schemes (Kass and Steffey, 1989). A key property of hierarchical models is their connection to parametric empirical Bayes (Efron and Morris, 1973): Consider the energy function implied by model above

$$U(u,t|\vartheta) = \ln p\left(\tilde{y}|u^{(1)},\vartheta\right) + \ln p\left(u^{(1)}|u^{(2)},\vartheta\right) + \ldots + \ln p\left(\tilde{v}^{(m)}\right) \qquad (24)$$

As with Eq. (19), the first and last terms have the usual interpretation of log-likelihoods and priors. However, the intermediate terms are ambiguous. One the one hand, they are components of the prior. On the other, they depend on quantities that have to be inferred; namely, supraordinate states; hence empirical Bayes. For example, the prediction $\tilde{g}(u^{(i)},\theta^{(i)})$ plays the role of a prior expectation on $\tilde{v}^{(i-1)}$. In short, a hierarchical form endows models with the ability to construct their own priors. This feature is central to many inference and estimation procedures ranging from mixed-effects analyses in classical covariance component analysis to automatic relevance detection in machine learning formulations of related problems (see Friston et al., 2002, 2007 for a fuller discussion of hierarchical models of static data).

The hierarchical forms for the states and predictions are

$$v = \begin{bmatrix} v^{(1)} \\ \vdots \\ v^{(m)} \end{bmatrix} \quad x = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(m)} \end{bmatrix} \quad f = \begin{bmatrix} f\left(x^{(1)}, v^{(1)}\right) \\ \vdots \\ f\left(x^{(m)}, v^{(m)}\right) \end{bmatrix} \quad g = \begin{bmatrix} g\left(x^{(1)}, v^{(1)}\right) \\ \vdots \\ g\left(x^{(m)}, v^{(m)}\right) \end{bmatrix} \qquad (25)$$

The prediction errors now encompass the hierarchical structure and priors on the causes. This means the prediction error on the response is supplemented with prediction errors on the causes

$$\varepsilon^v = \begin{bmatrix} y \\ v \end{bmatrix} - \begin{bmatrix} g \\ \eta \end{bmatrix} \qquad (26)$$

Note that the data y and prior expectations η enter the prediction error at only the lowest and highest level respectively; at intermediate

[5] When the states have a Markov blanket (*i.e.*, there are unknown parameters) the variational energy includes an additional mean-field term, $V(u,t) = U(u,t) + W(u,t)$ as described in Friston et al (2007, 2008).

levels the prediction error $v^{(i-1)}-g(x^{(i)},v^{(i)})$ mediates empirical priors on the causes. The forms of the derivatives of the prediction error with respect to the states are

$$\tilde{\varepsilon}_u = -\begin{bmatrix} I\otimes(g_v - D^T) & I\otimes g_x \\ I\otimes f_v & (I\otimes f_x) - D \end{bmatrix} \tag{27}$$

Comparison with Eq. (22) shows an extra D^T in the upper-right block; this reflects the fact that, in hierarchical models, causes also affect the prediction error within their own level, as well as the lower predicted level. We have presented $\tilde{\varepsilon}_u$ in this form to highlight the role of causes in linking successive hierarchical levels (the D^T matrix) and the role of hidden states in linking successive temporal derivatives (the D matrix). These constraints on the structural and dynamic form of the system are specified by the functions $g(x,v)$ and $f(x,v)$ respectively. The partial derivatives of these functions are assembled according to the structure of the model. Their key feature is a block-diagonal form, reflecting the hierarchical separability of the model

$$\begin{aligned} g_v &= \begin{bmatrix} g_v^{(1)} & & \\ 0 & \ddots & \\ & \ddots & g_v^{(m)} \\ & & 0 \end{bmatrix} \quad g_x = \begin{bmatrix} g_x^{(1)} & & \\ 0 & \ddots & \\ & \ddots & g_x^{(m)} \\ & & 0 \end{bmatrix} \\ f_v &= \begin{bmatrix} f_v^{(1)} & & \\ & \ddots & \\ & & f_v^{(m)} \end{bmatrix} \quad f_x = \begin{bmatrix} f_x^{(1)} & & \\ & \ddots & \\ & & f_x^{(m)} \end{bmatrix} \end{aligned} \tag{28}$$

Note that the partial derivatives of $g(x,v)$ have an extra row to accommodate the highest level.

The precisions and temporal smoothness

In hierarchical models, the precision at the first level encodes the precision of observation noise; at the last level, it is simply the prior precision of the causes, $\Pi^v=\Pi^{(m+1)z}$. The intermediate levels are empirical prior precisions on the causes of dynamics in subordinate levels. Independence assumptions about the innovations means their precisions have a block-diagonal form

$$\Pi^z = \begin{bmatrix} \Pi^{(1)z} & & & \\ & \ddots & & \\ & & \Pi^{(m)z} & \\ & & & \Pi^v \end{bmatrix} \quad \Pi^w = \begin{bmatrix} \Pi^{(1)w} & & \\ & \ddots & \\ & & \Pi^{(m)w} \end{bmatrix} \tag{29}$$

In generalised coordinates, precisions are the Kronecker tensor product of the precision of temporal derivatives, $S(\gamma)$ and the precision on each innovation

$$\tilde{\Pi}^z = S(\gamma)\otimes\Pi^z \tag{30}$$

Similarly for Π^w. This assumes the precisions can be factorised, into dynamic and innovation-specific parts. The dynamic part encodes the temporal dependencies among the innovations and can be expressed as a function of their autocorrelations

$$S(\gamma) = \begin{bmatrix} 1 & 0 & \ddot{\rho}(0) & \cdots \\ 0 & -\ddot{\rho}(0) & 0 & \\ \ddot{\rho}(0) & 0 & \ddot{\rho}(0) & \\ \vdots & & & \ddots \end{bmatrix}^{-1} \tag{31}$$

Here $\ddot{\rho}(0)$ is the second derivative of the autocorrelation function of the fluctuations, evaluated at zero. It is a ubiquitous measure of roughness in the theory of stochastic processes. See Cox and Miller (1965) for details.

Note that when the innovations are uncorrelated, the curvature (and higher derivatives) of the autocorrelation $\ddot{\rho}(0)\to\infty$ become large. In this instance, the precisions of the temporal derivatives fall to zero and the energy is determined by, and only by, the prediction error on the causes and the motion of the hidden states. This limiting case is the model assumed by state-space models used in conventional Bayesian filtering. $S(\gamma)$ can be evaluated for any analytic autocorrelation function. For convenience, we assume that the temporal correlations of all innovations have the same Gaussian form. This gives

$$S(\gamma) = \begin{bmatrix} 1 & 0 & -\frac{1}{2}\gamma & \cdots \\ 0 & \frac{1}{2}\gamma & 0 & \\ -\frac{1}{2}\gamma & 0 & \frac{3}{4}\gamma^2 & \\ \vdots & & & \ddots \end{bmatrix}^{-1} \tag{32}$$

Where γ is the precision parameter of a Gaussian $\rho(t)$ and increases with roughness. Clearly, the conditional density of the temporal hyperparameter $\gamma\subset\vartheta$ could be estimated. Here, for simplicity, we assume γ is known. Typically, $\gamma>1$, which ensures the precisions of higher-order derivatives converge quickly. This is important because it enables us to truncate the representation of generalised coordinates to a relatively low order. This is because high-order prediction errors have a vanishingly small precision. In Friston et al. (2008) we established that an embedding order of $n=6$ is sufficient in most circumstances (i.e., a representation of high-order derivatives up to sixth order).

From derivatives to sequences

Up until now we have treated the trajectory of the response $\tilde{y}(t)$ as a known quantity, as if data were available in generalised coordinates of motion; however, empirical data are usually measured discretely, as a sequence, $y=[y(t_1),\ldots,y(t_N)]^T$. This measurement or sampling is part of the generative process, which has to be accommodated in the first level of the model: A discrete sequence $g=[g(t_1),\ldots,g(t_N)]^T$ can be generated from the derivatives $\tilde{g}(t)$ using Taylor's theorem

$$g = \tilde{E}(t)\tilde{g}(t) \quad \tilde{E}(t) = E\otimes I \quad E_{ij} = \frac{(t_i - t)^{(j-1)}}{(j-1)!} \tag{33}$$

Provided $\tilde{E}(t)$ is invertible[6], we can use the linear bijective mapping $\tilde{E}(t)\tilde{y}(t)=y$ to evaluate generalised responses from local sequences (see Friston et al., 2008 for details).

Integrating the path of particles

Variational filtering integrates the paths of an ensemble of particles, $u^{[i]}$ according to Eq. (16b), so that their sample density at any time, approximates the conditional density on the states, $q(u,t)$. This entails integrating stochastic differential equations for each

[6] The number of elements in a local sequence equals the number of generalised coordinates.

particle, using an augmented system that includes the data and priors. This ensures that changes in the energy gradients are accommodated properly in the integration scheme. There are several ways to integrate these equations; we use a computationally intensive but accurate scheme (Ozaki, 1992) based on the matrix exponential of the system's Jacobian, $\Im(t)$. Ozaki (1992) shows the ensuing updates are consistent, coincide with the true trajectory (at least for linear systems) and retain the qualitative characteristics of the continuous formulation. For each particle, we update the states, over a time-step Δt (usually the time between observations) using

$$\begin{bmatrix} \Delta\widetilde{y} \\ \Delta u^{[i]} \\ \Delta\widetilde{\eta} \end{bmatrix} = (\exp(\Delta t\Im) - I)\Im^{-1}\begin{bmatrix} D\widetilde{y} \\ V\left(u^{[i]},t\right)_u + D\widetilde{\mu} \\ D\widetilde{\eta} \end{bmatrix} + \begin{bmatrix} 0 \\ \varsigma \\ 0 \end{bmatrix} \tag{34a}$$

where $\widetilde{\mu} = \langle u^{[i]} \rangle_i$ is the sample mean over particles. The Jacobian

$$\Im = \begin{bmatrix} D & 0 & 0 \\ V_{uy} & V_{uu} & V_{u\eta} \\ 0 & 0 & D \end{bmatrix} \tag{34b}$$

comprises the curvatures

$$\begin{matrix} V_{uu} = -\widetilde{\varepsilon}_u^T\widetilde{\Pi}\widetilde{\varepsilon}_u \\ V_{uy} = -\widetilde{\varepsilon}_u^T\widetilde{\Pi}\widetilde{\varepsilon}_y \\ V_{u\eta} = -\widetilde{\varepsilon}_u^T\widetilde{\Pi}\widetilde{\varepsilon}_\eta \end{matrix} \quad \widetilde{\varepsilon}_y = \begin{bmatrix} I\otimes\varepsilon_y^v \\ 0 \end{bmatrix} \quad \widetilde{\varepsilon}_\eta = \begin{bmatrix} I\otimes\varepsilon_\eta^v \\ 0 \end{bmatrix} \quad \varepsilon_y^v = \begin{bmatrix} I \\ 0 \end{bmatrix} \quad \varepsilon_\eta^v = -\begin{bmatrix} 0 \\ I \end{bmatrix}$$

The forms for the error derivatives ε_y^v *and* ε_η^v reflect the fact that data and priors only affect the prediction error at the first and last levels respectively. The stochastic term ς in Eq. (34a) is sampled from a unit normal distribution and scaled by the square root of its implicit covariance

$$\Sigma^\varsigma = \langle \varsigma\varsigma^T \rangle = \int_0^{\Delta t} \exp(tV_{uu})\Omega\exp(tV_{uu})^T dt \tag{35}$$

Where Ω=2I is the covariance of the underlying Langevin force, which is the same over all states and orders of motion. This can be computed fairly quickly as described in Appendix A. Note that when Δt is small, the covariance of the stochastic terms $\Sigma^\varsigma \approx \Omega\Delta t$. The form of Eq. (35) is explained in Appendix B.

For each particle and time-step, the prediction errors and ensuing gradients and curvatures are evaluated and the particle's position in generalised coordinates is updated according to Eq. (34a). The initial positions are drawn from a unit normal distribution. After the paths have been integrated to the end of the observed time series, their sample density constitutes an approximation to the time-varying conditional density on hidden states and causes. In most cases, one is interested in the marginal density on the values of the states (e.g., the conditional mean and covariance); however, the conditional density actually encodes a distribution on generalised states and implicitly their instantaneous trajectories. Note that unlike particle filtering or related sampling techniques, particles are not selected or destroyed. Furthermore, unlike Bayesian smoothing schemes, there is no need for forward and backward passes. Variational filtering uses a single pass, while conserving particles. See Eyink (2001) for a discussion of variational estimators that enjoy 'mean optimality'. These obtain from forward integration of a 'perturbed' Fokker-Planck equation and backward integration of an adjoint equation, related to the Pardoux–Kushner equation for optimal smoothing.

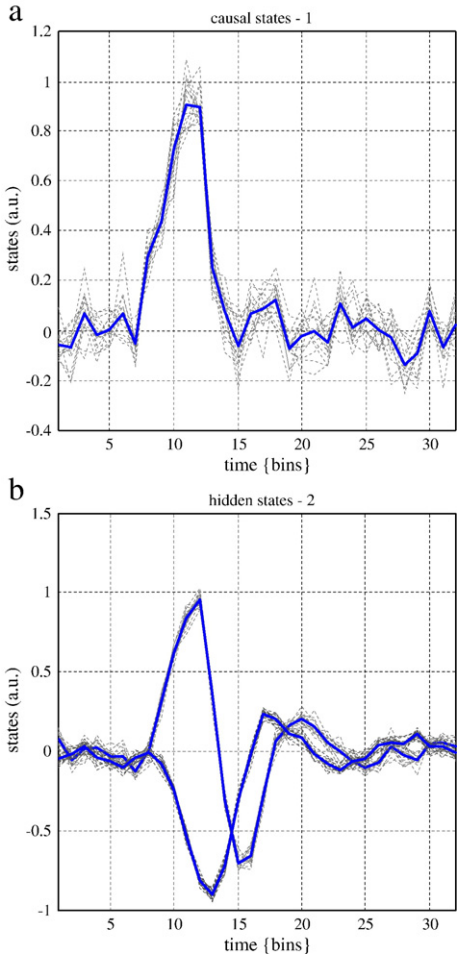

Fig. 3. Variational densities on the causal and hidden states of a linear convolution model. These plots show the trajectories or paths of sixteen particles tracking the mode of the input or cause (a) and two hidden states (b). The sample mean of this distribution is shown in blue over the 32 time bins, during which responses or data were inverted.

This concludes the theoretical background. In the next section, we examine the operational features of this inversion scheme.

Variational filtering of linear and nonlinear models

In this section, we focus on the functionality of variational filtering and how it compares with established schemes. This functionality is quite broad because the conditional density covers not only hidden states but also the causal states or inputs. This means we can infer on the inputs to a system. This is precluded in conventional filtering, which treat the inputs as noise. We consider Bayesian deconvolution of dynamic systems to estimate hidden and causal states, assuming the parameters and hyperparameters are known. We start with a simple linear model to outline the basic nature of variational filtering and then move on to nonlinear dynamic models that have been used previously for comparative studies of extended Kalman and particle filtering.

A linear convolution model

To compare free and fixed-form schemes, we start with a linear convolution or state-space model, under which the approximating

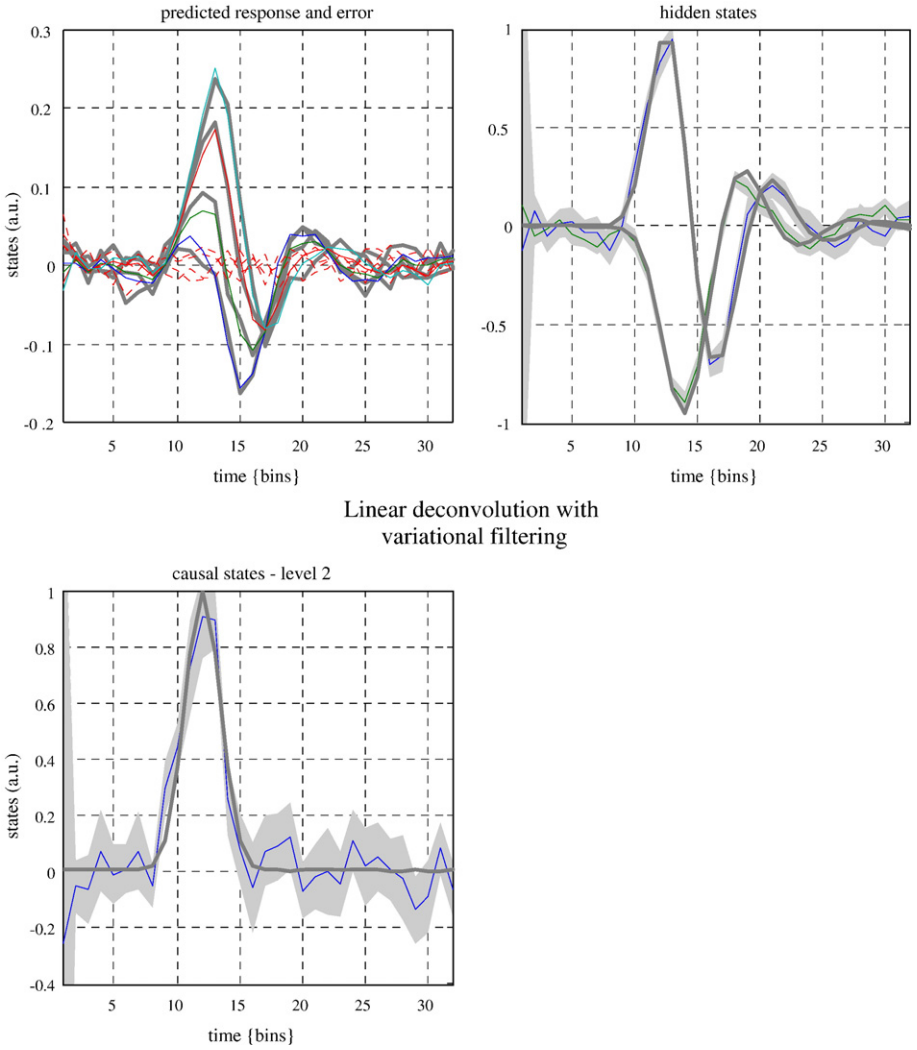

Fig. 4. Alternative representation of the sample density shown in the previous figure. This format will be used in subsequent figures and summarizes the predictions and conditional densities on the states of a hierarchical dynamic model. Each row corresponds to a level, with causes on the left and hidden states on the right. In this case, the model has just two levels. The first (upper left) panel shows the predicted response and the error 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 coloured line and the 90% conditional confidence intervals by the grey area. In this case, the confidence tubes were based on the sample density of the ensemble of particles shown in the previous figure. Finally, the thick grey lines depict the true values used to generate the response.

conditional densities should be the same. This model can be expressed as

$$\begin{aligned} y &= g(x,v) + z^{(1)} \\ \dot{x} &= f(x,v) + w^{(1)} \\ v &= \eta + z^{(2)} \end{aligned} \tag{36}$$

$$\begin{aligned} g(x,v) &= \theta_1 x \\ f(x,v) &= \theta_2 x + \theta_3 v \end{aligned}$$

We have omitted superscripts on the states because there is only one level of hidden states and one level of inputs. In this model, input perturbs hidden states, which decay exponentially to produce an output that is a linear mixture of hidden states. Our example uses a single input, two hidden states and four outputs. This is a single input-multiple output linear system, where

$$\theta_1 = \begin{bmatrix} 0.1250 & 0.1633 \\ 0.1250 & 0.0676 \\ 0.1250 & -0.0676 \\ 0.1250 & -0.1633 \end{bmatrix} \quad \theta_2 = \begin{bmatrix} -0.25 & 1.00 \\ -0.50 & -0.25 \end{bmatrix} \quad \theta_3 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{37}$$

This model is used to generate data for the examples below. This entails the integration of stochastic differential equations in generalised coordinates, which is relatively straightforward (see Appendix B of Friston et al., 2008). We generated data over 32 time bins, using innovations sampled from Gaussian densities with the following precisions[7]

Linear convolution model

Level	$g(x,v)$	$f(x,v)$	Π^z	Π^w	η
$m=1$	$\theta_1 x$	$\theta_2 x+\theta_3 v$	e^8	e^{16}	
$m=2$			*1*		*0*

When generating data, we used a deterministic Gaussian function $v = \exp\left(\frac{1}{4}(t-12)^2\right)$ centred on $t=12$. However, when

[7] Where scalar precisions scale the appropriate identity matrix.

inverting the model the cause is unknown and is subject to mildly informative shrinkage priors with zero mean and unit precision; $p(v)=N(0,1)$. We will use embeddings of $n=6$ with temporal hyperparameters, $\gamma=4$ for all simulations. This model specification enables us to evaluate the variational energy at any point in time and invert the model given an observed response.

Variational filtering and DEM

DEM approximates the density of an ensemble of solutions by assuming it has a Gaussian form. This assumption reduces the problem to finding the path of the mode, which entails integrating an ordinary differential equation that is identical to Eq. (16a) but without the random terms. The conditional covariance is then evaluated usingthe curvature of the variational energy at the mode. Variational filtering relaxes this fixed-form assumption and integrates the corresponding stochastic differential equations to furnish the paths of an ensemble and an approximating sample density. Here the conditional covariance is encoded in the dispersion of particles that is constrained by the curvature of the variational energy. We can compare the fixed-form density provided by DEM with the sample density from variational filtering. Generally, this is non-trivial because nonlinearities in the likelihood model render the true conditional non-Gaussian, even under Gaussian assumptions about the priors and innovations. However, with a linear convolution model in generalised coordinates, the Gaussian form is exact and we would expect a close correspondence between variational filtering and DEM.

Fig. 3 shows the trajectories or paths of sixteen particles tracking the mode of the cause (top) and two hidden states (bottom). The sample mean of this distribution is shown in blue. An alternative representation of the sample density is shown in Fig. 4. This format will be used in subsequent figures and summarizes the predictions and conditional densities on the states. Each row corresponds to a level in the model, with causes on the left and hidden states on the right. The first (upper left) panel shows the predicted response and the error on this response. For the hidden states (upper right) and

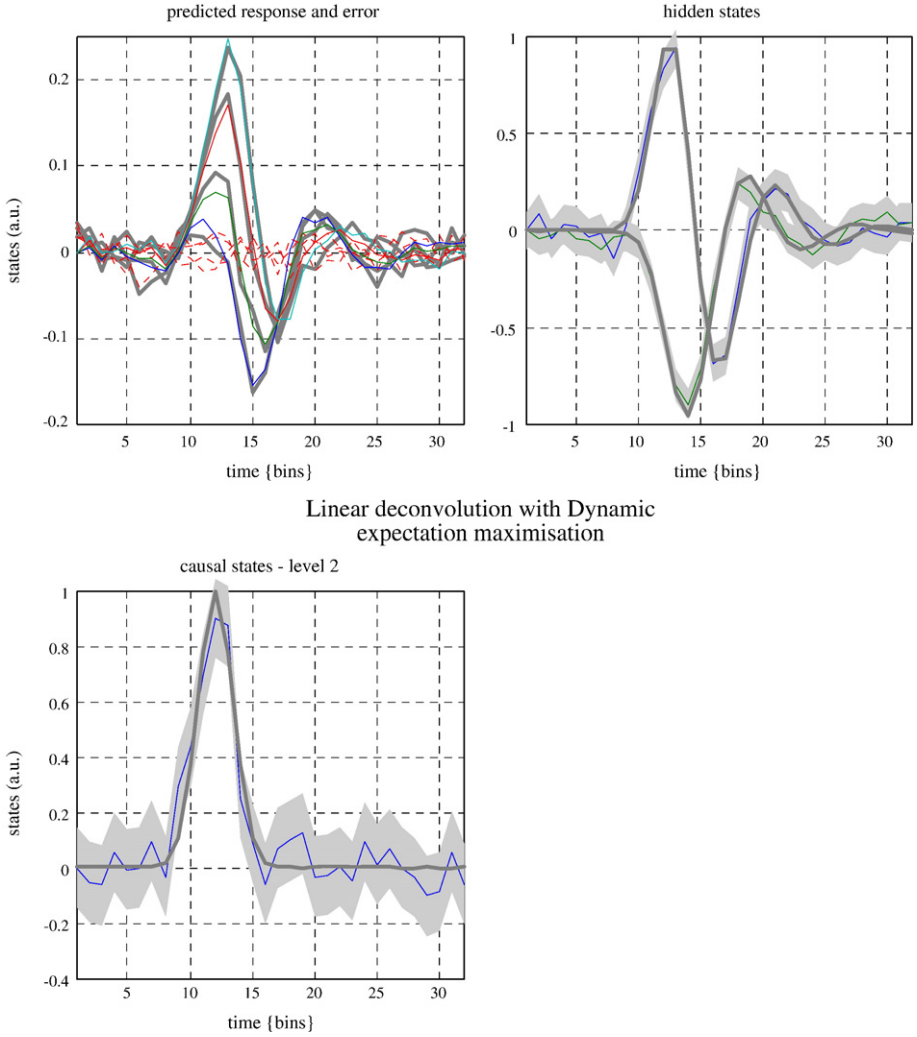

Fig. 5. This is exactly the same as the previous figure, summarising conditional inference on the states of a linear convolution model. The only difference is that here, we have used a Laplace approximation to the variational density and have integrated a single trajectory; that of the conditional mode. Note that the modes (blue lines) are indistinguishable from the variational filter modes (Fig. 6). The conditional variance on the causal and hidden states is very similar but with one key difference; in DEM the confidence tubes have the same width throughout. This is because we are dealing with a linear system. In contrast, the conditional density based on the variational filter shows an initial transient as particles converge to the mode, before attaining equilibrium in a moving frame of reference.

causes (lower left) the conditional mode is depicted by a coloured line and the 90% conditional confidence intervals by the grey area. These are sometimes referred to 'tubes'. Here, the confidence tubes are based upon the sample density of the ensemble shown in Fig. 3. It can be seen that there is a pleasing correspondence between the sample mean (blue) and veridical states (grey). Furthermore, the true values lie within the 90% confidence intervals.

We then repeated the inversion using exactly the same model and response using DEM. The results are shown in Fig. 5 using the same format as the previous figure. Critically, the ensuing modes (blue) are indistinguishable from those obtained with variational filtering (c.f., Fig. 4). The conditional variance on the causal and hidden states is again very similar but with one key difference; in DEM the conditional tubes have the same width throughout. This is because we are dealing with a linear system, where variations in the state have the same effect in measurement space at all points in time. In contrast, the conditional density based on the variational filter shows an initial transient as the particles converge on the mode, before attaining equilibrium in a moving frame of reference. The integration time for DEM is an order of magnitude faster than for the variational filter (about 1 s versus 10) because we only integrate the path of a single particle (the approximating mode) and eschew integration of stochastic differential equations.

In summary, there is an expected convergence between variational filtering and its fixed-form homologue, when the fixed-form assumptions are correct. In these cases, the fixed-form approximation is computationally more efficient. However, fixed-form assumptions are not always appropriate. In the next example, we consider a nonlinear system, whose conditional density is bimodal. In this case DEM fails completely, in relation to filtering.

A nonlinear convolution model

Here, we focus on the effect of nonlinearities with a model that has been used previously to compare extended Kalman and particle filtering (c.f., Arulampalam et al., 2002)

Nonlinear (double-well) convolution model

Level	$g(x,v)$	$f(x,v)$	Π^z	Π^w	η
$m=1$	$\frac{1}{16}x^2$	$\frac{2x}{1+x^2}-\frac{1}{16}x+\frac{1}{4}v$	e^2	e^{16}	
$m=2$			$\frac{1}{8}$		0

This is a particularly difficult system to invert for many schemes because the quadratic form of the observer function renders inference on the hidden states and their causes inherently ambiguous[8]. To make matters more difficult, the hidden states are deployed symmetrically about zero in a double-well potential. Transitions from one well to the other are caused by inputs or high amplitude fluctuations. Fig. 6 shows the phase-diagram of this system by plotting $f(x,0)$ against x (top) and the implicit potential (the negative integral of $f(x,0)$; bottom).

We drove this system with a slow sinusoidal input $v(t) = 8\sin\left(\frac{1}{16}\pi t\right)$ to generate data and then tried to invert the model, using only the response. Again, priors on the input were mildly informative with zero mean; $p(v)=N(0,8)$.

[8] Although one might hope its inversion is made much easier with access to the trajectory of the responses.

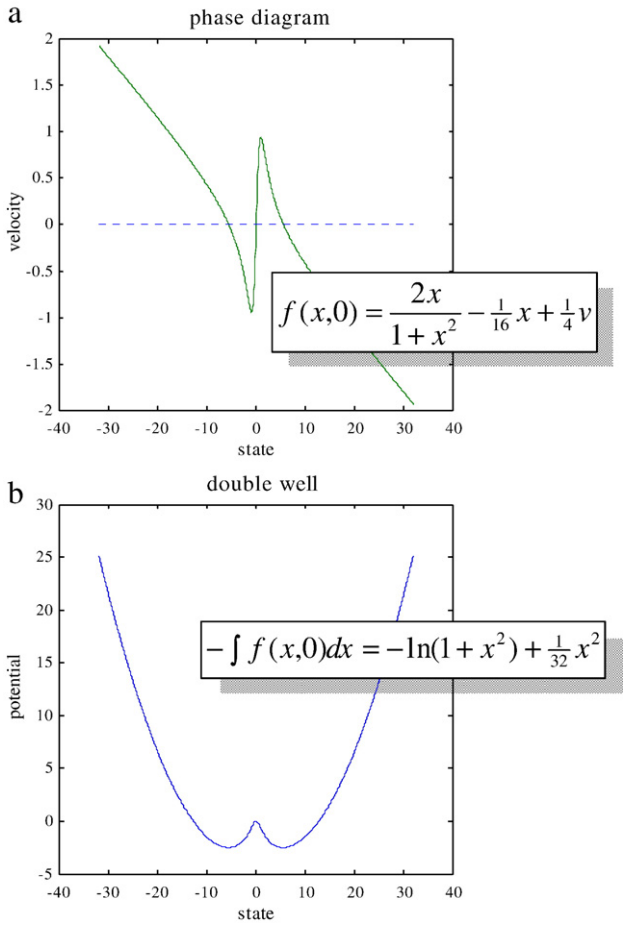

Fig. 6. Schematic detailing the nonlinear convolution model in which hidden states evolve in a double-well potential. (a): Plot of the velocity of states against states (in the absence of input). This shows how states converge on two fixed-point attractors in the absence of input or random fluctuations. These attractors correspond to the minima of the implicit potential field in (b).

Comparative evaluations

We generated a 64 time-bin response and inverted it using DEM. The results are shown in Fig. 7. As in previous figures, the blue lines represent the conditional estimate of hidden and causal states, while the grey lines depict the true values. It can be seen immediately that the inversion has failed to represent the ambiguity about whether hidden states are positive or negative. The fixed-form solution asserts, incorrectly, that the states are always positive with deleterious consequences for the conditional density on the inputs. It is interesting to note that in this nonlinear system, the confidence tubes on the hidden states are time-dependent; the conditional uncertainty increases markedly when the states approach zero (c.f., the fixed-width confidence intervals under linear deconvolution in Fig. 5). This is because changes in the states produce smaller changes in the response, at these low values.

Particle filtering

As demonstrated above, fixed-form schemes such as DEM and extended Kalman filtering fail to represent non-Gaussian (e.g., multi-

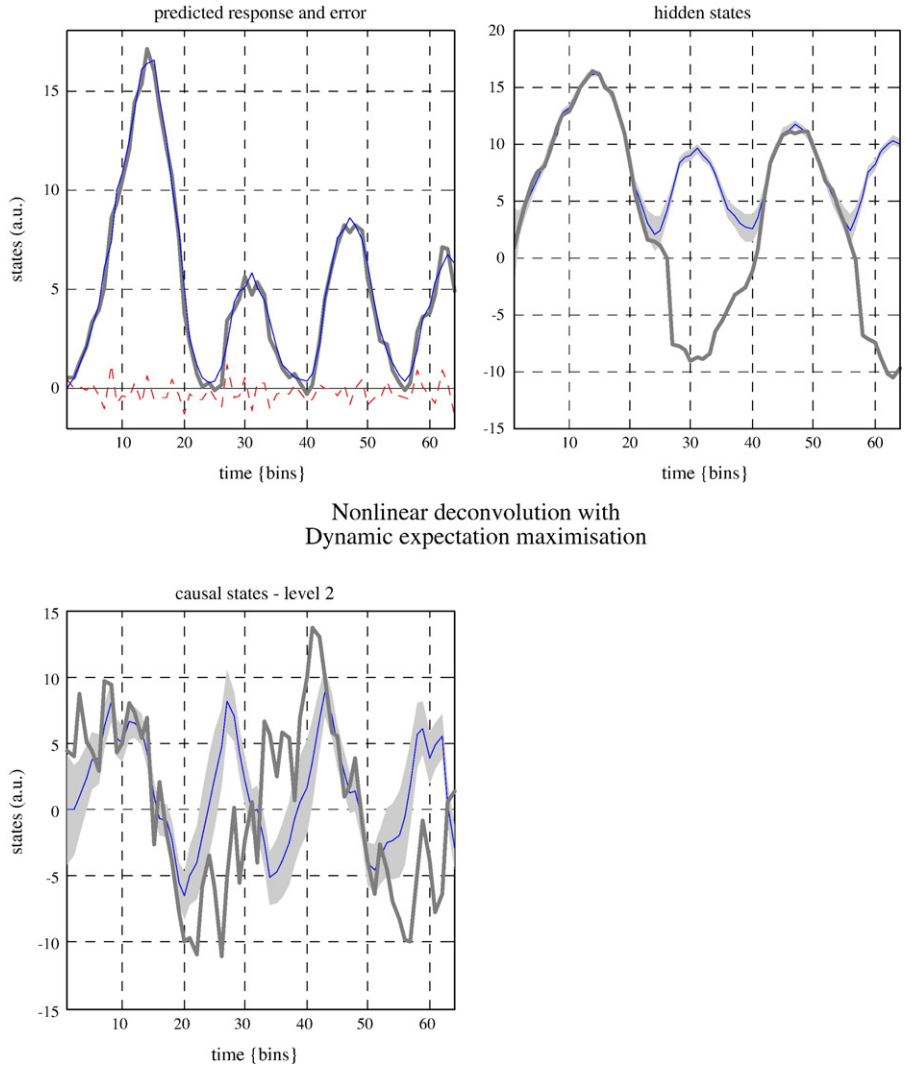

Fig. 7. An example of deconvolution with DEM using the nonlinear double-well convolution model described in the main text. In this case, the response is always positive. As in previous figures, the blue lines represent the conditional estimates of hidden and causal states, while the thick grey lines depict the true values.

modal) conditional densities required for accurate deconvolution. In this instance, particle filtering and related grid-based approximations provide solutions that allow for non-Gaussian posteriors on the hidden states. In these schemes, particles are subject to stochastic perturbations and re-sampling so that they come to approximate the conditional density. This approximation rests on which particles are retained and which are eliminated, where selection depends on the energy of each particle. See Appendix B for a description of particle filters for state-space models formulated in continuous time.

These sequential Monte-Carlo techniques should not be confused with the ensemble dynamics of variational filtering. In variational filtering particles are conserved and experience forces that depend on energy gradients. In sequential sampling methods the energy is used to select and eliminate particles. In relation to variational filtering, sequential sampling techniques appear unnecessarily complicated. Furthermore, they rely on some rather *ad hoc* devices to make them work (see Appendix B and var der Merwe et al., 2000). For these reasons, we will not provide any further background on particle filtering but simply use it as a reference for variational filtering.

Variational filtering

We next inverted the double-well model using variational and particle filtering. Fig. 8 (top) shows the trajectory of 32 particles using variational filtering and the true values of the hidden states. It is seen that the ensemble splits into two, reflecting the ambiguity about their positive or negative sign. The sample density (lower left) shows the resulting bimodal distribution nicely and is very similar to the corresponding density obtained with particle filtering (lower right). The key difference between variational and particle filtering is that variational filtering also furnishes an ensemble density on the inputs, whereas particle filtering does not. Fig. 9 shows $q(v,t)$ in terms of trajectories (top) and the sample density (bottom). It is evident that inference on the input is not as accurate as inference on hidden states, because inputs express themselves in measurement space vicariously though hidden states. However, there are two key things to note; first, the conditional density is not symmetric about zero. This reflects that fact that the hidden states are a nonlinear convolution of the inputs, which breaks the

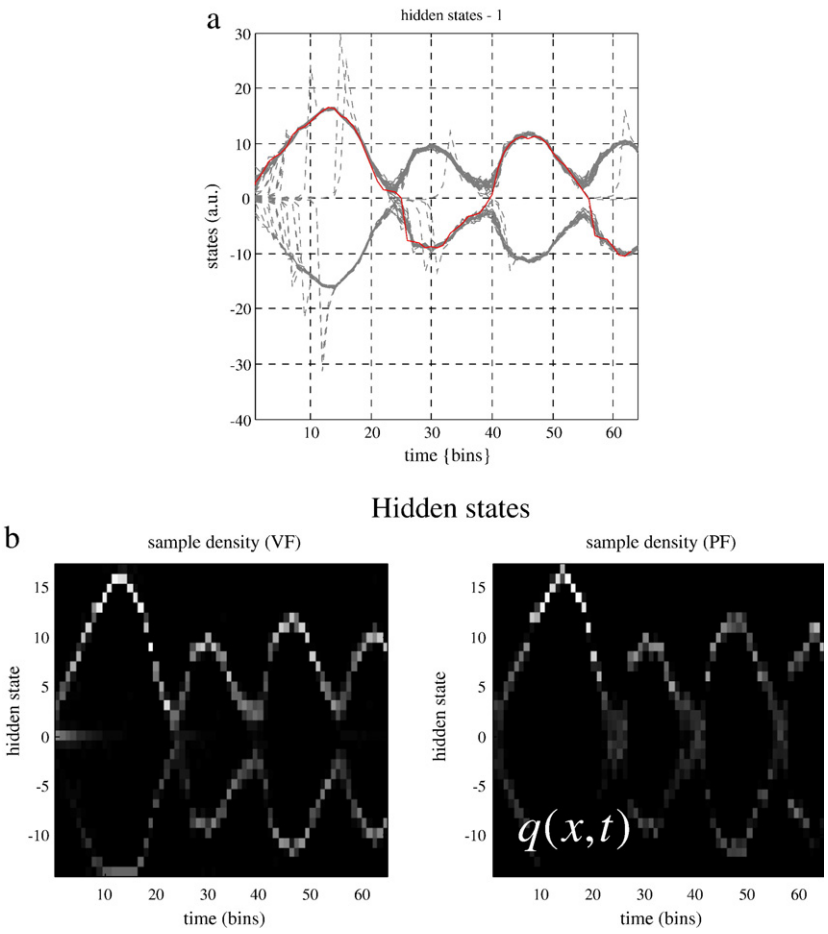

Fig. 8. (a) Trajectories of 32 particles from variational filtering, using the double-well model. The paths are shown for the hidden states with the true trajectory in red. (b): The same results but presented as a sample density in the space of hidden states for variational (left) and particle filtering (right).

symmetry. Second, the most precise conditional densities obtain when the mode and true inputs coincide (circled region).

Summary

These examples have shown that variational filtering provides veridical approximations to the conditional density on the states of dynamic models. When, models have a simple linear state-space form, DEM and variational filtering give the same results. For nonlinear models, in which the Laplace assumption of Gaussian posterior fails, variational filters give the same results as particle filtering. The principal advantage that variational filtering has over conventional schemes is that its conditional densities are on hidden states and their causes; both in generalised coordinates of motion. In the next section, we exploit inference on causes to infer the neuronal activity causing observed hemodynamics responses.

An empirical application

In this, the final section, we illustrate variational filtering by inverting a hemodynamic model of how neuronal activity in the brain generates data sequences in functional magnetic resonance imaging (fMRI). This example has been chosen because inference about brain states from non-invasive neurophysiologic observations is an important issue in cognitive neuroscience and functional imaging (e.g., Friston et al., 2003; Gitelman et al., 2003; Buxton et al., 2004; Riera et al., 2004; Sotero and Trujillo-Barreto, in press).

The hemodynamic model

The hemodynamic model has been described extensively in previous communications (Buxton et al., 1998; Friston, 2002). In brief, neuronal activity causes an increase in a vasodilatory signal h_1 that is subject to auto-regulatory feedback. Blood flow h_2 responds in proportion to this signal and causes changes in blood volume h_3 and deoxyhemoglobin content, h_4. The observed signal is a nonlinear function of volume and deoxyhemoglobin. These dynamics are modelled by the differential equations

$$\begin{aligned}\dot{h}_1 &= v - \kappa(h_1 - 1) - \chi(h_2 - 1)\\ \dot{h}_2 &= h_1 - 1\\ \dot{h}_3 &= \tau(h_2 - F(h_3))\\ \dot{h}_4 &= \tau(h_2 E(h_2) - F(h_3)h_4/h_3)\end{aligned} \tag{38}$$

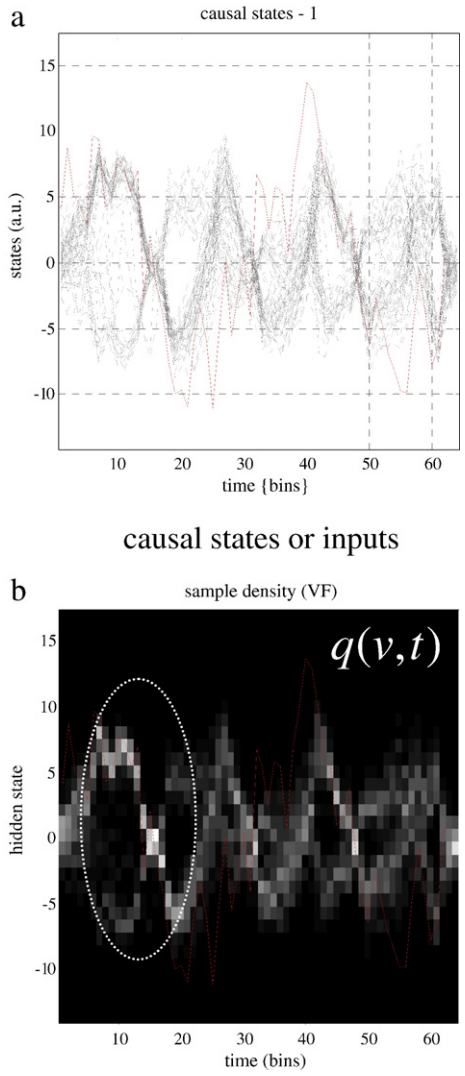

Fig. 9. (a) Trajectories of 32 particles from variational filtering using the double-well model. Here, the paths are shown for the cause or input with the true trajectory in red. (b): The same results presented as a sample density in image format. The circled region shows that the sample density is relatively precise (i.e., a peaked distribution) when and only when, its mode corresponds to the true and relatively unambiguous input.

In this model, changes in vasodilatory signal h_1 are elicited by neuronal input, v. Relative oxygen extraction $E(h_2) = \frac{1}{\varphi}(1 - (1-\varphi)^{1/h_2})$ is a function of flow, where φ is resting oxygen extraction fraction and outflow is a function of volume $F(h_3) = h_3^{1/\alpha}$, through Grubb's exponent α. A description of the parameters of this model and their assumed values are provided in Table 1.

All these hemodynamic states are nonnegative quantities. One can implement this formal constraint with the transformation, $x_i = \ln h_i \Leftrightarrow h_i = exp(x_i)$. Under this transformation the differential equations above can be written as

$$\dot{h}_i = \frac{\partial h_i}{\partial x_i}\frac{\partial x_i}{\partial t} = h_i \dot{x}_i = f_i(h, v) \tag{39}$$

Table 1
Biophysical parameters (state)

	Description	Value
κ	Rate of signal decay	1.2 s^{-1}
χ	Rate of flow-dependent elimination	0.31 s^{-1}
τ	Transit time	2.14 s
α	Grubb's exponent	0.36
φ	Resting oxygen extraction fraction	0.36
Biophysical parameters (observer)		
V_0	Blood volume fraction	0.04
K_1	Intravascular coefficient	7φ
K_2	Concentration coefficient	2
K_3	Extravascular coefficient	2φ–0.2

This allows us to formulate the model in terms of the hidden states $x_i = \ln h_i$ with unbounded support (i.e., the trajectories of particles can be positive or negative).

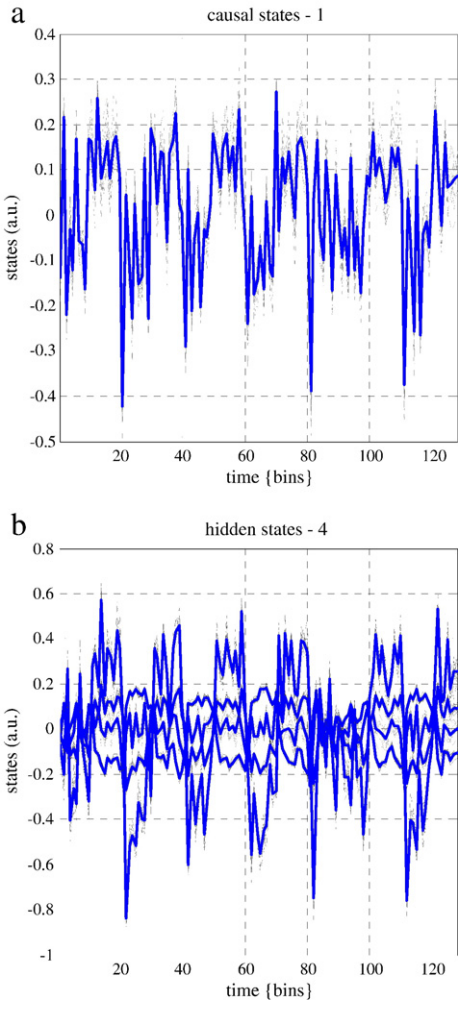

Fig. 10. Trajectories from variational filtering using real fMRI data and the hemodynamic model described in the main text. (a) 16 trajectories in the space of neuronal causes or activity, showing clear onset and offset transients with each new 10-bin experimental condition. (b) The same trajectories but now shown over the four hidden hemodynamic states. Each time bin corresponds to 3.22 s.

Hemodynamic convolution model

Level	$G(x,v)$	$f(x,v)$	Π^z	Π^w	η^v
$m=1$	$V_0(k_1(1-h_2)+$ $k_2(1-h_4/h_3)+$ $k_3(1-h_3))$	$\begin{bmatrix} v-\kappa(h_1-1)-\chi(h_2-1)/h_1 \\ (h_1-1)/h_2 \\ \tau(h_2-F(h_3))/h_3 \\ \tau(h_2E(h_2)-F(h_3)h_4/h_3)/h_4 \end{bmatrix}$	e^2	e^8	
$m=2$			1		0

This model represents a multiple-input, single-output model with four hidden states. In this example, we assume state noise has precision, e^8 which corresponds to random fluctuations with amplitudes of about 20% of the evoked changes in hidden states. The unknown cause has weakly informative shrinkage priors.

Data and pre-processing

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 multi-slice 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). 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°/s 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). A time-series from extrastriate cortex was summarised using the principal local eigenvariate of a region centred on the maximum of a contrast testing for the effect of visual motion. This regional response was used for deconvolution.

Variational filtering

Using the regional response, we attempted to deconvolve both the hidden states and neuronal input from the observed time-series.

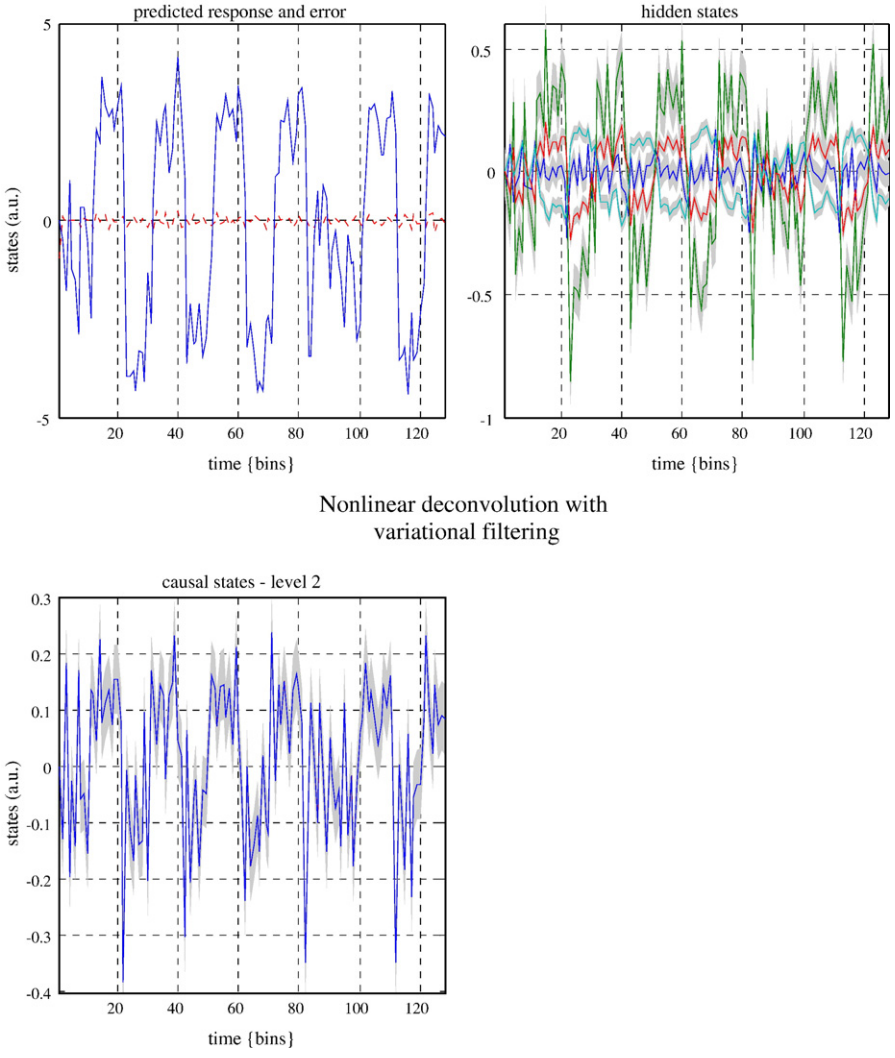

Fig. 11. These are the same results shown in the previous figure but presented in terms of conditional means and 90% confidence tubes (see Fig. 4 for details).

The trajectories of 16 particles over the first 120 scans are shown in Fig. 10 for the neuronal input (top) and hidden hemodynamic states (bottom). It is clear that the conditional density is unimodal. This means is sensible to display the densities in term of 90% confidence tubes as in Fig. 11. This unimodal density reflects the fact that the model is only weakly nonlinear and there are no severe indeterminacies. Indeed, very similar results were obtained under a fixed-form Laplace assumption using DEM (Fig. 12). This suggests that the conditional density is roughly Gaussian.

A summary of the hemodynamics is shown in the Fig. 13. This figure plots the hemodynamic states in terms of the conditional expectation of $h_i = exp(x_i)$; instead of x_i in Figs. 11 and 12). Each time bin corresponds to 3.22 s. In the upper panel, the hidden states are overlaid on periods (grey) of visual motion. These hidden states correspond to flow-inducing signal, flow, volume and deoxyhemoglobin (dHb). It can be seen that neuronal activity, shown in the lower panel, induces a transient burst of signal (blue), which is suppressed rapidly by the resulting increase in flow (green). The increase in flow dilates the venous capillary bed to increase volume (red) and dilute deoxyhemoglobin (cyan). The concentration of deoxyhemoglobin (involving volume and dHb) determines the measured response. Interestingly, the underlying neuronal activity appears to show an offset transient that is more pronounced than the onset transient. In either case, we can be almost certain that changes in visual stimulation are associated with changes in neuronal activity. The dynamics of inferred activity, flow and other biophysical states are physiologically plausible. For example, activity-dependent changes in flow are around 14%, producing about a 5% change in fMRI signal.

Summary

As noted in Friston et al 2008, "it is perhaps remarkable that so much conditional information about the underlying neuronal and hemodynamics can be extracted from a single scalar time-series, given only the functional form of its generation". This speaks to the power of generative modelling, in which constraints on the form of the model allow one to focus data on inferring 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

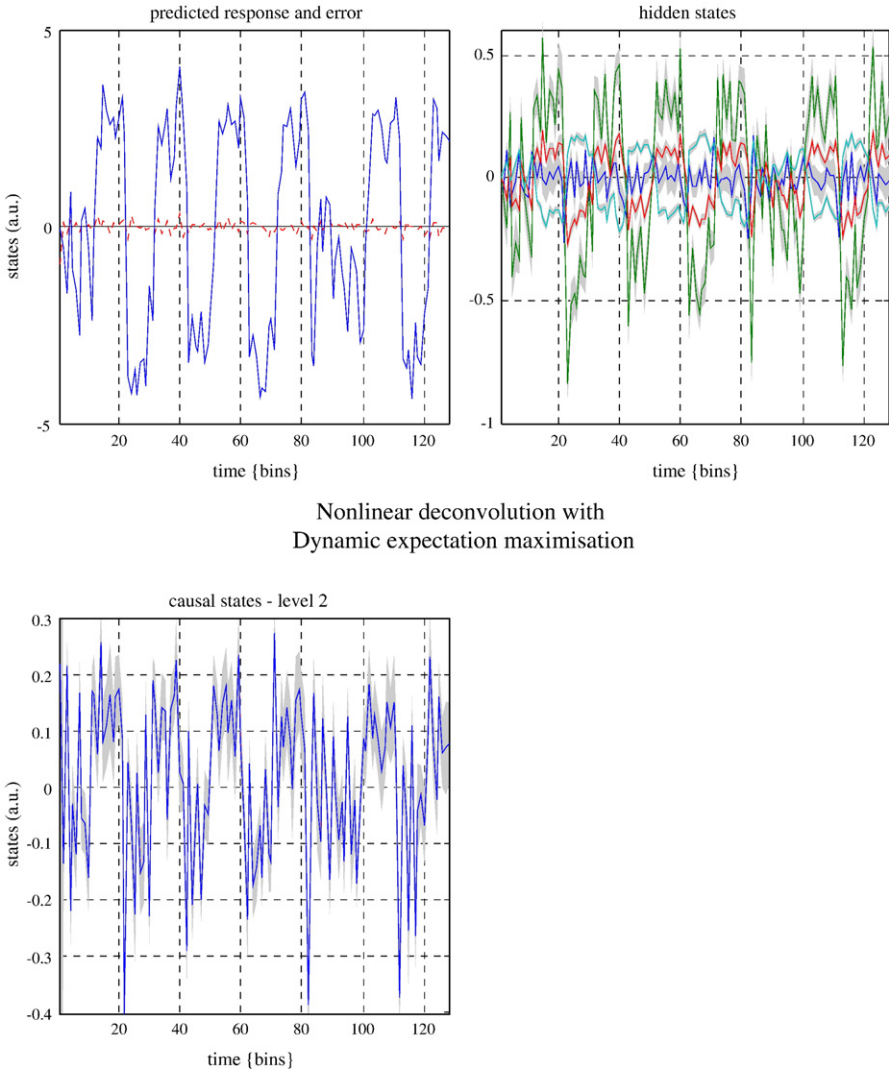

Fig. 12. The equivalent results for the hemodynamic deconvolution using DEM. These densities should be compared with those in Fig. 11 that were obtained using variational filtering.

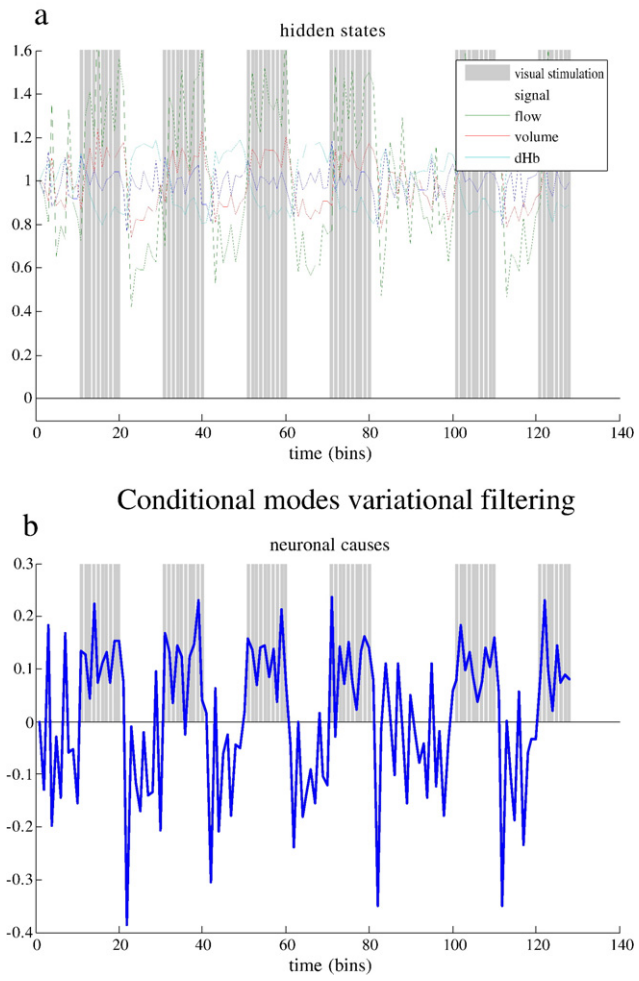

Fig. 13. These are the same results shown in Fig. 11 but focussing on the conditional expectations of the hidden states and neuronal causes. In the upper panel (a), the hidden states are overlaid on periods (grey bars) of visual motion. These hidden states correspond to flow-inducing signal, flow, volume and deoxyhemoglobin (dHb). It can be seen that neuronal activity, shown in the lower panel (b), 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 deoxyhemoglobin (cyan). The concentration of deoxyhemoglobin (involving volume and dHb) determines the measured response.

Riera et al., 2004 and Sotero and Trujillo-Barreto, in press for important exceptions). One of the motivations for the variational treatment presented in this paper was to develop an inference scheme that can deal with state-noise. Variational filtering may find a useful role in ensuring that fixed-form Laplace-based schemes are justified when using these nonlinear models.

Conclusion

We have presented a variational treatment of dynamic models that furnishes the time-dependent free-form conditional densities on a system's states by maximising their variational action. This action represents a lower-bound on the model's marginal likelihood or log-evidence, integrated over time. The approach rests on formulating the variational or ensemble density in generalised coordinates of motion. The resulting scheme can be used for online Bayesian inversion of stochastic dynamic causal models and eschews some limitations of alternative approaches, such as particle filtering. Critically, variational filtering provides conditional densities on both the hidden states and unobserved inputs to a system.

Variational vs. incremental approaches

The variational approach to dynamic systems presented here differs in several ways from incremental approaches such as extended Kalman and particle filtering. The first distinction relates to the nature of the generative models. The variational approach regards the generative model as mapping between the instantaneous trajectories of causes and responses. In contradistinction, incremental approaches consider the mapping to be between instantaneous quantities *per se*. In this sense, the variational treatment above can be regarded as a generalisation of model inversion to cover mappings between paths. Incremental approaches simply treat the response as an ordered sequence and infer the current state using previous estimates. However, the underlying causes and responses are analytic functions of time, which provide constraints on inversion that cannot be exploited by incremental schemes. For example, most incremental approaches assume uncorrelated random components (e.g., a Weiner process for system noise). However, in reality these random fluctuations are almost universally the product of ensemble dynamics that are smooth functions of time. The variational approach accommodates this easily with generalised coordinates of high-order notion and a parametric form for the associated temporal correlations.

The second key difference between conventional and variational filtering is the support of the ensemble density itself. In conventional procedures this covers only the hidden states, whereas the full variational density should cover both the hidden and causal states. This has a number of important consequences. Perhaps the simplest is that particle filtering cannot be used to deconvolve the inputs to a system (i.e., causes) from its responses.

Variational filtering relies on an ensemble of particles being drawn towards peaks on the variational energy landscape; so that their sample density approximates the conditional density we require. The coupling of high-order motion to lower orders (through mean-field effects) ensures this distribution is relatively stationary (in a moving frame of reference). This rests on the assumption that the variational energy manifold is changing slowly, in relation to the implicit diffusion of particles. Clearly, if a system changes quickly (i.e., shows bifurcations or chaotic itinerancy), it may take some time for equilibrium to be attained on a new variational energy manifold. This speaks to optimising the rate of ascent of the energy gradients. In the examples above, this was assumed to be one (i.e., there is no explicit rate constant in Eq. (6) or Eq. (12)). It may well be the case that higher values are required for dynamical systems showing exotic behaviours. This will be a focus of future work.

We envisage that variational filtering will find its principal role in validating fixed-from approximations to the conditional density using computationally more efficient approaches like DEM. Indeed the last section of this note could be used to motivate the Laplace assumption in the context of hemodynamic models.

Software note

The variational scheme above is implemented in Matlab code and is available freely from http://www.fil.ion.ucl.ac.uk/spm. A DEM

toolbox provides several demonstrations from a graphical user interface. These demonstrations reproduce the figures of this paper (see spm_DFP.m and ancillary routines).

Acknowledgments

The Wellcome Trust funded this work. We would like to acknowledge the very helpful discussions with members of the Theory and Methods Group at the Wellcome Trust Centre for Neuroimaging and John Shawe-Taylor, Centre for Computational Statistics and Machine Learning, UCL.

Appendix A

Covariance of stochastic terms:

An efficient way to compute

$$\Sigma^{\varsigma} = \int_0^{\Delta t} \exp(tV_{uu})\Omega\exp(tV_{uu})^T dt \tag{A1.1}$$

Is to pre-compute $A=exp(\tau V_{uu})$ where $N\tau=\Delta t$ and accumulate terms as in the Pade approximation

for i=1:N

$$\begin{aligned}\Sigma^{\varsigma} &:= \Sigma^{\varsigma} + \tau B\Omega B^T \\ B &:= BA\end{aligned} \tag{A1.2}$$

end

Stating with $\Sigma^{\varsigma}=0$ and $B=A$ and terminating if $|B\Omega B^T|$ falls below some tolerance.

Appendix B

Particle filtering:

This appendix provides a pseudo-code specification of particle filtering based on var der Merwe et al. (2000) and formulated for models of the form:

$$\begin{aligned}y &= g(x) + z \\ \dot{x} &= f(x,v) + w\end{aligned} \tag{A2.1}$$

This can be re-written, using local linearisation, as a discrete-time state-space model. This is the formulation treated in conventional Bayesian filtering procedures

$$\begin{aligned}y_t &= \mathrm{g}_x x_t + z_t \\ x_t &= \mathrm{f}_x x_{t-1} + w_{t-1} \\ \mathrm{g}_x &= g(x_t)_x \\ \mathrm{f}_x &= \exp\left(\Delta t f(x_t)_x\right) \\ z_t &= z(t) \\ w_{t-1} &= \int_0^{\Delta t} \exp(f_x\tau)(f_v v(t-\tau) + w(t-\tau))d\tau\end{aligned} \tag{A.2.2}$$

The key thing to note here is that process noise w_{t-1} is simply a convolution of the exogenous input, $v(t)$ and innovations, $w(t)$. This is relevant for Kalman filtering and related nonlinear Bayesian tracking schemes that assume w_{t-1} is a well-behaved noise sequence. We have used the term process noise to distinguish it from system noise, $w(t)$ in hierarchical dynamic models. This distinction does not arise in simple state-space models. The covariance of process noise is

$$\begin{aligned}\langle w_t w_t^T\rangle &= \int_0^{\Delta t} \exp(f_x\tau)\Omega\exp(f_x\tau)^T d\tau \approx \Omega\Delta t \\ \Omega &= f_v\Sigma^v f_v^T + \Sigma^w\end{aligned} \tag{A2.3}$$

assuming temporal correlations can be discounted and that the Lyapunov exponents of f_x are small relative to the time-step.

In this pseudo-code description, each particle is denoted by its state $x_t^{[i]}$. These states are updated stochastically from a proposal density, using a random variate $w^{[i]}$ and are assigned importance weights $q^{[i]}$ based on their likelihood. These weights are then used to re-sample the particles to ensure an efficient representation of the ensemble density.

for all t

Prediction step: for all i

$$\begin{aligned}x_t^{[i]} &= \mathrm{f}_x x_{t-1}^{[i]} + w^{[i]} \\ \xi &= y - g\left(x_t^{[i]}\right) \\ q^{[i]} &= \exp\left(-\frac{1}{2}\xi^T \Pi^z \xi\right)\end{aligned}$$

Normalise importance weights

$$q^{[i]} = \frac{q^{[i]}}{\sum_i q^{[i]}}$$

Selection step: for all i

$$x_t^{[i]} \leftarrow x_t^{[r]} \tag{A2.4}$$

end.

where $w^{[i]}$ is drawn from a proposal density $N(0,\Omega)$ and $x_t^{[i]}\leftarrow \mathrm{x}_t^{[r]}$ denotes sequential importance re-sampling. The indices r are selected on the basis of the importance weights. Π^z is the precision of observation noise. In our implementation (*spm_pf.m*) we use multinomial re-sampling based on a high-speed Niclas Bergman Procedure written by Arnaud Doucet and Nando de Freitas.

References

Archambeau, C., Cornford, D., Opper, M., Shawe-Taylor, J., 2007. Gaussian process approximations of stochastic differential equations. JMLR: Workshop and Conference Proceedings 1, pp. 1–16.

Arulampalam, S., Maskell, S., Gordon, N.J., Clapp, T., 2002. A tutorial on particle filters for on-line nonlinear/non-Gaussian Bayesian tracking. IEEE Trans. Signal Process. 50 (2), 174–188.

Attias, H., 2000. In: Leen, T., et al. (Ed.), A variational Bayesian framework for graphical models. Adv. Neur. Info. Proc. Sys., 12. MIT Press, Cambridge, MA.

Beal, M.J., Ghahramani, Z., 2003. The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In: Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D. (Eds.), Bayesian Statistics. AFM Smith and M West. Chapter 7. OUP, UK.

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

Buxton, R.B., Uludag, K., Dubowitz, D.J., Liu, T.T., 2004. Modeling the hemodynamic response to brain activation. NeuroImage 23 (Suppl 1), S220–S233.

Corduneanu, A., Bishop, C.M., 2001. Variational Bayesian model selection for mixture distributions. In: Jaakkola, T., Richardson, T. (Eds.), Artificial Intelligence and Statistics. Morgan Kaufmann, Los Altos, CA, pp. 27–34.

Cox, D.R., Miller, H.D., 1965. The theory of stochastic processes. Methuen, London.

Choudrey, R.A., Roberts, S.J., 2001. Variational mixture of Bayesian independent component analysers, Technical Report PARG-01–04, Department of Engineering Science. University of Oxford.

Efron, B., Morris, C., 1973. Stein's estimation rule and its competitors — an empirical Bayes approach. J. Am. Stats. Assoc. 68, 117–130.

Eyink, G.L., 1996. Action principle in nonequilibrium statistical dynamics. Phys. Rev. E 54, 3419–3435.

Eyink, G.L., 2001. A Variational Formulation of Optimal Nonlinear Estimation. Technical report: Report number: LA-UR00–5264 arXiv: physics/0011049v2 [physics.data-an].

Eyink, G.L., Restrepo, J., Alexander, F.J., 2004. A mean-field approximation in data, assimilation for nonlinear dynamics. Physica D 195, 347–368.

Feynman, R.P., 1972. Statistical mechanics. Benjamin, Reading, MA, USA.

Friston, K.J., 2002. Bayesian estimation of dynamical systems: an application to fMRI. NeuroImage 16, 513–530.

Friston, K.J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J., 2002. Classical and Bayesian inference in neuroimaging: theory. NeuroImage 16 (2), 465–483.

Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. NeuroImage 19, 1273–1302.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W., 2007. Variational free energy and the Laplace approximation. NeuroImage 34 (1), 220–234.

Friston, K., Trujillo-Barreto, N., Daunizeau, J. (2008). DEM: A variational treatment of dynamic systems. NeuroImage, under - review.

Gitelman, D.R., Penny, W.D., Ashburner, J., Friston, K.J., 2003. Modeling regional and psychophysiologic interactions in fMRI: the importance of hemodynamic deconvolution. NeuroImage 19 (1), 200–207.

Graham, R., 1978. Path integral methods in nonequilibrium thermodynamics and statistics. In: Garrido, L., Seglar, P., Shepherd, P.J. (Eds.), Stochastic Processes in Nonequilibrium Systems. Lecture Notes in Physics, vol. 84. Springer-Verlag, Berlin.

Hinton, G.E., von Cramp, D., 1993. Keeping neural networks simple by minimising the description length of weights. Proceedings of COLT-93, pp. 5–13.

Kass, R.E., Steffey, D., 1989. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). J. Am. Stat. Assoc. 407, 717–726.

Kerr, W.C., Graham, A.J., 2000. Generalised phase space version of Langevin equations and associated Fokker–Planck equations. Eur. Phys. J. B. 15, 305–311.

MacKay, D.J.C., 1995. Free-energy minimisation algorithm for decoding and cryptoanalysis. Electron. Lett. 31, 445–447.

Onsager, L., Machlup, S., 1953. Fluctuations and irreversible processes. Phys. Rev. 91, 1505–1512.

Ozaki, T., 1992. A bridge between nonlinear time-series models and nonlinear stochastic dynamical systems: a local linearization approach. Statistica Sin. 2, 113–135.

Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004. Comparing dynamic causal models. NeuroImage 22, 1157–1172.

Riera, J.J., Watanabe, J., Kazuki, I., Naoki, M., Aubert, E., Ozaki, T., Kawashima, R., 2004. A state-space model of the hemodynamic approach: nonlinear filtering of BOLD signals. NeuroImage 21 (2), 547–567.

Särkkä, Simo, 2007. On unscented Kalman filtering for state estimation of continuous-time nonlinear systems. IEEE Trans. Automat. Contr. 52 (9), 1631–1641.

Sotero, R.C., Trujillo-Barreto, N.J., 2007. Biophysical model for integrating neuronal activity, EEG, fMRI and metabolism, NeuroImage, in press. DOI: 10.1016/j.neuroimage.2007.08.001.

Trujillo-Barreto, N., Aubert-Vazquez, E., Valdes-Sosa, P., 2004. Bayesian model averaging. NeuroImage 21, 1300–1319.

var der Merwe, R., Doucet, A., de Freitas, N., Wan, E., 2000. The unscented particle filter. Technical Report CUED/F-INFENG/TR 380.

Weissbach, P., Pelster, A., Hamprecht, 2002. High-order variational perturbation theory for the free energy. Phys. Rev. Lett. 66, 036129.

Whittle, P., 1991. Likelihood and cost as path integrals. J. R. Stat. Soc. Series B (Methodological) 53 (3), 505–538.