Two problems with variational expectation maximisation for time-series models

Richard E. Turner, Pietro Berkes, and Maneesh Sahani
Gatsby Computational Neuroscience Unit
17 Alexandra House, Queen Square, London, WC1N 3AR, London
{turner, berkes, maneesh}@gatsby.ucl.ac.uk

Abstract

Variational methods are a key component of the approximate inference and learning toolbox. These methods fill an important middle ground, retaining distributional information about uncertainty in latent variables, unlike *maximum a posteriori* methods (MAP), and yet requiring fewer computational resources than Monte Carlo Markov Chain methods. In particular the variational Expectation Maximisation (vEM) and variational Bayes algorithms, both involving variational optimisation of a free energy, are widely used in time-series modelling. Here, we investigate the success of vEM in simple probabilistic time-series models. First we consider the inference step of vEM, and show that a consequence of the well-known compactness property is a failure to propagate uncertainty in time, thus limiting the usefulness of the retained distributional information. In particular, the uncertainty may appear to be smallest precisely when the approximation is poorest. Second, we consider parameter learning and analytically reveal systematic biases in the parameters found by vEM. Surprisingly, simpler variational approximations (such a mean-field) can lead to less bias than more complicated structured approximations.

1 The variational approach

We begin with a very brief review of vEM. The Expectation-Maximisation (EM) algorithm [1] is a standard approach to finding maximum likelihood (ML) parameters for latent variable models, including hidden Markov Models and linear or non-linear state space models (SSMs) for time-series. The algorithm can be re-formulated as a variational optimisation of a free-energy [2, 3]. Consider observations collected into a set Y, that depend on latent variables X and parameters θ. We seek to maximise $\log p(Y|\theta)$ with respect to θ. By introducing a new distribution over the latent variables $q(X)$, we can write

$$\log p(Y|\theta) = \log \int dX\ p(Y, X|\theta) = \log \int dX\ p(Y, X|\theta) \frac{q(X)}{q(X)}, \tag{1}$$

$$\geq \int dX\ q(X) \log \frac{p(Y, X|\theta)}{q(X)} = F(q(X), \theta). \tag{2}$$

This last quantity is the free energy. It is smaller than the log-likelihood by an amount equal to the Kullback-Leibler (KL) divergence between $q(X)$ and the posterior distribution on the latents $p(X|Y, \theta)$

$$F(q(X), \theta) = \log p(Y|\theta) - \mathrm{KL}(q(X)||p(X|Y, \theta)), \tag{3}$$

For fixed θ, the optimum value for q is clearly equal to $p(X|Y, \theta)$, at which point the KL divergence vanishes and the free energy equals the log-likelihood. Thus, alternate maximisation of $F(q, \theta)$

with respect to q (the E-step) and θ (the M-step) will eventually find parameters that maximise the likelihood.

In many models, calculation of this posterior is intractable. Thus, the vEM approach is to instead optimise q restricted to a class of distributions $\mathcal{Q}$, within which the minimum of the KL divergence can tractably be found. The optimal q is called the variational approximation to the posterior. Constrained optimisation of q now alternates with optimisation of θ to find a maximum of the free energy, though not necessarily the likelihood. The optimal parameters are taken to approximate the ML values.

Most often, the class $\mathcal{Q}$ is defined to contain all distributions that factor over disjoint sets C_i of the latent variables in the problem: $q(X) = \prod_{i=1}^{I} q_i(x_{C_i})$. For example, if each latent variable appears in a factor of it own, the approximation is called *mean-field*. Partial factorisations, which keep some of the dependencies between variables are called *structured approximations*. In both cases the q_i's are found iteratively, by repeating the following updates,

$$q(x_i) \propto \exp\left(\langle \log p(Y, X|\theta)\rangle_{\prod_{j\neq i} q_j(x_{C_j})}\right). \tag{4}$$

Here, we analyse the accuracy of vEM in two stages. We first look at the relationship between the true posterior distribution and the variational approximation. It is well known that variational methods tend to be compact [4]. For instance, a unimodal variational approximation to a multimodal distribution will match the largest mode [5], rather than averaging across all of them, and a spherical Gaussian variational approximation will match the shortest length-scale of a correlated Gaussian. We show that this compactness results in a complete failure to propagate uncertainty between time-steps, often making the variational approximation most over-confident exactly when it is poorest. We then consider the accuracy of the vEM parameter estimates. As the variational bound on the likelihood is parameter dependent, variational methods can be biased away from peaks in the likelihood, toward regimes where the bound in tighter. As a result, the best approximations for learning are not necessarily the tightest, but rather those that result in bounds which depend least on the parameters. Both of these properties are exemplified using simple time-series models, although the conclusions are likely to apply more generally.

2 Variational approximations do not propagate uncertainty

Fully factored variational approximations (so called mean-field approximations) have been used for inference in time-series models as they are fast and yet still return estimates of uncertainty in the latent variables [6]. Here, we show that in a simple model, the variational iterations fail to propagate uncertainty between the factors, rendering these estimates of uncertainty particularly inaccurate in time-series (see [7] for a related example).

We consider a time-series model with a single latent variable x_t at each time-step drawn from an AR(1) prior with coefficient λ and innovations variance σ^2,

$$p(x_t|x_{t-1}) = \mathrm{Norm}(\lambda x_{t-1}, \sigma^2). \tag{5}$$

The marginal mean of this distribution is zero and the marginal variance is $\sigma^2_\infty = \frac{\sigma^2}{1-\lambda^2}$. Typically the latent variables are assumed carry strong temporal correlations, so that λ is close to 1[1]. We consider arbitrary instantaneous likelihood functions, $p(y_t|x_t)$. Using an approximating distribution which is factored over time $q(x_{1:T}) = \prod_{t=1}^{T} q(x_t)$, the update for the latent variable at time t follows from Eq. 4,

$$q(x_t) = \frac{1}{Z} p(y_t|x_t) \exp(\langle \log p(x_t|x_{t-1}) p(x_{t+1}|x_t)\rangle_{q(x_{t-1})q(x_{t+1})}), \tag{6}$$

$$= \frac{1}{Z'} p(y_t|x_t) \mathrm{Norm}\left(\frac{\lambda}{1+\lambda^2}\left(\langle x_{t-1}\rangle + \langle x_{t+1}\rangle\right), \frac{\sigma^2}{1+\lambda^2}\right) = \frac{1}{Z'} p(y_t|x_t) q_{\mathrm{prior}}(x_t). \tag{7}$$

[1] In fact the effective time-scale of Eq.5 is $\tau_{eff} = -1/\log(\lambda)$ and so a change in λ from 0.9 to 0.99 is roughly equivalent to a change from 0.99 to 0.999. This is important when the size of the biases in the estimation of λ are considered.

That is, the variational update is formed by combining the likelihood with a variational prior-predictive $q_{\text{prior}}(x_t)$ that contains the contributions from the latent variables at the adjacent time-steps. This variational prior-predictive is interesting because it is identical to the true prior-predictive when there is no uncertainty in the adjacent variables. That is, *none* of the (potentially large) uncertainty in the value of the adjacent latent variables is propagated to $q(x_t)$, and the width of the variational predictive is consequently narrower than the width of state-conditional distribution $p(x_t|x_{t-1})$ (compare to Eq. 5)[2].

Temporally factored variational methods for time-series models will thus generally recover an approximation to the posterior which is narrower than the state-conditional distribution. As the whole point of time-series models is that there are meaningful dependencies in the latents, and therefore the state-conditional often has a small width, the variational uncertainties may be tiny compared to the true marginal probabilities. Thus, the mean-field approach essentially reduces to iterative MAP-like inference, except that we find the mean of the posterior rather than a mode. In the next section, it will be shown that this does have some advantages over the MAP approach, notably that pathological spikes in the likelihood can be avoided.

In conclusion, although variational methods appear to retain some information about uncertainty, they fail to propagate this information between variables. In particular, in time-series with strong correlations between latents at adjacent times, the variational posterior becomes extremely concentrated, even though it is least accurate. An ideal distributional approximation would perhaps behave in the opposite fashion, returning larger uncertainty when it is likely to be more inaccurate.

3 Variational approximations are biased

In the last section we showed that variational approximations under-estimate the uncertainties in inference. We now ask how these inaccuracies might affect the parameter estimates returned by vEM. This question is important in many contexts. For example, scientific enquiry is often concerned with the values of a parameter, to substantiate claims like "natural scenes vary slowly" or "natural sounds are sparse", for instance.

What makes for a good variational approximation in this case? The instant reaction is that the free-energy should be as close to the likelihood as possible. That is $\mathrm{KL}(q(X)||p(X|Y,\theta))$ should be as small as possible for all X. However, from the perspective of learning it is more important to be *equally tight everywhere*, or in other words it is more important for the KL-term to be as parameter-independent as possible: If $\mathrm{KL}(q(X)|p(X|Y,\theta))$ varies strongly as a function of the parameters, this can shift the peaks in the free-energy away from the peaks in the likelihood, toward the regions were the bound is tighter. (See [8] for a related example for variational Bayes in mixture models.)

We now illustrate this effect in a linear SSM. In particular, we show that the mean-field approximation can actually have less severe parameter-dependent biases than two structural approximations, and can therefore lead to better vEM parameter estimates, even though it is less tight everywhere.

Deriving the learning algorithms
In the following we first introduce an elementary SSM, for which we can find the exact likelihood ($\log p(y|\theta)$). We then examine the properties of a set of different variational learning algorithms. This set comprises a mean-field approximation, two different structural approximations, and zero-temperature EM. This final approximation can be thought of as vEM where the approximating distributions are delta functions centred on the *maximum a posteriori* (MAP) estimates [3]. The analysis of these schemes proceeds as follows: First the optimal E-Step updates for these approximations are derived; Second, it is shown that, as the SSM is a simple one, the free-energies and the zero-temperature EM objective function can be written purely in terms of the parameters. That is, $\max_{q(x)} F(\theta, q(x))$ and $\max_X \log p(Y, X|\theta)$ have closed form solutions, and do not require iterative updates to be computed as is usual. Thus, we can study the relationship between the peaks in the likelihood and the peaks in the free-energies and zero-temperature EM objective function, for any dataset. An outline of the derivation of these quantities is given here, but for more detail see the associated technical report [9].

[2] This problem only gets worse if the prior dynamics have longer dependencies, e.g. if $p(x_t|x_{t-1:t-\tau}) = \mathrm{Norm}(\sum_{t'=1}^{\tau} \lambda_{t'} x_{t-t'}, \sigma^2)$ then the variational prior-predictive has a variance, $\frac{\sigma^2}{1+\sum_{t'=1}^{\tau} \lambda_{t'}^2}$.

Consider an SSM which has two latent variables per time-step and two time-steps. We take the priors on the latent variables to be linear-Gaussian, and the observations are given by summing the latents at the corresponding time-step and adding Gaussian noise,

$$p(x_{k,1}) = \mathrm{Norm}\left(0, \frac{\sigma_x^2}{1-\lambda^2}\right), \tag{8}$$

$$p(x_{k,2}|x_{k,1}) = \mathrm{Norm}\left(\lambda x_{k,1}, \sigma_x^2\right), \tag{9}$$

$$p(y_t|x_{1,t}, x_{2,t}) = \mathrm{Norm}(x_{1t} + x_{2t}, \sigma_y^2). \tag{10}$$

This defines a joint Gaussian over the observations and latent variables. From this we can compute the likelihood exactly by marginalising,

$$p(y_1, y_2|\theta) = \mathrm{Norm}(0, \Sigma_Y), \quad \Sigma_Y = I\sigma_y^2 + 2\frac{\sigma_x^2}{1-\lambda^2}\begin{bmatrix} 1 & \lambda \\ \lambda & 1 \end{bmatrix}. \tag{11}$$

The posterior distribution over the latent variables is also Gaussian, and is given by, $p(\mathbf{x}|\mathrm{y}) = \mathrm{Norm}(\mu_{\mathbf{x}|\mathrm{y}}, \Sigma_{\mathbf{x}|\mathrm{y}})$, where $\mathbf{x} = [x_{11}, x_{21}, x_{12}, x_{22}]^T$. The covariance and mean are

$$\Sigma^{-1}_{\mathbf{x}|\mathrm{y}} = \begin{bmatrix} \frac{1}{\sigma_y^2} + \frac{1}{\sigma_x^2} & \frac{1}{\sigma_y^2} & -\frac{\lambda}{\sigma_x^2} & 0 \\ \frac{1}{\sigma_y^2} & \frac{1}{\sigma_y^2} + \frac{1}{\sigma_x^2} & 0 & -\frac{\lambda}{\sigma_x^2} \\ -\frac{\lambda}{\sigma_x^2} & 0 & \frac{1}{\sigma_y^2} + \frac{1}{\sigma_x^2} & \frac{1}{\sigma_y^2} \\ 0 & -\frac{\lambda}{\sigma_x^2} & \frac{1}{\sigma_y^2} & \frac{1}{\sigma_y^2} + \frac{1}{\sigma_x^2} \end{bmatrix}, \quad \mu_{\mathbf{x}|\mathrm{y}} = \frac{1}{\sigma_y^2}\Sigma_{\mathbf{x}|\mathrm{y}}\begin{bmatrix} y_1 \\ y_1 \\ y_2 \\ y_2 \end{bmatrix}. \tag{12}$$

The posterior is correlated through time because of the linear-Gaussian prior, and correlated across chains because of explaining away. The correlations through time increase as the prior becomes slower ($|\lambda|$ increases) and less noisy (σ_x^2 decreases). The correlations across chains increase as the observation noise (σ_y^2) decreases.

We now derive the optimal E-Step for four different approximations: The first three approximations provide uncertainty estimates and these are the fully factored mean-field approximation (q_1), factorisation over chains but not time (q_2), and factorisation over time but not chains (q_3), as shown in the following table: The optimal E-Step updates for these three distributions can be found by

	factored over time	unfactored over time
factored over chains	$q_1(\mathbf{x}) = q_{11}(x_{11})q_{12}(x_{12})q_{13}(x_{21})q_{14}(x_{22})$	$q_2(\mathbf{x}) = q_{21}(x_{11}, x_{12})q_{22}(x_{21}, x_{22})$
unfactored over chains	$q_3(\mathbf{x}) = q_{31}(x_{11}, x_{21})q_{32}(x_{12}, x_{22})$	$p(\mathbf{x}|\mathrm{y}) = q(x_{11}, x_{12}, x_{21}, x_{22})$

minimising the variational KL. Each factor is found to be Gaussian, with a mean and precision that match the corresponding elements in $\mu_{\mathbf{x}|\mathrm{y}}$ and $\Sigma^{-1}_{\mathbf{x}|\mathrm{y}}$. The fourth and final approximation is zero-temperature EM (q_4), for which the E-Step is given by the MAP estimate for the latent variables. As the posterior is Gaussian, the mode and the mean are identical and so the MAP estimates are identical to the variational values for the means.

The next step is to compute the free-energies. In the first three cases, the Gaussianity of the posterior as well as q_1, q_2, and q_3 makes it possible to compute the KL divergences analytically:

$$\mathrm{KL}_i\left(\prod_{a=1}^{A} q_{ia}(\mathbf{x}_a)||p(\mathbf{x}|\mathrm{y})\right) = \frac{1}{2}\log\frac{\prod\Sigma_{ia}}{\Sigma_{\mathbf{x}|\mathrm{y}}}. \tag{13}$$

Using this expression we find,

$$\mathrm{KL}_1 = \frac{1}{2}\log\frac{\left(\sigma_y^2 + \sigma_x^2\right)^4}{\sigma_y^4\gamma}, \ \mathrm{KL}_2 = \frac{1}{2}\log\frac{\left(\left(\sigma_y^2 + \sigma_x^2\right)^2 - \lambda^2\sigma_y^4\right)^2}{\sigma_y^4\gamma}, \tag{14}$$

$$\text{and } \mathrm{KL}_3 = \frac{1}{2}\log\frac{\left(\sigma_y^2 + 2\sigma_x^2\right)^2}{\gamma}, \tag{15}$$

where $\gamma = (1-\lambda^2)\left((2\sigma_x^2 + \sigma_y^2)^2 - \lambda\sigma_y^4\right)$. In the fourth approximation, the KL divergence between a Gaussian and a delta function is infinite. Therefore, the KL term is discarded for zero-temperature EM and the log-joint is used as a pseudo-free energy.

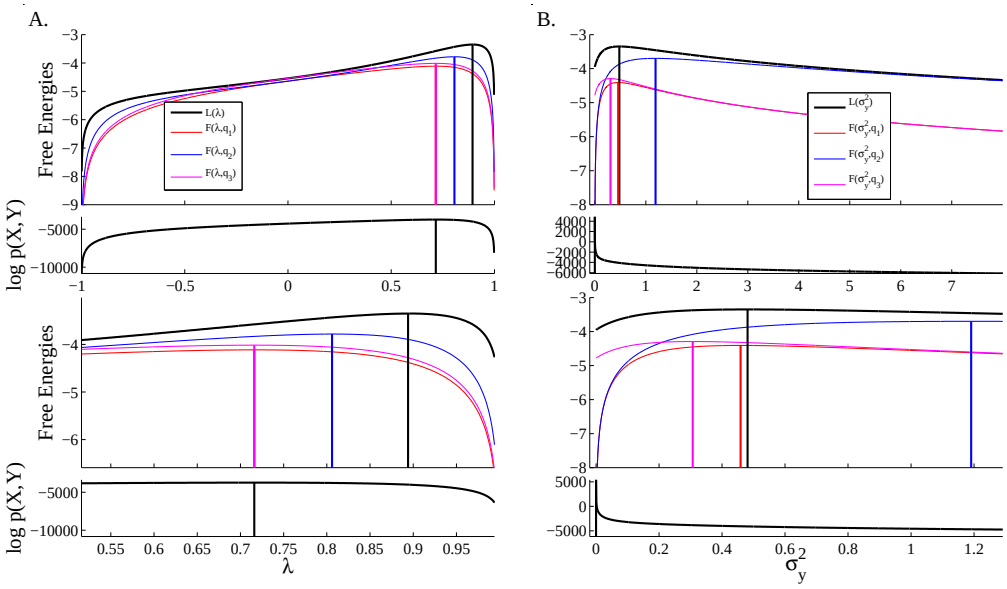

Figure 1: Biases in the Free-energies for a simple linear dynamical system. True/ML parameters are $\lambda = 0.9$, $\sigma_x^2 = 1 - \lambda^2 = 0.19$, and $\sigma_y^2 = 0.43$. In each case one parameter is learned and the others are set to their true/ML values. A. learning λ, B. learning σ_y^2. Large panels show the uncertainty preserving methods ($q_{1:3}$). Small panels show the zero-temperature EM approach (q_4). The bottom two panels show a zoomed in region of the top two panels.

General properties of the bounds: A sanity check

We now verify that these results match our intuitions. For example, as the mean field approximation is a subclass of the other approximations, it is *always* the loosest of the bounds, $\mathrm{KL}_1 > \mathrm{KL}_2, \mathrm{KL}_3 > 0$. Furthermore, approximation 3 (factorising over time) becomes looser than approximation 2 (which does not) when temporal correlations dominate over the correlations between chains. This is indeed the case as $\mathrm{KL}_3 > \mathrm{KL}_2$ when $r = \frac{\sigma_x^2}{|\lambda|\sigma_y^2} < 1$. Moreover, approximation 2 (which factorises over chains) is equivalent to the mean field approximation, $\mathrm{KL}_1 = \mathrm{KL}_2$, when there are no temporal correlations, $\lambda = 0$ or $\sigma_x^2 = \infty$, and in this case the true posterior matches approximation 3, $\mathrm{KL}_3 = 0$. Similarly, approximation 3 is equivalent to the mean-field approximation when the observation noise is infinity $\sigma_y^2 = \infty$, and here approximation 2 is exact $\mathrm{KL}_2 = 0$.

We can now consider how the maxima in the likelihood relate to the maxima in the Free-energies. Unfortunately, there is no closed form solution for these maxima, but in the simple examples which follow, the free-energies and likelihoods can be visualised. In general, we use as our data-set a large number of samples drawn from the forward model ($N > 10000$) and in all cases the ML parameters are essentially equal to the true parameters.

The model has a total of three parameters. We first consider learning just one of these parameters and set the others to the true/ML value. This will allow us to develop some intuition about the ways in which different approximations lead to different biases in the parameter estimates. In this case, the likelihood and free-energies are easy to visualise; some typical examples are shown in Fig. 1. We then consider how the bias changes as a function of the true/ML parameters, and observe that there is no universally preferred approximation, but instead the least biased approximation depends on the parameter that is being learned and on the value of the true/ML parameters. Finally, in we will study the bias when learning the dynamic parameter and the observation noise simultaneously.

Learning the dynamical parameter, λ

We begin by considering learning λ, with the other parameters fixed. As the magnitude of the dynamical parameter increases, so does the correlation in the posterior between successive latent variables in the same chain, that is $x_{k,1}$ and $x_{k,2}$. This means the factorisation over time results in looser bounds as the magnitude of λ increases (KL_3 increases, Eq. 3). Furthermore, as the

correlation between latents in the same chain increases, ($x_{k,1}$ and $x_{k,2}$), so does the correlation between x_{11} and x_{22} (propagated by the explaining away). This means, somewhat surprisingly, that the approximation which does not factorise over time, but over chains, also becomes looser as the magnitude of λ increases. That is, KL_2 increases with the magnitude of λ. Due to the fact that both bounds become less tight as λ increases, the free-energies peak at lower values of λ than the likelihood does, and therefore yield under-estimates (see [10] for a similar result).

The mean-field approximation suffers from both of the aforementioned effects, and it is therefore looser than both. However, with regard to their dependence on λ, KL_1 and KL_3 are equivalent. This means that the mean field approximation and the approximation that factors over time recover identical values for the dynamical parameter, even though the former is looser. Curiously, the solution from zero-temperature EM is also *identical to the mean-field (q_1) and temporally factored (q_3) solutions*. One of the conclusions to draw from this is that most severe approximation need not necessarily yield the most biased parameter estimates.

Learning the observation noise, σ_y^2, and the dynamical noise, σ_x^2

Next we consider learning σ_y^2, with the other parameters fixed to their true values. Due to explaining away, decreasing the observation noise increases the correlation between variables at the same time step, i.e., between x_{1t} and x_{2t}. This means that the approximation that factors over chains, becomes worse as σ_y^2 decreases, and therefore KL_2 is an increasing function of σ_y^2 . In contrast, the approximation that factorises over time, but not over chains, becomes tighter as σ_y^2 decreases i.e. KL_3 is a decreasing function of σ_y^2. As the mean-field approximation shares both of these effects it lies somewhere between the two, depending on the settings of the parameters. This means that whilst approximation 3 under-estimates the observation noise, and approximation 2 over-estimates it, the loosest approximation of the three, the mean field approximation, can actually provide the best estimate, as its peak lies in between the two. The purpose of the next section is to characterise the parameter regime over which this occurs.

In contrast to the situation with the dynamical parameter, the zero-temperature EM objective behaves catastrophically as a function of the observation noise, σ_y^2. This is caused by a narrow spike in the likelihood-surface at $\sigma_y^2 = 0$. At this point the latent variables arrange themselves to explain the data perfectly, and so there is no likelihood penalty (of the sort $-\frac{1}{2\sigma_y^2}(y_t - x_{1,t} - x_{2,t})^2$). In turn, this means the noise variance can be shrunk to zero which maximises the remaining terms ($\propto -\log \sigma_y^2$). The small cost picked up from violating the prior-dynamics is no match for this infinity.

This is not a very useful solution from either the perspective of learning or inference. It is a pathological example of overfitting[3]: There is an infinitesimal region of the likelihood-posterior surface with an infinite peak. By integrating over the latent variables, in a variational method for example, the problem vanishes as the peak has negligible mass and so makes only a small contribution. So, although variational methods often do not preserve as much uncertainty information as we would like, and are often biased, by recovering means and not modes they provide better joint estimates than the catastrophic zero-temperature EM approach.

Learning the dynamical noise σ_x^2 with the other parameters fixed at their true values results in a very similar situation: approximation 2 under-estimates σ_x^2, and approximation 3 over-estimates it, while the mean-field approximation returns a value in between. Once again the MAP solution suffers from an overfitting problem whereby the inferred value of σ_x^2 is driven to zero.

Characterising the space of solutions

In the previous section we found that for a particular setting of the true/ML parameter, the mean-field approximation was the most unbiased (see Fig. 1). How typical is this scenario? One way of answering this question is to evaluate the bias in the parameters learned using the four approximation schemes for many different data-sets each with different maximum-likelihood parameters. In practice three methods are used to find the optimal settings of the parameters. The first is to perform a grid based search, the second is to perform direct gradient ascent on the free-energy and the third is to run vEM. All three methods return identical results up to experimental error.

As a typical example, we show the bias in inferring λ for many different maximum-likelihood settings of σ_y^2 and λ in Fig. 2A. In each case σ_x^2 was set to the ML value, which was close to the

[3]This is the SSM analogue to Mackay's so-called KABOOM! problem in soft K-means [4]

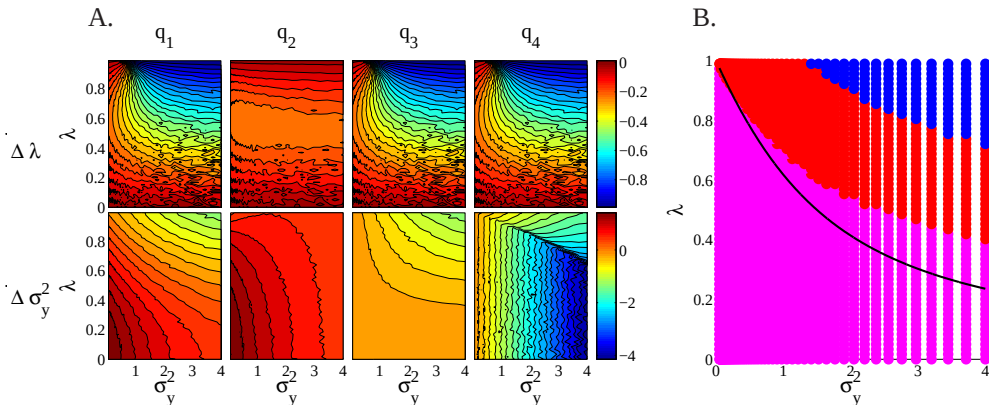

Figure 2: A. Biases for infering a single parameter as a function of σ_y^2 and λ. For all points $\sigma_x^2 = 1 - \lambda^2$. Bias is defined as $\Delta\Theta = \Theta_{INF} - \Theta_{ML}$ so that over-estimation results in a positive bias. Columns correspond to the four approximations. Top Row: Bias in λ. Bottom Row: Bias in σ_y^2.B. The best approximation for finding σ_y^2 indicated by color (q_1 red, q_2 blue and q_3 magenta). The black solid line is $r = \sigma_x^2/|\lambda|\sigma_y^2 = 1$ and below it approximation 3 is tightest, and above it approximation 2 is tightest.

true value of $1 - \lambda^2$. The parameter is under-estimated in all cases, often by a substantial amount (e.g. for approximations 1,3, and 4, at high σ_y^2 and λ values, the bias is almost one). The bias from using approximation 2 is always smaller than that from using the others, and it is to be preferred everywhere. However, this does not generalise for other parameters. The bias for inferring σ_y^2 is shown in Fig. 2B. As noted in the previous section, approximation 2 over-estimates the observation noise, whilst approximation 3 and 4 under-estimate it. The mean-field approximation combines the behaviours of approximation 2 and 3 and therefore under-estimates in regions where λ and σ_y^2 are small, and over-estimates in regions where they are large. In the intermediate region, these effects cancel and this is the region in which the mean-field approximation is the best. This is shown in Fig. 2C which indicates the best approximation to use for inferring the observation noise at different parts of the space. The mean-field solution is to be preferred over a fairly large part of the space.

Which is the best approximation therefore depends not only on which parameter has to be learned, but also on the ML value of parameters.

Simultaneous inference of pairs of parameters

So far we have considered estimating a single parameter keeping the others at their true values. What happens when we infer pairs of parameters at once? Consider, for instance, inferring the dynamical parameter λ and the observation noise σ_y^2 with σ_x^2 held at its ML/true value (see Fig. 3). As before, three methods are used to find the optimal parameter settings (gridding, gradient ascent and vEM). In a small minority of cases the objective functions are multi-modal, in which case the agreement between the methods depends on the initialisation. In order to avoid this ambiguity, the gradient based methods were initialised at the values returned from the method of gridding the space. This procedure located the global optima. The most striking feature of Fig. 3A. is that the biases are often very large (even in regimes where the structural approximations are at their tightest). Moreover, as there is a many to one mapping between the true parameters and the inferred parameters this indicates that it is impossible to simply correct for the variational bias by looking at the inferences.

Fig. 3B. shows that, in contrast to the case where only one parameter is inferred at a time, the mean-field solution is no-longer superior to the structural approximations. It also indicates that whilst tightness is a guide for choosing the best approximation, it is not very accurate. It is also notable that when all three parameters are inferred together (data not shown), the biases become even larger.

Finally, we consider the relevance of this toy example, and in particular what happens in longer time-series ($T > 2$) with more hidden variables ($K > 2$). In general both of these changes result in posterior distributions that have richer correlational structure. (That is, the posterior covariance matrix has more off-diagonal terms.) The variational approximations thus ignore larger parts of this structure and therefore the KL terms and associated biases will become correspondingly larger.

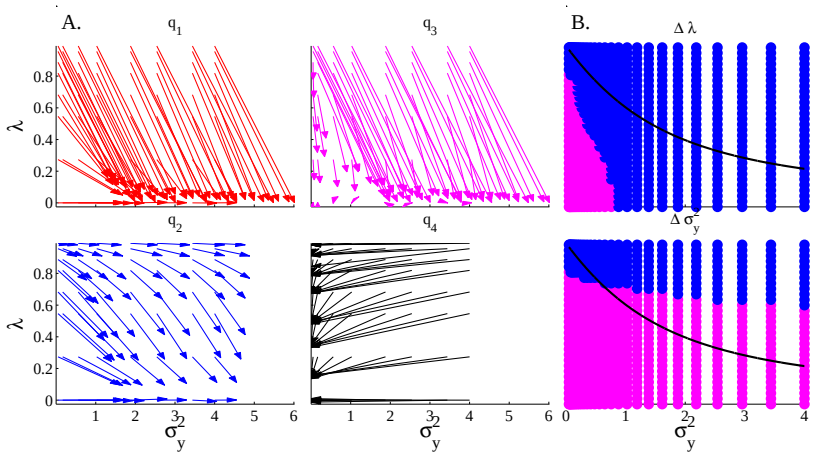

Figure 3: Simultaneous inference of λ and σ_y^2 with biases shown as a function of the true/ML settings of the parameters. A. For each approximation ($q_{1:4}$) a number of simulations are run and each is represented by an arrow. The arrow begins at the true/ML setting of the parameters and the tip ends at the inferred value. Ideally the arrows would be very short, but in fact they are often very large. B. The best uncertainty preserving approximation ($q_{1:3}$) for finding λ (Top) and σ_y^2 (Bottom) indicated by color (q_1 red, q_2 blue and q_3 magenta). The black solid line is $r = \sigma_x^2/|\lambda|\sigma_y^2 = 1$ and below it approximation 3 is tightest, and above it approximation 2 is tightest.

4 Conclusion

We have discussed two problems in the application of vEM to time-series models. First, the compactness property of variational inference leads to a failure to propagate posterior uncertainty through time. Second, the dependence of the variational lower bound on the model parameters often leads to strong biases in parameter estimates. We found that the relative bias of different approximations depended not only on which parameter was sought, but also on its true value. Moreover, tightest bound did not always yield the smallest bias: in some cases, structured approximations were more biased than the mean-field approach. Variational methods did, however, avoid the over fitting problem which plagues MAP estimation. Despite these shortcomings, variational methods remain a valid, efficient alternative to computationally costly Markov Chain Monte Carlo methods. However, the choice of the variational distribution should be complemented with an analysis of the dependency of the variational bound on the model parameters. Hopefully, these examples will inspire new algorithms that pool different variational approximations in order to achieve better performance.

Acknowledgments

We thank David Mackay for inspiration. Supported by the Gatsby Charitable Foundation.

References

[1] A. Dempster, N. Laird, and D. Rubin. *J. of the Royal Stat. Society: B*, 39:1–38, 1977.
[2] R. Hathaway. *Statistics & Probability Letters*, 4(2):53–56, 1986.
[3] R. Neal and G. Hinton. In M. I. Jordan, ed., *Learning in Graphical Models*, pp. 355–370. Kluwer Academic Press, 1998.
[4] D. Mackay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.
[5] B. C.M.M. *Pattern Recognition and Machine Learning*. Springer, 2006.
[6] M. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.
[7] J. Winn and T. Minka. Expectation propogation and variational message passing: a comparison using infer.net. http://videolectures.net/abi07_winn_ipi/, 2007. NIPS 2007 Workshop Inference in continuous/hybrid models.
[8] D. Mackay. A problem with variational free energy minimization. 2001.
[9] R. Turner and M. Sahani. Technical report: Two problems with variational expectation maximisation for time-series models. Technical report, Gatsby Computational Neuroscience Unit, 2008. Report.
[10] B. Wang and D. Titterington. *Neural Proc. Lett*, 20(3):151–170, 2004.