NeuroImage

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

Technical Note

Variational filtering

K.J. Friston*

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

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

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

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

Introduction

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

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

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

Notation

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

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

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

Variational Bayes and ensemble learning

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

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

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

Variational Bayes

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

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

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

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

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

$$q(\vartheta) = \prod_i q(\vartheta^i) \quad (2)$$

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

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

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

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

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

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

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

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