Bayesian State Estimation Using Generalized Coordinates

Bhashyam Balaji[a], and Karl Friston[b]

[a]Radar Systems Section, Defence Research and Development Canada–Ottawa,
3701 Carling Avenue, Ottawa, ON, Canada K1A 0Z4
[b]The Wellcome Trust Centre for Neuroimaging, Institute of Neurology, UCL, 12 Queen
Square, London, WC1N 3BG, UK

ABSTRACT

This paper reviews a simple solution to the continuous-discrete Bayesian nonlinear state estimation problem that has been proposed recently. The key ideas are analytic noise processes, variational Bayes, and the formulation of the problem in terms of generalized coordinates of motion. Some of the algorithms, specifically dynamic expectation maximization and variational filtering, have been shown to outperform existing approaches like extended Kalman filtering and particle filtering. A pedagogical review of the theoretical formulation is presented, with an emphasis on concepts that are not as widely known in the filtering literature. We illustrate the appliction of these concepts using a numerical example.

Keywords: Variational Filtering, Continuous-Discrete Filtering, Kolmogorov equation, Fokker-Planck equation, Dynamical Causal Modelling, Hierarchical dynamical models

1. INTRODUCTION

The continuous-discrete Bayesian filtering problem is to estimate some state given the measurements, where the state is assumed to evolve according to a continuous-time stochastic process and the measurements are samples of a discrete-time stochastic process.[1] The conditional probability density function provides a complete probabilistic solution to the problem, and can be used to compute state estimators such as the conditional mean.

Several approaches have been proposed in the literature. The standard extended Kalman filter is the benchmark nonlinear filtering algorithm. It is based on the application of the linear Kalman filter to the model obtained via linearization of the nonlinear state and measurement models. Another related approach is the unscented Kalman filter.[2] Such approaches often work well for practical problems. However, they are not general solutions; for instance, they cannot model formally multi-modal posterior distributions.

A more general solution is provided by particle filters.[3–5] They are based on sequential importance sampling based Monte-Carlo approximations based on point mass, or particle, representation of the probability densities. In principle, they provide a more general solution than the EKF or the UKF; for instance, they can describe multi-modal densities. However, the basic particle filter often requires too many particles, i.e., it succumbs to the "curse of dimensionality", even for relatively benign models (e.g., linear model with unstable plant noise that is easily tackled using the KF).[6]

Another approach proposed to tackling the continuous-discrete and continuous-continuous filtering problems is based on Feynman path integral methods that are used in quantum field theory.[7–9] The simplest path-integral approximation, the Dirac-Feynman approximation, has been shown to be sufficiently accurate for solving challenging problems.

In this paper, we provide a pedagogical review of a novel Bayesian state estimation scheme recently proposed by Friston and collaborators.[10–12] The key concepts underlying this approach rests on an analytic noise process (rather than Wiener process), variational Bayes,[13, 14] and the formulation of the problem in terms of generalized coordinates. It has been shown to be very versatile and robust, and has also been successfully applied to simulated models, as well as real data in the problem of deconvolving hemodynamic states and neuronal activity

Further author information: (Send correspondence to Bhashyam Balaji)
E-mail: Bhashyam.Balaji@drdc-rddc.gc.ca, Telephone: 1 613 998 2215

Signal Processing, Sensor Fusion, and Target Recognition XX, edited by Ivan Kadar,
Proc. of SPIE Vol. 8050, 80501Y · © 2011 SPIE · CCC code: 0277-786X/11/$18 · doi: 10.1117/12.883513

from functional MRI responses in the brain. It has also led to a remarkably simple and useful model for inference and learning in the brain.[15]

A review of the some key concepts in variational Bayes are presented in Sections 2 and 3. The dynamic causal models and generalized coordinates are briefly reviewed in Section 4. The resulting Bayesian state estimation algorithms, Dynamic Expectation Maximization (DEM) and variational filtering (VF), are reviewed in Section 6. The formalism is applied to real data in Section 7.

2. VARIATIONAL BAYES FOR STATIC SYSTEMS

2.1 Log-Evidence, Free Energy and the Kullback-Leibler Divergence

There are a few cosntructs that prove to be very important in Variational Bayes approaches, that are well-known in statistical physics and machine learning but are used less widely known to conventional Bayesian filtering practitioners. A brief review of some of the relevant concepts and results are presented.

Let y be the measurement data and m be the model assumptions. The quantity $p(y|m)$, i.e., the conditional probablity of the data given the measurements, is referred to as the *evidence*. The logarithm of the evidence, $\ln p(y|m)$ is termed the *log-evidence*.

Let $q(\theta)$ be a probability density function over parameters θ. The *entropy* $H(q)$ is defined as

$$\begin{aligned} H(q) &\equiv -\int q(\theta)\ln q(\theta), \\ &\equiv -\langle \ln q(\theta)\rangle_{q(\theta)}. \end{aligned} \tag{1}$$

The *internal energy* is defined as

$$\begin{aligned} G(y,m,q) &\equiv \int d\theta q(\theta)\ln p(y,\theta|m), \\ &= \langle U(y,\theta|m)\rangle_{q(\theta)}, \qquad U(y,\theta|m) \equiv \ln p(y,\theta|m), \end{aligned} \tag{2}$$

and $U(y,\theta|m)$ is a *Gibbs energy function*. The sum of the entropy and the internal energy is termed the *free-energy*, i.e.,

$$\begin{aligned} F(y,m,q) &= G(y,m,q) + H(q), \\ &= \left\langle \ln \frac{p(y,\theta|m)}{q(\theta)} \right\rangle_{q(\theta)} \end{aligned} \tag{3}$$

Finally, the *Kullback-Leibler (KL) cross-entropy* or the *KL divergence term* is defined as

$$\begin{aligned} D_{KL}(q(\theta||p(\theta|y,m)) &\equiv \int d\theta q(\theta)\ln\frac{q(\theta)}{p(\theta|y,m)}, \\ &\equiv \left\langle \ln\frac{q(\theta)}{p(\theta|y,m)} \right\rangle_{q(\theta)}. \end{aligned} \tag{4}$$

The following straightforward result plays a major role in the subsequent discussion:

LEMMA 2.1. *The log-evidence is equal to the free-energy plus the K-L Divergence:*

$$\ln p(y|m) = F(y,m,q) + D_{KL}(q(\theta||p(\theta|y,m)). \tag{5}$$

Proof. The proof is straightforward and follows from the definition of the quantities as noted below:

$$\begin{aligned} F(y,m,q) + D_{KL}(q(\theta||p(\theta|y,m)) &= \int d\theta q(\theta)\left[\ln\frac{q(\theta)}{p(\theta|y,m)} + \ln\frac{p(y,\theta|m)}{q(\theta)}\right], \qquad (6)\\ &= \int d\theta q(\theta) \ln\frac{p(y,\theta|m)}{p(\theta|y,m)},\\ &= \int d\theta q(\theta)\ln\frac{p(y,\theta,m)p(y,m)}{p(m)p(\theta,y,m)},\\ &= \int d\theta q(\theta)\ln p(y|m),\\ &= \ln p(y|m). \end{aligned}$$

□

2.2 Lower bound on the log-evidence

Observe that the left-hand side of Equation 5 is independent of the density function q, while both the terms on the right-hand side depend on q. In other words, the $q-$dependence cancels. The great significance of Lemma 2.1 arises because of the following lemma that we state without proof:[14]

LEMMA 2.2. *Let $P(x)$ and $Q(x)$ be the probability density functions. The KL divergence is always positive, i.e.,*

$$D_{KL}(Q||P) \geq 0 \quad (\textit{Gibb's inequality}), \qquad (7)$$

with the equality when $Q = P$.

Therefore, Lemmas 2.1 and 2.2 imply that the free energy $F(y,q)$ *furnishes a lower bound on the log-evidence, because the K-L term $D_{KL}(q(\theta)|p(\theta|y,m)$ is always positive.*

2.3 Conditional Probability Density

The lowest value $D_{KL}(q(\theta)|p(\theta|y,m)$ can take is 0. Now, if the approximating density $q(\theta)$ is true posterior density $p(\theta|y,m)$, then the KL divergence is zero, and the free energy is exactly the log-evidence.

Of course, we do not know the true posterior density $p(\theta|y,m)$. However, we can turn the argument on its head: *if we can find the density $q(\theta)$ that maximizes the free-energy, then the approximate density $q(\theta)$ is the true posterior density $p(\theta|y,m)$*!

In summary, the method of obtaining the posterior density and log-evidence of the model is as follows. First, determine the $q(\theta)$ that maximizes the free-energy of the model, where the free-energy is given by

$$F(y,m,q) = \int d\theta q(\theta)\log\frac{p(y,\theta|m)}{q(\theta)}. \qquad (8)$$

The lower-bound approximation to the log-evidence is simply given by the maximum of the free-energy. Since maximizing the free-energy minimizes the KL divergence, the variational density $q(\theta)$ is approximately the desired posterior density, i.e., $q(\theta) \approx p(\theta|y,m)$. This can then be used for inference on the parameters of the model selected.

2.4 Mean-Field Approximation

The introduction of the variational density $q(\theta)$ has done something very significant. It has converted the difficult problem of integration

$$p(y|m) \equiv \int d\theta p(y,\theta|m), \qquad (9)$$

over the unknown parameters θ to compute the evidence $p(y|m)$ into an easier optimization problem via induction of a bound that can be optimized with respect to $q(\theta)$:

$$p(y|m) = \max_{q(\theta)} \left\langle \frac{\ln p(y,\theta|m)}{\ln q(\theta)} \right\rangle_{q(\theta)} \tag{10}$$

The problem is simplified if one can induce a bound that can be optimized with respect to $q(\theta)$. Often, one assumes that $q(\theta)$ factorizes over a partition of the parameters:

$$q(\theta) = \prod_{i=1}^{P} q(\theta^i), \tag{11}$$

where $\theta \equiv \left\{\theta^1, \theta^2, \ldots, \theta^p\right\}$. The parameters θ^i are a partition of θ, i.e., $\theta^i \cap \theta^j = \{\}$, when $i \neq j$. A convenient choice of factorization is often dictated by a separation of temporal scales or some other heuristic that ensures strong correlations are retained within each subset and discounts weak correlations between them. In classical statistical physics, this is referred to as the *mean-field approximation*. Finally, the *Markov blanket* of θ^i, written as $\theta^{\backslash i}$ is the set of parameters not in θ^i.

2.5 Variational Density

The following lemma shows that the variational density has a rather simple form.

LEMMA 2.3. *The free-energy is maximized with respect to* $q(\theta)$ *when*

$$q(\theta^i) = \frac{1}{Z^{(i)}} \exp(V(\theta^i)), \qquad V(\theta^i) \equiv \langle U(\theta) \rangle_{q(\theta^{\backslash i})}, \tag{12}$$

where $Z^{(i)}$ *is the partition function normalization constant.*

Proof. Recall that the free-energy is given by

$$\begin{aligned} F(y,\theta) &= \int d\theta q(\theta) \log \frac{p(y,\theta|m)}{q(\theta)}, \\ &= \int d\theta^i f^i, \end{aligned} \tag{13}$$

where

$$f^i \equiv \int d\theta^{\backslash i} q(\theta) q(\theta^{\backslash i}) \ln p(y,\theta|m) - \int d\theta^{\backslash i} q(\theta) q(\theta^{\backslash i}) (\ln q(\theta^i) + \ln q(\theta^{\backslash i})) \tag{14}$$

The variation of the free-energy w.r.t. $q(\theta^i)$ yields

$$\begin{aligned} \delta_{q(\theta^i)} F &= \int d\theta^{\backslash i} q(\theta^{\backslash i}) \ln p(y,\theta|m) - \int d\theta^{\backslash i} q(\theta^{\backslash i})(\ln q(\theta^i) + \ln q(\theta^{\backslash i})) - \int d\theta^{\backslash i} q(\theta) q(\theta^{\backslash i}) (\frac{1}{\ln q(\theta^i)}), \\ &= V(\theta^i) - \ln q(\theta^i) - \ln Z^{(i)}, \end{aligned}$$

where $Z^{(i)}$ is the combination of the terms that are independent of θ. It therefore follows that

$$\begin{aligned} \delta_{q(\theta^i)} F &= 0, \\ &= V(\theta^i) - \ln q(\theta^i) - \ln Z^{(i)}, \end{aligned} \tag{15}$$

or

$$q(\theta^i) = \frac{1}{Z^{(i)}} \exp(V(\theta^i)), \qquad V(\theta^i) \equiv \langle U(\theta) \rangle_{q(\theta^{\backslash i})}. \tag{16}$$

□

The quantity $V(\theta^i)$ is also referred to as the *variational energy.*

Observe that the mode of the ensemble density, i.e., value of θ^i that maximizes $q(\theta^i)$, maximizes the variational energy. Finally, note that when there is only one set, the variational density reduces to the Boltzmann distribution:

$$q\left(\theta\right) = \frac{1}{Z}\exp\left(V(\theta)\right). \tag{17}$$

2.6 The Fokker-Planck-Kolmogorov forward equation (FPKfe) and the Ensemble Density

The relationship between the Langevin equation and the FPKfe is next exploited to provide an ensemble representation of the variational density.

Lemma 2.4. *Suppose the particles in the $i-th$ parameter space are propagated using the Langevin equation, to wit,*

$$\frac{d}{dt}\theta^i = \nabla_{\theta^i}V(\theta^i) + \Gamma(t), \tag{18}$$

where $V(\theta^i) \equiv \langle U(\theta)\rangle_{q(\theta^{\setminus i})}$ and $\Omega \equiv \langle \Gamma(t)\Gamma(t)^T\rangle = 2I$. Then, the stationary solution for the ensemble density $p(t,\theta^i)$ is the same as the variational density, i.e.,

$$q(\theta^i) = \frac{1}{Z^{(i)}}\exp(V(\theta^i)). \tag{19}$$

Proof. The FPKfe corresponding to the Langevin Equation is[16]

$$\begin{aligned}\frac{\partial p}{\partial t}(t,\theta^i) &= -\nabla_{\theta^i}[\cdot\nabla_{\theta^i}V(\theta^i)p(t,\theta^i)] + \frac{1}{2}\Omega\nabla_{\theta^i}\cdot\nabla_{\theta^i}p(t,\theta^i),\\ &= \nabla_{\theta^i}\cdot\left[\nabla_{\theta^i}p(t,\theta^i) - p(t,\theta^i\nabla_{\theta^i}V(\theta^i)\right],\end{aligned} \tag{20}$$

where $p(t,\theta^i)$ is the ensemble density function. Since

$$\nabla_{\theta^i}\exp\left(V(\theta^i)\right) = \left[\nabla_{\theta^i}V(\theta^i)\right]\exp\left(V(\theta^i)\right), \tag{21}$$

the RHS of Equation 21 vanishes, implying that stationary solution for the ensemble density is the variational density. □

To reiterate, one can obtain samples, or "particles", from the desired ensemble density (Equation 19) by simply simulating the Langevin stochastic differential equation (Equation 18). Since the variational density is the stationary solution to a density on an ensemble of solutions, the variational density is also referred to as the *ensemble density.*

3. VARIATIONAL BAYES FOR DYNAMIC SYSTEMS

So far only the static case has been considered. In the dynamic case, some parameters (or "states" $u(t)$) may change with time, and the remaining parameters θ may be constant. This leads to a natural partition into states and parameters, or $\theta \rightarrow u(t), \theta$, and the natural mean-field approximation for the variational density $q = q(u(t))q(\theta)$, and the associated energies are now a function of time.

The variational Bayes analysis can be carried out in analogy with the time-independent case. Specifically, for the time-dependent case, the natural quantity to consider is the integral of the log-evidence over time.

$$\int dtp(y(t)|m). \tag{22}$$

If the time-series is uncorrelated, this is simply the log-evidence of the time-series.

Similarly, the quantity analogous to the free-energy (termed the *free-action*), energy function and the internal energy can be defined as

$$\begin{aligned}\bar{U}(y,u,\theta|m) &\equiv \int dt \ln p(y(t),u(t),\theta|m) \\ \bar{G}(y,u,\theta|m) &\equiv \int dt \langle U(y,u(t),\theta|m)\rangle_{q(u(t))} \\ \bar{F}(y,m,q) &\equiv \int dt \langle U(u,t|\theta)\rangle_{q(u,t)} - \int dt \langle \ln q(u,t)\rangle_{q(u(t))}.\end{aligned} \tag{23}$$

For simplicity, consider the case where the parameters θ are known. Then, as in the static case, the variational energy is simply the internal energy, i.e., $V(u(t)) = U(u(t))$. Therefore, the variational density is simply

$$q(u(t)) = \frac{1}{Z}\exp\left(V(u(t))\right). \tag{24}$$

The discussion in Section 2 suggests that the variational density can also be interpreted as an ensemble density. However, the density of an ensemble in the variational energy manifold is now time-dependent. Since the ensemble represent the variational density, the particles in the ensemble are such that the free-action is maximized. Since it is not time-independent, a stationary solution is not available. However, it is expected that the ensemble density will be (nearly) stationary in a frame of reference that moves with the manifold's topology, provided that it does not change too rapidly. A key feature of the generalized coordinates is that they realize this stationarity in a rather simple and elegant manner.

4. DYNAMIC CAUSAL MODELS AND GENERALIZED COORDINATES

The state space or dynamic causal models (DCMs) we consider are defined as follows

$$\begin{aligned}\dot{x}(t) &= f(x(t),\nu(t)) + v(t), \\ y(t) &= h(x(t),\nu(t)) + w(t)\end{aligned} \tag{25}$$

The first set of equations, the *state equations*, implying a coupling between neighboring orders of motion of the hidden states and confer a memory on the system.

Although they are similar in form to the usual state-space models, there is a crucial difference between the DCMs and the usual state-space models studied in the filtering literature. Recall that the noise process in the state-space models are assumed to be Weiner processes, and so are not analytic. In contrast, the noise processes in DCMs are analytic. This is a crucial and important difference with significant ramifications, and central in the use of the generalized coordinates in solving filtering problems.

Consider the state model of the state-space model in Equation 26. The analyticity of the noise process can be exploited by recursively differentiating the state equation with respect to time to obtain the following set of equations:

$$\begin{aligned}\frac{dx}{dt}(t) &= f(x(t),\nu(t)) + v(t), \\ \frac{d^2x}{dt^2}(t) &= \sum_{i=1}^{N}\frac{\partial f}{\partial x_i}(x(t),\nu(t))\frac{dx_i}{dt}(t) + \sum_{i=1}^{N_\nu}\frac{\partial f}{\partial \nu_i}(x(t),\nu(t))\frac{d\nu_i}{dt} + \frac{dv}{dt}(t), \\ \frac{d^3x}{dt^3}(t) &= \sum_{i,j=1}^{N}\frac{\partial^2 f_i}{\partial x_i \partial x_j}(x(t),\nu(t))\frac{dx_i}{dt}(t) + \sum_{i=1}^{N}\frac{\partial f_i}{\partial x_i}(x(t),\nu(t))\frac{d^2x_i}{dt^2}(t) + \\ &\qquad \sum_{i,j=1}^{N_\nu}\frac{\partial^2 f}{\partial \nu_i \partial \nu_j}(x(t),\nu(t))\frac{d^2\nu_i}{dt^2} + \sum_{i=1}^{N_\nu}\frac{\partial f}{\partial \nu_i}(x(t),\nu(t))\frac{d^2\nu_i}{dt^2} + \frac{d^2v}{dt^2}(t), \\ &\vdots\end{aligned} \tag{26}$$

There are infinitely many equations thus available, and it is clear that this expansion can become complicated and unwieldy fairly quickly. However, a great simplification arises when one retains only those terms that are linear in the partial derivatives. This approximation is exact when the state model is linear (as the higher-order derivatives vanish). Then, Equation 27 now becomes

$$\begin{aligned}
\frac{dx}{dt}(t) &= f(x(t),\nu(t)) + v(t), \qquad (27)\\
\frac{d^2x}{dt^2}(t) &= \sum_{i=1}^{N} \frac{\partial f}{\partial x_i}(x(t),\nu(t))\frac{dx_i}{dt}(t) + \sum_{i=1}^{N_\nu} \frac{\partial f}{\partial \nu_i}(x(t),\nu(t))\frac{d\nu_i}{dt} + \frac{dv}{dt}(t),\\
\frac{d^3x}{dt^3}(t) &\approx \sum_{i=1}^{N} \frac{\partial f}{\partial x_i}(x(t),\nu(t))\frac{d^2x_i}{dt^2}(t) + \sum_{i=1}^{N_\nu} \frac{\partial f}{\partial \nu_i}(x(t),\nu(t))\frac{d^2\nu_i}{dt^2} + \frac{d^2v}{dt^2}(t),\\
&\vdots
\end{aligned}$$

In the following, these approximations are treated as equalities: the derivatives are evaluated at each time instant and the linear approximation is local to the current state.

In the continuous-discrete filtering problem, the measurements are given at discrete time instants. Let t_i be a time instant for which measurements are available and let $\tilde{x}(t) \equiv \begin{bmatrix} x(t) & x'(t) & x''(t) & \cdots \end{bmatrix}^T \equiv \begin{bmatrix} x & x' & x'' & \cdots \end{bmatrix}^T$ and $\tilde{y}(t) \equiv \begin{bmatrix} y(t) & y'(t) & y''(t) & \cdots \end{bmatrix}^T \equiv \begin{bmatrix} y & y' & y'' & \cdots \end{bmatrix}^T$. The $\tilde{x}(t)$ and $\tilde{y}(t)$ are referred to as the *generalized coordinates* and the *generalized measurements*, respectively. Then (using subscripts to denote derivatives),

$$\begin{aligned}
x' &= f(x,\nu) + v, & y &= h(x,\nu) + w, \qquad (28)\\
x'' &= f_x(x,\nu)x' + f_\nu(x,\nu)\nu' + v', & y' &= h_x(x,\nu)x' + h_\nu(x,\nu)\nu' + w',\\
x' &= f_x(x,\nu)x'' + f_\nu(x,\nu)\nu'' + v'', & y'' &= h_x(x,\nu)x'' + h_\nu(x,\nu)\nu'' + w'',\\
&\vdots & &\vdots
\end{aligned}$$

The point $\tilde{x}$ can be regarded as encoding the instantaneous trajectory of $x(t)$ at time t. The measurement (observer) equations reveal that the generalized states are needed to generate a generalized response that encodes a path or trajectory.

This formulation can be summarized very compactly as follows (suppressing dependence on x, ν):

$$\begin{aligned}
D\tilde{x} &= \tilde{f} + \tilde{v}, \qquad (29)\\
\tilde{y} &= \tilde{g} + \tilde{w},
\end{aligned}$$

where D is a matrix with whose first-leading diagonal contains identity matrices (of dimension of ν), and

$$\begin{aligned}
\tilde{f} &= \begin{bmatrix} f & f_x x' + f_\nu \nu' & f_x x'' + f_\nu \nu'' & \cdots \end{bmatrix}^T, \qquad (30)\\
\tilde{g} &= \begin{bmatrix} g & g_x x' + g_\nu \nu' & g_x x'' + g_\nu \nu'' & \cdots \end{bmatrix}^T,\\
\tilde{v} &= \begin{bmatrix} v & \dot{v} & \ddot{v} & \cdots \end{bmatrix}^T,\\
\tilde{w} &= \begin{bmatrix} w & \dot{w} & \ddot{w} & \cdots \end{bmatrix}^T.
\end{aligned}$$

5. ENSEMBLE DYNAMICS IN GENERALIZED COORDINATES OF MOTION

In order to construct a scheme based on ensemble dynamics as in the static case, we require the equations of motion for an ensemble whose variational density is stationary in a frame of reference that moves with the mode. This is accomplished by coupling different orders of motion through mean-field effects.

Let $u = \{\nu, \nu'\}$ so that $V(u(t)) := V(\nu, \nu')$ and the induced variational density in generalized coordinates is $q(u(t)) := q(\nu, \nu')$. The following lemma[10] forms the basis of variational filtering and DEM.

Lemma 5.1. *The variational density $q(t,u) = \frac{1}{Z}\exp(V(u(t)))$ is the stationary solution in a moving frame of reference for an ensemble whose equations of motion and ensemble dynamics are*

$$\begin{aligned}\dot{\nu}(t) &= \nabla_\nu V(u(t)) + \mu' + \Gamma(t), \\ \dot{\nu}'(t) &= \nabla_\nu V(u(t)) + \Gamma(t),\end{aligned} \tag{31}$$

where μ' is the mean velocity over the ensemble (a mean field effect).

Proof. Following the steps in Lemma 2.3, the FPKfe reduces to

$$\dot{p}(t,u) = \mu' \nabla_\nu q(u). \tag{32}$$

Under the coordinate transformation $v = \nu - \mu' t$, the change in the ensemble density is zero because

$$\begin{aligned}p(t,v,\nu') &= p(t,\nu - \mu' t, \nu'), \\ \dot{p}(t,v,\nu') &= \dot{p}(t,\nu,\nu') - \mu' \nabla_\nu q(\nu,\nu') = 0.\end{aligned} \tag{33}$$

□

A nice physical interpretation is as follows.[10] The motion of particles is coupled through the mean of the ensemble's velocity. In this moving frame of reference, the particles experience two forces–a deterministic force due to energy gradients that drive the particles to the peak, and the random forces which disperse the particles. The interesting aspect is that the gradients and the peak move with the same velocity and are stationary in the moving frame of reference. This enables particles driven by mean-field effects to easily track the peak.

6. DYNAMIC EXPECTATION MAXIMIZATION AND VARIATIONAL FILTERING

In this section, we conclude by presenting the Bayesian state estimation schemes, DEM and VF.

6.1 Precisions

The temporal dependencies among the random fluctuations are encoded by their temporal precision which can be expressed as a function of their autocorrelations as follows:[17]

$$S(\gamma) = \begin{bmatrix} 1 & 0 & \frac{d^2\rho}{dt^2}(0) & \cdots & \\ 0 & -\frac{d^2\rho}{dt^2}(0) & 0 & \cdots & \\ \frac{d^2\rho}{dt^2}(0) & \cdots & 0 & \frac{d^4\rho}{dt^4}(0) & \cdots \\ \vdots & \vdots & \vdots & \ddots & \end{bmatrix}^{-1} \tag{34}$$

Physically, $\ddot{\rho}(0)$ is a measure of roughness, and the $\ddot{\rho}(0) \to \infty$ corresponds to the state-space model case (Wiener process).

The temporal precision $S(\gamma)$ can be evaluated for any analytic autocorrelation function. When the temporal correlations have the same Gaussian form

$$S(\gamma) = \begin{bmatrix} 1 & 0 & -\frac{1}{2}\gamma & \cdots \\ 0 & \frac{1}{2}\gamma & 0 & \cdots \\ -\frac{1}{2}\gamma & 0 & \frac{3}{4}\gamma^2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{35}$$

where γ is the precision parameter of a Gaussian $\rho(t)$. It is also possible to consider other processes (e.g., $1/f$ noise). Typically, $\gamma > 1$, which ensures that the precisions of the higher-order derivatives converge fairly quickly. For instance, in many cases, an embedding order of $n = 6$ is adequate. In other words, we only consider generalised motion up to order n, because higher orders have nearly zero precision.

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

$$\tilde{\Pi}^v = S(\gamma) \otimes \Pi^v, \tag{36}$$

and likewise for $\tilde{\Pi}^w$–the inverse of $\tilde{\Sigma}^w$.

6.2 Energy Functions: Log-Likelihoods and Prior

Since we have assumed that the parameters are known, the variational energy is the same as the internal energy. The quantity of interest is the energy function $U(t) = \ln p(y, \tilde{x}, \tilde{\nu}|\theta)$. Since

$$\begin{aligned} p(y, \tilde{x}, \tilde{\nu}|\theta) &= \frac{p(y, \tilde{x}, \tilde{\nu}, \theta)}{p(\tilde{x}, \tilde{\nu}, \theta)} \frac{p(\tilde{x}, \tilde{\nu}, \theta)}{p(\tilde{\nu}, \theta)} \frac{p(\tilde{\nu}, \theta)}{p(\theta)}, \\ &= p(\tilde{y}|\tilde{x}, \tilde{\nu}, \theta) p(\tilde{x}|\tilde{\nu}, \theta) p(\tilde{\nu}), \\ &= N(\tilde{y} - \tilde{g}, \tilde{\Sigma}^z) \times N(D\tilde{x} - \tilde{f}, \tilde{\Sigma}^w) p(\tilde{\nu}). \end{aligned} \tag{37}$$

Thus, $\ln p(y, \tilde{x}, \tilde{\nu}|\theta)$ is given by

$$V(u) = -\frac{1}{2} \begin{bmatrix} D\tilde{x} - \tilde{f} & \tilde{y} - \tilde{h} \end{bmatrix} \begin{bmatrix} \tilde{\Pi}^v & \\ & \tilde{\Pi}^w \end{bmatrix} \begin{bmatrix} D\tilde{x} - \tilde{f} \\ \tilde{y} - \tilde{h} \end{bmatrix} \tag{38}$$

6.3 Converting Discrete Measurement Data to Generalized Measurements

A lacuna in our description so far is that we are assuming that the generalized measurements $\tilde{y}$ are available. However, the data is not available in the generalized coordinates of motion; rather only discrete data measurements are available. This impasse is resolved by (yet again!) exploiting analyticity; (local) discrete measurements are generated by the observation function in Equation 30, using the generalised motion of hidden states and a Taylor series.[10]

6.4 DEM and VF

The formalism of Section 5 focused on first order motion but can be extended easily to cover arbitrarily high order motion: the ensuing ensemble dynamics in generalized coordinates $u = \tilde{\nu} = \begin{bmatrix} \nu & \nu' & \nu'' & \cdots \end{bmatrix}^T$ are

$$\dot{u} = \nabla_u V(u) + D\tilde{\mu} + \Gamma(t), \tag{39}$$

where $V(u)$ is given by Equation 38.

Variational filtering simply entails integrating the paths of the multiple particles according to the stochastic differential equations in Equation 39. Note that unlike particle filtering, there is no resampling; all particles are preserved.

DEM (with known parameters and hyperparameters) is the fixed-form homologue of VF. Specifically, DEM approximates the ensemble density by assuming Gaussian form. As a result, this assumption reduces the problem to finding the path of the mode, which entails integrating an ODE that is identitical to Equation 39 but without the random term. The resulting generalised gradient ascent then becomes the D-step of DEM. The conditional covariance follows (analytically) from the curvature of the variational energy.

In this paper, it has been assumed that the parameters and the hyperparameters are known. If not, one can estimate them using the mean-field approximation, in the E- and M-steps of DEM. These are so-called by analogy with the equivalent steps of the Expectation-Maximization algorithm.

7. A RADAR TRACKING EXAMPLE

In previous publications, it has been shown that DEM is capable of tracking states in models that are highly nonlinear (even chatotic), but only when the posterior was unimodal.[11] It was also demonstrated that the VF could handle the multi-modal case.[10]

The model considered here is ubiquitous in the radar tracking literature.[18] The state follows a continuous white noise acceleration model with process noise intensity $\tilde{q}$

$$\begin{bmatrix} x_1(t+T) \\ x_2(t+T) \\ x_3(t+T) \\ x_4(t+T) \end{bmatrix} = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(1) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{bmatrix} + v(t), \tag{40}$$

where the covariance of the process noise is

$$E\left\{v(k)v(k)^T\right\} = \begin{bmatrix} \frac{1}{3}T^3 & \frac{1}{2}T^2 & 0 & 0 \\ \frac{1}{2}T^2 & T & 0 & 0 \\ 0 & 0 & \frac{1}{3}T^3 & \frac{1}{2}T^2 \\ 0 & 0 & \frac{1}{2}T^2 & T \end{bmatrix} \tilde{q} \tag{41}$$

The range and angle measurements are assumed available:

$$\begin{aligned} y_1(t) &= \sqrt{x_1^2 + x_2^2} + w_1(t), \\ y_2(t) &= \tan^{-1}\left(\frac{x_2(t)}{x_1(t)}\right) + w_2(t). \end{aligned} \tag{42}$$

The measurement noises w_1 and w_2 are assumed to be zero mean white Gaussian with standard deviation [10 m,1 m/s]. Since the posterior distribution is uni-modal, only the results for the DEM are shown here.

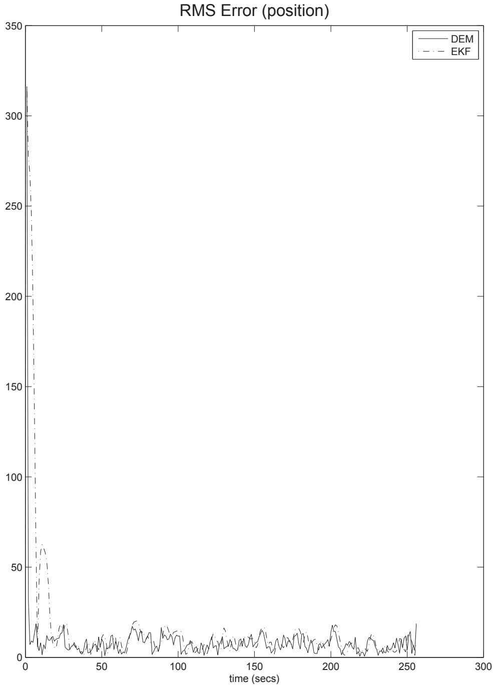

Figure 1. Position RMS Error

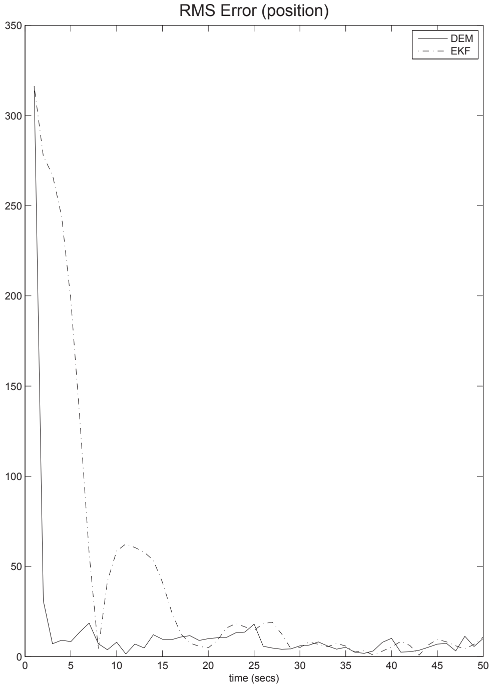

Figure 2. Position RMS Error (zoomed plot)

Figure 1 shows the simulation results for RMS position errors for the two filters, and Figure 2 shows the zoomed version of the same plot. It is clear that at later times the EKF and DEM give essentially identical performance. However, in the initial period, the DEM performance is much better than the EKF—the DEM appears to converge much more quickly than the EKF. The DEM and the KF were initialized in an identical manner.

The normalized estimation error squared (NEES) is also shown in Figure 3. The NEES is a measure of the consistency of the state estimator. It is noted that the NEES is significantly more consistent than the EKF.

Note that the initial condition was not chosen to be consistent, and so the KF performance would be better if the initialization were improved. Howver, the point is that the DEM is very robust, and the consistency is not as significantly impacted. This shows yet another aspect of the superiority of the DEM over the EKF.

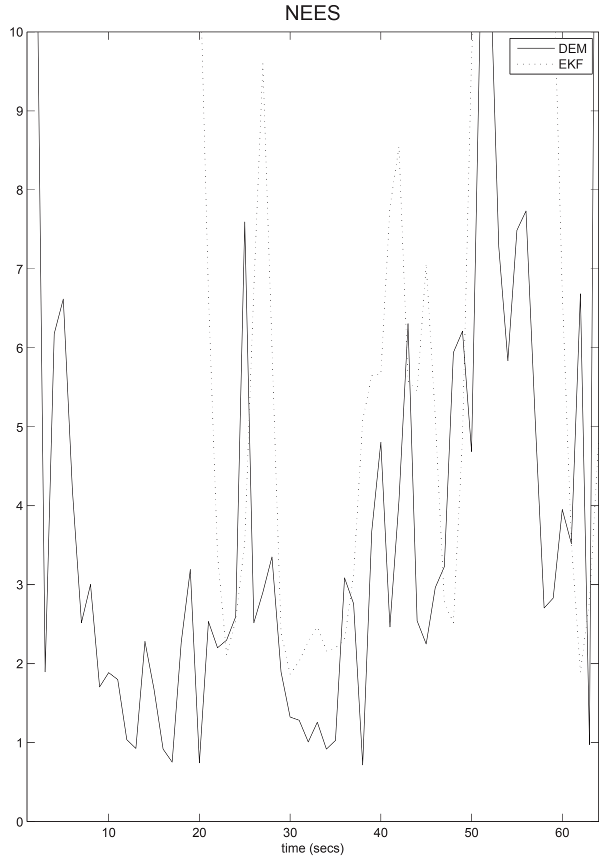

Figure 3. NEES

8. CONCLUSION AND FUTURE WORK

In this paper, a recently proposed novel approach to Bayesian state estimation was reviewed and applied to a radar tracking problem. The relevant concepts and results in variational Bayes were reviewed.

There is a lot of scope for future work. The algorithms based on generalized coordinates and variational Bayes, such as DEM, VF and generalized filtering, need to be compared in terms of performance (and relative to the posterior Cramér-Rao lower bound) for radar tracking problems. The real-time implementation and more accurate numerical implementations also need to be investigated. This shall be reported in future publications.

9. ACKNOWLEDGEMENTS

This work was supported in part by a DRDC Technology Investment Fund (TIF). The DEM toolbox in MATLAB$^{©}$ was used in this paper. It is available freely from `http://www.fil.ion.ucl.ac.uk/spm`.

REFERENCES

[1] Jazwinski, A. H., [*Stochastic Processes and Filtering Theory*], Dover Publications (2007).

[2] Sarkka, S., "On unscented kalman filtering for state estimation of continuous-time nonlinear systems," *Automatic Control, IEEE Transactions on* **52**(9), 1631–1641 (Sept. 2007).

[3] Gordon, N., Salmond, D., and Smith, A., "Novel approach to nonlinear/non-gaussian bayesian state estimation," *Radar and Signal Processing, IEE Proceedings F* **140**, 107–113 (April 1993).

[4] Moral, P. D., [*Feynman-Kăc Formulae*], Springer-Verlag (March 2004).

[5] Bain, A. and Crisan, D., [*Fundamentals of Stochastic Filtering*], Springer-Verlag (2009).

[6] Daum, F. and Huang, J., "Curse of dimensionality and particle filters," in [*Aerospace Conference, 2003. Proceedings. 2003 IEEE*], **4**, 4–1979–4–1993 (8-15, 2003).

[7] Balaji, B., "Estimation of indirectly observable Langevin states: path integral solution using statistical physics methods," *Journal of Statistical Mechanics: Theory and Experiment* **2008**(01), P01014 (17pp) (2008).

[8] Balaji, B., "Universal nonlinear filtering using path integrals II: The continuous-continuous model with additive noise," *PMC Physics A* **3:2** (10 February 2009).

[9] Balaji, B., "Continuous-discrete path integral filtering," *Entropy* **11**(3), 402–430 (2009).

[10] Friston, K., "Variational filtering," *NeuroImage* **41**(3), 747 – 766 (2008).

[11] Friston, K., Trujillo-Barreto, N., and Daunizeau, J., "DEM: A variational treatment of dynamic systems," *NeuroImage* **41**(3), 849 – 885 (2008).

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

[13] Feynman, R. P. and Hibbs, A. R., [*Quantum Mechanics and Path Integrals*], McGraw-Hill Book Company (1965).

[14] MacKay, D. J., [*Information Theory, Inference, and Learning Algorithms*], Cambridge University Press (2003).

[15] Friston, K., "Hierarchical models in the brain," *PLoS Comput Biol* **4**, e1000211 (11 2008).

[16] Risken, H., [*The Fokker-Planck Equation: Methods of Solution and applications*], Springer-Verlag, 2$^{\text{nd}}$ ed. (1999).

[17] Cox, D. R. and Miller, H. D., [*The Theory of stochastic processes*], Methuen, London, UK (1965).

[18] Bar-Shalom, Y., Li, X. R., and Kirubarajan, T., [*Estimation with Applications to Tracking and Navigation*], John Wiley and Sons Inc. (2001).