ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 41 (2008) 849–885

DEM: A variational treatment of dynamic systems

K.J. Friston,[a,*] N. Trujillo-Barreto,[b] and J. Daunizeau[a]

[a]*The Wellcome Department of Imaging Neuroscience, University College London, United Kingdom*
[b]*Cuban Neuroscience Centre, Havana, Cuba*

Received 5 October 2007; revised 14 January 2008; accepted 25 February 2008
Available online 10 March 2008

This paper presents a variational treatment of dynamic models that furnishes time-dependent conditional densities on the path or trajectory of a system's states and the time-independent densities of its parameters. These are obtained by maximising a variational action with respect to conditional densities, under a fixed-form assumption about their form. The action or path-integral of free-energy represents a lower bound on the model's log-evidence or marginal likelihood required for model selection and averaging. This approach rests on formulating the optimisation dynamically, in generalised coordinates of motion. The resulting scheme can be used for online Bayesian inversion of nonlinear dynamic causal models and is shown to outperform existing approaches, such as Kalman and particle filtering. Furthermore, it provides for dual and triple inferences on a system's states, parameters and hyperparameters using exactly the same principles. We refer to this approach as dynamic expectation maximisation (DEM).

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

Introduction

This paper presents a variational treatment of dynamic causal models formulated as differential equations. We have referred to this scheme briefly, in the context of how the brain might make inferences about sensory data (Friston, 2005; p825). It arose while pursuing an agenda established by von Helmholtz, who sought a basis for neurological energy in his work on conservation laws in physics (Helmholtz, 1860). The treatment presented here focuses on statistical fundaments and applications. In brief, the scheme generalises established approaches to static models using the Laplace approximation (*e.g.*, Friston et al., 2007). The key aspect of this generalisation is the solution of time-dependent conditional trajectories or paths. The equations of motion of these trajectories ensure that their free-energy path-integral (*i.e.*, action) is stationary. This means that the ensuing trajectories encode the time-varying conditional density of the system's state. Using a generative model, in generalised coordinates of high-order motion, finesses temporal dependencies among the states, lead to a relatively fast analytic scheme.

This scheme supports inference on the hidden states of dynamical systems, their parameters and hyperparameters. It goes beyond conventional Bayesian filtering and dual-estimation schemes to provide conditional densities over states, parameters prescribing the nonlinear mixing of states and hyperparameters governing random fluctuations. Furthermore, it operates online and may represent the kind of inferential processes operating in systems like the brain. The applications of this scheme are diverse and will be pursued in a series of subsequent papers. Furthermore, unlike standard variational schemes, the optimisation does not need closed-form updates (*i.e.*, conjugate priors) and can be applied to any model. In this paper, we concentrate on technical aspects and theory. This treatment is rather long and dense; however, it introduces a single scheme that can be implemented with one routine,[1] which grandfathers most approaches to inverting parametric models with continuous variables.

The derivations in this paper involve a fair amount of differentiation. 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; $\tilde{x} = (x, x', x'', \ldots)$. These comprise high-order time derivates and can be regarded as the instantaneous trajectory of a variable. Note that in generalised coordinates, $\dot{x} \equiv x'$ is not necessarily true.

This paper comprises six sections. The first reviews variational approaches to ensemble learning under the Laplace approximation, starting with static models and generalising to dynamic systems.

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

1053-8119/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2008.02.054

[1] **spm_DEM.m** available from http://www.fil.ion.ucl.ac.uk/spm; see software note.

The second describes the variational steps for model inversion. In the third section, we look at a generic hierarchical dynamic model and the update equations it entails. In the fourth section, we demonstrate Bayesian inversion of some nonlinear, dynamic systems to compare their performance with standard Bayesian filtering and, in the fifth section, system identification techniques. In the final section, we provide an illustrative application, in an empirical setting, by deconvolving neuronal activity from observed hemodynamic responses in the brain.

Variational Bayes and ensemble learning

Variational Bayes 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. In addition, it 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 the parameters of any given model and on the model itself. Variational methods for approximating densities in statistical physics were introduced by Feynman (1972) within the path-integral formulation and appeared in the statistical literature in the form of ensemble learning (Hinton and von Cramp, 1993; MacKay, 1995; Neal and Hinton, 1998, Friston et al., 2007). In ensemble learning, an ensemble or variational density $q(\vartheta)$ optimises a free-energy bound on the log-evidence to provide an approximate posterior or conditional density.

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 reprise 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 (*i.e.*, inference on functions of time). Archambeau et al. (2007) motivate the importance of this inference for models based on stochastic differential equations and presents 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.

Variational Bayes for static models

The log-evidence for any parametric model can be expressed in terms of the free-energy and a 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 an energy function of the data, $G(y)$, corresponding to the Gibb's or internal energy, $U(y,\vartheta)=\ln p(y,\vartheta)$ expected under the ensemble density and the entropy, $H(\vartheta)_q$, which is a measure of uncertainty on that density. In this paper, energies are the negative of the corresponding quantities in physics; this ensures that the free-energy increases with log-evidence. Eq. (1) indicates that $F(y,q)$ is a lower bound on the log-evidence because the cross-entropy or divergence term, $D(q(\vartheta)||p(\vartheta|y,m))$ is always positive.

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 an arbitrary density, $q(\vartheta)$ 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, one usually assumes $q(\vartheta)$ factorises over a partition[2] of the parameters

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

This factorization usually appeals to separation of temporal scales, or the distinction between parameters underlying deterministic and stochastic effects. In statistical physics this is called a mean-field approximation. Under this approximation, it is relatively simple 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). *The free-energy is maximised with respect to $q(\vartheta^i)$ when*

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

where Z^i is a normalisation constant (i.e., partition function). We will call $V(\vartheta^i)$ the variational energy, noting that its expectation under $q(\vartheta^i)$ is the expected internal 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 or variational marginal maximises its variational energy.

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

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

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

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

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

If the analytic form of Eq. (3) was tractable (through the use of conjugate priors), $q(\vartheta^i)$ could be optimised directly, by iterative solution of the self-consistent nonlinear equations Eq. (3) represents. This is known as variational Bayes; see Beal and Gha-

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

hramani (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 stationary solution in the limit, t→∞. This rests on 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)[3] 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

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

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.

Variational Bayes 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, such that we now have states and parameters; $\vartheta \to u,\vartheta$. Later, we will consider two parameter sets, $\vartheta = \theta,\lambda$; corresponding to parameters and hyperparameters, which specify the deterministic and random fluctuations of the generative process respectively.

In a dynamic setting, the ensemble or variational density $q = q(u,t)q(\vartheta)$ and associated energies become functionals of time. By analogy with Lagrangian mechanics, this induces the notion of *action*; the time-integral (or, more exactly, anti-derivative) of energy. We will denote action with a bar over the corresponding energy; *i.e.*, $\bar{F}$ $\bar{U}$ and $\bar{V}(\vartheta^i)$ for the free, internal and variational action respectively. The free-action can be expressed as

$$\begin{aligned} \bar{F} &= \int dt\langle U(u,t|\vartheta)\rangle_{q(u,t)q(\vartheta)} - \int dt\langle \ln q(u,t)\rangle_{q(u,t)} + \bar{F}(0) \\ \bar{F}(0) &= \langle U(\vartheta)\rangle_{q(\vartheta)} - \langle \ln q(\vartheta)\rangle_{q(\vartheta)} \end{aligned} \tag{9}$$

where $\partial_t\bar{F} = F$. Here, $U(u,t|\vartheta) = \ln p(y(t),u(t)|\vartheta)$ is the instantaneous energy conditioned on the parameters and $U(\vartheta) = \ln p(\vartheta)$ is the prior energy of the parameters. The constant of integration, $\bar{F}(0)$ corresponds to the free-energy before seeing any data. This is also the (negative) divergence between the conditional and prior densities on the parameters and, in the absence of data, is maximised when $q(\vartheta) = p(\vartheta)$.

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).

Here, 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 $q(u,t)$ and $q(\vartheta^i)$ which maximise action[4]. By the fundamental Lemma, action is maximised with respect to the variational marginals when, and only when

$$\begin{aligned} &\delta_{q(u,t)}\bar{F} = 0 \Leftrightarrow \partial_{q(u,t)}f^u = 0 \\ &\int du f^u = \partial_t\bar{F} = F \\ &\delta_{q(\vartheta^i)}\bar{F} = 0 \Leftrightarrow \partial_{q(\vartheta^i)}f^i = 0 \\ &\int d\vartheta^i f^i = \bar{F} \end{aligned} \tag{10}$$

The solution for the states is the same as in the static case; implying that the ensemble density of the states remains a functional of their variational energy $V(u,t)$

$$\begin{aligned} q(u,t) &= \frac{1}{Z^i}\exp\left(V(u,t)\right) \\ V(u,t) &= \langle U(u,t|\vartheta)\rangle_{q(\vartheta)} \end{aligned} \tag{11}$$

where $\partial_t\bar{V}(u) = V(u,t)$. However, following the derivations in Lemma 1, variational action replaces variational energy for the parameters

$$\begin{aligned} q(\vartheta^i) &= \frac{1}{Z^i}\exp\left(\bar{V}(\vartheta^i)\right) \\ \bar{V}(\vartheta^i) &= U(\vartheta) + \int dt\langle U(u,t|\vartheta)\rangle_{q(u,t)q(\vartheta^{\setminus i})} \end{aligned} \tag{12}$$

These equations are intuitively sensible, because the conditional density of the states should reflect the instantaneous energy, whereas the conditional density of the parameters can only be determined after all the data have been observed. In other words, the conditional or variational energy involves the prior energy and an integral of time-dependent energy.

Consider the density of an ensemble that flows on the variational energy manifold. Because this manifold evolves with time, the ensemble will deploy itself in a time-varying way that optimises free-energy and its action. Unlike the static case, it will not

[3] *i.e.*, a random fluctuation, whose variance scales linearly with time; in statistical thermodynamics and simulated annealing, this corresponds to a temperature of one.

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

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 topology of the manifold (assuming its form does not change quickly). This stationarity arises by 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 context, the ensemble density $q(u,t)$ now evolves in a changing variational energy field, $V(u,t)$, which is generally a function of the states and their motion[5]; 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. In essence, 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 high to low-order motion through mean-field effects.

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

$$\begin{aligned}\dot{v} &= V(u,t)_v+\mu' + \Gamma(t)\\ \dot{v}' &= V(u,t)_{v'}+\Gamma(t)\end{aligned}$$
$$\dot{q}(u,t) = \nabla_v\cdot q(u)\mu' + \nabla_u\cdot\left[\nabla_u q(u) - q(u)\nabla_u V(u,t)\right] \tag{13}$$

where μ' *is the mean velocity over the ensemble (i.e., a mean-field effect).*

Proof. Substituting $q(u,t)=\frac{1}{Z^u}\exp\left(V(u,t)\right)$ and its derivatives into Eq. (13) gives

$$\dot{q}(u,t) = \nabla_v\cdot q(u)\mu' \tag{14}$$

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

$$\begin{aligned}\upsilon &= v-\mu' t\\ q(\upsilon,v',t) &= q(v-\mu' t,v',t)\\ \dot{q}(\upsilon,v',t) &= \dot{q}(v,v',t)-\nabla_v\cdot q(u)\mu' = 0\end{aligned} \tag{15}$$

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

The mean velocity μ' is simply the average flow of particles. It is a mean-field quantity in the sense that each particle's position is affected by the mean velocity of all particles. Heuristically, in a frame of reference that moves with this velocity, the only forces acting on particles are the deterministic effects exerted by the gradients of the variational energy, which drive particles towards its peak and random forces, which disperse particles. Critically, the gradients and peak are stationary in the moving frame of reference, enabling particles to converge on a 'moving target'. This is because the mean velocity μ' is also the velocity of the mode (see below).

For a related example in statistical physics, see Kerr and Graham (2000) who use ensemble dynamics in generalised coordinates to provide a generalised phase-space version of Langevin and associated Fokker–Planck equations: Langevin equations for mechanical systems with canonical position and momentum usually limit noise to the equations for the momentum (motion). Kerr and Graham derive Fokker–Planck equations for mechanical systems that include noise in the equations of motion for all the canonical variables. This affords a more general model of systems in contact with a heat bath that, for example, can model the rate of approach to thermal equilibrium. See also Weissbach et al. (2002) for an example of variational perturbation theory for the free-energy.

The path of the conditional mode

In static systems, the mode of the conditional density maximises variational energy (Lemma 1). Similarly, in dynamic systems, the trajectory of the conditional mode, $\mu^u(t)=\tilde{\mu}=\mu,\mu'$ maximises variational action. This can be seen easily by noting the gradient of the variational energy at the mode is zero

$$\begin{aligned}\partial_u V(\mu^u,t)=0 &\Leftrightarrow \delta_u\overline{V}(\mu^u)=0\\ \partial_t\overline{V}(u) &= V(u,t)\end{aligned} \tag{16}$$

This is sufficient for the mode to maximise variational action (by the Fundamental Lemma of variational calculus). This analysis says that changes in variational action, $\overline{V}(u)$, with respect to variations of the path of the mode are zero (*c.f.*, Hamilton's principle of stationary action). Intuitively, it means that 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. In brief, coupling the motion of states and their velocity 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.

The path of stationary action

Above, we assumed that the variational energy was a function of only 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 high-order motion to give ensemble dynamics in generalised coordinates[6] $u=\tilde{v}=v,v',v'',\ldots$

$$\begin{aligned}\dot{v} &= V(u,t)_v+\mu' + \Gamma(t) & \dot{\mu} &= \mu'\\ \dot{v}' &= V(u,t)_{v'}+\mu'' + \Gamma(t) \quad\Rightarrow & \dot{\mu}' &= \mu'' = \ddot{\mu}\\ \dot{v}'' &= \ldots & \mu'' &= \ldots\end{aligned} \tag{17}$$

where the mode $\tilde{\mu}=\mu,\mu',\mu'',\ldots$satisfies $V(\tilde{\mu},t)_u=0$. Eq. (17) could form 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. (17) and using their sample distribution to approximate $q(u,t)$. We will refer to this as *variational filtering* and consider an example in the next section. In this paper, we focus on *fixed-form*

[5] 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.

[6] Introducing $u=\tilde{v}$ may seem redundant but later $u=\tilde{x},\tilde{v}$ will cover states with distinct roles in generating data.

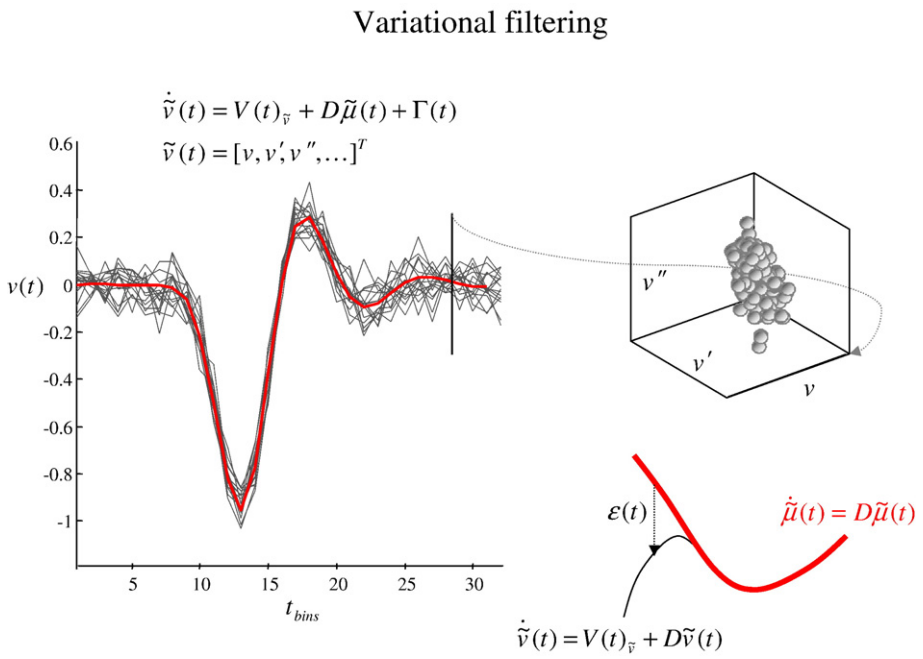

Fig. 1. Schematic illustrating the nature of variational filtering. The left panel shows the evolution of 32 particles over time as they negotiate a changing variational energy landscape. The peak or mode of this landscape is depicted by the red line. Particles flow, deterministically towards this mode; while, at the same time, they 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. 5). Here we focus on one state in six generalised coordinates of motion, three of which are shown in the insert. The trajectories below the insert show a path corresponding to the conditional mode (red) and the trajectory of an approximating particle (black) that is exempt from random forces.

approximations to the ensemble density, which require only the path of the mode. This can by computed easily by integrating the path of a particle that converges to the path of the true mode:

Lemma 3. (Trajectory following). *For large κ, the path of a particle, whose motion in generalised coordinates conforms to*

$$\begin{aligned}\dot{v} &= \kappa V(u,t)_v + v'\\ \dot{v}' &= \kappa V(u,t)_{v'} + v''\\ \dot{v}'' &= \ldots\end{aligned} \tag{18}$$

converges exponentially to the mode at a rate proportional to the constant κ.

Proof. We can express the motion of this particle in terms of a derivative operator D

$$\dot{u} = \kappa V(u,t)_u + Du \tag{19}$$

This matrix operator is a block matrix, whose first leading-diagonal contains identity matrices I

$$D = \begin{bmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix} \otimes I$$

The Kronecker Tensor product, $\otimes$ replaces each element of the first matrix with that element multiplied by the second matrix. Because the mode satisfies the extremal condition; $V(\tilde{\mu},t)_u = 0$, the first-order expansion of the variational energy gradient, around its mode is

$$\begin{aligned}V(u,t)_u &= V(\tilde{\mu},t)_u + V(\tilde{\mu},t)_{uu}\varepsilon\\ &= V(\tilde{\mu},t)_{uu}\varepsilon\\ \varepsilon &= u - \tilde{\mu}\end{aligned} \tag{20}$$

This expansion enables us to characterise the local evolution of ε, the difference between the location of the approximating particle and the mode in generalised coordinates (*c.f.*, a linear stability analysis; see inset in Fig. 1). Substituting Eq. (20) into Eq. (19) and using $\dot{\tilde{\mu}} = D\tilde{\mu}$ from Eq. (17) gives

$$\begin{aligned}\dot{\varepsilon} &= \Im(t)\varepsilon\\ \Im(t) &= \kappa V(u,t)_{uu} + D.\end{aligned} \tag{21}$$

It is easy to see that the ε will decay exponentially to zero, for suitably large values of κ. This is because the Jacobian $\Im(t)$ is dominated by the negative-definite curvature term. The rate of convergence to the mode is proportional to the negative real eigenvalues of $\Im(t)$ (*c.f.*, Lyapunov exponents). □

Summary

In this section, we have seen that inference on both models and their 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 a corresponding variational energy. This energy is simply the internal energy $\ln(p(y,\vartheta|m))$ expected under the Markov Blanket of each set of parameters. For dynamic systems, we introduce time-varying states and replace energies with actions to optimise 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. Alternatively, one can focus on the conditional mode and ingrate the path of a single particle that maximises variational action. This particle behaves in the same way as any other member of the ensemble except that it is not subject to random fluctuations. We now consider the most common fixed-form approximation.

The Laplace approximation

As mentioned above, variational filtering (*i.e.*, integrating an ensemble of solutions) can approximate conditional densities on the states with any form (see Friston, 2008). However, we will adopt a simpler approach by assuming that the ensemble density has a fixed Gaussian form. This reduces the problem of integrating the paths of an entire ensemble, using stochastic differential equations (Eq. (17)), to the much simpler problem of integrating the deterministic motion of the conditional mode (Eq. (18)). We can then use analytic results for the covariance, to obtain the sufficient statistics of the variational density. These results follow from the Laplace approximation, in which the precision (inverse covariance) is simply the negative curvature of the internal energy at the mode. We now outline these results for general static and dynamic cases.

Static models

Under the Laplace approximation, the marginals of the ensemble density assume a Gaussian form $q(\vartheta^i)=N(\vartheta^i:\mu^i,\Sigma^i)$ with variational parameters μ^i and Σ^i, corresponding to the mode and conditional covariance of the i-th parameter set. In an attempt to keep notation consistent, we will use μ^i for the conditional expectation or mean of the i-th set of unknowns and η^i for their prior expectation. Similarly, we will use Σ^i and C^i for the conditional and prior covariances and Π^i and P^i for the corresponding inverses (*i.e.*, precisions).

The advantage of the Laplace assumption is that the conditional covariance can be evaluated very simply: under this fixed-form assumption the free, internal and variational energies of static systems are

$$\begin{aligned} F &= U(\mu) + \sum_i W^i + H \\ U(\vartheta) &= \ln\ p(y,\vartheta) \\ V(\vartheta^i) &= U(\vartheta^i,\mu^{\setminus i}) + \sum_{j\neq i} W^j \\ H &= \tfrac{1}{2}\sum_i \left(\ \ln\ |\Sigma^i| + p^i\ \ln\ 2\pi e\right) \\ W^i &= \tfrac{1}{2} tr\left(\Sigma^i U_{\vartheta^i\vartheta^i}\right) \end{aligned} \tag{22}$$

$p^i=\dim(\vartheta^i)$ is the number of parameters in the i-th set and $U(\mu)$ is the internal energy evaluated at the conditional mode of all sets. As noted in Lemma 1, the conditional modes are simply those parameters that maximise variational energy. The conditional precisions are obtained as an analytic function of the modes by differentiating Eq. (22) with respect to the covariances and solving for zero

$$\begin{aligned} &F_{\Sigma^i} = \tfrac{1}{2} U_{\vartheta^i\vartheta^i} + \tfrac{1}{2}\Pi^i = 0 \Rightarrow \\ &\Pi^i = -U_{\vartheta^i\vartheta^i} \end{aligned} \tag{23}$$

This is the negative curvature of the internal energy at the conditional mode. Note that this solution does not depend on the mean-field approximation but only on the Laplace approximation. Substitution into Eq. (22) means $W^i = \frac{1}{2}p^i$ and[7]

$$\begin{aligned} F &= U(\mu) + H \\ H &= \tfrac{1}{2}\sum_i \left(\ \ln\ |\Sigma^i| + p^i\ \ln\ 2\pi\right) \end{aligned} \tag{24}$$

This gives a compact and simple form for the free-energy and conditional precisions; and reduces the problem of inference to finding the conditional modes of the variational energy. This generally proceeds in a series of iterated steps, in which the mode of each parameter set is updated. These updates optimise the variational energy in Eq. (22) with respect to μ^i, using the sufficient statistics $\mu^{\setminus i}$ and $\Sigma^{\setminus i}$ of the other sets. It is evident that the quantities W^i represent the contribution to the variational energy of other parameter sets. We will refer to these as mean-field terms. We have discussed special cases of this fixed-form scheme previously and have shown how iterative optimisation reduces to expectation maximisation (EM; Dempster et al., 1977) and restricted maximum likelihood (ReML; Harville, 1977) for linear models (Friston et al., 2007; see also Fahrmeir and Tutz 1994 for a related discussion in the context of generalised linear models).

The price paid, when using the Laplace approximation, is that we cannot represent multimodal or discrete distributions. However, one can represent non-Gaussian densities through non-linear transformations (we will see examples of this when representing non-negative states and scale-parameters in the final section). We now reprise the derivations above for dynamic models.

Dynamic models

For dynamic models one follows the same treatment but replacing the free-energy with action. For simplicity, we will deal explicitly with two set of parameters, which we will call parameters and hyperparameters, $\vartheta=\theta,\lambda$. Unless stated otherwise, all quantities and their derivatives are evaluated at the

[7] In which we have removed constant terms in the entropy that do not contribute to the free-energy.

conditional mode. Under the Laplace assumption, the free, internal and variational actions are (*c.f.*, Eq. (22) and using $U(t):=U(u,t|\theta,\lambda)$)

$$\begin{aligned}\overline{F} &= \overline{U}(\mu) + \overline{H}\\ \overline{U} &= \int U(t)dt + U(\theta) + U(\lambda)\end{aligned}$$

$$\begin{aligned}\overline{V}(u) &= \int U\left(u,t|\mu^{\theta},\mu^{\lambda}\right) + W(t)^{\theta} + W(t)^{\lambda}dt\\ \overline{V}(\theta) &= \int U\left(\mu^{u},t|\theta,\mu^{\lambda}\right) + W(t)^{u} + W(t)^{\lambda}dt + U(\theta)\\ \overline{V}(\lambda) &= \int U\left(\mu^{u},t|\mu^{\theta},\lambda\right) + W(t)^{u} + W(t)^{\theta}dt + U(\lambda)\end{aligned}$$

$$\begin{aligned}\overline{H} &= \int \tfrac{1}{2}\ \ln\ |\Sigma(t)^{u}|dt + \tfrac{1}{2}\ \ln\ |\Sigma^{\theta}| + \tfrac{1}{2}\ \ln\ |\Sigma^{\lambda}|\\ &\quad + \tfrac{1}{2}\left(Np^{u} + p^{\theta} + p^{\lambda}\right)\ \ln\ 2\pi\\ W(t)^{u} &= \tfrac{1}{2}tr\left(\Sigma^{u}U(t)_{uu}\right)\\ W(t)^{\theta} &= \tfrac{1}{2}tr\left(\Sigma^{\theta}U(t)_{\theta\theta}\right)\\ W(t)^{\lambda} &= \tfrac{1}{2}tr\left(\Sigma^{\lambda}U(t)_{\lambda\lambda}\right)\end{aligned} \tag{25}$$

$\Sigma(t)^{u}$ is the conditional covariance of the states at time $0\le t\le N$. Following the arguments for static models, the conditional precisions are the negative curvatures of $\bar{U}$, the internal action evaluated at the conditional mode

$$\begin{aligned}\Pi(t)^{u} &= -\overline{U}_{uu} = -U(t)_{uu}\\ \Pi^{\theta} &= -\overline{U}_{\theta\theta} = -\int U(t)_{\theta\theta}dt - U(\theta)_{\theta\theta}\\ \Pi^{\lambda} &= -\overline{U}_{\lambda\lambda} = -\int U(t)_{\lambda\lambda}dt - U(\lambda)_{\lambda\lambda}\end{aligned} \tag{26}$$

Notice that the precisions of the parameters and hyperparameters increase with the number of observations, as we would expect. To evaluate these precisions we need only the modes, which maximise variational action.

In line with conventional variational schemes, we can update the modes of our three parameter sets in three distinct steps. However, the step dealing with the state (D-step) must integrate its conditional mode over time and accumulate the quantities necessary for updating the parameters (E-step) and hyperparameters (M-step). We now consider optimising the modes or conditional expectations in each of these steps.

Dynamic expectation maximisation

The D-step

Eq. (19) prescribes the (approximate) trajectory of the conditional mode, which can be realised with a local linearisation (following Ozaki, 1992) by integrating over Δt to recover the motion of the particle tracking the mode:

$$\begin{aligned}\Delta\tilde{\mu} &= \left(\exp(\Delta t\Im) - I\right)\Im^{-1}\dot{\tilde{\mu}}\\ \dot{\tilde{\mu}} &= \kappa V(\tilde{\mu},t)_{u} + D\tilde{\mu} \Rightarrow\\ \Im &= \kappa V(\tilde{\mu},t)_{uu} + D\end{aligned} \tag{27}$$

Ozaki (1992) shows that these updates are consistent, coincide with the true trajectory (at least for linear systems) and retain the qualitative characteristics of the continuous formulation. Indeed, this local linearisation is the basis of the innovation approach to time-series modelling (see below and Ozaki and Iino 2001). For simplicity, we have suppressed the dependency of $V(u,t)$ on the data. However, it is generally necessary to augment Eq. (27) with any time-varying quantities that affect the variational energy. This ensures that the forces that act on the mode change appropriately over the integration time. The form of the ensuing Jacobian $\Im(t)$ is described in the next section.

The updates in Eq. (27) provide the conditional trajectory $\tilde{\mu}(t)$ at each time point. Usually, Δt is the time between observations but could be smaller, if nonlinearities in the model render local linearity assumptions untenable (we will see an example of this later, when illustrating nonlinear deconvolution). Note that when there are no variational influences and $V_u = V_{uu} = 0$ this update reduces to a Taylor expansion of the mode's motion; *i.e.* $\tilde{\mu}(t+\Delta t) = \exp(\Delta tD)\tilde{\mu}(t)$.

The E- and M-steps

Exactly the same scheme can be used for the E- and M-steps. However, in this instance there are no generalised coordinates to consider. Furthermore, we can set the interval between updates to be arbitrarily long because the parameters are updated after the time-series has been integrated. If $\Delta t\to\infty$ is sufficiently large, the matrix exponential in Eq. (27) disappears[8] giving

$$\begin{aligned}\Delta\mu^{\theta} &= -\Im(\theta)^{-1}\dot{\mu}^{\theta}\\ \dot{\mu}^{\theta} &= \kappa\overline{V}(\theta)_{\theta} \Rightarrow\\ \Im(\theta) &= \kappa\overline{V}(\theta)_{\theta\theta}\end{aligned} \tag{28}$$

similarly for the hyperparameters. Eq. (28) is a conventional Gauss–Newton update scheme, in which κ disappears. In this sense, the D-Step can be regarded as a generalisation of classical ascent schemes to generalised coordinates that cover dynamic systems. In practice, we retain the matrix exponential because it provides a graceful regularisation of the ascent (see Friston et al., 2007 for details). This can be useful when dealing with highly nonlinear dynamic models that exhibit structural instabilities[9].

These updates furnish a variational scheme under the Laplace approximation. In practice, we find that $\kappa = 1$ is sufficient large to ensure convergence in the majority of situations. To further simplify things, we will assume $\Delta t = 1$; *i.e.*, sampling intervals serve as units of time during model specification. With these simplifications, the DEM scheme can be summarised as iterating until convergence

D-step (states)

for $t = 1{:}N$

$$\begin{aligned}\Im &= V(\tilde{\mu},t)_{uu} + D\\ \Delta\tilde{\mu} &= \left(\exp\left(\Im\right) - I\right)\Im^{-1}\left(V(\tilde{\mu},t)_{u} + D\tilde{\mu}\right)\\ \Pi(t)^{u} &= -U(t)_{uu}\end{aligned}$$

end

E-step (parameters)

$$\begin{aligned}\Delta\mu^{\theta} &= -\overline{V}(\theta)_{\theta\theta}^{-1}\overline{V}(\theta)_{\theta}\\ \Pi^{\theta} &= -\overline{U}(\theta)_{\theta\theta}\end{aligned}$$

[8] Because the curvature of the Jacobian is negative definite.

[9] For example, if the parameters change and the system looses a fixed-point attractor, the states generally diverge exponentially. This corresponds to a phase-transition in the free-energy landscape, which can confound simple ascent schemes. To counter this, we halve Δt when the objective function fails to increase and double it otherwise.

M-step (hyperparameters)

$$\begin{aligned}\Delta\mu^{\lambda} &= -\overline{V}(\lambda)^{-1}_{\lambda\lambda}\overline{V}(\lambda)_{\lambda}\\ \Pi^{\lambda} &= -\overline{U}(\lambda)_{\lambda\lambda}\end{aligned} \tag{29}$$

where the integrals in Eq. (25) are approximated with the appropriate sums.

We will call these three updates D-, E- and M-steps to highlight their connection with expectation maximisation (EM; Dempster et al., 1977). Provided the internal energy is linear in the hyperparameters, DEM is exact and there is no formal distinction between the E- and M-steps (see Friston et al., 2007). The D-step can be construed as a dynamic step that effectively deconvolves states from data. The reason we call it DEM as opposed to a dynamic variational scheme is that the update rules optimise the variational action explicitly, using a coordinate ascent. In a standard variational scheme, the updates would be based on analytic solutions to Eqs. (11) and (12), which optimise free-action implicitly. However, these closed-form solutions have to be derived for each model and often entail assumptions of convenience (*e.g.*, conjugate priors). DEM does not require these closed-form solutions (or conjugate priors) and requires only the gradients and curvatures of the internal energy. Having said this, for simple models, many of the DEM and standard variational updates are formally identical, particularly in the D- and E-steps.

Summary

In this section, we have seen how the inversion of dynamic models can be formulated as an optimisation of free-action. This action comprises the path-integral of free-energy associated with changing states and a constant (of integration) corresponding to the prior energy of time-invariant parameters (see Eq. (25)). By assuming a fixed-form (Laplace) approximation to the conditional density, one can reduce optimisation to finding the conditional modes of unknown quantities, because their conditional covariance is simply the curvature of the internal action (evaluated at the mode). The conditional modes of (mean-field) marginals optimise variational action, which can be framed in terms of gradient ascent. For the states, this entails finding a path or trajectory with stationary variational action. This path can be tracked using a fast gradient ascent that supplements the flow with the conditional mode of generalised motion.

To implement this scheme we need the gradients and curvatures of the internal energy, which are defined by the generative model implicit in; $\overline{U}(u,\vartheta)=\int U(u,t|\vartheta)dt+U(\vartheta)$. Next, we consider generative models for dynamic systems and the variational steps they entail.

Nonlinear dynamic causal models

In this section, we apply the results of the previous section to an input-state-output model with additive noise. This is a general model that 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 (see Efron and Morris, 1973 for a discussion of empirical priors in the context of static models). 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. If motion is discounted completely, the model reduces to conventional nonlinear models under parametric assumptions; and the scheme becomes formally identical to expectation maximisation.

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

$$\begin{aligned}y &= g(x,v)+z\\ \dot{x} &= f(x,v)+w\end{aligned} \tag{30}$$

The continuous nonlinear functions f and g of the states are parameterised by θ. 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 usually observed directly. We assume that 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 Wiener 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

$$\begin{aligned} y &= g(x,v)+z & x' &= f(x,v)+w\\ \dot{y} &= g_x x'+g_v v'+\dot{z} & x'' &= f_x x'+f_v v'+\dot{w}\\ \ddot{y} &= g_x x''+g_v v''+\ddot{z} & x''' &= f_x x''+f_v v''+\ddot{w}\\ &\vdots & &\vdots \end{aligned} \tag{31}$$

The first (observer) equation show 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; $q(u,t)=q(\tilde{v},\tilde{x},t)$. The second (state) equations enforce a coupling between low and high-order motion of the hidden states $\tilde{x}$ and confer memory to the system.

The conditional energy function associated with this system; $U(t)=\ln p(\tilde{y}|u,\theta,\lambda)+\ln p(u)$ comprises a log-likelihood and prior. Gaussian assumptions about the fluctuations $p(\tilde{z})=N(\tilde{z}:0,\tilde{\Sigma}^z)$ furnish the likelihood, $p(\tilde{y}|\vartheta)=N(\tilde{y}:\tilde{g},\tilde{\Sigma}^z)$. This is because $\tilde{y}=\tilde{g}+\tilde{z}$, where the predicted response $\tilde{g}=g,g',g'',\ldots$ comprises the derivatives

$$\begin{aligned} g &= g(x,v) & f &= f(x,v)\\ g' &= g_x x'+g_v v' & f' &= f_x x'+f_v v'\\ g'' &= g_x x''+g_v v'' & f'' &= f_x x''+f_v v''\\ &\vdots & &\vdots \end{aligned} \tag{32}$$

Similarly, Gaussian assumptions about state noise $p(\tilde{w})=N(\tilde{w}:0,\tilde{\Sigma}^w)$ furnish a prior $p(u)=p(\tilde{x}|\tilde{v})p(\tilde{v})$ in terms of predicted motion, where $p(\tilde{x}|\tilde{v})=N(D\tilde{x}:\tilde{f},\tilde{\Sigma}^w)$. This is because the motion of the hidden states is $D\tilde{x}=\tilde{f}+\tilde{w}$, where the predicted motion is $\tilde{f}=f,f',f'',\ldots$

We will assume Gaussian priors, $p(\tilde{v})=N(\tilde{v}:\tilde{\eta}^v,\tilde{C}^v)$, where $\tilde{\eta}^v$ are the prior expectations of the generalised causes. To simplify things, we will assume these are flat and re-instate informative empirical priors with hierarchical models later. We assume the same form for priors on the parameters, $p(\theta)=N(\theta:\eta^\theta,C^\theta)$, with prior precision P^θ (similarly for the hyperparameters). Note that Gaussian priors are not restrictive because, $\tilde{g}(u,\theta)$, $\tilde{f}(u,\theta)$, $\tilde{\Sigma}(u,\lambda)^z$ and $\tilde{\Sigma}(u,\lambda)^w$ can all be nonlinear functions that embody probability integral transforms (*i.e.*, can implement a re-parameterisation in terms of non-Gaussian processes). We will illustrate this in the last section.

Generally, the covariances of the fluctuations $\tilde{\Sigma}(u,\lambda)^z$ and $\tilde{\Sigma}(u,\lambda)^w$ can be functions of the states that allow for state-dependent changes in the probabilistic context in which responses are generated. However, we will deal with state-dependent covariances elsewhere and assume they depend only on hyperparameters in this paper. 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 priors.

Energy functions

For these generative models, the actions associated with the free $\bar{F}$ and internal $\bar{U}$ energy are

$$\begin{aligned}
\bar{F} &= \bar{U}(\mu)+\bar{H}\\
\bar{U} &= \textstyle\sum_t U(t)+\tfrac{1}{2}\ \ln\ |P^\theta|+\tfrac{1}{2}\ \ln\ |P^\lambda|-\tfrac{1}{2}\varepsilon^{\theta T}P^\theta\varepsilon^\theta-\tfrac{1}{2}\varepsilon^{\lambda T}P^\lambda\varepsilon^\lambda\\
\bar{H} &= \tfrac{1}{2}\sum_t\ \ln\ |\Sigma(t)^u|+\tfrac{1}{2}\ \ln\ |\Sigma^\theta|+\tfrac{1}{2}\ \ln\ |\Sigma^\lambda|\\
U(t) &= \tfrac{1}{2}\ln|\tilde{\Pi}|-\tfrac{1}{2}\tilde{\varepsilon}^T\tilde{\Pi}\tilde{\varepsilon}-\frac{p}{2}\ \ln\ 2\pi
\end{aligned} \tag{33}$$

$$\tilde{\Pi}=\begin{bmatrix}\tilde{\Pi}^z & \\ & \tilde{\Pi}^w\end{bmatrix}\quad \tilde{\varepsilon}(t)=\begin{bmatrix}\tilde{\varepsilon}^v=\tilde{y}-\tilde{g}\\ \tilde{\varepsilon}^x=D\tilde{x}-\tilde{f}\end{bmatrix}\quad \begin{matrix}\varepsilon^\theta=\mu^\theta-\eta^\theta\\ \varepsilon^\lambda=\mu^\lambda-\eta^\lambda\end{matrix}$$

where, $p=rank(\tilde{\Pi})$ and we have replaced integrals with summations over time bins $t=1,\ldots,N$. Constant terms in the entropy have been omitted (*c.f.*, Eq. (25)) because they are cancelled by identical terms from the Gaussian priors in the free-action. The auxiliary variables $\tilde{\varepsilon}(t)$ comprise prediction errors for the response and generalised motion of hidden states, where $\tilde{g}(t)$ and $\tilde{f}(t)$ are the respective predictions. The precision of these predictions is encoded by $\tilde{\Pi}$, which depends on the magnitude of the random effects. The use of prediction errors simplifies exposition and may

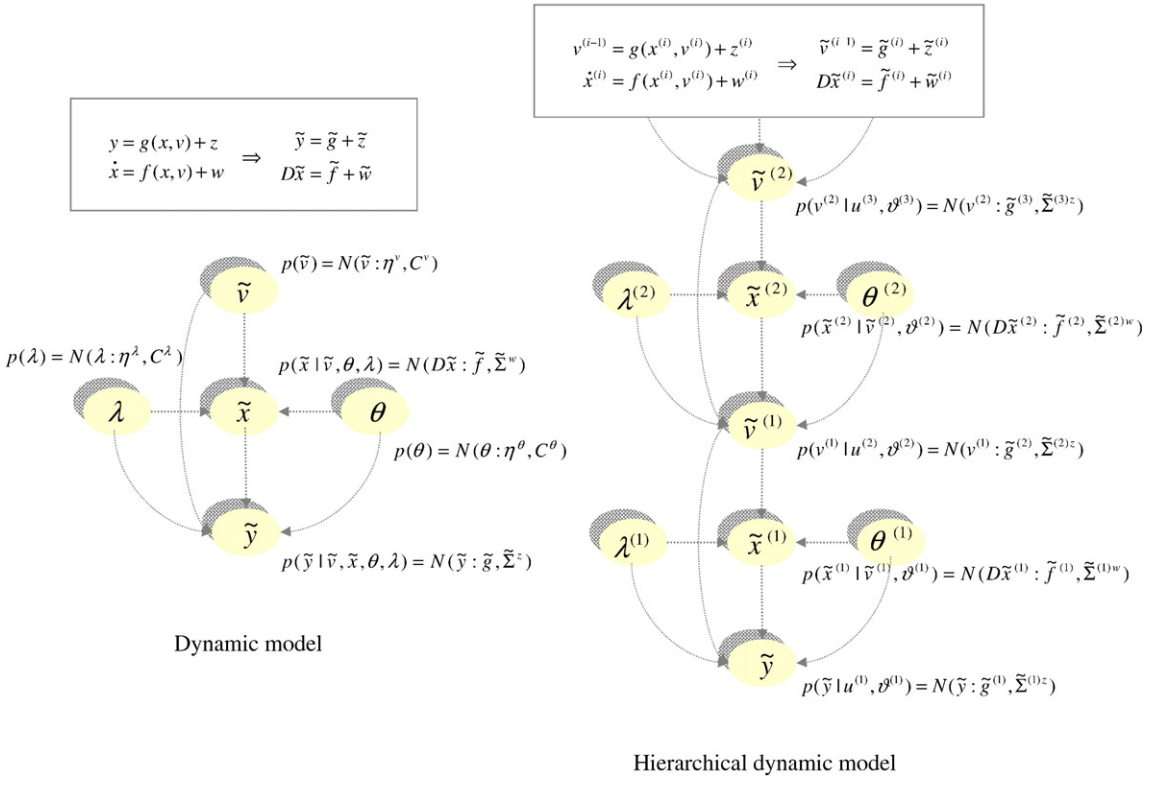

Fig. 2. Conditional dependencies of a dynamic (left) and hierarchical dynamic (right) model, shown as directed Bayesian graphs. The nodes of these graphs correspond to quantities in the model and the responses they generate. The arrows or edges indicate conditional dependencies between these quantities. 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). In the hierarchical dynamic model, priors on the causes are replaced by empirical priors, which depend on states and parameters in the level above.

be used by neurobiological implementations of this scheme (*i.e.*, encoded explicitly in the brain; see Friston 2005; Friston et al., 2006).

Conditional precisions

As established in the previous section, the conditional precisions are the curvatures of the internal energy

$$\begin{aligned}\Pi(t)^u &= -U(t)_{uu}\\ \Pi^\theta &= -\textstyle\sum_i U(t)_{\theta\theta} + P^\theta\\ \Pi^\lambda &= -\textstyle\sum_i U(t)_{\lambda\lambda} + P^\lambda\end{aligned}$$

$$\begin{aligned}U(t)_u &= -\tilde{\varepsilon}_u^T\tilde{\Pi}\tilde{\varepsilon} \quad U(t)_\theta = -\tilde{\varepsilon}_\theta^T\tilde{\Pi}\tilde{\varepsilon} \quad U(t)_{\lambda i} = -\tfrac{1}{2}tr\left(Q_i\left(\tilde{\varepsilon}\tilde{\varepsilon}^T - \tilde{\Sigma}\right)\right)\\ U(t)_{uu} &= -\tilde{\varepsilon}_u^T\tilde{\Pi}\tilde{\varepsilon}_u \; U(t)_{\theta\theta} = -\tilde{\varepsilon}_\theta^T\tilde{\Pi}\tilde{\varepsilon}_\theta \; U(t)_{\lambda\lambda ij} = -\tfrac{1}{2}tr\left(Q_i\tilde{\Sigma}Q_j\tilde{\Sigma}\right)\end{aligned} \tag{34}$$

where the covariance, $\tilde{\Sigma}$ is the inverse of $\tilde{\Pi}$. The i-th element of the energy gradient; $U(t)_{\lambda i}=\partial_{\lambda i}U(t)$ is the derivative with respect to the i-th hyperparameter (similarly for the curvatures). We have assumed that the precision of the random fluctuations is linear in the hyperparameters, where $Q_i=\partial_{\lambda i}\tilde{\Pi}$, and $\partial_{\lambda\lambda}\tilde{\Pi}=0$. This is an important assumption that simplifies things considerably (see Friston et al., 2007 for details). This makes the E-step exact because conditional uncertainty about the hyperparameters encoding precisions does not affect the variational energy of the states or parameters.

Conditional modes

To evaluate the conditional modes, we will need the derivatives of the prediction error with respect to the unknowns. Under local linearity assumptions, these have the following form[10]

$$\tilde{\varepsilon}_u = \begin{bmatrix}\tilde{\varepsilon}_v^v & \tilde{\varepsilon}_x^v\\ \tilde{\varepsilon}_v^x & \tilde{\varepsilon}_x^x\end{bmatrix} = -\begin{bmatrix}I\otimes g_v & I\otimes g_x\\ I\otimes f_v & I\otimes f_x - D\end{bmatrix} \quad \tilde{\varepsilon}_{\theta_i} = \tilde{\varepsilon}_{u\theta_i}u \quad u = \begin{bmatrix}\tilde{v}\\ \tilde{x}\end{bmatrix} \tag{35}$$

The form of our generative model (Eq. (31)) 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 that the motion of the hidden states conforms to the dynamics entailed by the state equation. $\tilde{\varepsilon}_{\theta i}$ is the change in prediction error with the i-th parameter (*i.e.*, the i-th column of $\tilde{\varepsilon}_\theta$). In computing $\tilde{\varepsilon}_\theta$ have used second-order terms mediating an interaction between the parameters and states

$$\tilde{\varepsilon}_{u\theta_i} = -\begin{bmatrix}I\otimes g_{v\theta_i} & I\otimes g_{x\theta_i}\\ I\otimes f_{v\theta_i} & I\otimes f_{x\theta_i}\end{bmatrix} \quad \tilde{\varepsilon}_{\theta_i u} = \tilde{\varepsilon}_{u\theta_i}^T \tag{36}$$

These second-order terms are needed to quantify how the states and parameters affect each other through mean-field effects (see below). It is relatively simple to supplement Eq. (36) with second-order terms involving f_{uu} and g_{uu} but in practice this is not necessary, provided that the integration intervals are sufficiently small to conform to local linearity assumptions. In what follows, we place the derivatives above into the three variational steps of the previous section.

The D-step

As mentioned above, when integrating the path of the approximate mode, it is necessary to augment the states to cover time-dependent variables that affect the variational energy; in this case the data. This augmented system and its Jacobian are

$$\begin{bmatrix}\dot{\tilde{y}}\\ \dot{u}\end{bmatrix} = \begin{bmatrix}D\tilde{y}\\ V(t)_u + Du\end{bmatrix} \quad \Rightarrow \quad \Im = \begin{bmatrix}D & 0\\ V(t)_{uy} & V(t)_{uu}+D\end{bmatrix} \tag{37}$$

Note that when the path of this particle converges to the mode, the extremal conditions $V(t)_u=0$ are met and the motion of the highest derivatives are zero. From Eq. (29) the update is

$$\begin{aligned}\begin{bmatrix}\Delta\tilde{y}\\ \Delta u\end{bmatrix} &= (\exp(\Delta t\Im) - I)\Im^{-1}\begin{bmatrix}D\tilde{y}\\ V(t)_u + Du\end{bmatrix}\\ V(t) &= -\tfrac{1}{2}\tilde{\varepsilon}^T\tilde{\Pi}\tilde{\varepsilon} + W(t)^\theta\\ V(t)_u &= -\tilde{\varepsilon}_u^T\tilde{\Pi}\tilde{\varepsilon} + W(t)_u^\theta\\ V(t)_{uu} &= -\tilde{\varepsilon}_u^T\tilde{\Pi}\tilde{\varepsilon}_u + W(t)_{uu}^\theta\\ V(t)_{uy} &= -\tilde{\varepsilon}_u^T\tilde{\Pi}\tilde{\varepsilon}_y\end{aligned} \tag{38}$$

$$\begin{aligned}W(t)^\theta &= -\tfrac{1}{2}tr\left(\Sigma^\theta\tilde{\varepsilon}_\theta^T\tilde{\Pi}\tilde{\varepsilon}_\theta\right)\\ W(t)_{u_i}^\theta &= -\tfrac{1}{2}tr\left(\Sigma^\theta\tilde{\varepsilon}_{\theta u_i}^T\tilde{\Pi}\tilde{\varepsilon}_\theta\right)\\ W(t)_{uu_{ij}}^\theta &= -\tfrac{1}{2}tr\left(\Sigma^\theta\tilde{\varepsilon}_{\theta u_i}^T\tilde{\Pi}\tilde{\varepsilon}_{\theta u_j}\right)\end{aligned}$$

Notice that the mean-field term, $W(t)^\lambda$ does not contribute to the D-step because it is not a function of the states (or parameters). This means that uncertainly about the hyperparameters does not affect the update for the states (or parameters), which is a result of the linear hyperparameterisation of the precision. In this non-hierarchical model $\tilde{\varepsilon}_y=I$ but takes a more structured form in hierarchical models (see below).

The E- and M-steps

A similar but simpler derivation follows for the E- and M-step updates

$$\begin{aligned}\Delta\mu^\theta &= -\overline{V}(\theta)_{\theta\theta}^{-1}\overline{V}(\theta)_\theta\\ \overline{V}(\theta) &= \textstyle\sum_t\left(U(t) + W(t)^u\right) - \tfrac{1}{2}\varepsilon^{\theta T}P^\theta\varepsilon^\theta\\ \overline{V}(\theta)_\theta &= \textstyle\sum_t\left(U(t)_\theta + W(t)_\theta^u\right) - P^\theta\varepsilon^\theta\\ \overline{V}(\theta)_{\theta\theta} &= \textstyle\sum_t\left(U(t)_{\theta\theta} + W(t)_{\theta\theta}^u\right) - P^\theta\end{aligned} \tag{39}$$

$$\begin{aligned}W(t)^u &= -\tfrac{1}{2}tr\left(\Sigma_t^u\tilde{\varepsilon}_u^T\tilde{\Pi}\tilde{\varepsilon}_u\right)\\ W(t)_{\theta_i}^u &= -\tfrac{1}{2}tr\left(\Sigma_t^u\tilde{\varepsilon}_{u\theta_i}^T\tilde{\Pi}\tilde{\varepsilon}_u\right)\\ W(t)_{\theta\theta_{ij}}^u &= -\tfrac{1}{2}tr\left(\Sigma_t^u\tilde{\varepsilon}_{u\theta_i}^T\tilde{\Pi}\tilde{\varepsilon}_{u\theta_j}\right)\end{aligned}$$

[10] In practice, the first block of $\tilde{\varepsilon}_\theta$ can be replaced by $-g_\theta$ to eschew linearity assumptions.

Similarly for the hyperparameters

$$\begin{aligned}
\Delta\mu^{\lambda} &= -\overline{V}(\lambda)^{-1}_{\lambda\lambda}\overline{V}(\lambda)_{\lambda}\\
\overline{V}(\lambda) &= \sum_t\Big(U(t)+W(t)^{u}+W(t)^{\theta}\Big)-\tfrac{1}{2}\varepsilon^{\lambda T}P^{\lambda}\varepsilon^{\lambda}\\
\overline{V}(\lambda)_{\lambda} &= \sum_t\Big(U(t)_{\lambda}+W(t)^{u}_{\lambda}+W(t)^{\theta}_{\lambda}\Big)-P^{\lambda}\varepsilon^{\lambda}\\
\overline{V}(\lambda)_{\lambda\lambda} &= \sum_t U(t)_{\lambda\lambda}-P^{\lambda}\\
W(t)^{u}_{\lambda_i} &= -\tfrac{1}{2}tr\Big(\Sigma^{u}_{i}\widetilde{\varepsilon}^{iT}_{u}Q_i\widetilde{\varepsilon}^{i}_{u}\Big)\\
W(t)^{\theta}_{\lambda_i} &= -\tfrac{1}{2}tr\Big(\Sigma^{\theta}\widetilde{\varepsilon}^{iT}_{\theta}Q_i\widetilde{\varepsilon}^{i}_{\theta}\Big)
\end{aligned}\tag{40}$$

Although uncertainty about the hyperparameters does not affect the updates for the states and parameters, uncertainty about both the states and parameters enters the hyperparameter update. However, the simplification afforded by a linear hyperparameterisation of the precision means that $V(\lambda)_{\lambda\lambda}=U(\lambda)_{\lambda\lambda}$ is simply the negative precision of the hyperparameters.

These steps represent a full variational scheme. A simplified version, which discounts uncertainty about the parameters and states in the D- and E-steps, would be the analogue of an EM scheme. This simplification is easy to implement by removing $W(t)^{\theta}$ and $W(t)^{u}$ from the D- and E-steps respectively. We will pursue this in the context of neurobiological implementations elsewhere. Removing the mean-field effects $W(t)^{i}$ from all steps would reduce the scheme to an iterated conditional mode (ICM) version of DEM, in which conditional uncertainty is ignored completely. We now have the results necessary for a variational inversion of dynamic systems. However, before demonstrating the scheme we will consider an important generalisation that augments empirical priors on the dynamics of hidden states with empirical priors on the causes.

Hierarchical nonlinear dynamic models

Hierarchical dynamic models (HDM) are important because they subsume many other models. In fact (with the exception of mixture models), they cover most parametric models that one could conceive of; from independent components analysis to generalised convolution models. The range and relationships among these special cases are themselves a large area, to which we will devote a subsequent paper. Here we simply describe the general form of these models and their inversion.

HDMs have the following form, which generalises the ($m=1$) DCM above

$$\begin{aligned}
y &= g\big(x^{(1)},v^{(1)}\big)+z^{(1)}\\
\dot{x}^{(1)} &= f\big(x^{(1)},v^{(1)}\big)+w^{(1)}\\
&\vdots\\
v^{(i-1)} &= g\big(x^{(i)},v^{(i)}\big)+z^{(i)}\\
\dot{x}^{(i)} &= f\big(x^{(i)},v^{(i)}\big)+w^{(i)}\\
&\vdots\\
v^{(m)} &= \eta^{v}+z^{(m+1)}
\end{aligned}\tag{41}$$

$f^{(i)}$ and $g^{(i)}$ are continuous nonlinear functions of the states. The innovations $z^{(i)}$and $w^{(i)}$ are conditionally independent fluctuations that enter 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 means that the HDM has a Markov property over levels, which simplifies the architecture of attending inference schemes. See Kass and Steffey (1989) for a discussion of approximate Bayesian inference in conditionally independent hierarchical models of static data. A key property of these hierarchical models is their connection to parametric empirical Bayes (Efron and Morris, 1973): consider the conditional energy function implied by the HDM above, in generalised coordinates $u^{(i)}=v^{(i)},v'^{(i)},\ldots,x^{(i)},x'^{(i)},\ldots$

$$\begin{aligned}
U(u,t|\vartheta) = \ &\ln\ p(\widetilde{y}|u^{(1)},\vartheta)+\ \ln\ p(u^{(1)}|u^{(2)},\vartheta)\\
&+\ldots+\ \ln\ p(\widetilde{v}^{(m)})
\end{aligned}\tag{42}$$

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 hand, they depend on quantities that have to be inferred; namely, supraordinate states and parameters. For example, the prediction $g\sim(u^{(i)},\theta^{(i)})$ plays the role of a prior expectation on $v\sim^{(i-1)}$, hence empirical Bayes; similarly for the hidden states. 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 determination in machine learning formulations of related problems (see Friston et al., 2002, 2007 for a fuller discussion of hierarchical models of static data).

HDMs are inverted in exactly the same way as above, with the following forms for the states and predictions

$$v=\begin{bmatrix}v^{(1)}\\ \vdots\\ v^{(m)}\end{bmatrix}x=\begin{bmatrix}x^{(1)}\\ \vdots\\ x^{(m)}\end{bmatrix}f=\begin{bmatrix}f\big(x^{(1)},v^{(1)}\big)\\ \vdots\\ f\big(x^{(m)},v^{(m)}\big)\end{bmatrix}g=\begin{bmatrix}g\big(x^{(1)},v^{(1)}\big)\\ \vdots\\ g\big(x^{(m)},v^{(m)}\big)\end{bmatrix}\tag{43}$$

The implicit 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

$$\widetilde{\varepsilon}=\begin{bmatrix}\widetilde{\varepsilon}^{v}\\ \widetilde{\varepsilon}^{x}\end{bmatrix}\quad \varepsilon^{v}=\begin{bmatrix}y\\ v\end{bmatrix}-\begin{bmatrix}g\\ \eta^{v}\end{bmatrix}\tag{44}$$

Note that the data and priors only enter the prediction error at the lowest and highest level respectively. At intermediate levels the prediction errors $v^{(i-1)}-g(x^{(i)},v^{(i)})$ mediate empirical priors on the causes. In other words, the causes are themselves predicted by supraordinate levels. This prediction and the ensuing constraints are the central feature of hierarchical models. The forms of the derivatives of the prediction error with respect to the states are[11]

$$\widetilde{\varepsilon}_{u}=-\begin{bmatrix}I\otimes(g_v-D^{T}) & I\otimes g_x\\ I\otimes f_v & (I\otimes f_x)-D\end{bmatrix}\tag{45}$$

A comparison with Eq. (36) shows an extra D^{T} matrix in the upper-left block; this reflects the fact that, in hierarchical models, causes also affect the prediction error within their own level, as

[11] The form of the partial derivatives with respect to the parameters is unchanged.

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. The key feature of these derivatives is their block-diagonal form, reflecting the hierarchical separability of the model

$$g_v = \begin{bmatrix} g_v^{(1)} & & \\ 0 & \ddots & \\ & \ddots & g_v^{(m)} \\ & & 0 \end{bmatrix} \quad g_x = \begin{bmatrix} g_x^{(1)} & & \\ 0 & \ddots & \\ & \ddots & g_x^{(m)} \\ & & 0 \end{bmatrix}$$
$$f_v = \begin{bmatrix} f_v^{(1)} & & \\ & \ddots & \\ & & f_v^{(m)} \end{bmatrix} \quad f_x = \begin{bmatrix} f_x^{(1)} & & \\ & \ddots & \\ & & f_x^{(m)} \end{bmatrix} \tag{46}$$

Note that the partial derivatives of $g(x,v)$ have an extra row to accommodate the highest hierarchical level. Finally, for hierarchical models, it is necessary to augment the D-step update (*c.f.*, Eq. (38)) with the priors

$$\begin{bmatrix} \dot{\tilde{y}} \\ \dot{u} \\ \dot{\tilde{\eta}}^v \end{bmatrix} = \begin{bmatrix} D\tilde{y} \\ V(t)_u + Du \\ D\tilde{\eta} \end{bmatrix} \Rightarrow \Im(t) = \begin{bmatrix} D & 0 & 0 \\ V(t)_{uy} & V(t)_{uu} + D & V(t)_{u\eta} \\ 0 & 0 & D \end{bmatrix} \tag{47}$$

This is because the priors, like the data, are time-varying quantities that affect the variational energy and its gradients. $V(t)_{uy} = -\tilde{\varepsilon}^T{}_u \tilde{\Pi} \tilde{\varepsilon}_y$ and $V(t)_{u\eta} = -\tilde{\varepsilon}^T{}_u \tilde{\Pi} \tilde{\varepsilon}_\eta$ where

$$\tilde{\varepsilon}_y = \begin{bmatrix} I \otimes \varepsilon_y^v \\ 0 \end{bmatrix} \quad \tilde{\varepsilon}_\eta = \begin{bmatrix} I \otimes \varepsilon_\eta^v \\ 0 \end{bmatrix} \quad \varepsilon_y^v = \begin{bmatrix} I \\ 0 \end{bmatrix} \quad \varepsilon_\eta^v = -\begin{bmatrix} 0 \\ I \end{bmatrix} \tag{48}$$

These forms reflect the fact that data and priors only affect the prediction error at the first and last levels respectively.

The precisions and temporal smoothness

So far, we have not considered the precisions $\tilde{\Pi}^z(\lambda)$ and $\tilde{\Pi}^w(\lambda)$ in any detail, other than to note that they are functions of the hyperparameters. 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. Independence assumptions about the innovations at each level means that their precisions have the following block-diagonal form

$$\Pi^z = \begin{bmatrix} \Pi^{(1)z} & & & \\ & \ddots & & \\ & & \Pi^{(m)z} & \\ & & & P^v \end{bmatrix} \quad \Pi^w = \begin{bmatrix} \Pi^{(1)w} & & \\ & \ddots & \\ & & \Pi^{(m)w} \end{bmatrix} \tag{49}$$

The intermediate levels are empirical prior precisions on the causes of dynamics in subordinate levels. The latter comprises a mixture of precision components that are functions of the hyperparameters. Usually, each component is a sparse matrix with leading-diagonal entries, encoding the precisions of the corresponding innovations (*e.g.*, all the innovations within one level of a hierarchy). To form the precisions in generalised coordinates, we take the Kronecker tensor product of a temporal precision, $S(\gamma)$, encoding temporal dependencies and the precision over innovations

$$\begin{aligned} \tilde{\Pi}^z &= S(\gamma) \otimes \Pi^z \\ \Pi^z &= \textstyle\sum_i \Pi(\lambda_i^z) \end{aligned} \tag{50}$$

similarly for Π^w. This assumes that the precisions (and covariances) can be factorised, into temporal and stationary innovation-specific parts. The matrices $Q_i = \partial_{\lambda i} \tilde{\Pi}$ needed for the M-step are simply

$$Q_i^z = \begin{bmatrix} S \otimes \partial_{\lambda_i^z} \Pi(\lambda_i^z) & \\ & 0 \end{bmatrix} \quad Q_i^w = \begin{bmatrix} 0 & \\ & S \otimes \partial_{\lambda_i^w} \Pi(\lambda_i^w) \end{bmatrix} \tag{51}$$

See Appendix A for the particular hyperparameterisation used in this paper.

Temporal correlations

The temporal precision encodes temporal dependencies among the innovations and can be expressed as a function of their autocorrelations

$$S(\gamma) = \begin{bmatrix} 1 & 0 & \ddot{\rho}(0) & \cdots \\ 0 & -\ddot{\rho}(0) & 0 & \\ \ddot{\rho}(0) & 0 & \dddot{\rho}(0) & \\ \vdots & & & \ddots \end{bmatrix}^{-1} \tag{52}$$

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 $\ddot{\rho}(0) \to \infty$ becomes large. In this instance, the precisions of the temporal derivatives fall to zero and the variational energy is determined by, and only by, the magnitude of the prediction errors on the causes and the first-order motion of the hidden states. This limiting case is assumed by conventional state-space models used in Bayesian filtering; it corresponds to the assumption that the fluctuations or innovations are independent (*c.f.*, a Wiener process or random walk). Although, this is a convenient assumption for conventional schemes and appropriate for physical systems with Brownian processes, it is less plausible for biological and other systems, where random fluctuations are themselves the product of other dynamical systems.

$S(\gamma)$ can be evaluated for any analytic autocorrelation function. For convenience, we assume that the temporal correlations of all innovations have the same Gaussian form[12]. This gives

$$S(\gamma) = \begin{bmatrix} 1 & 0 & -\frac{1}{2}\gamma & \cdots \\ 0 & \frac{1}{2}\gamma & 0 & \\ -\frac{1}{2}\gamma & 0 & \frac{3}{4}\gamma^2 & \\ \vdots & & & \ddots \end{bmatrix}^{-1} \tag{53}$$

where γ is the precision parameter of a Gaussian $\rho(t)$ and increases with roughness. Clearly, the conditional density of the temporal

[12] This assumption is relaxed easily by using state and level-specific temporal precisions in Eq. (49).

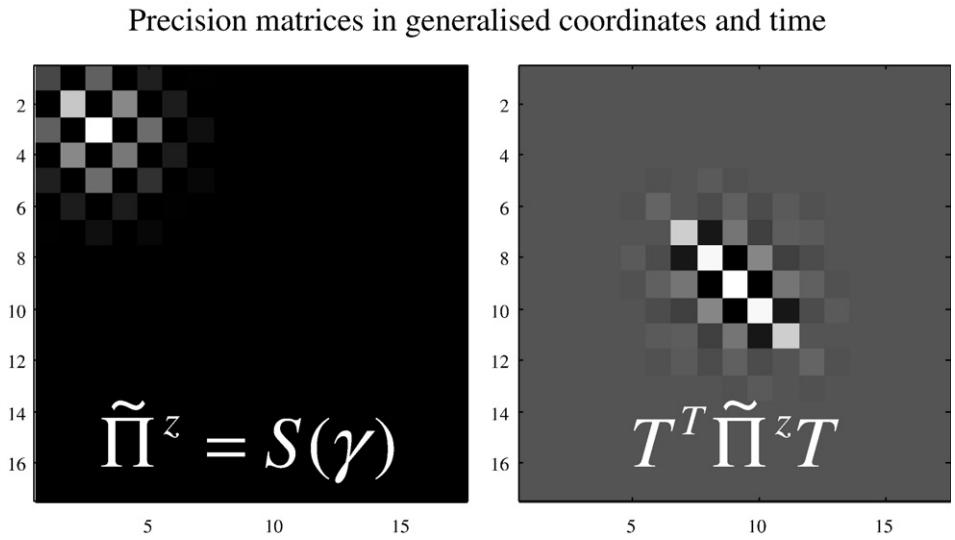

Fig. 3. Image representations of the precision matrices encoding temporal dependencies among a random fluctuation or innovation. The precision in generalised coordinates (left) and over discrete samples in time (right) is shown for a roughness $\gamma=4$ and seventeen observations (or an embedding order of $n=16$). This corresponds roughly to an autocorrelation function whose width is half a time bin. With this degree of temporal correlation only a few (*i.e.*, five) discrete observations are specified with any precision.

hyperparameter γ could be estimated along with the other hyperparameters. We will discuss this elsewhere. Here, for simplicity, we will assume that 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 in generalised coordinates to a relatively low order. This is because high-order prediction errors have a vanishingly small precision. In the next section, we will see that an embedding order[13] of $n=6$ is sufficient for most systems (*i.e.*, a representation of high-order derivatives up to sixth order). In fact, one can truncate the representation of the causes even further (*e.g.*, second-order; $d=2$) without loosing accuracy. This can be thought of as a further approximation, in addition to the Laplace and mean-field approximations, under which high-order derivatives of $\tilde{v}$ have a point mass at zero. For example, when $d=2$, the variational density $q(u,t)=q(v,v'v'',x,x',x''\ldots,t)$ provides an analytic approximation to the time-varying conditional density that is differentiable to first order. However, this level of approximation cannot be applied to the hidden states, because they can exhibit motion to arbitrarily high order (see Eq. (33)).

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. This is fine for analogue data (*e.g.*, in electrical or biophysical systems); however, empirical data are often measured discretely, as a sequence, $y=[y(1),\ldots,y(N)]^T$. 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 $g=[g(1),\ldots,g(N)]^T$ can be generated from the derivatives $\tilde{g}(t)$ using Taylor's theorem

$$g=\tilde{E}(t)\tilde{g}(t)\quad \tilde{E}(t)=E\otimes I\quad E_{ij}=\frac{(i-t)^{(j-1)}}{(j-1)!} \tag{54}$$

Similarly for a sequence of innovations $z=\tilde{E}(t)\tilde{z}(t)$ with covariance $\tilde{E}^T\tilde{\Sigma}^z\tilde{E}^T$ and precision $T^T\tilde{\Pi}^z T$, where $T=\tilde{E}^{-1}$. Under discrete sampling, the internal energy becomes

$$\begin{aligned} U(t) &= \frac{1}{2}\ \ln\ |T^T\tilde{\Pi}^z T| - \frac{1}{2}(y-g)^T T^T\tilde{\Pi}^z T(y-g)+\ldots \\ &= \frac{1}{2}\ \ln\ |\tilde{\Pi}^z| - \frac{1}{2}\tilde{\varepsilon}^{v^T}\tilde{\Pi}^z\tilde{\varepsilon}^v+\ldots \end{aligned} \tag{55}$$

This energy function is exactly the same as Eq. (33)[14], provided $\tilde{y}(t)=T(t)y$. This is assured if $\tilde{E}(t)$ is invertible; *i.e.*, the number of elements in the generalised response is equal to the length of the sequence; $n+1=N$. In short, discrete sequences are treated in exactly the same way as generalised responses, by substituting $\tilde{y}(t)=T(t)y$.

It is interesting to consider the precision, $T(t)^T\tilde{\Pi}^z T(t)$ whose non-stationary form renders precision local in time. In other words, a dynamic model, formulated in generalised coordinates of motion, specifies the time-series in a local sense. A typical example is shown in Fig. 3. This means that the D-step needs only local sequences and can operate 'on-line'. More formally, the precision of measurements in the past or future falls quickly and they do not contribute to the variational energy or its optimisation. This means $\tilde{y}(t)=T(0)\left[y\left(t-\frac{n}{2}\right);\ldots;y\left(t+\frac{n}{2}\right)\right]$ can be evaluated using the $n+1=N$ samples closest to the current time bin.

[13] We use 'embedding order' by analogy with attractor reconstruction and lags in autoregressive modelling.

[14] To within an additive constant.

Summary

At this point, the reader must be getting a bit overwhelmed with equations. However, this section has shown how the variational principles of the previous section can be applied to a specific model with fairly simple linear algebra. Critically, this model is about as complicated as one could imagine; it comprises causes and hidden states, whose dynamics can be coupled with arbitrary (analytic) nonlinear functions. Furthermore, these states can have random fluctuations with unknown amplitude and arbitrary (analytic) autocorrelation functions. This means one can invert nearly any model, given just its likelihood function and priors. There is no need to derive bespoke update rules or use conjugate priors, all that the scheme requires are the functions $f^{(i)}$ and $g^{(i)}$. These functions are then differentiated numerically or analytically to give all the quantities needed for inversion. A key aspect of the model considered in this section is its hierarchical form, which induces empirical priors on the causes. These recapitulate the constraints on hidden states, furnished by the hierarchy implicit in generalised motion. This concludes the theoretical background. In the next section, we examine the operational features of this inversion scheme.

Variational inversion of dynamic models

In the remaining sections, we focus on the implementation of DEM, its functionality and how it compares with established schemes. This functionality is quite broad because the conditional density covers not only hidden and causal states but also the parameters, and hyperparameters, mediating interactions among states. This means it is a multiple estimation or inference scheme of the sort used in blind deconvolution. Deconvolution schemes, such as Bayesian filtering, infer only hidden states, assuming that the parameters and covariances are known. Other schemes, employed in system identification, estimate model parameters through generalised kernels or transfer functions by treating the inputs and outputs as known (to first or second-order statistics). Blind deconvolution entails inference on both states and parameters. These schemes solve a dual-estimation problem; for example, Valpola and Karhunen (2002) describe a detailed and general approach to ensemble learning for nonlinear state-space models. Nonlinear mappings in this model are represented using multilayer perceptron networks. Honkela et al. (2006) extend this approach to cover continuous-time formulations, using a variational approach with similarities to the approach advocated here. See also Wang and Titterington (2004) for a careful analysis of variational Bayes for continuous linear dynamical systems and Sørensen (2004) for a review of the statistical literature on continuous nonlinear dynamical systems. In general, these treatments belong to the conventional class of schemes that assume Wiener or diffusion processes for state noise and, unlike DEM, do not consider generalised motion.

DEM treats all model quantities as unknown. Knowledge about any set of parameters ϑ^i is implemented with precise priors so that deconvolution and system identification proceed using exactly the same algorithm and code (this also applies to the identification of static models with DEM). Strictly speaking, DEM solves triple inference problems because it covers states, parameters and hyperparameters.

In this section, we focus on the Bayesian deconvolution of dynamic systems to estimate hidden and causal states, assuming that the parameters and hyperparameters are known. We start with a simple linear dynamic model to outline the basic nature of variational inversion and then move on to nonlinear dynamic models that have been used previously for comparative studies of extended Kalman and particle filtering. We then turn to autonomous nonlinear systems and show how DEM can be used for attractor reconstruction, even with chaotic systems. In the next section, we focus on dual and triple estimations, when the parameters and hyperparameters are unknown. These scenarios are not covered by Bayesian filtering but, if we assume that the causes are known, we can compare dual estimation of the parameters and hyperparameters with the identification of deterministic dynamic models using expectation maximisation (EM). To conclude, we

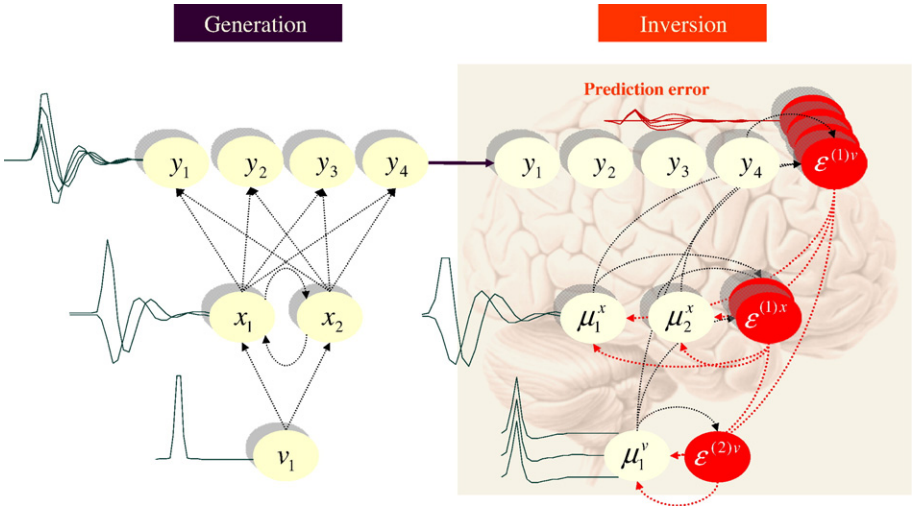

Fig. 4. This is a schematic showing the linear convolution model used in the demonstrations of DEM in subsequent figures. This summary corresponds to a directed Bayesian graph, in which the nodes are connected by arrows, depicting conditional dependencies. In this model, a simple Gaussian 'bump' function acts as a cause to perturb dynamics among two coupled hidden states. These dynamics are then projected to four response variables, whose time courses are cartooned on the left. This figure also summarises the architecture of the implicit inversion scheme, in which prediction errors drive the conditional modes to optimise variational action. Critically, the prediction errors propagate their effects up the hierarchy (*c.f.*, Bayesian belief propagation or message passing), whereas the predictions are passed down the hierarchy. This sort of scheme can be implemented easily in neural networks and may be used by the brain (see Friston et al., 2006 for a neurobiological treatment of this architecture).

consider the triple estimation of states, parameters and hyperparameters using the simple convolution model of this section. In the final section, we illustrate triple estimation with a nonlinear convolution model, in an empirical setting, using a hemodynamic model of brain responses evoked by attention to visual motion. These examples were chosen to show that DEM outperforms conventional approaches, when they exist. We will try to explain why it is superior and disclose key operational aspects of DEM.

A linear convolution model

In the examples below we will use the same model in a number of different contexts. This linear convolution is summarised in Fig. 4 and can be written as

$$\begin{aligned} y &= g(x,v) + z^{(1)} \\ \dot{x} &= f(x,v) + w^{(1)} \\ v &= \eta^v + z^{(2)} \end{aligned} \tag{56}$$

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

We have omitted superscripts on the states because the models considered here are single-level models. In this model, after causes or inputs arrive, the ensuing perturbation decays exponentially to produce an output that is a linear mixture of hidden states. Our example uses a single input, conforming to a Gaussian bump function, two hidden states and four outputs. This is a single input-multiple output linear system, where

$$\theta_1 = \begin{bmatrix} 0.1250 & 0.1633 \\ 0.1250 & 0.0676 \\ 0.1250 & -0.0676 \\ 0.1250 & -0.1633 \end{bmatrix} \quad \theta_2 = \begin{bmatrix} -0.25 & 1.00 \\ -0.50 & -0.25 \end{bmatrix} \quad \theta_3 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{57}$$

These parameters were 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). The model can be specified in terms of the likelihood functions and priors at each level[15]

Linear convolution model

Level	$g(x,v)$	$f(x,v)$	Π^z	Π^w	η^v
$m=1$	$\theta_1 x$	$\theta_2 x+\theta_3 x$	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 inverting the model the cause is unknown and was subject to mildly informative shrinkage priors with zero mean and unit precision.

Unless stated otherwise, we will use embedding orders of $n=6$ and $d=2$, with temporal hyperparameters, $\gamma=4$ for all our simulations. We will usually generate data over 32 time bins, using innovations sampled from Gaussian densities. Note there are no priors on the parameters or hyperparameters because we treat them as known for the present. This model specification enables us to evaluate the variational energy at any point in time and invert the model, given any response data.

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

Deconvolution: inference on states

We start with a validation of DEM using variational filtering, based on the ensemble dynamics of the first section. We then examine how the accuracy of DEM changes with embedding order and conclude with an evaluation of DEM in relation to established Bayesian filtering techniques.

Variational filtering and DEM

Recall that DEM approximates the density of an ensemble of solutions by assuming that it has a Gaussian form; this reduces the

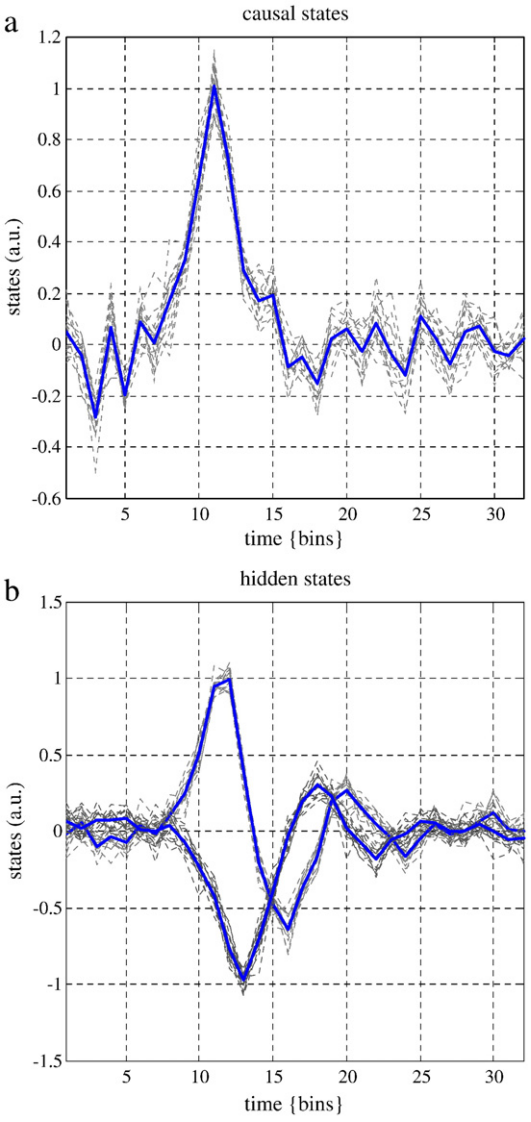

Fig. 5. Variational densities on the causal and hidden states of the linear convolution model of the previous figure. These show the trajectories or paths of sixteen particles tracking the mode of the single cause (top) and two hidden states (bottom). The sample mean of this distribution is shown in blue over the 32 time bins, during which responses or data were inverted.

problem to finding the path of the mode. Variational filtering relaxes this fixed-form assumption and integrates the paths of an ensemble to furnish an approximating sample density. We can compare the fixed-form density provided by DEM with the sample density from variational filtering to ensure that the Gaussian assumption is appropriate. Generally, this is non-trivial because non-linearities in the likelihood model render the true conditional non-Gaussian, even under Gaussian assumptions about the priors and innovations. In our case, in generalised coordinates and with a linear convolution model, the Gaussian form is exact and we would expect a close correspondence between DEM and variational filtering.

Given an observed response and model specification above, we can evaluate the variational energy $V(t)=-\tilde{\varepsilon}^T\tilde{\Pi}\tilde{\varepsilon}+W(t)^\theta+W(t)^\lambda$ (see Eq. (38)) at any point in time and perform variational filtering by integrating several particles according to $\dot{u}=V(u,t)_u+D\mu+\Gamma(t)$ (see Eq. (17)). The details of this integration and the ramifications of variational filtering will be dealt with in a separate paper

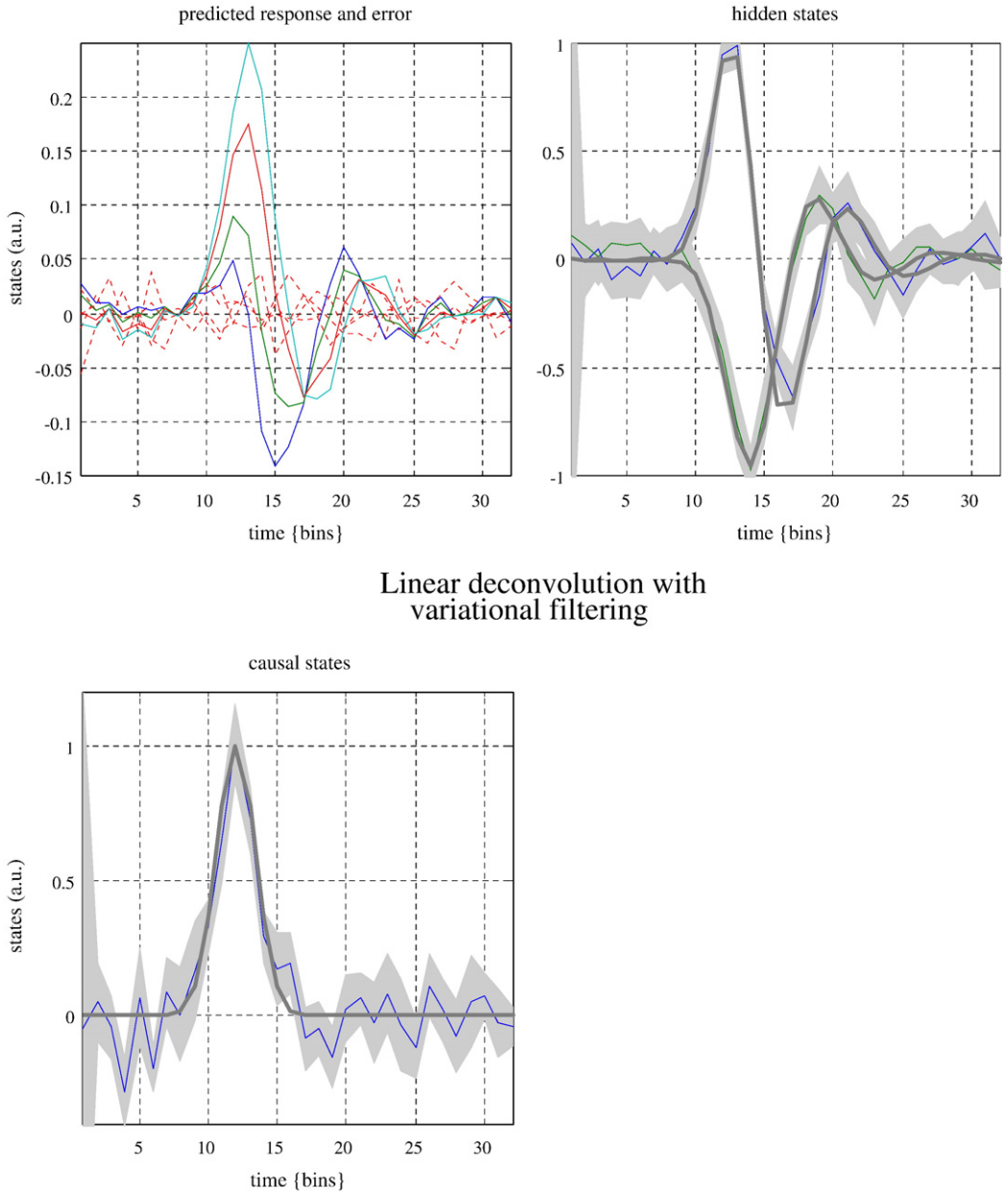

Fig. 6. Alternative representation of the sample density shown in the previous figure. This format will be used in subsequent figures and summarises the predictions and conditional densities on the states of a hierarchical dynamic model. Each row corresponds to a level, with causes on the left and hidden 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 the 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. These are sometimes referred to as "tubes". 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.

(Friston, 2008). Here, we use variational filtering just for cross-validation. Fig. 5 shows the trajectories or paths of sixteen particles tracking the mode of the single 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. 6. This format will be used in subsequent figures and summarises 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 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. 5. It can be seen that there is a pleasing correspondence between the sample mean (blue) and veridical states (grey). Furthermore, the true values lie largely within the 90% confidence intervals.

We then repeated the inversion using exactly the same model and response variable using DEM. The results are shown in Fig. 7

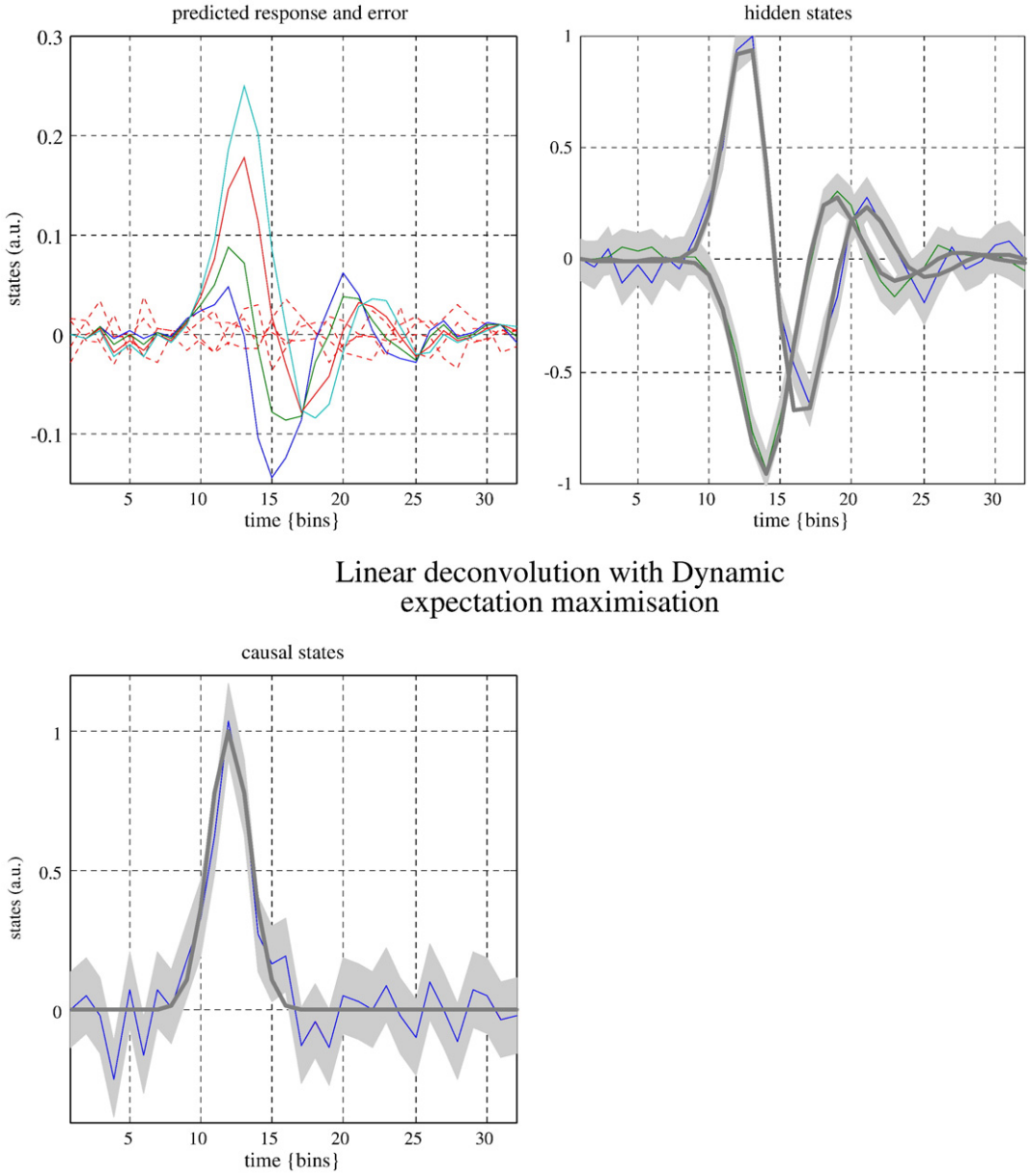

Fig. 7. This is exactly the same as the previous figure, summarising conditional inference on the states of the linear convolution model of Fig. 4. The only difference is that here we have used a Laplace approximation to the variational density and have integrated a single trajectory; that of the conditional 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 and variations in the state have the same effect in measurement or observation space, at all points in time. 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.

using the same format as the previous figure. Critically, the ensuing modes (blue) are indistinguishable from those obtained with variational filtering (*c.f.*, Fig. 6). The conditional variance on the causal and hidden states is 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 the integration of stochastic differential equations.

Embedding orders

Next, we examine the dependency of inversion accuracy on the embedding order using the same linear convolution model and response above. Fig. 8 shows a numerical analysis of accuracy as a function of the number of generalised coordinates of motion; $n=1, \ldots,13$. The upper panel shows accuracy in terms of the sum of squared error on the expected causes, relative to their true values. The embedding order of these causes was fixed at $d=2$. This means that the increase in accuracy with n is mediated by a more complete representation of the motion of the hidden states. It can be seen that the accuracy increases as the order increases up until about six. The lower panels show the conditional expectation (black line) and the true cause (grey line) for $n=1$ (left) and $n=7$ (right). A useful heuristic for these results is obtained from Fig. 3, which shows that the precisions of the sixth and higher derivatives are essentially zero.

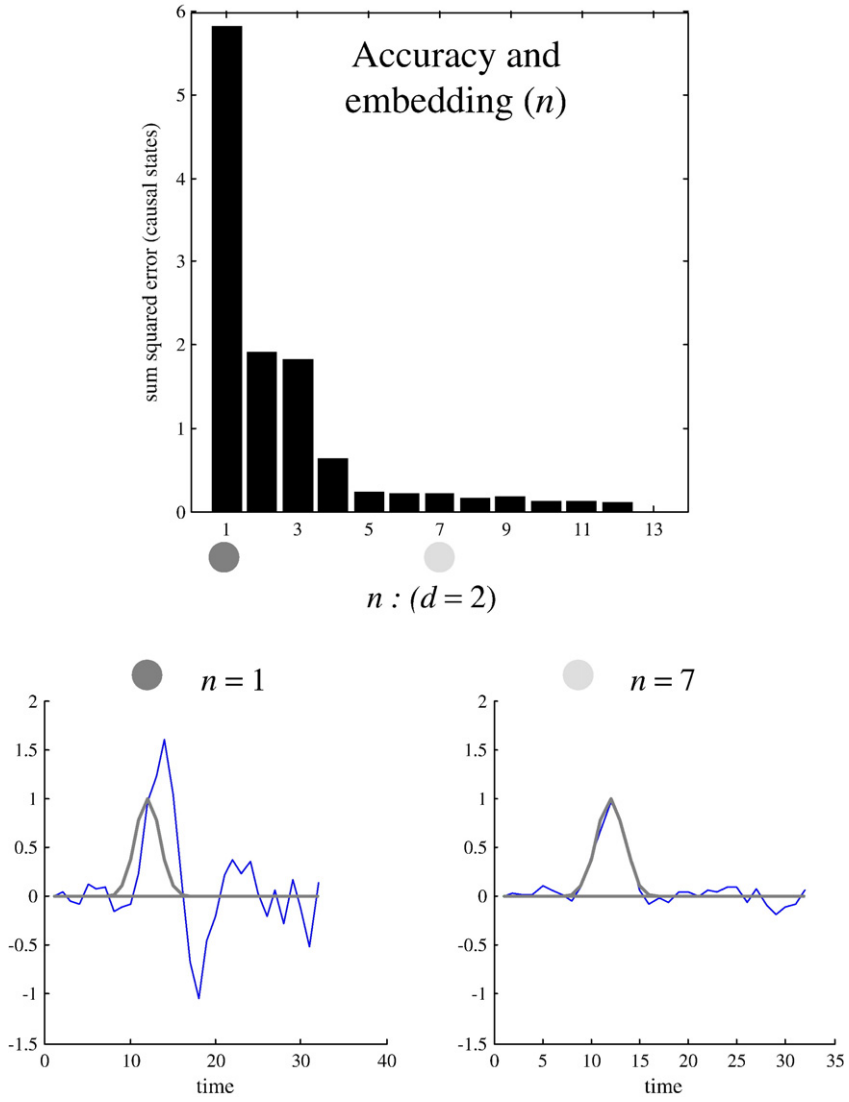

Fig. 8. A numerical analysis of accuracy as a function of the embedding dimension or number of generalised coordinates of motion ($n=1,\ldots,13$). The upper panel shows accuracy in terms of the sum of squared difference between the predicted and true causes at the second level. Note that the embedding dimension of these causes was fixed at, $d=2$. This means that increases in accuracy are mediated solely by a more accurate representation of the motion of the hidden states. These results were obtained with a roughness of $\gamma=4$ using the linear convolution model of the previous figures. The lower panels show the conditional expectation of the cause (black line) and the true values (grey line) for $n=1$ (left) and $n=7$ (right).

This means that they do not contribute in any substantial way to the free-energy and can be discounted with impunity.

Fig. 9 shows the effects of changing the embedding dimension for the causes; $d=0,\ldots,5$ on the sum of squared error for the hidden states. These results were obtained using an embedding dimension of $n=15$ for the hidden states and response. As in the previous figure, the embedding of the hidden states did not change but they are estimated more accurately if the embedding dimension of the causes is increased. In this example, an embedding dimension of $d=3$ is sufficient to maximise accuracy. In the lower panels the true (grey) and predicted (black) hidden states are shown for $d=0$ (left) and $d=3$ (right). These correspond to modelling just the amplitude of the cause and its generalised motion to third order respectively. In this case, we can also evaluate the free-energy bound on the log-evidence (top-right panel), which reflects both accuracy and complexity. We can do this because changing the embedding order of the causes does not affect the embedding of the responses. This means that we can compare $F=\ln p(\tilde{y}|d)$ in a meaningful way.

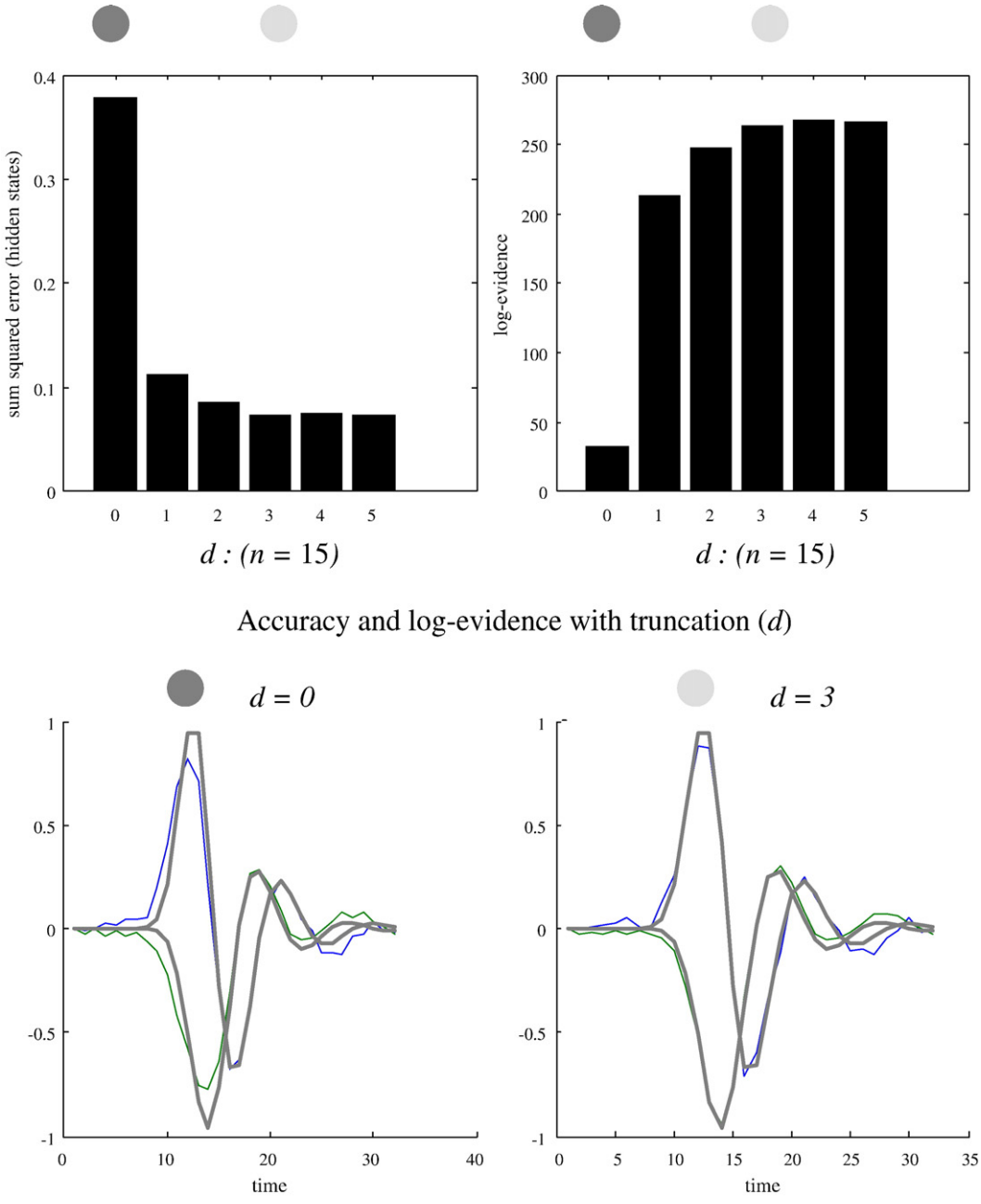

Fig. 9. As for the previous figure but here detailing the effects of changing the embedding dimension for the causes ($d=0,\ldots,5$) on the sum of squared error for the hidden states (top left) and log-evidence (top right). These results were obtained for the linear convolution model of the previous figures, using an embedding dimension of $n=15$ for the hidden states and data. As in the previous figure, the embedding of the hidden states did not change but they are estimated more accurately, if the embedding dimension of the causes is increased. In this example, an embedding dimension of $d=3$ is sufficient to maximise accuracy. The lower panels show the true (grey) and predicted (black) hidden states for $d=0$ (left) and $d=3$ (right).

Bayesian filtering and DEM

In this subsection, we compare DEM with Bayesian filtering, both conceptually and quantitatively. We start with Kalman filtering, which is appropriate for our linear convolution model. We then turn to nonlinear convolution models, which usually call for extended Kalman filtering or, in the case of highly non-Gaussian conditional densities, particle filtering. To place these comparative analyses in context, we first review extended Kalman and particle filtering. See Arulampalam et al. (2002) for a useful introduction to Kalman and particle filters for online Bayesian tracking. Our implementations follow var der Merwe et al. (2000) (see Appendices C and D).

Kalman filtering

Bayesian inversion in the D-step is related to Bayesian belief update procedures (*i.e.*, incremental or recursive Bayesian filters). The conventional approach to online Bayesian tracking of states in nonlinear or non-Gaussian systems employs extended Kalman filtering or sequential Monte-Carlo methods such as particle filtering. These Bayesian filters approximate the conditional densities of hidden states in a recursive and computationally

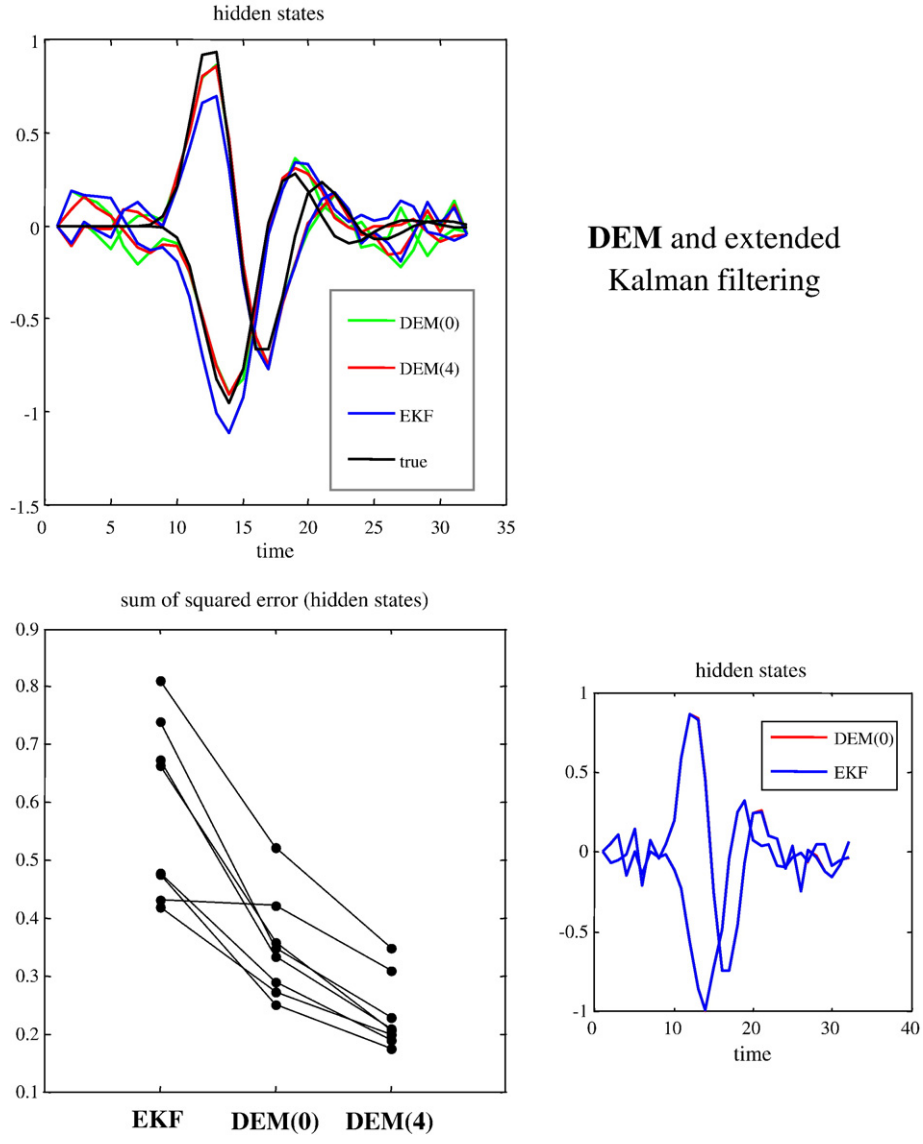

Fig. 10. A comparative evaluation of DEM and extended Kalman filtering. These results summarise the conditional predictions of the hidden states after deconvolution with DEM using two levels of smoothness $1/\gamma=0$ and $\gamma=4$; corresponding to DEM(0) and DEM(4). The top panel shows the results of a single realisation for both DEM deconvolutions, extended Kalman filtering and the true values. It is immediately apparent that both DEM schemes provide more accurate predictions than the extended Kalman filter. Furthermore, the DEM with smoothness constraints afforded a smoother prediction and is more in line with the true values. This is because the true response was generated with relatively smooth innovations; $\gamma=4$. The lower left panel shows the results of repeating these realisations eight times. The sum of squared error (on the hidden states) over realisations is shown for the extended Kalman filter and both DEM schemes. Note the marked improvement of DEM over Kalman filtering and the further improvement that obtains when temporal correlations are properly accommodated. We can discount the benefits of DEM by removing causes from the model and inverting it under zero smoothness (infinite roughness). In this instance, the results are indistinguishable from an extended Kalman filter; both being optimal under the linear model used to generate the data (lower right insert).

expedient fashion, assuming that the parameters and hyperparameters of the system are known. These schemes deal with systems of the form

$$\begin{aligned} y &= g(x) + z \\ \dot{x} &= f(x, v) + w \end{aligned} \tag{58}$$

This is a simple one-level HDM in which exogenous causes or inputs can enter nonlinearly at the level of the hidden states. The extended Kalman filter is a generalisation of the Kalman filter, in which the operators of a linear state equation are replaced by the partial derivatives of $f(x,v)$ with respect to the states (see Appendix C). Kalman filtering proceeds recursively in two steps; prediction and update. The prediction uses the Chapman–Kolmogorov equation to compute the density of the hidden states conditioned on the response up to, but not including, the current observation $y_{\to t-1}$.

$$p(x_t|y_{\to t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{\to t-1})dx_{t-1} \tag{59}$$

This conditional density is then treated as a prior on the next observation and Bayes rule is used to compute the conditional density of the states, conditioned upon all observations, $y_{\to t}$. This gives the Bayesian update

$$q(x_t) = p(x_t|y_{\to t}) \propto p(y_t|x_t)p(x_t|y_{\to t-1}) \tag{60}$$

Critically, the conditional density covers only the hidden states. This is important because it precludes inference on causes and the ability to recover inputs from outputs. This is a key limitation of Bayesian filtering, in relation to DEM. The second key difference is that Bayesian filtering assumes that state or process noise is an uncorrelated Wiener process. In contrast, DEM handles smooth fluctuations gracefully because it represents the states in generalised coordinates. The impact of these differences is easy to demonstrate using the linear convolution model above.

Fig. 10 shows a comparative evaluation of DEM and Kalman filtering[16]. These results summarise the conditional predictions of the hidden states for deconvolution with DEM using two levels of temporal correlations $1/\gamma = 0$ and $\gamma = 4$; corresponding to DEM(0) and DEM(4). The first corresponds to the assumption of uncorrelated (infinitely rough) innovations and the second to the true correlations. The top panel shows the results of a single deconvolution with DEM and Kalman filtering and the true values. It is apparent that both DEM schemes provide more accurate predictions than the extended Kalman filter. Furthermore, DEM with appropriate temporal constraints furnishes a smoother prediction that is closer to veridical values, whereas the Kalman filter over-fits the hidden states and provides a sub-optimal solution. This is because the true response was generated with relatively smooth innovations; $\gamma = 4$. The lower panel shows the results of repeating these realisations eight times. The sum of squared error (on the hidden states) over realisations is shown for the extended Kalman filter and both DEM schemes. Note the marked improvement of DEM over Kalman filtering and the further improvement that obtains when temporal correlations are included. These improvements reflect the fact that DEM represents the causes and the form of their influence on the hidden states, whereas the Kalman filter does not. The subsequent improvement, when smoothness is introduced, reflects the fact that this is a better model for the smooth innovations used to generate the data.

Despite these results, Kalman filtering provides the optimum solution, in a maximum likelihood sense, when the assumptions of the underlying model hold and one is not interested in causes or inputs. If this is the case, extended Kalman filtering and DEM should give the same results for the linear dynamic system considered above, provided we remove the empirical priors on the states implicit in their causes. The inset in Fig. 10 shows that they do; the conditional expectations of the hidden states from both schemes are indistinguishable. This convergence rests on making the models used by DEM and extended Kalman filtering the same, by removing serial correlations (*i.e.*, making $\gamma = \infty$) and the causes by using the model below for inversion

Linear state-space model

Level	$g(x,v)$	$f(x,v)$	Π^z	Π^w
m=1	$\theta_1 x$	$\theta_2 x$	e^8	e^0

These examples illustrate the relationship between DEM and extended Kalman filtering. It is not appropriate to think of Kalman filtering and related approaches as special cases of DEM because they have distinct derivations. DEM does not derive priors on the present from the past, nor does it require a backwards pass as in Bayesian smoothing. However, DEM produces the same results as Bayesian filtering, in the special cases that Bayesian filtering is exact (*i.e.*, works). We now consider cases in which extended Kalman filtering does not work.

Particle filtering

Prediction and update proceed under Gaussian assumptions in both Kalman filtering and extended variational filtering. This is fine for linear systems. However, in nonlinear systems the extended Kalman filter may fail to represent non-Gaussian (*e.g.*, multimodal) conditional densities required for accurate recursive filtering. In this instance, particle filtering and related grid-based approximations provide solutions that allow for non-Gaussian posteriors on the hidden states. This is usually achieved by point-mass (or particle) representations of the ensemble density as in variational filtering. These 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[17].

These sequential Monte-Carlo techniques should not be confused with the ensemble dynamics of variational filtering. In variational filtering the 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 comparison with variational filtering, sequential sampling techniques appear unnecessarily complicated. Furthermore, they rely on some rather *ad hoc* devices to make them work (see Appendix D and var der Merwe, 2000). For these reasons, we will not provide any further background on them but focus on why they are used: like variational filtering, particle filtering allows a free-form approximation to the conditional density. This can be particularly useful in nonlinear systems that typically have non-Gaussian posteriors.

[16] Under this linear model, extended Kalman filtering reduces to Kalman filtering.

[17] In Bayesian filtering there is only one set of unknown parameters (*i.e.*, the hidden states) and the variational energy reduces to the internal energy.

A nonlinear convolution model

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

Nonlinear convolution model

Level	$g(x,v)$	$f(x,v)$	Π^z	Π^w	η^v
$m=1$	$\frac{1}{5}x^2$	$e^v - x \ln 2$	e^4	e^{16}	
$m=2$			2		$\frac{1}{2} + \sin\left(\frac{1}{16}\pi t\right)$

This system comprises a slow sinusoidal input or cause that excites increases in a single hidden state. The response is a quadratic function of the hidden states. Similar (double-well) systems have often been used to study bimodal posteriors, because the quadric observer induces ambiguity about the sign of the hidden state. However, in this model, the state is always positive because the cause enters through the exponential $e^v > 0$. Here our focus is on the effects of the nonlinear observer: when the input is negative and $x \to 0$ decays to small values, the conditional density becomes very broad. This is because variations in x produce very small variations in the response. Conversely, when x is large, its conditional variance is small. These nonlinear effects can confound extended Kalman filtering because it uses local linearity assumptions to update its predictions. DEM can finesse these effects by integrating

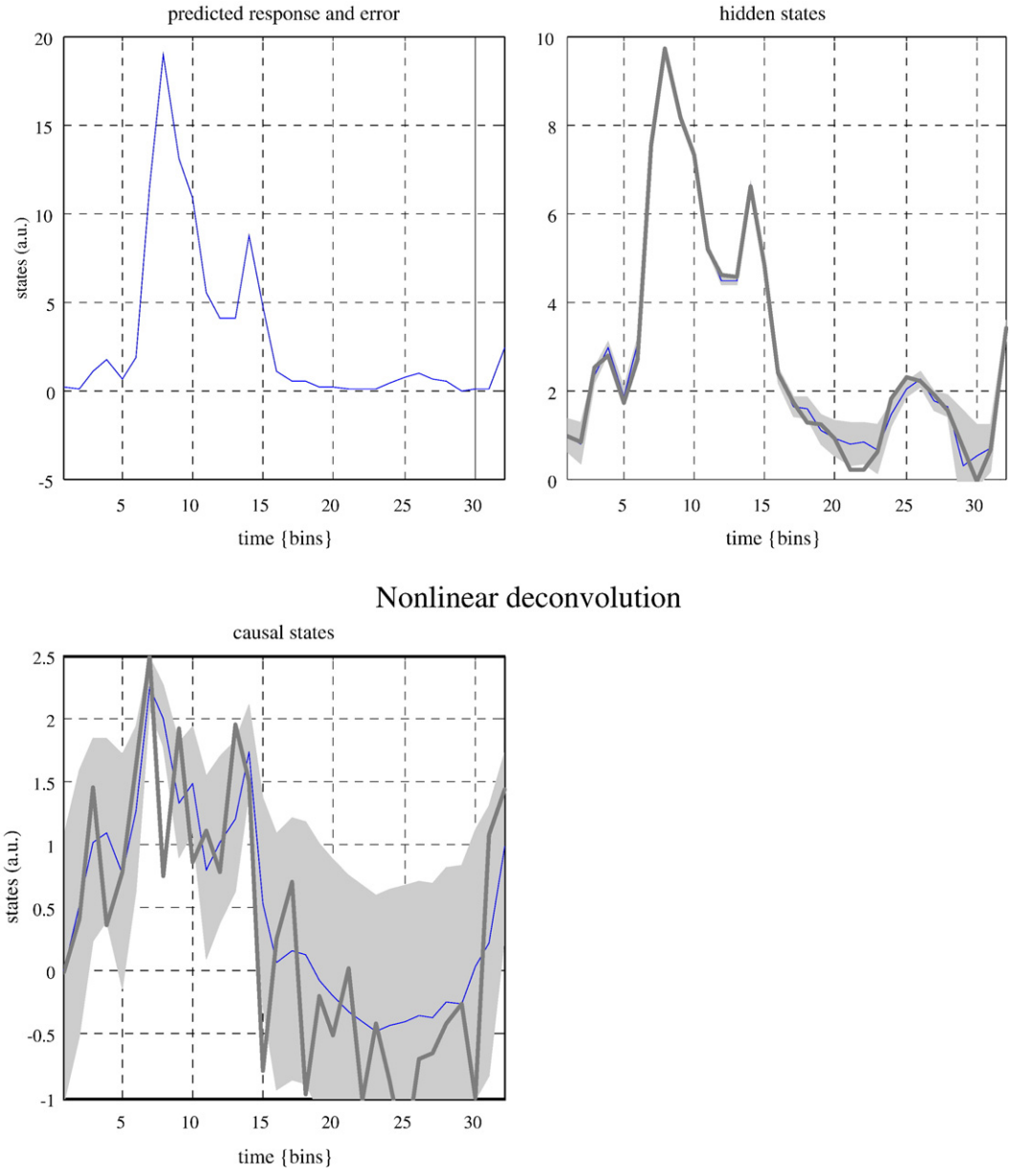

Fig. 11. An example of deconvolution using the nonlinear model described in the main text. In this case, the response is always positive. As in previous figures, the blue lines represent the conditional estimate of hidden and causal states, while the thick grey lines depict the true values. Note that in all cases the true response lies within the 90% confidence intervals (grey area). The key thing to observe here is that when the hidden and causal states approach zero the conditional uncertainty increases markedly. This is because variations in the states are expressed with smaller amplitude in observation space; this is a direct reflection of the nonlinear form of this model (*c.f.*, the fixed-width confidence intervals under linear deconvolution in Fig. 7).

the conditional trajectory using much smaller time steps. DEM can do this because its generative model generates paths, which can be sampled at arbitrary times by the observation process. In contrast, extended Kalman filtering is based on a model that generates sparse data sequences at fixed sampling intervals.

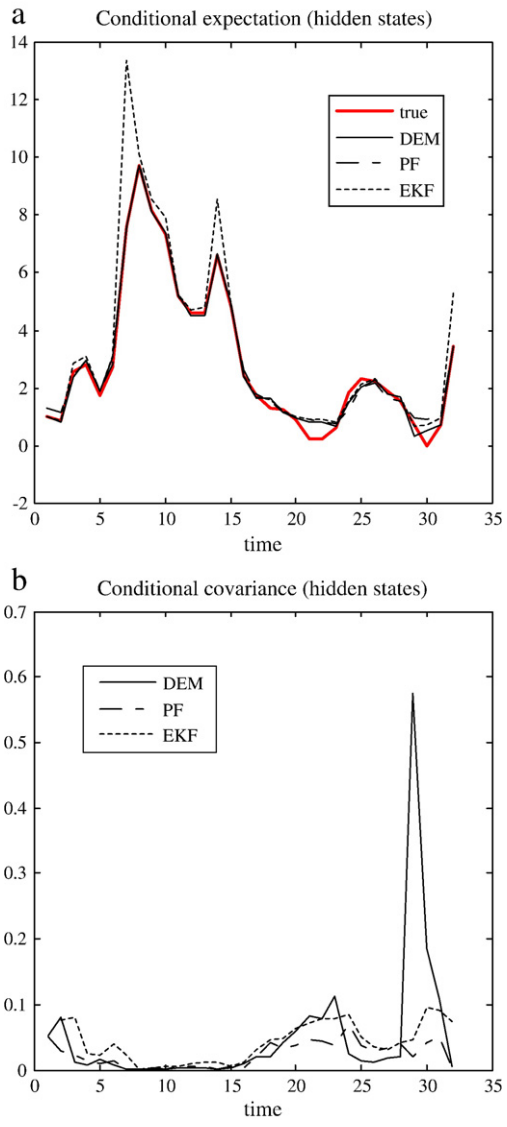

Fig. 12. This reproduces the results of the previous figure, for the conditional density of the hidden states, in terms of the time-dependent mode (upper panel) and conditional variance (lower panel). Here, we provide a comparative evaluation with extended Kalman and particle filtering. It can be seen that all three techniques properly show a non-stationary conditional covariance. The key differences among the three schemes are expressed in the conditional modes. It can be seen that particle filtering and DEM deliver almost indistinguishable estimates and that these are veridical; with maximum departures from the true values when the states are small. Critically, the extended Kalman filter is unable to deal with the nonlinearities in this model and overestimates the values of the hidden states when they are large. This is because the implicit integration between one observation and the next does not have access to generalised motion (DEM) or the free-form density associated with particle filtering.

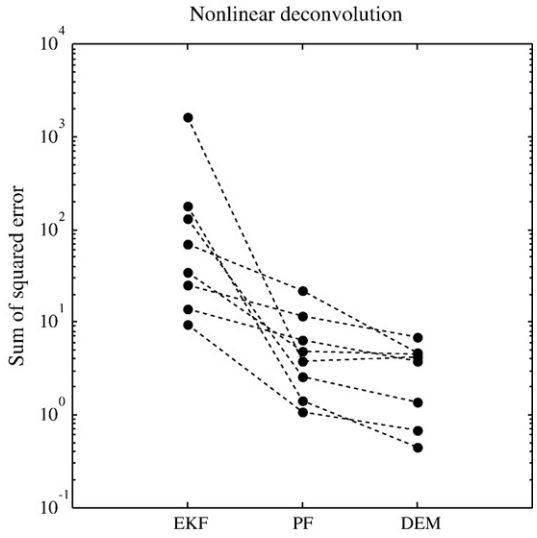

Fig. 13. Sum of squared errors following nonlinear deconvolution of the model considered in the previous figure. The accuracy here was assessed in terms of sum of squared error between the true and predicted hidden states, for three schemes (extended Kalman filter-EKF, particle filtering-PF and dynamic expectation maximisation-DEM). These results show that the relative failure of extended Kalman filtering is seen over multiple realisations. The performance of DEM was slightly better than that of particle filtering. Both were very substantially better, on average, than extended Kalman filtering; note that the sum of squared error is plotted on a log scale.

Comparative evaluations

We generated a 32 time-bin response, using the nonlinear convolution model above and inverted it using DEM. For these simulations we used nearly independent innovations; $\gamma = 1024$ and four time bins per sample; *i.e.*, $\Delta t = \frac{1}{4}$. The results are shown in Fig. 11. As in previous figures, the blue lines represent the conditional estimate of hidden and causal states, while the grey lines depict the true values. In all cases the true values lie within the 90% confidence intervals (grey area). The key thing to observe here is that when the hidden and causal states approach zero, the conditional uncertainty increases markedly (*c.f.*, the fixed-width confidence intervals under linear deconvolution in Fig. 7).

We then inverted the same model and response using extended Kalman and particle filtering (as described in Appendices C and D). Fig. 12 shows the ensuing conditional densities of the hidden states, in terms of time-dependent modes (upper panel) and conditional variances (lower panel). It can be seen that all three schemes properly identify a non-stationary conditional covariance, with DEM exhibiting more sensitivity to inflated conditional uncertainty, when the hidden states approach zero. The key differences among the three schemes are evident in the conditional modes. It can be seen that particle filtering and DEM deliver almost indistinguishable estimates and that these are veridical; with maximum departures from the true values when the states are small. Critically, the extended Kalman filter is unable to deal with nonlinearities and overestimates the values of the hidden states, when they are large. This is because the implicit integration between one observation and the next does not have access to generalised motion (DEM) or a free-form density (particle filtering).

We repeated this whole procedure eight times to assess the relative accuracy of the three schemes over multiple realisations. Accuracy was assessed in terms of the sum of squared error between the true and predicted hidden states. The results in Fig. 13 show that the performance of DEM was slightly better than that of particle filtering. Both were substantially better, on average, than extended Kalman filtering. In fact, to plot the differences clearly, we had to use a log scale. In the next section, we turn to systems whose evolution is governed entirely by nonlinear interactions among the hidden states. Here, representing generalised motion becomes critical and in this example both extended Kalman filtering and particle filtering fail completely, in comparison to DEM.

An autonomous nonlinear model

In the final example of inference on states, we consider deconvolving the hidden states from the output of an autonomous, nonlinear dynamical system specified as

Autonomous nonlinear model

Level	$g(x,v)$	$f(x,v)$	Π^z	Π^w
$m=0$	$x_1+x_2+x_3$	$\begin{bmatrix} 18x_2 - 18x_1 \\ 46.92x_1 - 2x_3x_1 - x_2 \\ 2x_1x_2 - 4x_3 \end{bmatrix}$	e^0	e^{16}

This is an instance of the famous Lorenz attractor and exhibits deterministic chaos as the path of the hidden states diverges

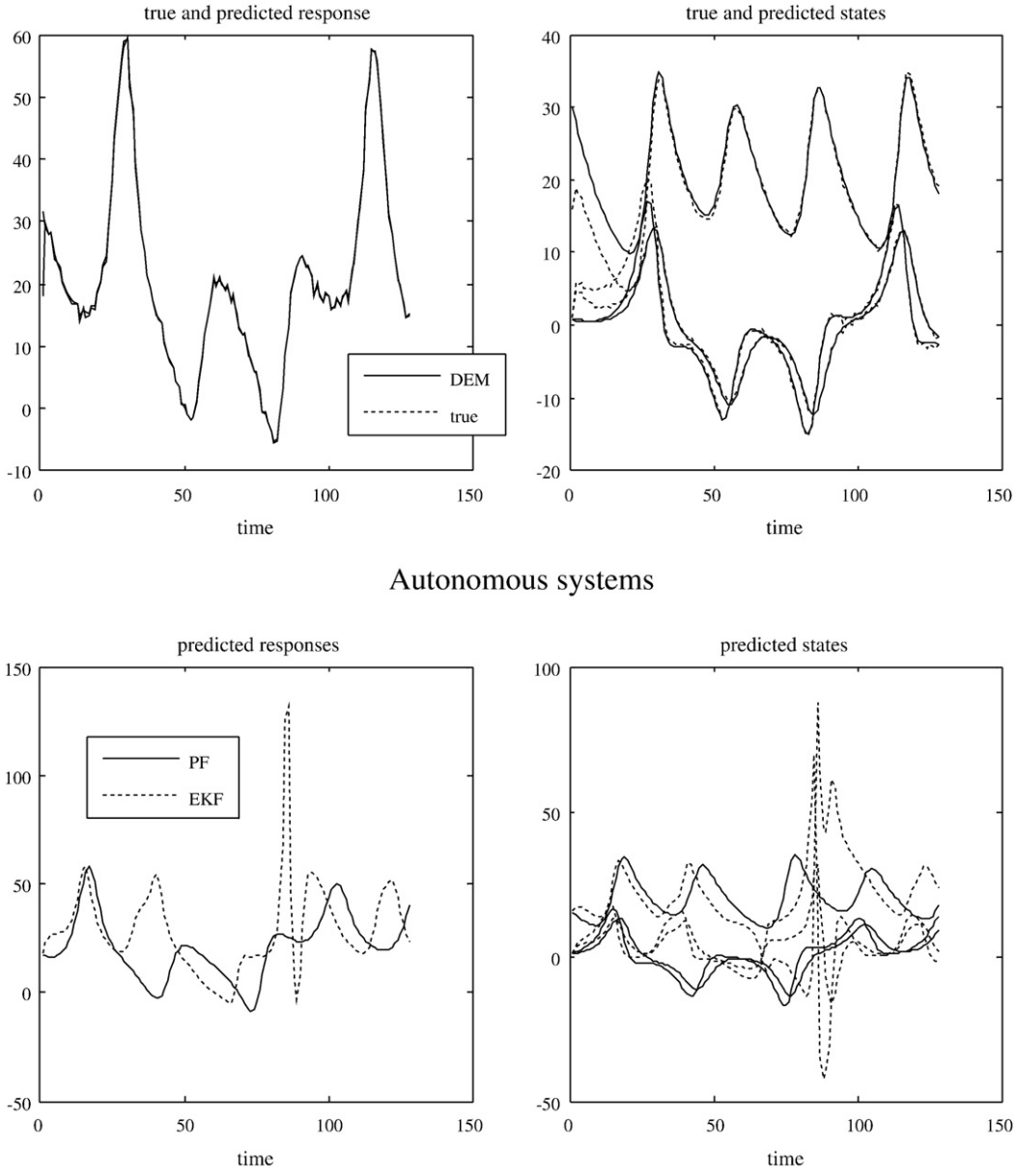

Fig. 14. Deconvolution of an autonomous (Lorenz) system. The upper panels show the true (dashed line) and predicted (solid line) responses (left) and hidden states (right) for a realisation of a system based upon a Lorenz attractor. The DEM scheme used starting values of the hidden states that were different from those used to generate the data. Despite this, after a few time steps, the estimated trajectory converges to the true trajectory. In contrast, when the starting conditions are randomized for the equivalent deconvolution with both particle filtering and extended Kalman filtering (bottom panels) there is a complete failure to track the trajectories: although particle filtering (solid line) appears to 'hang-on' longer than extended Kalman filtering (dashed line).

exponentially on a butterfly-shaped manifold, embedded in three dimensional state-space. There are no inputs in this system; the dynamics are autonomous, being generated by nonlinear interactions among the states and their motion. In this example, the outputs are simply the sum over the states at any point in time, plus an innovation with unit precision and $\gamma=64$. We specified a small amount of noise on the states but this was negligible in relation to the flow induced by the state equation.

We generated 128 samples from this model using $\Delta t = \frac{1}{32}$ and initial conditions $x=[0.9,0.8,30]^T$. We then tried to recover the hidden states from the univariate response using DEM, extended Kalman and particle filtering. Critically, we used an initial condition of $x=[1,1,16]^T$ that differed from the true starting value. Chaotic systems of this sort show a sensitivity to initial conditions, which presents an interesting challenge for inversion schemes, when the initial conditions are unknown. The upper panels of Fig. 14 show the true (dashed line) and predicted (solid line) responses (left) and hidden states (right) using DEM. After a few time steps, the estimated trajectory converges to the true trajectory. In contrast, both particle filtering and extended Kalman filtering

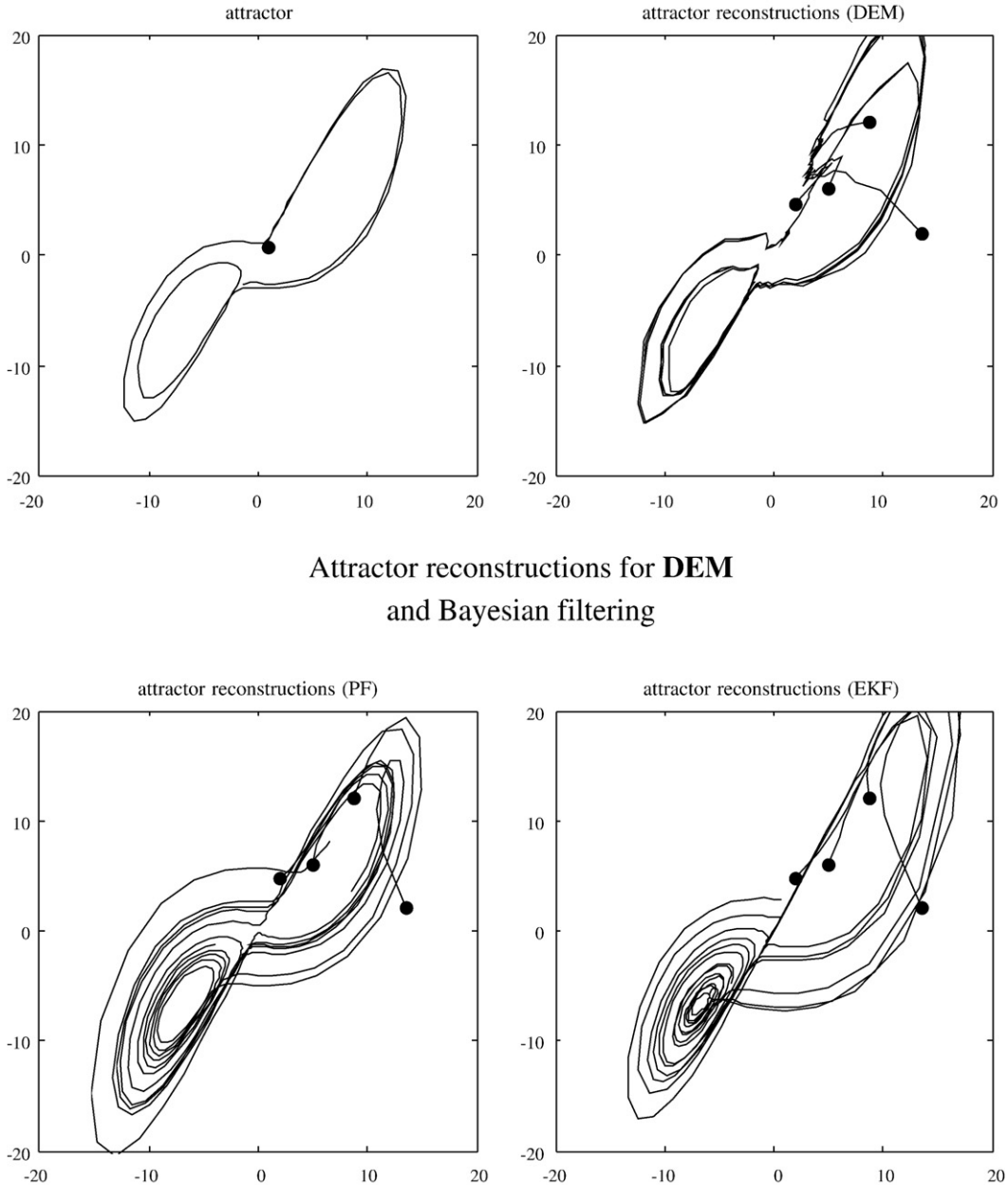

Fig. 15. The simulations of the previous figure were repeated four times, using different starting conditions and different random innovations. The ensuing responses were deconvolved using DEM, particle and extended Kalman filtering. We summarised the resulting trajectories in terms of the first two hidden states and plotted the trajectories against each other in their corresponding state-space. The true trajectories are shown on the upper left and the corresponding predictions from DEM are shown on the upper right. Again, one can see that despite perturbations to the initial conditions, the estimated trajectories converge quickly to the true trajectories. This is not the case for particle filtering or extended Kalman filtering (bottom panels), where the conditional trajectories fail to track the two trajectories and succumb to the systems attractor.

(bottom panels) fail to track the true trajectories; although particle filtering (solid line) appears to 'hang-on' longer than extended Kalman filtering (dashed line).

The simulations in the previous figure were repeated four times using different initial conditions and different innovations. We summarised the resulting trajectories in terms of the first two hidden states and plotted their trajectories against each other in their corresponding state-space. The true trajectories are shown on the upper left and the corresponding predictions from DEM are shown on the upper right (Fig. 15). Again, one can see that, despite perturbations on the initial conditions, the estimated trajectories converge quickly to the true trajectories. This is not the case for particle filtering or extended Kalman filtering (bottom panels), where the conditional trajectories fail to track the two trajectories and succumb to the system's attractor.

It may seem gratuitous to include this example; however, the ability to invert models with autonomous dynamics is relevant when dealing with nonlinear neuronal-mass models that can exhibit bifurcations and deterministic chaos. For example, these models are prevalent in theoretical studies of epilepsy (Breakspear et al., 2006) and may furnish useful forward or generative models of empirical data (*e.g.*, seizure activity in the EEG).

Summary

These examples have shown that DEM provides veridical approximations to the conditional density on the states of dynamic models; even under nonlinearities (*i.e.*, nonlinear convolution models and those showing chaotic dynamics). When models have a simple linear state-space form with uncorrelated innovations, DEM and Kalman filtering give the same results. For nonlinear models, in which extended Kalman filtering fails, DEM gives the same results as particle filtering. DEM can even cope with autonomous systems were both conventional fixed-form and free-form Bayesian filters fail. The principal advantage that DEM has, over conventional schemes, is that it uses conditional densities on hidden states producing responses *and their causes* and both are in *generalised coordinates of motion*.

Beyond deconvolution

This section considers parameter and hyperparameter estimation. We will focus on the linear convolution model of the previous section and ask whether it is possible to estimate the states and parameters (and hyperparameters) simultaneously, knowing only the functional form of the model generating data. There are relatively few examples of inversion schemes that cover three sets of unknowns; a prominent exception is the Variational Kalman Smoother (Ghahramani and Beal, 2001), which uses a mean-field approximation, not dissimilar to DEM. Another important approach is the innovation method (Ozaki and Iino, 2001). These schemes furnish maximum likelihood estimates of the parameters and states using local linearization (LL) methods; LL finesses the integration of stochastic differential equations and is not unrelated to the Kalman filter. Specifically, the LL method has two components: i) local discretisation of continuous stochastic differential equations over time steps (which can be variable) and ii) inference on parameters using the likelihood, which is obtained via the innovations from a Kalman filtering step. Jimenez and Ozaki (2006) extend the original innovation approach of Ozaki for nonlinear cases. A good review of LL-based inference and identification methods, including a comparative study, can be found in Jimenez et al. (2006). Unlike DEM, the Variational Kalman Smoother and innovation methods assume uncorrelated (Wiener process) innovations and do not cover generalised coordinates. Before dealing with triple estimation problems and joint inference on parameters and states, we will compare DEM with the equivalent scheme for parameter estimation in the absence of uncertainty about the states; namely EM.

Dual estimation

In EM there are only two sets of unknown quantities. When identifying a system through its parameters, this means that the causes must be known and the dynamics must be deterministic. In other words, the state noise has to be sufficiently small to enable prediction of the hidden states from the known causes. It is fairly straightforward to estimate the parameters and hyperparameters of deterministic dynamical systems with known inputs. This is because one can treat the time-series as a finite-length data sequence that is completely specified by the inputs and parameters. In Friston (2002) we described such a scheme for deterministic systems that is almost identical to DEM but eschews the D-step (and generalised coordinates).

To compare DEM and EM we generated data as before with a Gaussian bump function input, minimal state noise ($\Pi^{w(1)}=e^{16}$) and moderate levels of observation noise ($\Pi^{z(1)}=e^{8}$). We then inverted the linear convolution model below, using precise and veridical priors on the causes to suppress uncertainty about the states.

Linear convolution model for dual estimation

Level	$g(x,v)$	$f(x,v)$	Π^z	Π^w	η^v	η^θ	P^θ	η^λ	P^λ
$m=1$	$\theta_1 x$	$\theta_2 x+\theta_3 v$	$\exp(\lambda^z)$	$\exp(\lambda^w)$		0	e^{-8}	0	e^{-16}
$m=2$			e^{-16}		$\exp\left(\frac{1}{4}(t-12)^2\right)$				

Note that the model now has priors on the parameters and hyperparameters because these are unknown quantities. The priors on the parameters are uninformative shrinkage priors, with a small precision. The priors on the hyperparameters, sometimes referred to as hyperpriors are similarly uninformative. These Gaussian hyperpriors effectively place lognormal hyperpriors on the precisions of the innovations because the precisions are $\exp(\lambda^z)$ and $\exp(\lambda^w)$. See Appendix A for more details. At the second level, the causes are dominated by the Gaussian prior because the stochastic component has very low variance (*i.e.*, $\Pi^{(2)z}=e^{16}$).

For reasons that will be clear later, we only treated two of the parameters as unknown; one parameter from the observation function (the first) and one from the state equation (coupling the first hidden state to the second). These parameters had true values of 0.125 and −0.5 respectively (Eq. (57)). Fig. 16 summarises the results after convergence of DEM (about sixteen iterations). Note that the causes have very tight confidence tubes because we used very informative priors. Although we used virtually no system noise, its precision was estimated to be about $\Pi^{(1)w}=e^{4}$, which subtends a narrow confidence tube (*c.f.*, Fig. 7).

Comparative evaluation

The same dual estimation was implemented using expectation maximisation (EM), which effectively uses the true states because it used the true cause under deterministic dynamics. The true and conditional estimates of the state and observer parameters are

summarised in the lower left of Fig. 16 in terms of their expectation and conditional 90% confidence intervals (red bars) for DEM (grey) and EM (white). The EM estimates are slightly overconfident in that true values lie outside the 90% confidence intervals. It can be seen that the DEM estimates have a greater conditional variability and are closer to the true values. This is a systematic difference between DEM and EM: we repeated the simulations for eight independent realisations. The resulting conditional expectations are displayed over the true values in Fig. 17 and speak to a bias-variance trade-off between the two schemes. DEM provides more variable conditional estimates but they are unbiased.

Note that EM does not have a mean-field partition that covers the states. This induces conditional dependencies between the two parameters which may explain its bias and shrinkage towards the prior expectation of zero. In DEM, the parameters of the observer and state equations are conditionally independent. This is because

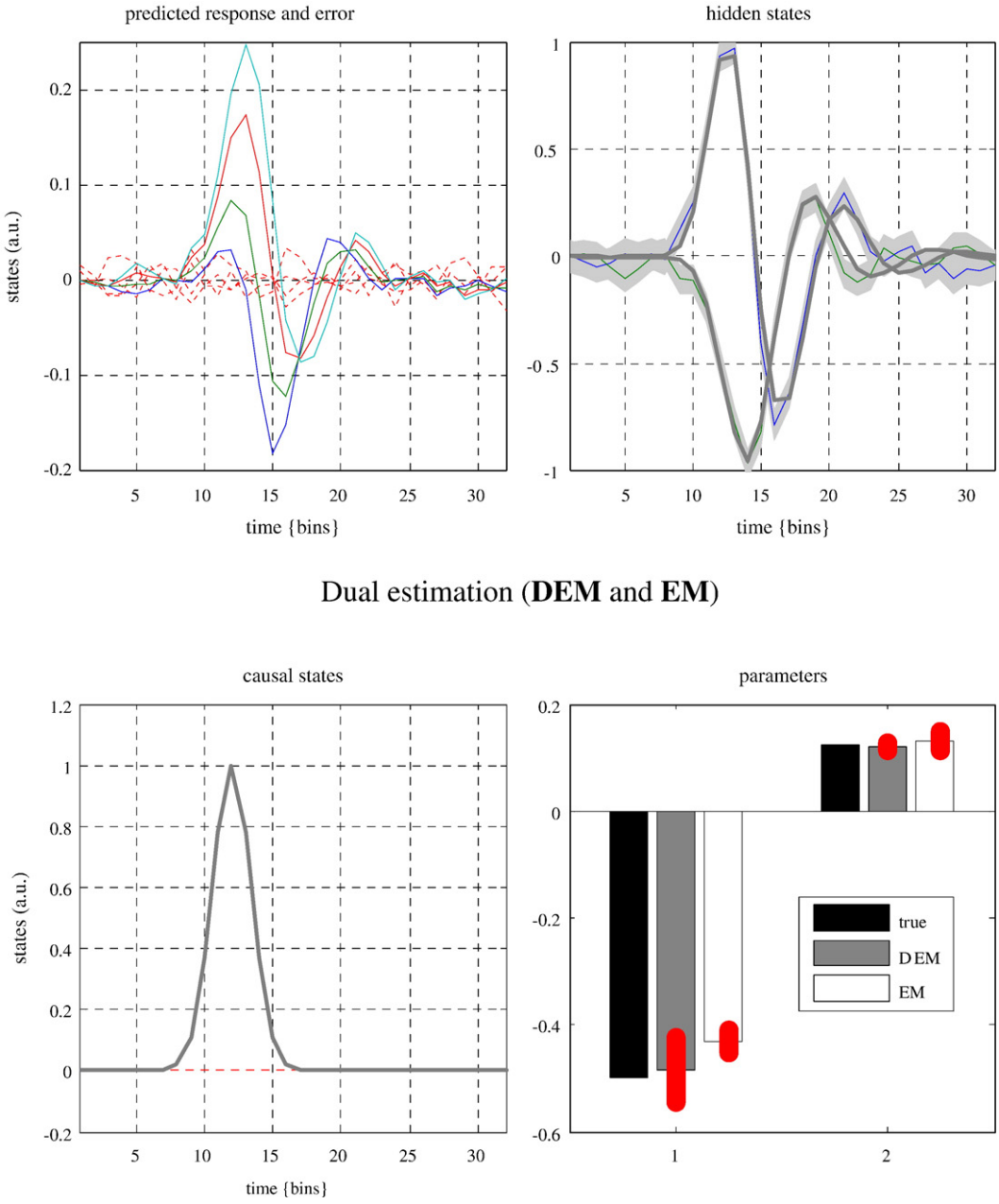

Fig. 16. As for Fig. 7; summarising the result of dual estimation using the linear convolution model. In this case, we used the known cause as a prior to suppress uncertainty about the hidden states. This allowed us to estimate the parameters and hyperparameters of the system with both DEM and EM. The hyperparameters correspond to the precision of random fluctuations in the response and the hidden states. The free parameters correspond to a single parameter from the state equation and one from the observer equation that govern the dynamics of the hidden states and the response respectively. For a comparative evaluation, the same dual estimation was implemented using expectation maximisation (EM). The true and conditional estimates are summarised on the lower left in terms of their expectation and conditional 90% confidence intervals (red lines). It can be seen that the DEM estimates have a greater conditional variability but are closer to the true values.

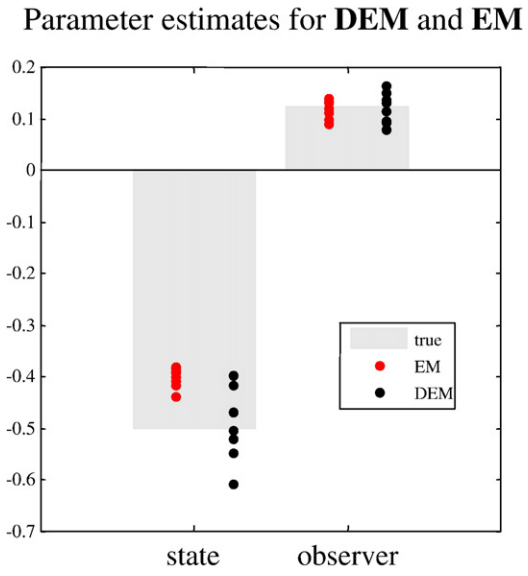

Fig. 17. We repeated the analysis described in the previous figure for eight independent realisations of the linear convolution model. This figure summarises the conditional expectations of the parameters of the state and observer functions for EM and DEM. These results are displayed over the true values (bars). These results suggest that there is a bias-variance trade-off when using EM and DEM. DEM provides more variable conditional estimates but they are much less biased than those obtained using EM. Note that EM does not have a mean-field partition that covers the states; this induces conditional dependencies between the two parameters which may, in part, explain this bias in the estimators and shrinkage towards the prior expectation of zero.

the random effects on the states and causes are conditionally independent under the mean-field approximation. The state-parameters only affect the prediction error on the motion of the states, while the observer parameters affect only the prediction error on the responses. In contrast, in EM, changes in the state-parameters are expressed in the response through the hidden states and induce conditional dependencies among all parameters. In this case, there was a negative conditional correlation between the two parameters.

Triple estimation

In our final simulations, we combine inference on the states, parameters and hyperparameters in a triple estimation procedure that exploits all three DEM steps. We generated data from the linear convolution model as above and inverted the following model

Linear convolution model for triple estimation

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

This is exactly the same as the previous model but the tight informative priors on the causes have been replaced with uninformative shrinkage priors. Fig. 18 shows the conditional densities on the states and parameters after convergence of the DEM scheme (after about 32 iterations). Remarkably, the inversion has recovered the states and parameters. EM cannot do this; therefore, the true and conditional estimates of the parameters are provided for DEM only (lower right). It can be seen that the true value of the causal state lies within the 90% confidence interval and that we could infer with substantial confidence that the cause was non-zero, when it occurs.

Summary

There are no generic schemes that can solve this triple estimation problem; although one could argue that bespoke variational update schemes based on Eq. (11) and Eq. (12) could be derived (*c.f.*, the treatment provided in Beal and Ghahramani 2003 for static systems). However, these would not be generic in the same sense as DEM. This is because DEM does not need model-specific updates (or the integrals implicit in Eq. (11) and Eq. (12)) to optimise the free-energy bound. DEM uses a coordinate ascent, under a mean-field approximation, that can be applied to any model. This point can be reiterated by noting that all the examples in this paper were inverted using just one routine, whose arguments are the model specification and the response (see software note). In fact, this routine can be used to invert any dynamic or static model under parametric assumptions. We will deal with this in another paper on a hierarchical model specification and show that many conventional analyses, ranging from ordinary least squares to independent component analysis, can be formulated as special cases of DEM. We now close with an example that gets closer to the sort of application we had in mind.

An empirical application to hemodynamics

In this, the final section, we illustrate DEM 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. Furthermore, generalisations of this model, to cover multiple brain regions, are used routinely in the dynamic causal modelling of neuronal interactions. The inversion of these models enables inference about parameters that control coupling among brain areas (see Friston et al., 2003; Penny et al., 2005). Even the inversion of single-area models has received considerable attention (*e.g.*, Gitelman et al., 2003; Buxton et al., 2004). There are several compelling applications of the innovation method in neuroimaging that have focussed on neural mass models of the electroencephalogram (Valdes et al., 1999; Sotero et al., 2007) and, more recently, hemodynamic time-series (Riera et al., 2004; Sotero and Trujillo-Barreto, in press). See also Deneux and Faugeras (2006) who present a careful analysis of identifiability, using the nonlinear model of fMRI data described below.

As noted by Johnston et al. (2006): “The most comprehensive model to date of the BOLD effect is formulated as a continuous-time system of nonlinear stochastic differential equations”. Johnston et al. (2006) present a particle filtering method for the analysis of the BOLD system and demonstrate it to be both accurate and robust in estimating the hidden physiological states. Conversely, Jacobsen et al. (2006) provide a careful and thorough inference on the model parameters, using a Metropolis–Hastings algorithm for sampling their posterior distribution. We combine both objectives, using DEM to estimate not only the states but the parameters and hyperparameters of the system generating those states.

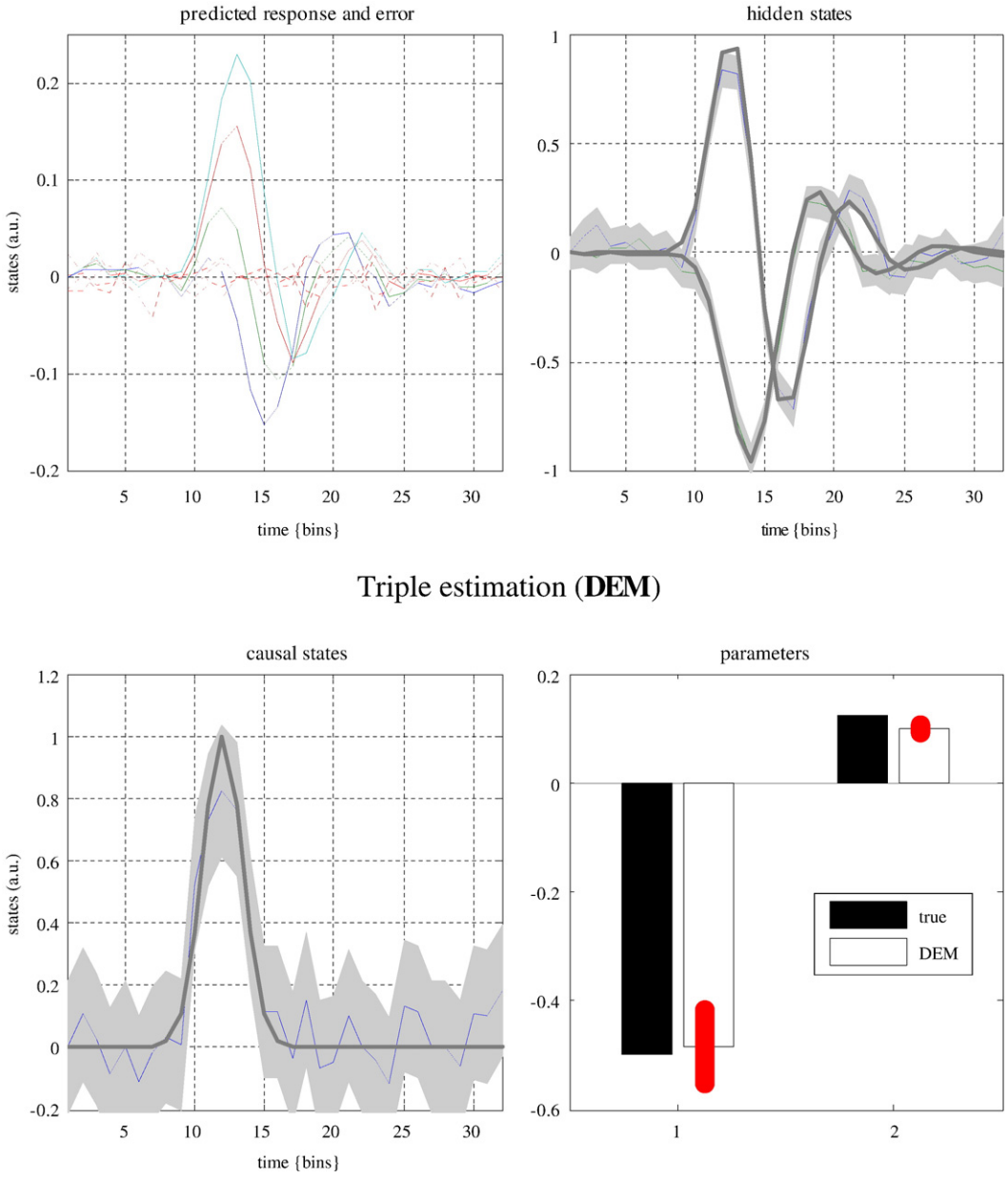

Fig. 18. As for the previous figure but in this instance we estimated the parameters, hyperparameters and the cause by treating it as an unknown quantity. This is an example of triple estimation, where we are trying to infer on the states of the system as well as the parameters governing its causal architecture. EM cannot do this; therefore the true and conditional estimates of the parameters are provided for the DEM scheme only (lower right). It can be seen that the true value of the causal state lies within the 90% confidence interval and that we could infer with substantial confidence that the cause was non-zero, when it occurs.

The hemodynamic model

The hemodynamic model has been described extensively in previous communications (*e.g.*, Friston 2002). It completes the Balloon–Windkessel model (Buxton et al., 2004; Mandeville et al., 1999) and is based on a large amount of neurophysiology and biophysics. 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 change 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

$$\begin{aligned}\dot{h}_1 &= \sum \theta_i v_i - \kappa(h_1 - 1) - \chi(h_2 - 1)\\ \dot{h}_2 &= h_1 - 1\\ \dot{h}_3 &= \tau(h_2 - F(h_3))\\ \dot{h}_4 &= \tau(h_2 E(h_2) - F(h_3)h_4/h_3)\end{aligned} \tag{61}$$

In this model, the vasodilatory signal is elicited by a mixture of neuronal inputs, v_i scaled by parameters, θ_i which encode their relative contribution to hemodynamic signals. Outflow is related

Table 1
Priors on free biophysical parameters

	Description		Prior
κ	Rate of signal decay/s =	$0.65\exp(\theta_\kappa)$	$p(\theta_\kappa) = N(0, \frac{1}{16})$
χ	Rate of flow-dependent elimination =	$0.41\exp(\theta_\chi)$	$p(\theta_\chi) = N(0, \frac{1}{16})$
τ	Rate hemodynamic transit/s =	$1.02\exp(\theta_\tau)$	$p(\theta_\tau) = N(0, \frac{1}{16})$
α	Grubb's exponent =	$0.32\exp(\theta_\alpha)$	$p(\theta_\alpha) = N(0, \frac{1}{16})$
φ	Resting oxygen extraction fraction =	$0.34\exp(\theta_\varphi)$	$p(\theta_\varphi) = N(0, \frac{1}{16})$
θ_i	Effect of the i-th neuronal input =	θ_i	$p(\theta_i) = N(0, 1)$

Fixed biophysical parameters

	Description	
V_0	Blood volume fraction	0.04
k_1	Intravascular coefficient	7φ
k_2	Concentration coefficient	2
k_3	Extravascular coefficient	2φ–0.2

to volume $F(h_3)=h_3^{1/\alpha}$ through Grubb's exponent α. The relative oxygen extraction $E(h_2) = \frac{1}{\varphi}(1-(1-\varphi)^{1/h_2})$ is a function of flow, where φ is a resting oxygen extraction fraction. A description of the parameters of this model is provided in Table 1.

There are many ways to formulate and use this model. Typically, if one was interested in the biophysical parameters, known experimental inputs could enter as informative priors on the causes to enable efficient estimation of the parameters. Conversely, if one was interested in the underlying neuronal and hemodynamic states, one might use fixed parameter values to assure an efficient deconvolution. To make things interesting, we will infer on both parameters and states by reprising the triple estimation of the previous section but now in the context of a nonlinear (*i.e.*, generalised) and more complicated convolution model. We will model experimentally controlled causes, v_i explicitly and use the known experimental design to place priors on them.

We will also use this model to illustrate how non-Gaussian densities can be modelled under the Laplace approximation, through nonlinear transformations: all the hemodynamic states are clearly non-negative quantities. One can accommodate this by assuming that $\ln h_i$ has a normal posterior; *i.e.*, the states have non-negative lognormal densities. This is easy to implement with the transformation; $x_i=\ln h_i \Leftrightarrow h_i=\exp(x_i)$. Under this transformation the differential equations above can be written

$$\dot{h}_i = \frac{\partial h_i}{\partial x_i}\frac{\partial x_i}{\partial t} = h_i\dot{x}_i = f_i(h,v) \tag{62}$$

This allows us to formulate the model in terms of the hidden states $x_i=\ln h_i$ under Gaussian assumptions

Hemodynamic convolution model

Level	$g(x,v)$	$f(x,v)$	Π^z	Π^w	η^λ	P^λ
$m=1$	$V_0(k_1(1-h_2)+k_2(1-h_4/h_3)+k_3(1-h_3))+c(\beta,t)$	$\begin{bmatrix}\sum\theta_i v_i-\kappa(h_1-1)-\chi(h_2-1)/h_1\\(h_1-1)/h_2\\\tau(h_2-F(h_3))/h_3\\\tau(h_2E(h_2)-F(h_3)h_4/h_3)/h_4\end{bmatrix}$	$\exp(\lambda^z)$	e^4	0	e^{-16}
$m=2$			1			

The priors on the parameters are detailed in Table 1. We also use lognormal priors on the hemodynamic parameters because they are non-negative rate-constants or fractions. For example, although the conditional density of $\theta_\kappa=\ln\kappa$ is Gaussian, the conditional density of κ is lognormal.

The model above represents a multiple-input, single-output model with four hidden states and unknown parameters. Five parameters control the hemodynamics, while the remaining parameters control the effect of causes on the vasodilatory signal. In this example, we treat the precision of observation noise as unknown but assume that state noise has precision, e^4, which corresponds roughly to random fluctuations with the same magnitude of evoked changes in hidden states (*i.e.*, a standard deviation of $e^{-2}\approx 14\%$). Priors on the causes v_i were relatively informative (unit precision; $e^0=I$) with expectations η_i^v based on experimental manipulations (see below).

This example also illustrates how confounds or time-varying effects, $c(\beta,t)$ can be modelled. These response components are not functions of the states but are parameterised. In this example, $c(\beta,t)=\beta C(t)$, where $C(t)$ corresponded to a discrete cosine temporal basis-set modelling instrumental drifts in the data that were of no interest. We used eight cosine components to remove or model low-frequency drifts over the 360 sample time-series. The parameters of these effects β were estimated in the E-step as described in Appendix E. To illustrate the ensuing inversion, DEM was applied to a single time-series from a visually responsive brain region, measured during a visual attention study with three experimental causes.

Data and results

Data were acquired from a normal subject at 2 T using a Magnetom VISION (Siemens, Erlangen) whole body MRI system. 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 the 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 last condition, subjects viewed stationary dots. The order of the conditions alternated between *Fixation* and photic 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) and 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. 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 on the three causes in the model. The associated parameters, θ_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 θ_2 exceeds zero would be large.

Conditional expectations and 90% confidence regions for the causal and hidden states are shown in Fig. 19. The dynamics of inferred activity, flow and other biophysical states are physiologically plausible. For example, activity-dependent changes in flow

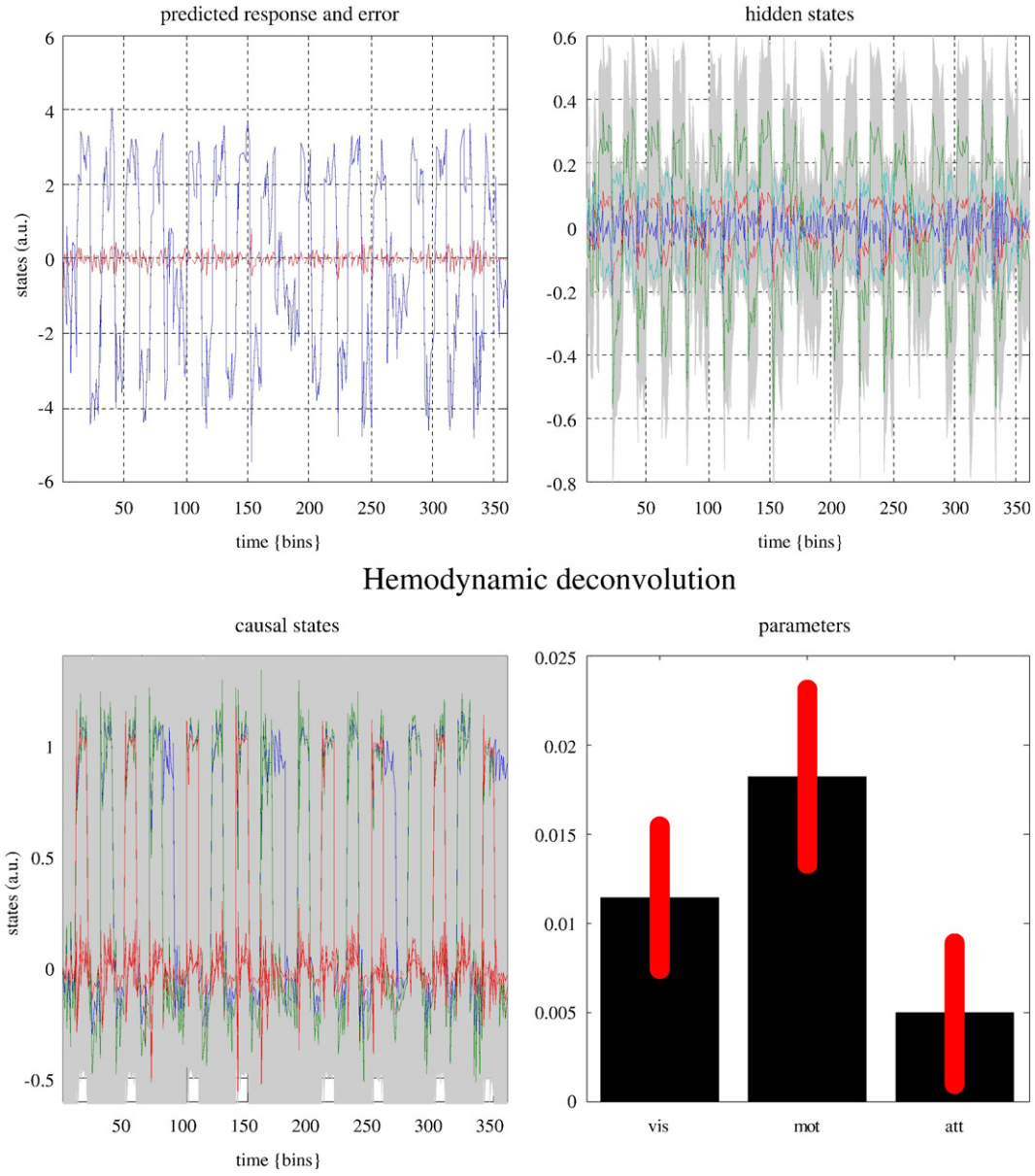

Fig. 19. As for the previous figure but now detailing the identification of a much more complicated convolution model. This nonlinear deconvolution was based upon empirical fMRI data caused by varying the cognitive set and visual stimuli a subject was exposed to. In this instance, there is one response (upper left), four state variables (upper right) and three causes (lower left). We used the experimental design to place priors on the causes but also allowed them to have stochastic components. The parameters, whose conditional estimates are shown on the lower right, couple the causes to the first hidden state and reflect how each induces a hemodynamic response. A more detailed summary of the hemodynamics is shown in the next figure.

are around 14%, producing about a 4% change in fMRI signal. The conditional estimates of the parameters, $\theta_i: i=1,\ldots 3$, are shown on the lower right. As anticipated, we can be almost certain that neuronal activity encoding motion induces a response. Interestingly, both visual stimulations *per se* (and perhaps attention) elicit responses in this area but not to the same degree.

A more detailed summary of the hemodynamics is shown in Fig. 20. This figure focuses on the first 120 time bins and plots the hemodynamic states in terms of $h_i = \exp(x_i)$ instead of x_i. Each time bin corresponds to 3.22 s. In the upper panel (a), the hidden states are shown, 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 (a mixture; $\Sigma\theta_i v_i$ of the three causes of the previous figure), shown in the lower panel (b), 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 onset transient that is rapidly corrected to give a rebound response a few seconds later. This seems to be a systematic feature that is evident in nearly all the epochs shown. Note that the conditional densities of the hemodynamic states and parameters are non-Gaussian (*i.e.*, lognormal), despite the Laplace assumption entailed by DEM. This is an example of the how the

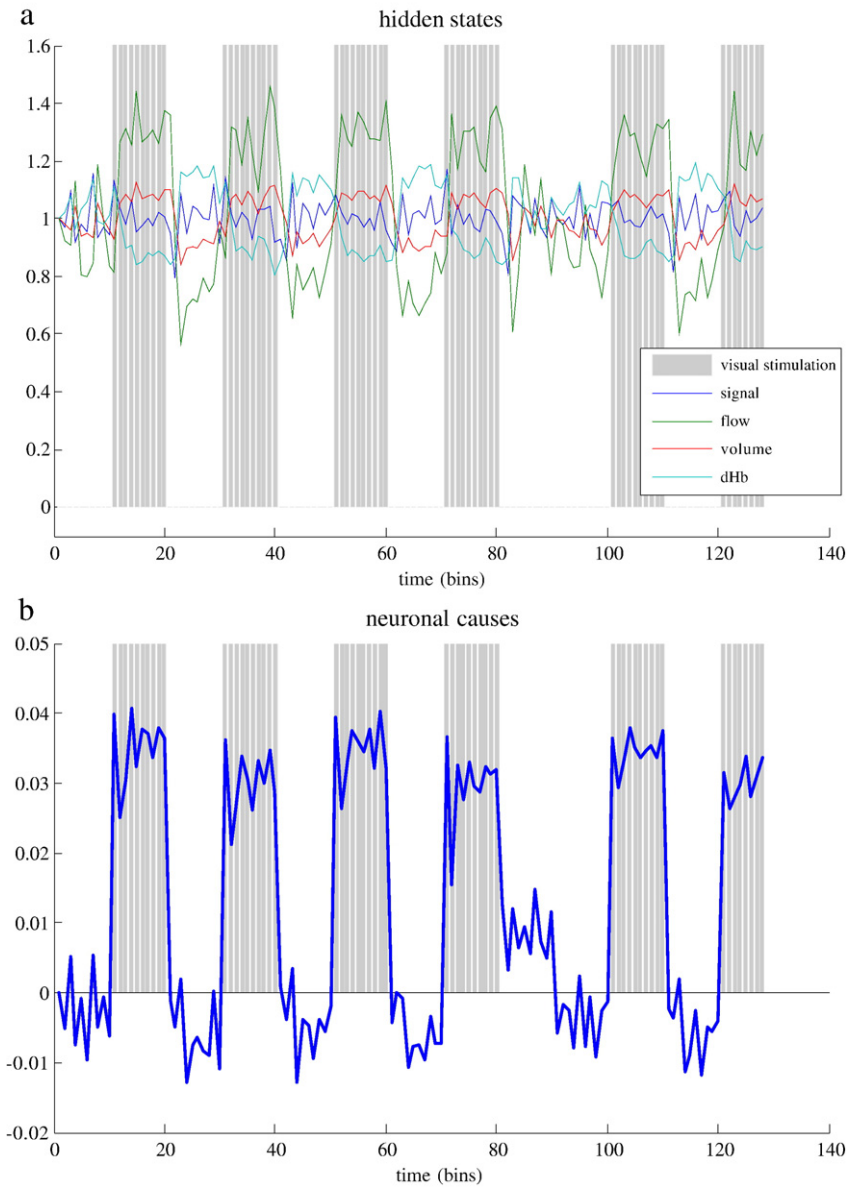

Fig. 20. These are the same results shown in the previous figure but focussing on the first 120 time bins. Each time bin corresponds to 3.22 s. In the upper panel (a), the hidden states are shown, overlaid on periods (grey bars) of visual stimulation. These hidden states correspond to flow-inducing signal, flow, volume and deoxyhemoglobin (dHb). It can be seen that neuronal activity (a mixture of the three causes of the previous figure, weighted by their corresponding parameter estimates), 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 state variables volume and dHb) determines the measured response. Interestingly, the underlying neuronal activity appears to show an onset transient that is rapidly corrected to give a rebound response a few seconds later. This seems to be a systematic feature that is evident in nearly all the epochs shown.

inversion of nonlinear models can be used to invert models with non-Gaussian densities.

Summary

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 and other biophysically informed priors allow one to focus data on interesting model parameters or hidden states. This focus is enabled by inversion schemes of the sort introduced in this paper. One might argue that the functional form itself is the most interesting aspect of study. In a subsequent paper, we will illustrate model selection with DEM.

This allows one to explore model space by using the free-action in Eq. (33) as a lower-bound approximation to the time-integrated log-evidence or marginal likelihood for each model (Berger, 1985; see also Penny et al., 2004). This instance of Bayesian model selection provides a principled way to optimise the model and its functional form *per se* and test hypotheses about one model, in relation to another, using their relative marginal likelihoods.

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 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 for both the parameters of neuronal systems and their states. The blind deconvolution of a single region above is a simple example of this and will be extended to cover networks of multiple regions in subsequent work. Another interesting approach to modelling neuronal dynamics rests on coupling different scales in wavelet space. In these models, the dynamics at each scale are determined by a coupled ensemble of nonlinear oscillators, which embody the principal scale-specific neurobiological processes (Breakspear and Stam, 2005). These models may be usefully treated within the inversion framework discussed above.

Conclusion

We have presented a variational treatment of dynamic models that furnishes the time-dependent conditional densities of a system's states and the time-independent densities of its parameters. These obtain by maximising the variational free-energy and action of a system respectively. The ensuing action represents a lower-bound approximation to the model's marginal likelihood or log-evidence, integrated over time. This is required for model selection and averaging. The approach rests on formulating the optimisation in term of stationary action, in generalised coordinates of motion. The resulting scheme can be used for online Bayesian inversion of nonlinear dynamic causal models and eschews some limitations of previous approaches, such as Kalman and particle filtering. We refer to this approach as dynamic expectation maximisation (DEM) because it uses a coordinate ascent on the action, which, by appeal to the same arguments used for expectation maximisation, is guaranteed to converge. This explicit optimisation of a bound avoids difficult integrations associated with conventional variational schemes and allows one to use models that do not have closed-form updates.

Variational vs. incremental approaches

The variational approach to dynamic systems presented here differs in several ways from incremental approaches such as extended Kalman filtering and related approaches. The first distinction relates to the nature of the generative models. The variational approach regards the generative model as mapping between causes and responses, which are functions of time. In contradistinction, incremental approaches consider the mapping to be between scalar quantities. In this sense, the variational treatment above can be regarded as a generalisation of model inversion to cover mappings between functions, where the functions happen to be of time. Incremental approaches simply treat the response as an ordered sequence and infer the current state using previous estimates. This creates a problem for incremental approaches because the underlying causes and responses are analytical functions of time, which provide constraints on inversion that cannot be exploited. For example, most incremental approaches assume uncorrelated random components (*e.g.*, a Wiener 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 motion and a parametric form for the temporal correlations.

The second key difference between conventional filtering techniques and DEM is the form 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 extended Kalman filtering cannot be used to deconvolve the inputs to a system (*i.e.*, causes) from its responses. A more subtle difference is that Kalman filtering cannot handle non-Gaussian causes gracefully. Conversely, non-Gaussian innovations are modelled simply in DEM with a nonlinear transformation of Gaussian causes. For example, in the nonlinear convolution model above, the causes $\exp(v)$ are effectively lognormal perturbations.

A third difference between variational and incremental treatments lies in the nature of causal inference that is supported. DEM inverts generative models that are formulated as causal systems using differential equations. Causal here means that the response is a function of states in the past; more exactly the system's generalised convolution kernels have support on, and only on, the past. This means that the conditional density pertains to states that cause responses in a control theory sense. This is why we refer to DEM as a scheme for the inversion of dynamic causal models. This is not the case for dual-estimation procedures based on state-space models parameterised in terms of transition matrices (Appendix C)

$$\mathbf{f}_x = \exp(\Delta t f_x) \tag{63}$$

See Büchel and Friston (1998) for an example in neuroimaging, based on variable parameter regression and Kalman filtering. The variational density on the parameters of $f(\vartheta)$ ensures that inference is on a causal model. However, inference on the parameters of $\mathbf{f}_x$ does not. This is because there are no constraints on these parameters which enforce causality. A simple heuristic here is that if $\mathbf{f}_x$ is negative, the underlying dynamic causal model is not defined, because $\ln \mathbf{f}_x = \Delta t f_x$ does not exist. This is important because there are many schemes that employ state-space models (*e.g.*, extended Kalman filtering, multivariate autoregression models, hidden Markov models, *etc.*) that use causal rhetoric (*e.g.*, Granger causality), which should not be confused with true causality. A further disadvantage of conventional state-space models is that the causal system may have only one free parameter but the number of elements required to specify $\mathbf{f}_x$ can be arbitrarily large, depending on the number of states.

Beyond filtering

Clearly, DEM is more than just a fixed-form filtering scheme; it applies to models with unknown states and parameters. Critically, the principle of stationary action that governs the evolution of the conditional mode of the states is recapitulated for all quantities responsible for generating data. These include parameters and hyperparameters. We have illustrated this using expectation maximisa-

tion as a reference. Although we have focused on three mean-field marginals, it is easy to extend the scheme to cover any mean-field partition. The distinction between an M-step and an E-step is that the latter ignore mean-field effects from other sets of parameters. In our case, we ignored uncertainty in the hyperparameters when estimating the conditional density of the parameters (but we did not ignore uncertainty about the states).

We have not been able to address many important issues in this (long) paper. Key omissions include the nature and implementation of variational filtering, which DEM approximates under the Laplace assumption; the inversion of hierarchal static models with DEM; the use of DEM for model selection, automatic relevance determination and exploration of model space; and biophysical implementations of DEM and extensions to cover action on data. The latter is an interesting issue and considers the optimisation of free-energy through active re-sampling of data (Friston et al., 2006). This may involve extending the notion of generalised coordinates to cover space as well as time. This extension is also relevant for the analysis of data that are functions of space; for example, continuous images. In this case, the generalised response becomes a tensor with rank greater than one and dimensions corresponding to the embedding order; *i.e.*

$$\tilde{y}(t) = \begin{bmatrix} y \\ \partial_t y \\ \vdots \end{bmatrix} \rightarrow \tilde{y}(x,t) = \begin{bmatrix} y & \partial_x y & \cdots \\ \partial_t y & \partial_t \partial_x y & \\ \vdots & & \ddots \end{bmatrix} \tag{64}$$

for data that are functions with domain, x. These issues will be dealt with in subsequent communications (although some are covered in software demonstrations; see software note).

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_DEM.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. Finally, we would like to thank our reviewers for their scholarly guidance.

Appendix A. Hyperparameterisation and hyperpriors

Our demonstrations used implicit lognormal hyperpriors. This was implemented by parameterising the precisions under Gaussian hyperpriors using

$$\Pi(\lambda_i^z) = \exp(\lambda_i^z)\Omega_i^z \tag{A.1}$$

similarly for $\Pi(\lambda_i^w)$. Here, the leading-diagonal elements of the user-specified components Ω_i^z encode which innovations share the same precision. Under this form, the hyperparameters, λ_i^z and λ_i^w become log-precisions. This hyperparameterisation ensures positive semi-definitive precisions because the implicit scale-parameters $\exp(\lambda_i^z)>0$ are non-negative. This enables us to retain the computational simplicity of the Laplace approximation, while precluding improper precisions. This is a further example of how nonlinear parameterisations can convert a Gaussian model into a non-Gaussian model. Strictly speaking, extra terms are entailed in the D- and E-steps because $\partial_{\lambda\lambda}\Pi \neq 0$ (see the appendix of Friston et al., 2007). However, these terms are small and are omitted for simplicity.

Appendix B. Integrating a stochastic DCM in generalised coordinates

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

$$\begin{array}{ll} y = g(x,v) + z & x' = f(x,v) + w \\ \dot{y} = g_x x' + g_v v' + \dot{z} & x'' = f_x x' + f_v v' + \dot{w} \\ \ddot{y} = g_x x'' + g_v v'' + \ddot{z} & x''' = f_x x'' + f_v v'' + \ddot{w} \\ \vdots & \vdots \end{array} \tag{A.2}$$

or, more compactly, $D\tilde{y} = D\tilde{g} + D\tilde{z}$ and $D\tilde{x} = \tilde{f} + \tilde{w}$. This allows us to integrate the system, under local linearity assumptions, given initial values for the states and innovations in generalised coordinates.

$$\begin{bmatrix} \dot{\tilde{v}} \\ \dot{\tilde{x}} \\ \dot{\tilde{z}} \\ \dot{\tilde{w}} \end{bmatrix} = \begin{bmatrix} D\tilde{g}_v & D\tilde{g}_x & D & \\ \tilde{f}_v & \tilde{f}_x & & I \\ & & D & \\ & & & D \end{bmatrix} \begin{bmatrix} \tilde{v} \\ \tilde{x} \\ \tilde{z} \\ \tilde{w} \end{bmatrix}$$

$$\tilde{g}_v = I \otimes g_v \quad \tilde{f}_v = I \otimes f_v$$
$$\tilde{g}_x = I \otimes g_x \quad \tilde{f}_x = I \otimes f_x$$

$$g_v = \begin{bmatrix} 0 & g_v^{(1)} & & \\ & \ddots & \ddots & \\ & & & g_v^{(m)} \\ & & & 0 \end{bmatrix} \quad g_x = \begin{bmatrix} g_x^{(1)} & & & \\ & \ddots & & \\ & & & g_x^{(m)} \\ & & & 0 \end{bmatrix} \quad v = \begin{bmatrix} y \\ v^{(1)} \\ \vdots \\ v^{(m)} \end{bmatrix}$$

$$f_v = \begin{bmatrix} 0 & f_v^{(1)} & & \\ & \ddots & \ddots & \\ & & & f_v^{(m)} \\ & & & 0 \end{bmatrix} \quad f_x = \begin{bmatrix} f_x^{(1)} & & & \\ & \ddots & & \\ & & & f_x^{(m)} \\ & & & 0 \end{bmatrix} \quad x = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(m)} \\ 0 \end{bmatrix} \tag{A.3}$$

In principle, we could generate time-series of any length given a sufficient embedding order and initial values. However, in practice, it is more expedient to use a low-order embedding (say $n=6$) and replace the trajectories of the innovations at each time-interval with $\tilde{z}(t) = T(t)z_{1:N}$, where $z_{1:N}$ is a random sequence of innovations that has been convolved with the appropriate smoothing kernel (similarly for $\tilde{w}(t)$).

Appendix C. Extended Kalman filtering

This appendix provides a pseudo-code specification of the extended Kalman filter based on var der Merwe (2000) and formulated for models of the form:

$$\begin{array}{l} y = g(x) + z \\ \dot{x} = f(x,v) + w \end{array} \tag{A.4}$$

Eq. (A.4) can be re-written, using local linearisation, as a discrete-time state-space model. This is the formulation treated in Bayesian filtering procedures (assuming $\Delta t=1$)

$$\begin{aligned} y_t &= \mathrm{g}_x x_t + z_t \\ x_t &= \mathrm{f}_x x_{t-1} + w_{t-1} \\ \\ \mathrm{g}_x &= g(x_t)_x \\ \mathrm{f}_x &= \exp\left(f(x_t)_x\right) \\ \\ z_t &= z(t) \\ w_{t-1} &= \int \exp\left(f_x\tau\right)\left(f_v v(t-\tau) + w(t-\tau)\right)d\tau \end{aligned} \tag{A.5}$$

The key thing to note here is that process noise w_{t-1} is simply a convolution of the exogenous causes, $v(t)$ 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

$$\langle w_t w_t^T \rangle = \int \exp(f_x\tau)\Omega\exp(f_x\tau)^T d\tau \approx \Omega \tag{A.6}$$

assuming that temporal correlations can be discounted and that the Lyapunov exponents of f_x are small relative to the time-step. Here $\Omega=f_v\Sigma^{(2)z}f_v^T+\Sigma^{(1)w}$ is the covariance of the underlying fluctuations from exogenous inputs and system noise. Note, system noise $w(t)$ on the states enters as an independent component of process noise. We use this in our implementation of extended Kalman filtering presented here in pseudo-code and implemented in *spm_ekf.m* (see software note)

for all t

Prediction step

$$\begin{aligned} x_t &= \mathrm{f}_x x_{t-1} \\ \Sigma_t^x &= \Omega + \mathrm{f}_x \Sigma_{t-1}^x \mathrm{f}_x^T \end{aligned}$$

Update or correction step

$$\begin{aligned} K &= \Sigma_t^x \mathbf{g}_x^T \left(\Sigma + \mathbf{g}_x \Sigma_t^x \mathbf{g}_x^T\right)^{-1} \\ x_t &\leftarrow x_t + K(y - g(x_t)) \\ \Sigma_t^x &\leftarrow (I - K\mathbf{g}_x)\Sigma_t^x \end{aligned} \tag{A.7}$$

end

where $\Sigma=\Sigma^{(1)z}$ is the covariance of observation noise. The Kalman gain matrix, K is used to update the prediction of future states, x_t and their conditional covariance, Σ_t^x given each new observation. We actually use $x_t=x_{t-1}+(\mathbf{f}_x-I)f_x^{-1}f(x_{t-1})$. This is a slightly more sophisticated update that uses the current state as the expansion point for the local linearisation (see main text and Ozaki (1992)).

Appendix D. Particle filtering

This appendix provides a pseudo-code specification of particle filtering based on var der Merwe et al. (2000) and formulated for state-space models described above (Appendix C). 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

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

Normalise importance weights

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

Selection step: for all i

$$x_t^{[i]} \leftarrow x_t^{[r]} \tag{A.8}$$

end

where $w^{[i]}$ is drawn from a proposal density $N(0,\Omega)$ and $x_t^{[i]}\leftarrow x_t^{[r]}$ denotes sequential importance re-sampling. The indices r are selected on the basis of the importance weights. $\Sigma=\Sigma^{(1)z}$ and $\Omega=f_v\Sigma^{(2)z}f_v^T+\Sigma^{(1)w}$ are defined in Appendix C. 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.

Appendix E. Confounds

It is simple to model confounds by adding them to the first level of the model

$$\begin{aligned} y &= g(x,v) + c(\beta,t) + z \\ \dot{x} &= f(x,v) + w \end{aligned} \tag{A.9}$$

For example, $c(\beta,t)$ could represent drift terms with a known form but unknown amplitude. These could be modelled with $c(\beta,t)=c\beta(t)$ where $C(t)$ is a suitable basis set encoding the temporal form. The time-invariant parameters β are generally treated as fixed-effects (*i.e.*, with uninformative priors). These parameters are absorbed into the system parameters by simply augmenting the error derivatives so that $\Delta\mu^\theta\rightarrow\Delta\mu^\theta,\Delta\mu^\beta$ in the E-step (Eq. (39)), where

$$\tilde{\varepsilon}_\beta = \begin{bmatrix} \tilde{\varepsilon}_\beta^v = -T\partial_\beta c_{0:n} \\ \tilde{\varepsilon}_\beta^x = 0 \end{bmatrix} \quad \begin{matrix} \tilde{\varepsilon}_\theta \rightarrow [\tilde{\varepsilon}_\theta, \tilde{\varepsilon}_\beta] \\ \tilde{\varepsilon}_{\theta u} \rightarrow [\tilde{\varepsilon}_{\theta u}, 0] \end{matrix} \quad P^\theta \rightarrow \begin{bmatrix} P^\theta & \\ & 0 \end{bmatrix} \tag{A.10}$$

and $c_{0:n} = \left[c\left(\beta, t-\frac{n}{2}\right); \ldots; c\left(\beta, t+\frac{n}{2}\right)\right]$ denotes a local vector of confounds evaluated at μ^β. Note that in hierarchical models only the first level error derivatives are affected by confounds. Uncertainty about β does not enter inference on the states in the D-step because the states and confounds do not interact. However, the conditional expectation of the confounds is removed from the response variable, before computing the prediction errors and their derivatives

$$\tilde{y} = T(0)(y_{0:n} - c_{0:n}) \tag{A.11}$$

otherwise, the DEM scheme is identical.

References

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

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., Smith, A.F.M., West, M. (Eds.), Bayesian Statistics. OUP, UK. Chapter 7.

Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis, 2nd Ed. Springer.

Breakspear, M., Stam, C.J., 2005. Dynamics of a neural system with a multiscale architecture. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 360 (1457), 1051–1074.

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.

Büchel, C., Friston, K.J., 1998. Dynamic changes in effective connectivity characterised by variable parameter regression and Kalman filtering. Hum. Brain Mapp. 6, 403–408.

Breakspear, M., Roberts, J.A., Terry, J.R., Rodrigues, S., Mahant, N., Robinson, P.A., 2006. A unifying explanation of primary generalized seizures through nonlinear brain modeling and bifurcation analysis. Cereb. Cortex 16 (9), 1296–1313.

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

Dempster, A.P., Laird, N.M., Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. Series B 39, 1–38.

Deneux, T., Faugeras, O., 2006. Using nonlinear models in fMRI data analysis: model selection and activation detection. NeuroImage 32 (4), 1669–1689.

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.

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

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.J., 2005. A theory of cortical responses. Philos. Trans. R. Soc. Lond. B Biol. Sci. 360, 815–836.

Friston, K., Kilner, J., Harrison, L., 2006. A free energy principle for the brain. J. Physiol. Paris 100 (1–3), 70–87.

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.J., 2008. Variational filtering. NeuroImage 41, 747–766.

Ghahramani, Z., Beal, M.J., 2001. The Variational Kalman Smoother. Technical Report. University College London. citeseer.ist.psu.edu/ghahramani01variational.html.

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.

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

Helmholtz, H., 1860/1962. In: Southall, J.P.C. (Ed.), Handbuch der physiologischen optik, Vol. 3. Dover, New York. English trans.

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.

Honkela, M., Tornio, Raiko, T., 2006. Variational Bayes for continuous-time nonlinear state-space models. NIPS*2006 Workshop on Dynamical Systems, Stochastic Processes and Bayesian Inference. Whistler, B.C., Canada.

Jacobsen, D.J., Madsen, K.H., Hansen, L.K., 2006. Biomedical imaging: macro to nano. 3rd IEEE Int. Symp. Page(s) 952–955.

Jimenez, J.C., Ozaki, T., 2006. An approximate innovation method for the estimation of diffusion processes from discrete data. J. Time Series Analysis 27, 77–97.

Jimenez, J.C., Biscay, R.J., Ozaki, T., 2006. Inference methods for discretely observed continuous-time stochastic volatility models: a commented overview. Asia-Pac. Financ. Mark. 12, 109–141.

Johnston, L.A., Duff, E., Egan, G.F., 2006. Particle filtering for nonlinear BOLD signal analysis. Lect. Notes Comput. Sci. 4191 (LNCS - II), 292–299.

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.

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

Neal, R.M., Hinton, G.E., 1998. A view of the EM algorithm that justifies incremental sparse and other variants. In: Jordan, M.I. (Ed.), Learning in Graphical Models. Kulver Academic Press.

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.

Ozaki, T., Iino, M., 2001. An innovation approach to non-Gaussian time series analysis. J. Appl. Prob. 38A, 78–92.

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

Penny, W.D., Trujillo-Barreto, N.J., Friston, K.J., 2005. Bayesian fMRI time series analysis with spatial priors. NeuroImage 24, 350–362.

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.

Sørensen, H., 2004. Parametric inference for diffusion processes observed at discrete points in time: a survey. Int. Stat. Rev. 72 (3), 337–354.

Sotero, R.C., Trujillo-Barreto, N.J., Iturria-Medina, Y., Carbonell, F., Jimenez, J.C., 2007. Realistically coupled neural mass models can generate EEG rhythms. Neural Comput. 19 (2), 478–512.

Sotero R.C., Trujillo-Barreto N.J., in press. Biophysical model for integrating neuronal activity, EEG, fMRI and metabolism, NeuroImage, DOI: 10.116/j.neuroimage.2007.08.001.

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

Valdes, P.A., Jimenez, J.C., Riera, J., Biscay, R., Ozaki, T., 1999. Nonlinear EEG analysis based on neural mass models. Biol. Cybern. 81 (5–6), 415–424.

Valpola, H., Karhunen, J., 2002. An unsupervised ensemble learning method for nonlinear dynamic state-space models. Neural Comput. 14 (11), 2647–2692.

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

Wang, B., Titterington, D.M., 2004. Variational Bayesian Inference for Partially Observed Diffusions. Technical Report 04-4. University of Glasgow. http://www.stats.gla.ac.uk/Research/-TechRep2003/04-4.pdf.

Weissbach, F., 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., B 53 (3), 505–538.