Variational Bayesian inference for partially observed stochastic dynamical systems

B Wang¹ and D M Titterington²

¹ Department of Mathematics, University of York, Heslington, York YO10 5DD, UK
² Department of Statistics, University of Glasgow, Glasgow G12 8QQ, UK.

E-mail: ¹bw527@york.ac.uk ²mike@stats.gla.ac.uk

Abstract. In this paper the variational Bayesian approximation for partially observed continuous time stochastic processes is studied. We derive an EM-like algorithm and describe its implementation. The variational Expectation step is explicitly solved using the method of conditional moment generating functions and stochastic partial differential equations. The numerical experiments demonstrate that the variational Bayesian estimate is more robust than the EM algorithm.

1. Introduction

For most models of interest involving missing data a full Bayesian analysis is computationally complex because complicated multiple integrations are involved. Markov chain Monte Carlo for numerical integration helps to side-step this problem, but it is clearly time-consuming, samples of parameter values have to be stored and it may not be clear whether or not convergence has occurred.

Recently, a deterministic approximate approach to intractable Bayesian learning problems, the variational Bayesian approximation, has been introduced in the neural-computing literature, for instance in [1 - 3], and is widely recognised to be effective [4 - 9]. Variational Bayes draws together variational ideas from the analysis of intractable latent variable models [10] and from Bayesian inference [11, 12]. This framework facilitates calculation of approximations to posterior distributions over the hidden variables, parameters and structures. They are computed via an iterative algorithm that is closely related to the Expectation-Maximisation (EM) algorithm and its convergence is guaranteed in an analogous way.

Suppose that $\theta$ represents the set of parameters in the model and that $x$ and $y$ denote respectively the missing and observed data. Bayesian inference centres on the posterior distribution of $\theta$, given $y$, which may in theory be obtained as the marginal distribution from the joint distribution of $\theta$ and $x$, given $y$. In the variational approach, this joint posterior density is approximated by a density $q_y(x, \theta)$, chosen to maximise

$$\int q_y(x, \theta) \log \frac{p(x, y, \theta)}{q_y(x, \theta)} dx d\theta,$$

which is equivalent to minimising the Kullback-Leibler divergence between the exact and approximate joint distributions of $\theta$ and $x$, given $y$. To be at all useful, the approximation
some or even all of the dependences between the variables. Typically \( q_\theta(x, \theta) \) is assumed to take a factorised form, one factor involving only \( x \) and the other involving only \( \theta \). Given a suitable prior on \( \theta \), the factors corresponding to the parameters turn out to have the same distributional form as the conjugate family that would obtain were there no missing data, if such a conjugate family exists; the appropriate values of the hyperparameters are obtained by optimisation.

The literature concerning the variational Bayesian method has been dominated by contexts involving discrete-time models, such as hidden Markov model [1], graphical models [2, 3], mixture models [5, 8, 9] and state space models [6, 7]. In this paper we apply the variational Bayesian approach to linear partially observed continuous-time stochastic processes. This kind of model is widely used in the fields of signal filtering, prediction and control. The EM algorithm for calculating the Maximum Likelihood estimate or Maximum A-Posteriori estimate for partially observed diffusions was developed in [13]. In [14] this algorithm was compared with the direct maximisation method for the MLE of parameters in these models. In [15, 16] the implementation of the E-step was simplified by the use of finite-dimensional filters for integral processes. The purpose of this paper is to derive the variational Bayesian algorithm for continuous-time linear Gaussian systems. In this setting the variational posterior densities for the unobserved states have to be properly replaced by Radon-Nikodym derivatives. We shall see that the resulting algorithm preserves the basic features of the variational methods in discrete-time contexts. Moreover, as in the discrete-time case, there are similarities between the variational Bayesian method and the EM algorithm for the partially observed diffusions model [13]. For example, the Expectation steps of both algorithms normally involve Kalman filtering and smoothing. In this paper we employ the method of conditional moment generating functions for integrals and stochastic integrals to deal with the variational Expectation step without using Kalman smoothing.

Section 2 presents the statistical model and the prior distributions of the parameters. In Section 3 we derive the EM-like algorithm, and we show how to implement it in Section 4. Subsection 4.1 implements the variational maximisation step, and Subsection 4.2 describes the variational expectation step, although only sketchily; full details are available in a technical report by the authors. Some numerical experiments are reported in Section 5, and we conclude in Section 6 with some discussion.

2. Statistical model

On a measurable space \((\Omega, \mathcal{A}, P)\) the following are given:

(a) a family of probability measures \(\mathcal{M} = \{P_\theta, \theta \in \Theta\}\) depending on a parameter \(\theta \triangleq (A, C)\);
(b) a pair of stochastic processes \(X = \{X_t, 0 \leq t \leq T\}\) and \(Y = \{Y_t, 0 \leq t \leq T\}\) taking values in \(\mathbb{R}^m\) and \(\mathbb{R}^d\), respectively, such that, under \(P_\theta\),

\[
\begin{align*}
    dX_t &= AX_t dt + BdW_t, & X_0 &= \xi, \\
    dY_t &= CX_t dt + DdV_t, & Y_0 &= 0,
\end{align*}
\]

where \(\{W_t, 0 \leq t \leq T\}\) and \(\{V_t, 0 \leq t \leq T\}\) are respectively \(m\)-dimensional and \(d\)-dimensional independent standard Brownian motions, and \(\xi\) is an \(\mathbb{R}^m\)-valued Gaussian variable, independent of the Brownian motions, with mean \(\lambda_0\) and covariance matrix \(\Pi_0\), and with density denoted by \(p_0(x)\).

Unlike in the case of discrete-time state-space models, the probability measures induced by stochastic differential equations driven by Brownian motions with different covariance structures are not mutually absolutely continuous, so it is not possible to estimate \(B\) and \(D\) by the maximum likelihood method. In fact these parameters can be estimated in terms of the quadratic variations...
of the observation and filtered state processes; see, for example, [15, 17]. Therefore we assume that \( B \) and \( D \) are known and there is no loss of generality in taking them to be identity matrices.

Assume that \( \Omega \) is the canonical space \( C([0, T]; \mathbb{R}^{m+d}) \), in which case \( X \) and \( Y \) are the canonical processes on \( C([0, T]; \mathbb{R}^m) \) and \( C([0, T]; \mathbb{R}^d) \), respectively, and \( P_\theta \) is the probability law of \( (X, Y) \). Here \( X = \{X_t, 0 \leq t \leq T\} \) is the state process, which is not directly observed; rather, the information about its evolution is obtained through the noisy observed process \( Y = \{Y_t, 0 \leq t \leq T\} \).

Let \( \mathcal{Y}_T \) denote the \( \sigma \)-algebra generated by the process \( Y \), and let \( P^Y_\theta \) denote the restriction of \( P_\theta \) to \( \mathcal{Y}_T \). The framework defined in (1) and (2) ensures that the probability measures in \( \mathcal{M} \) are mutually absolutely continuous. Thus, if we let \( \theta_0 = (A_0, C_0) \) be the reference parameter and write \( P_{\theta_0} \) as \( P_0 \), according to [18], the likelihood function for estimating the parameter \( \theta \) on the basis of a given observation path \( \{Y_t, 0 \leq t \leq T\} \) can be expressed as

\[
L(\theta) \triangleq \frac{dP^Y_\theta}{dP^0_0} = \mathbb{E}_0(Z^\theta_T|\mathcal{Y}_T),
\]

where \( \mathbb{E}_0 \) denotes the expectation under \( P_0 \), and \( Z^\theta_T \) is the Radon-Nikodym derivative of \( P_\theta \) with respect to \( P_0 \); that is

\[
Z^\theta_T \triangleq \frac{dP_\theta}{dP_0} = \exp\left\{ \int_0^T (AX_t - A_0X_t)\top dW_t - \frac{1}{2} \int_0^T |AX_t - A_0X_t|^2 dt \right. \\
+ \left. \int_0^T (CX_t - C_0X_t)\top dV_t - \frac{1}{2} \int_0^T |CX_t - C_0X_t|^2 dt \right\} \\
= \exp\left\{ \int_0^T (AX_t - A_0X_t)\top dW_t - \frac{1}{2} \int_0^T |AX_t - A_0X_t|^2 dt \right. \\
+ \left. \int_0^T (CX_t - C_0X_t)\top dY_t - \frac{1}{2} \int_0^T X_t\top (C - C_0)\top (C + C_0)X_t dt \right\}, \tag{3}
\]

which is also called the “complete data” likelihood [13].

Using the Bayesian approach, we assign a prior density \( p_0(\theta) \) with the following structure to the unknown parameters: the \( i \)th row vector of the \( A \) matrix, denoted by \( a_i^\top \), is given a Gaussian prior with mean \( \mu_0^i \) and covariance matrix equal to a diagonal matrix \( \alpha_0 \); the \( i \)th row vector of \( C \), denoted by \( c_i^\top \), is given a Gaussian prior with mean \( \nu_0^i \) and covariance matrix equal to a diagonal matrix \( \beta_0 \); and all these row and column vectors are assumed independent. Thus the posterior probability distribution of the parameter \( \theta \) is defined by Bayes’ Theorem as

\[
p(\theta|Y) = \frac{L(\theta)p_0(\theta)}{\int L(\theta)p_0(\theta)d\theta} = \frac{\mathbb{E}_0(Z^\theta_T|\mathcal{Y}_T)p_0(\theta)}{\int \mathbb{E}_0(Z^\theta_T|\mathcal{Y}_T)p_0(\theta)d\theta}.
\]

As pointed out in [4, 7], the exact Bayesian treatment would require us to compute marginals of the posterior distribution over all the unknown parameters and hidden states. This involves the cross-integration terms of up to fourth order; for example, expression (3) contains terms in the exponent of the form \( |CX_t - C_0X_t|^2 \). Therefore, integrating over the parameters and the hidden states would be time-consuming. Now we consider the variational approximation for this continuous-time partially-observed diffusion model, in which the approximate posteriors are computed via an iterative algorithm.
3. The variational Bayesian treatment
The basic idea of the variational Bayesian method is simultaneously to approximate the
intractable joint distribution of both hidden states and parameters with a simpler distribution,
usually of a factorised form, corresponding to assuming that the hidden states and parameters
are independent. In the case of the continuous-time model, the idea is the same but we need to
deal with a Radon-Nikodym derivative rather than a density for the variational posterior of the
hidden state.
Denote by $P_0^X\mid Y$ the probability law of $X$ conditional on the observed process $Y$ with the
fixed parameter $\theta_0$. The ensemble loglikelihood can thus be expressed as
\[
\log \int L(\theta)p_0(\theta)d\theta = \log \int \mathbb{E}_0(Z_T^0|Y)p_0(A)p_0(C)dAdC
= \log \int Z_T^0 \cdot p_0(A)p_0(C)P_0^{X|Y}(dX)dAdC.
\]
We use an approximating conditional distribution $\tilde{Q}(X, A, C)$, which factorises as follows:
\[
\int \Phi(X, A, C)d\tilde{Q}(X, A, C) = \int \Phi(X, A, C)Q_A(A)Q_C(C)Q_X(dX)dAdC
\]
for any function of interest $\Phi$, where $Q_X(\cdot)$ is a probability measure on the space $C([0, T]; \mathbb{R}^m)$.

By Jensen’s inequality we have
\[
\log \int L(\theta)p_0(\theta)d\theta
= \log \int Z_T^0 \cdot \frac{dP_0^{X|Y}}{dQ_X} \cdot \frac{p_0(A)}{Q_A(A)} \cdot \frac{p_0(C)}{Q_C(C)} \cdot Q_A(A)Q_C(C)Q_X(dX)dAdC
\geq \int \log \left\{ \frac{Z_T^0 \cdot p_0(A)p_0(C)}{\frac{dQ_X}{dP_0^{X|Y}}} \cdot Q_A(A)Q_C(C) \right\} Q_A(A)Q_C(C)Q_X(dX)dAdC
\]
\[\triangleq \mathcal{F}(Q_X(X), Q_A(A), Q_C(C)).\]

The function $\mathcal{F}$ is called free energy. For variational Bayesian learning the free energy $\mathcal{F}$
is the key quantity with which we work. Learning proceeds with iterative updates of the
variational posteriors. The optimum forms of these approximate posteriors can be found by
taking functional derivatives of $\mathcal{F}$ with respect to each distribution. This results in the following
theorem.

**Theorem 1.** The free energy $\mathcal{F}$ is maximised by the following distributions:
(i) the optimal variational posteriors of the parameters are multivariate Gaussian and satisfy
\[
Q_A(A) \propto \exp\{\log Z_T^0\}_X p_0(A), \quad (4)
Q_C(C) \propto \exp\{\log Z_T^0\}_X p_0(C), \quad (5)
\]
where $\langle \cdot \rangle_X$ denotes expectation under $Q_X(X)$;
(ii) the optimal variational posterior of the unobserved state satisfies
\[
\frac{dQ_X}{dP_0^{X|Y}} \propto \exp\{\log Z_T^0\}_\theta,
\]
where $\langle \cdot \rangle_\theta$ denotes expectation with respect to $Q_A(A)$ and $Q_C(C)$. 


Proof. The proof of (i) is straightforward if we solve the following variational equations associated with the functional $F$:

$$\frac{\delta F}{\delta Q_A(A)} = 0 \quad \text{and} \quad \frac{\delta F}{\delta Q_C(C)} = 0.$$ 

For (ii), the functional $F$ can be rewritten as

$$F = \int \left[ \langle \log Z^\theta_T \rangle \theta - \log \frac{dQ_X}{dP_X^X|Y} \right] Q(dX) + \text{constant independent of } Q(X)$$

$$= \int \left[ \langle \log Z^\theta_T \rangle \theta - \log \frac{dQ_X}{dP_X^X|Y} \right] \cdot \frac{dQ_X}{dP_X^X|Y} \cdot P_X^X|Y(dX) + \text{constant}.$$ 

Noting that $P_X^X|Y(\cdot)$ is a reference probability measure and is independent of $Q(X)$, we take the functional derivatives of $F$ with respect to $dQ_X/dP_X^X|Y$, and (ii) follows.

Theorem 1 suggests that we can maximise the free energy numerically by iterating between the variational posteriors given by (i) and (ii). Therefore, we obtain the following EM-like algorithm.

VE Step: compute $\langle \log Z^\theta_T \rangle_X$, and obtain the variational posteriors of $A$ and $C$ according to (i): $a_i \sim N(\mu_i, \alpha_i)$ and $c_i \sim N(\nu_i, \beta_i)$.

VM Step: compute $\langle \log Z^\theta_T \rangle_\theta$, and obtain the variational posterior of the hidden state according to (ii), namely, the Radon-Nikodym derivative of the variational distribution with respect to the reference measure $P_X^X|Y$.

4. Implementation of the variational approximation

Practical implementation of the iterative algorithm provided in the previous section is not simple, mainly because the VE step involves the conditional expectation under a measure on the space of continuous functions on $[0, T]$. In general, this kind of expectation results in nonlinear filtering and smoothing problems which do not have any analytical form. However, under our linear setting, finite-dimensional solutions exist, as demonstrated in the sequel.

4.1. VM step

Since the variational posteriors of $A$ and $C$ are Gaussian the computation of the VM step is straightforward. From (6) one has

$$\frac{dQ_X}{dP_X^X|Y} \propto \exp\{ \langle \log Z^\theta_T \rangle \theta \}$$

$$= \exp\{ \int_0^T X_t^\top \langle (A - A_0)^\top \rangle_A dW_t - \frac{1}{2} \int_0^T X_t^\top \langle (A - A_0)^\top (A - A_0) \rangle_A X_t dt \}
\quad + \int_0^T X_t^\top \langle (C - C_0)^\top \rangle_C dY_t - \frac{1}{2} \int_0^T X_t^\top \langle (C - C_0)^\top (C + C_0) \rangle_C X_t dt \}
\triangleq \Gamma_T.$$ 

Define $\bar{U}^\top = (\mu_1, \cdots, \mu_m)$, $\bar{V}^\top = (\nu_1, \cdots, \nu_m)$ and

$$F_1 = \bar{U} - A_0, \quad H_1 = (\bar{U} - A_0)^\top (\bar{U} - A_0) + \text{diag}(\text{tr}(\alpha_1), \cdots, \text{tr}(\alpha_m)).$$
\( F_2 = \bar{V} - C_0, \quad H_2 = (\bar{V} - C_0)\top (\bar{V} - C_0) + \text{diag}(\text{tr}(\beta_1), \ldots, \text{tr}(\beta_d)). \)

Then \( \Gamma_T \) can be rewritten as

\[
\Gamma_T = \exp\left\{ \int_0^T X_t\top F_1 \, dW_t - \frac{1}{2} \int_0^T X_t\top H_1 X_t \, dt \right. \\
+ \left. \int_0^T X_t\top F_2 \, dY_t - \frac{1}{2} \int_0^T X_t\top H_2 X_t \, dt \right\}.
\]

Therefore, we obtain \( dQ_X / dP_0^{X|Y} = K \Gamma_T \). Since

\[
\int dQ_X / dP_0^{X|Y} \cdot P_0^{X|Y} (dX) = 1,
\]

it follows that

\[
K = \left( \int \Gamma_T \cdot P_0^{X|Y} (dX) \right)^{-1} = \left( \mathbb{E}_0 [\Gamma_T | Y_T] \right)^{-1}.
\]

This Radon-Nikodym derivative implies that \( \{X_t, 0 \leq t \leq T\} \) is no longer a diffusion process under the variational distribution \( Q_X \); that is, it does not satisfy any stochastic differential equation under \( Q_X \).

4.2. VE step

Given an observation path \( \{Y_t, 0 \leq t \leq T\} \), \( Z^\theta_T \) is a functional on the space \( C([0,T]; \mathbb{R}^m) \) of \( \mathbb{R}^m \)-valued continuous functions on \([0,T] \). Now we consider the expectation of \( \log Z^\theta_T \) under the probability measure \( Q_X \) on \( C([0,T]; \mathbb{R}^m) \).

From the VM step, we have

\[
\langle \log Z^\theta_T \rangle_X = \int \log Z^\theta_T \cdot Q(dX) = \int \log Z^\theta_T \cdot \frac{dQ_X}{dP_0^{X|Y}} \cdot P_0^{X|Y} (dX) = \mathbb{E}_0 \left[ \log Z^\theta_T \cdot \frac{dQ_X}{dP_0^{X|Y}} | Y_T \right].
\]

Hence, from (4), (5) and (3), we obtain, for each \( i \),

\[
Q_A(A) \propto \exp \left\{ \mathbb{E}_0 \left[ \log Z^\theta_T \cdot \frac{dQ_X}{dP_0^{X|Y}} | Y_T \right] \right\} p_0(A)
\]

\[
\propto \exp \left\{ \mathbb{E}_0 \left[ \int_0^T X_t\top (A - A_0)^\top dW_t \cdot \frac{dQ_X}{dP_0^{X|Y}} | Y_T \right] \\
- \frac{1}{2} \mathbb{E}_0 \left[ \int_0^T X_t\top (A - A_0)^\top (A - A_0) X_t dt \cdot \frac{dQ_X}{dP_0^{X|Y}} | Y_T \right] \right\} p_0(A),
\]

which gives

\[
Q_A(A_i^\top) \sim \mathcal{N}((S_i + a_{0,i}^\top \tilde{S} + \mu_0 \chi_{0^{-1}})[\alpha_0^{-1} + \tilde{S}]^{-1}, [\alpha_0^{-1} + \tilde{S}]^{-1}),
\]
then the conditional expectations (7)-(9) can be computed as

\begin{align}
\tilde{S} & = \mathbb{E}_0 \left[ \int_0^T X_t X_t^\top dt \cdot \frac{dQ_X}{dP_0^X|\mathcal{Y}_T} \right], \\
S_i & = \mathbb{E}_0 \left[ \int_0^T X_t^\top dW_t^i \cdot \frac{dQ_X}{dP_0^X|\mathcal{Y}_T} \right],
\end{align}

where \{W_t^i\} is the \(i\)th component of \{W_t\}.

Similarly,

\[ Q_C(c_t^\top) \sim \mathcal{N}((\tilde{S}_t + a_{0,t}^\top \tilde{S} + \nu_0^\top \beta_0^{-1})[\beta_0^{-1} + \tilde{S}]^{-1}, [\beta_0^{-1} + \tilde{S}]^{-1}) , \]

where

\[ \tilde{S}_t = \mathbb{E}_0 \left[ \int_0^T X_t^\top dY_t^i \cdot \frac{dQ_X}{dP_0^X|\mathcal{Y}_T} \right], \]

in which \{Y_t^i\} is the \(i\)th component of \{Y_t\}.

Therefore we see that, similarly to the discrete-time case [4, 6, 7], in general, the VE step involves filtered estimates of the (stochastic) integrals of the unobserved state and their exponents that normally have to be computed via a Kalman smoothing procedure. These do not have any analytical form and require large memory in any numerical implementation. In laying the groundwork for their version of the EM algorithm, Elliott and Krishnamurthy [15] developed the method of conditional moment generating functions for integrals and stochastic integrals in order to compute the estimates such as \( \mathbb{E}_0[\int_0^T X_t^\top dY_t^i|\mathcal{Y}_T] \), \( \mathbb{E}_0[\int_0^T X_t^\top dW_t^i|\mathcal{Y}_T] \) and \( \mathbb{E}_0[\int_0^T X_t^\top dY_t^i|\mathcal{Y}_T] \). Next we employ this technology to obtain the explicit form of the VE step without using Kalman smoothing but using the Kalman filter, which is much simpler than the former.

Instead of using a different group of stochastic partial differential equations for calculating the conditional expectation of each integral process as presented in [15, 16], we use a method of ensemble computation in which only a single group of stochastic partial differential equations is needed with all the conditional expectations (7)-(9).

In fact, if we write

\[ \mathcal{Y}_T = \int_0^T X_t^\top R_1 X_t dt + \int_0^T X_t^\top R_2 dW_t + \int_0^T X_t^\top R_3 dY_t , \]

then the conditional expectations (7)-(9) can be computed as

\[ \mathbb{E}_0 \left[ \mathcal{Y}_T \cdot \frac{dQ_X}{dP_0^X|\mathcal{Y}_T} \right] , \]

for suitable matrices \( R_1 \in \mathbb{R}^{m \times m} \), \( R_2 \in \mathbb{R}^{m \times m} \) and \( R_3 \in \mathbb{R}^{m \times d} \).

After considerable algebra, details of which are available in the accompanying technical report, we finally achieve that

\[ \frac{\mathbb{E}_0}{\mathbb{I}} \mathcal{Y}_T \cdot \frac{dQ_X}{dP_0^X|\mathcal{Y}_T} \]

\[ = \int_0^T \left[ \text{tr}(Q_0^0 K_0 + P_0^0 K_0) + (r_0^0)^\top \left( \frac{1}{2} G_0^\top G_0 - F_0 \right) \hat{X}_t^0 \\
+ (\tilde{X}_t^0)^\top \left( \frac{1}{2} G_0^\top G_0 - F_0 \right) r_0^0 + (\tilde{X}_t^0)^\top \left( \frac{1}{2} G_0^\top G_0 + \frac{1}{2} G_0^\top G_0 - F_0 \right) \hat{X}_t^0 \right] dt \\
+ \int_0^T (\tilde{H}_t \hat{X}_t^0 + H_0 r_0^0)^\top dY_t . \]
In the above,

\[ F_0 = \frac{1}{2} G_0^T C_0 + \frac{1}{2} H_1 + \frac{1}{2} H_2, \quad G_0 = F_1, \]

\[ H_0 = F_2 + C_0, \quad K_0 = \frac{1}{2} (G_0^T G_0 + H_0^T H_0 - 2F_0), \]

\[ \bar{F}_0 = -R_1, \quad \bar{G}_0 = R_2, \quad \bar{H}_0 = R_3, \]

\[ K_0 = \frac{1}{2} (G_0^T G_0 + G_0^T G_0 + H_0^T H_0 + H_0^T H_0 - 2F_0). \]

In addition, \( \hat{X}_t^0 \) and \( P_t^0 \) satisfy respectively the stochastic and ordinary differential equations

\[
\begin{align*}
d\hat{X}_t^0 &= (A_0 + 2P_t^0 K_0 + G_0 - P_t^0 H_0^T H_0) \hat{X}_t^0 dt + \bar{F}_0^T dY_t, \quad \hat{X}_0^0 = \lambda_0, \\
\dot{P}_t^0 &= (A_0 + G_0) P_t^0 + P_t^0 (A_0 + G_0)^T - P_t^0 (H_0^T H_0 - 2K_0) P_t^0 + 1, \quad P_0^0 = \Pi_0,
\end{align*}
\]

and \( r_t^0 \) and \( \Sigma_t^0 \) satisfy

\[
\begin{align*}
dr_t^0 &= (A_0 + P_t^0 G_0^T G_0 - 2P_t^0 F_0 + G_0) r_t^0 dt + (\Sigma_t^0 H_0^T + P_t^0 H_0^T) dY_t \\
&+ \Sigma_t^0 (G_0^T G_0 - 2F_0) + G_0 \hat{X}_t^0 dt, \\
\dot{\Sigma}_t^0 &= (A_0 + G_0 + P_t^0 G_0^T G_0 - 2P_t^0 F_0) \Sigma_t^0 \\
&+ \Sigma_t^0 (A_0 + G_0 + P_t^0 G_0^T G_0 - 2P_t^0 F_0)^T \\
&+ P_t^0 G_0^T + \bar{G}_0 P_t^0 + P_t^0 (G_0^T G_0 + G_0^T G_0 - 2\bar{F}_0) P_t^0,
\end{align*}
\]

with initial conditions \( r_0^0 = \Sigma_0^0 = 0. \)

Thus the integrals (7), (8) and (9) can be computed correspondingly for suitable \( R_1, R_2 \) and \( R_3 \), and the VE step is implemented explicitly.

5. Numerical experiments

In this section, we look at the special case of the one-dimensional autoregressive system,

\[
\begin{align*}
dX_t &= AX_t dt + dW_t, \quad X_0 = 0, \\
dY_t &= hX_t dt + dV_t, \quad Y_0 = 0,
\end{align*}
\]

where \( A \) is an unknown parameter which is assigned a normal prior distribution with mean \( \mu_0 \) and variance \( \alpha_0 \), and \( h \) is a known gain. We seek the variational posterior of \( A \). The reference parameter \( A_0 \) is taken to be 0.

VM step: Since, according to the variational posterior for \( A, A \sim \mathcal{N}(\mu, \alpha) \), we have

\[
\frac{dQ_A}{dP_0^{XY}} \propto \exp\left\{ \int_0^T \mu X_t dW_t - \frac{1}{2} \int_0^T (\mu^2 + \alpha) X_t^2 dt \right\}. \tag{13}
\]

VE step: for our special model we have

\[
Q_A(A) \sim \mathcal{N}\left( \frac{\mu_0 + b\alpha_0}{1 + a\alpha_0}, \frac{\alpha_0}{1 + a\alpha_0} \right), \tag{14}
\]

with

\[
\begin{align*}
a &= \mathbb{E}_0 \left[ \int_0^T X_t^2 dt \cdot \frac{dQ_A}{dP_0^{XY}[Y_T]} \right], \\
b &= \mathbb{E}_0 \left[ \int_0^T X_t dW_t \cdot \frac{dQ_A}{dP_0^{XY}[Y_T]} \right].
\end{align*}
\]
We take $R_3 = 0$ in (10), so that

$$F_0 = \frac{1}{2}(\mu^2 + \alpha + h^2), \quad \bar{F}_0 = -R_1,$$

$$G_0 = \mu, \quad \bar{G}_0 = R_2, \quad H_0 = h, \quad \bar{H}_0 = 0,$$

$$K_0 = \frac{1}{2} \alpha, \quad \bar{K}_0 = R_1 + R_2 \mu.$$

The corresponding $\hat{X}_t^0, P_t^0, r_t^0$ and $\Sigma_t^0$ satisfy the equations

$$d\hat{X}_t^0 = [\mu - (h^2 + \alpha)P_t^0] \hat{X}_t^0 dt + hP_t^0 dY_t,$$

$$dP_t^0 = 2\mu P_t^0 - (h^2 + \alpha)(P_t^0)^2 + 1,$$

$$dr_t^0 = [\mu - (h^2 + \alpha)P_t^0] r_t^0 dt + h\Sigma_t^0 dY_t$$

$$+ [R_2 + 2(R_1 + R_2 \mu)P_t^0 - (h^2 + \alpha)\Sigma_t^0] \hat{X}_t^0 dt,$$

$$d\Sigma_t^0 = 2[\mu - (h^2 + \alpha)P_t^0] \Sigma_t^0 + 2R_2 P_t^0$$

$$+ 2(\mu R_2 + R_1) (P_t^0)^2,$$

with initial conditions $\hat{X}_0^0 = P_0^0 = r_0^0 = \Sigma_0^0 = 0$.

Thus, we have

$$E_0 \left[ Y_T \cdot \frac{dQ_X}{dP_0^X} | Y_T \right]$$

$$= \int_0^T \left[ -\frac{1}{2} \Sigma_s^0 \alpha + P_s^0 (R_1 + R_2 \mu) - (h^2 + \alpha) r_s^0 \hat{X}_s^0 + (R_1 + R_2 \mu) (\hat{X}_s^0)^2 \right] ds$$

$$+ \int_0^T h r_s^0 dY_s.$$

To compute $a$ and $b$, we take $R_1 = 1$, $R_2 = 0$ and $R_1 = 0$, $R_2 = 1$ respectively in the last expression.

We choose $T = 20$ and discretise time into steps of $\Delta t = 0.01$. The SDE’s (11) and (12) are simulated by the Euler time-discretisation scheme. For example, (11) is approximated by

$$x_{n+1} = x_n + Ax_n \Delta t + \sqrt{\Delta t} \cdot w_n,$$

in which $\{w_n\}$ is a standard Gaussian white noise sequence. Solution of equations (15)-(18) is described in our technical report; alternatively they can be sampled by a discretisation scheme. We generate a sample of size $T / \Delta t$ using $A = -1$ and $h = 10$ so that our results are comparable with [13], and then compute the mean and variance of $A$’s variational posterior by iterating (13) and (14) using different hyperparameters ($\mu_0, \alpha_0$) in the prior distributions. Some of the experimental results are summarised in Table 1.

The parameter $A$ can also be estimated by the maximum likelihood method, and the estimate can be computed using the EM algorithm; see [13, 15]. In Table 2 we list the computational results using the method of [15] for different initial guesses $\mu_0$.

The experiments show that the VEM and EM algorithms have almost the same computational burden, though Bayesian inference needs one to evaluate the normalising constant more often than in the maximum likelihood method, and both become close to the true value quickly, even from initial guesses that are far from the true values. However, from Tables 1 and 2 we see that, if it converges, the EM algorithm always converges to the same point for different initial values, whereas the stationary points of the VEM algorithm differ a little, though all of them
Table 1. Experimental results for the VEM algorithm, displayed as posterior means, with variances in parentheses. Results are shown for the first two iterations and then iterations at which the various means have converged to 4 decimal places. The estimates are not displayed after convergence has occurred.

| Prior | -1.0 (4.0) | -10.0 (100.0) | 10.0 (100.0) | -50.0 (100.0) | 20.0 (100.0) | 25.0 (100.0) |
|-------|------------|---------------|-------------|--------------|-------------|-------------|
| Iteration 1 | -1.0156 (0.0950) | -5.6210 (0.5512) | -6.5221 (0.5695) | -47.6119 (4.5237) | -28.9921 (1.8395) | -91.0143 (4.5063) |
| 2 | -0.9518 (0.0889) | -1.5001 (0.1457) | -1.6849 (0.1674) | -44.0303 (4.1832) | -20.8580 (2.1206) | -80.3540 (8.2243) |
| 3 | -0.9502 (0.0887) | -0.9764 (0.0929) | -0.965 (0.0938) | -40.0878 (3.8082) | -11.7174 (1.1898) | -71.2609 (7.2935) |
| 5 | -0.9574 (0.0906) | -0.9389 (0.0906) | -30.4072 (2.8875) | -1.1521 (0.1145) | -55.9601 (5.7266) |
| 8 | -7.8165 (0.7401) | -0.9296 (0.0906) | -35.5292 (3.6338) | NaN (0.0906) | 97.4681 (0.0906) |
| 13 | -0.9946 (0.0908) | -1.0685 (0.1063) | 83.1805 (0.1063) | 112.2916 (0.1063) | 115.8845 (0.1063) |

Table 2. EM experimental results. ‘NaN’ means ‘not a number’ and indicates that convergence has failed. Results are shown for the first iteration and then iterations at which the various means have converged to 4 decimal places. The estimates are not displayed after convergence has occurred.

| Initial Value | -1.0 | -10.0 | 10.0 | -50.0 | 20.0 | 25.0 |
|--------------|------|------|------|------|------|------|
| Iteration 1  | -0.9487 | -2.9777 | -4.4837 | -46.3949 | -257.7984 | 71.1369 |
| 2            | -0.9475 | -1.0688 | -1.2629 | -42.3395 | NaN | 74.4490 |
| 5            | -0.9475 | -0.9475 | -25.6663 | -57.9601 | 83.1805 |
| 11           | -0.9475 | -0.9475 | -25.6663 | 83.1805 | 97.4681 |
| 16           | -0.9475 | -0.9475 | -25.6663 | -57.9601 | 112.2916 |
| 17           | -0.9475 | -0.9475 | -25.6663 | -57.9601 | 115.8845 |

are close to the true value; this is a consequence of the different values taken by the prior’s hyperparameters. On the other hand, when the initial values are too far away from the true values (as in the cases $\mu_0 = 20.0, 25.0$ in Tables 1 and 2), the EM algorithm no longer converges but the VEM algorithm may still converge to a point close to the true value. This indicates that the VEM algorithm is more robust than the EM algorithm.

6. Discussion
If data are observed only at discrete times, a similar variational algorithm can be developed along the above lines and using the technology introduced in [19]. This algorithm will be the same as that presented in [6, 7] for state-space models, but the implementation is different. The
former uses only the Kalman filter, whereas both the Kalman filter and smoother are needed for the latter.

Our method can also be applied to certain more general cases than linear models. If the systems are linear in the parameters and nonlinear in the state, similar algorithms can be developed and explicit implementation exists for certain models, such as the Benes processes [20], as well as the systems discussed in [21] and in [22, 23]. The VE step cannot be computed explicitly for general nonlinear models, but the Galerkins approximation can be used to compute the conditional expectations involved in the VE step; see for example [17].

Variational inference for stochastic dynamical systems is receiving more and more attention because of its potentially wide applications. While we were finishing this paper, we became aware of a project entitled “Variational Inference in Stochastic Dynamic Environmental Models” (VISDEM) [24]. One of the aims of this project is to study systematically the application of variational Bayesian methods to inference in stochastic dynamical systems and to investigate their applications in weather and climate forecasting.

Acknowledgments
This work was supported by a grant from the UK Science and Engineering Research Council. This version was compiled while the second author was a Visiting Fellow on the ‘Statistical Theory and Methods for Complex High-Dimensional Data’ Research Programme at the Isaac Newton Institute for Mathematical Sciences in Cambridge.

References
[1] MacKay D J C 1997 Cavendish Laboratory Report, University of Cambridge
[2] Attias H 1999 Proc. 15th Conf. on Uncertainty in Artificial Intelligence (Stockholm), ed H. Prade and K Laskey (Morgan Kaufmann) pp 21–30
[3] Attias H 2000 Advances in Neural Information Processing Systems vol 12, ed S Solla, T Leen and K R Muller (Cambridge MA: MIT Press) pp 209–15
[4] Beal M J 2003 PhD thesis, University College London
[5] Corduneanu A and Bishop C M 2001 Proc. 8th Int. Conf. on Artificial Intelligence and Statistics, ed T Richardon and T. Jaakkola (Morgan Kaufmann) pp 27–34
[6] Ghahramani Z and Beal M J 2001a Advances in Neural Information Processing Systems vol 13, ed T Leen, T Dietterich and V Tresp (MIT Press) pp 507–13
[7] Ghahramani Z and Beal M J 2001b Technical report, University College London
[8] Humphreys K and Titterington D M 2000 COMPSTAT2000, ed J G Bethlehem and P G M van der Heijden (Heidelberg: Physica-Verlag) pp 331–6
[9] Penny W D and Roberts S J 2000 Technical Report PARG-2000-01, Oxford University
[10] Saul L K, Jaakkola T and Jordan M I 1996 J. Artific. Intell. Res. 4 61–76
[11] Waterhouse S, MacKay D and Robinson T 1996 Advances in Neural Information Processing Systems vol 8, ed D S Touretzky, M Mozer and M Hasselmo (Cambridge MA: MIT Press) pp 351–7
[12] Hinton G E and van Camp D 1993 Proc. 6th Annual ACM Conf. on Computational Learning Theory (Santa Cruz CA) (New York: ACM Press) pp 5–13
[13] Dembo A and Zeitouni O 1986 Stoch. Process. Appl. 23 91–113
[14] Campillo F and LeGland F 1989 Stoch. Process. Appl. 33 245–74
[15] Elliott R J and Krishnanurthy V 1997 SIAM J. Control Optim. 35 1908–23
[16] Charalambous C D, Elliott R J and Krishnanurthy V 2003 SIAM J. Control Optim. 42 1578–603
[17] Charalambous C D and Logothetis A 2000 IEEE Trans. Automat. Control 5 928–34
[18] Liptser R S and Shirayayev A N 1977 Statistics of Random Processes: I, II (New York: Springer-Verlag)
[19] Charalambous C D and Logothetis A 1998 Proc. 37th IEEE Conf. on Decision and Control (Tampa FL)
[20] Benes V 1981 Stochastics 5 65–92
[21] Haussmann U and Pardoux E 1988 Stochastics 23 241–75
[22] Charalambous C D and Elliott R J 1997 IEEE Trans. Automat. Control 42 482–97
[23] Charalambous C D and Elliott R J 1998 SIAM J. Control Optim. 36 542–78
[24] Cornford D 2006 VISDEM: Variational inference in stochastic dynamic environmental models
   http://www.ncrg.aston.ac.uk/˜cornfosd/VISDEM/