Chapter 1

Two problems with variational expectation maximisation for time-series models

Richard Eric Turner and Maneesh Sahani[1]

1.1 Introduction

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 generally requiring less computational time 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 of variational inference 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.

1.2 The variational approach

We begin this chapter with a brief theoretical review of variational Expectation Maximisation algorithm, before illustrating the important concepts with a simple example in the next section. The vEM algorithm is an approximate version of the Expectation-Maximisation (EM) algorithm (Dempster, 1977). EM 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 relationship between EM and vEM is revealed when EM is formulated as a variational optimisation of a free-energy (Hathaway, 1986; Neal and Hinton, 1998). Consider observations collected into a set Y, that depend on latent variables X and parameters θ. We seek to maximise the likelihood of the parameters, $\log p(\mathrm{Y}|\theta)$. By introducing a new distribution over the latent variables,

[1] Gatsby Computational Neuroscience Unit, Alexandra House, 17 Queen Square, London, WC1N 3A

$q(\mathrm{X})$, we can form a lower bound on the log-likelihood using Jensen's inequality,

$$\log p(\mathrm{Y}|\theta) = \log \int d\mathrm{X}\, p(\mathrm{Y},\mathrm{X}|\theta) = \log \int d\mathrm{X}\, p(\mathrm{Y},\mathrm{X}|\theta) \frac{q(\mathrm{X})}{q(\mathrm{X})}, \tag{1.1}$$

$$\geq \int d\mathrm{X}\, q(\mathrm{X}) \log \frac{p(\mathrm{Y},\mathrm{X}|\theta)}{q(\mathrm{X})} = \mathcal{F}(q(\mathrm{X}),\theta). \tag{1.2}$$

The lower bound is called the free-energy. The free-energy is smaller than the log-likelihood by an amount equal to the Kullback-Leibler (KL) divergence between $q(\mathrm{X})$ and the posterior distribution of the latent variables, $p(\mathrm{X}|\mathrm{Y},\theta)$

$$\mathcal{F}(q(\mathrm{X}),\theta) = \int d\mathrm{X}\, q(\mathrm{X}) \log \frac{p(\mathrm{X}|\mathrm{Y},\theta)p(\mathrm{Y}|\theta)}{q(\mathrm{X})} \tag{1.3}$$

$$= \log p(\mathrm{Y}|\theta) - \int d\mathrm{X}\, q(\mathrm{X}) \log \frac{q(\mathrm{X})}{p(\mathrm{X}|\mathrm{Y},\theta)} \tag{1.4}$$

$$= \log p(\mathrm{Y}|\theta) - \mathrm{KL}(q(\mathrm{X})||p(\mathrm{X}|\mathrm{Y},\theta)). \tag{1.5}$$

This expression shows that, for fixed θ, the optimum value for q is equal to $p(\mathrm{X}|\mathrm{Y},\theta)$, at which point the KL divergence vanishes and the free-energy equals the log-likelihood. Thus, alternate maximisation of $\mathcal{F}(q,\theta)$ with respect to q (the E-step) and θ (the M-step) will eventually find parameters that maximise the likelihood locally.

The EM algorithm is widely used to find ML parameter estimates, however, in many models calculation of this posterior is intractable. For example, it is often impossible to find an analytic form for $p(\mathrm{X}|\mathrm{Y})$ because the normalising constant involves an intractable integral. Another common source of intractability arises in models in which the number of latent variables is very large. For instance, a model with K binary latent variables generally requires a posterior distribution over all 2^K possible states of those variables. For even moderately large K this results in a computational intractability.

One possible method of side-stepping these intractabilities is to use the vEM approach (Jordan et al., 1999) which 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[2]

$$q_{\mathrm{vEM}}(\mathrm{X}) = \underset{q(\mathrm{X})\in\mathcal{Q}}{\arg\min}\, \mathrm{KL}(q(\mathrm{X})||p(\mathrm{X}|\mathrm{Y},\theta)) \tag{1.6}$$

$$= \underset{q(\mathrm{X})\in\mathcal{Q}}{\arg\min} \int d\mathrm{X}\, q(\mathrm{X}) \log \frac{q(\mathrm{X})}{p(\mathrm{X}|\mathrm{Y},\theta)}. \tag{1.7}$$

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.

There are two main ways in which q can be restricted to a class of tractable distributions $\mathcal{Q}$. The first method is to specify a parameteric form for the approximating distribution, $q(\mathrm{X}) = q_\gamma(\mathrm{X})$. A common choice is a Gaussian in which case the variational parameters, γ, are the mean and the covariance. The E-Step of vEM

[2] Other variational bounds may also be used in learning (e.g., Jaakkola and Jordan 2000). However the term variational EM is generally reserved for the free-energy bound that we discuss in this chapter.

then amounts to minimising the KL divergence with respect to the parameters of the approximating distribution,

$$q_{\mathrm{vEM}} = \arg\min_{\gamma} \mathrm{KL}(q_{\gamma}(\mathrm{X}) || p(\mathrm{X}|\mathrm{Y}, \theta)). \tag{1.8}$$

The second method is to define the class $\mathcal{Q}$ to contain all distributions that factor over disjoint sets C_i of the latent variables in the problem,

$$q(\mathrm{X}) = \prod_{i=1}^{I} q_i(\mathrm{x}_{C_i}). \tag{1.9}$$

For example, if each latent variable appears in a factor of it own, the approximation is called *mean-field*,

$$q_{\mathrm{MF}}(\mathrm{X}) = \prod_{i=1}^{I} q_i(\mathrm{x}_i). \tag{1.10}$$

Partial factorisations, which keep some of the dependencies between variables are called *structured approximations*. Generally, these methods which rely on factored classes may be more powerful than using a pre-specified parametric form, as the optimal analytic form of the factors may often be obtained by direct optimisation of the free-energy. To find this form we solve for the stationary points of a Lagrangian that combines the free-energy with the constraint that each factor is properly normalised. With respect to a factor $q_i(\mathrm{x}_{C_i})$ we have

$$\frac{\delta}{\delta q_i(\mathrm{x}_{C_i})} \left(\mathcal{F}(q(\mathrm{X}), \theta) - \sum_{i=1}^{I} \lambda_i \left(\int d\mathrm{x}_{C_i} q_i(\mathrm{x}_{C_i}) - 1 \right) \right) = 0, \tag{1.11}$$

where the λ_i are the Lagrange multipliers. Taking the functional derivative, and solving, we obtain

$$q_i(\mathrm{x}_{C_i}) \propto \exp \left(\langle \log p(\mathrm{Y}, \mathrm{X}|\theta) \rangle_{\prod_{j \neq i} q_j(\mathrm{x}_{C_j})} \right). \tag{1.12}$$

This set of equations, one for each factor, may be applied iteratively to increase the free-energy. The procedure is guaranteed to converge as the free-energy is convex in each of the factors $q_i(\mathrm{x}_{C_i})$ (Boyd and Vandenberghe, 2004).

1.2.1 A motivating example

Let us illustrate the EM and vEM algorithms described in the previous section by applying them to a simple model. The same example will also then serve to motivate the problems which are addressed later in this chapter. In the model a one-dimensional observation y, is generated by adding a zero-mean Gaussian noise variable with variance σ_{y}^2 to a latent variable, x, itself drawn from a zero mean Gaussian, but with unit variance:

$$p(\mathrm{x}) = \mathrm{Norm}(\mathrm{x}; 0, 1), \quad \text{and} \quad p(\mathrm{y}|\mathrm{x}, \sigma_{\mathrm{y}}^2) = \mathrm{Norm}(\mathrm{y}; \mathrm{x}, \sigma_{\mathrm{y}}^2). \tag{1.13}$$

The model may be viewed a very simple form of factor analysis with a one-dimensional observation and one factor. There is only one parameter to learn: the observation noise, σ_{y}^2. For tutorial purposes, we consider exact maximum-likelihood learning of this parameter from a single data-point. In fact, it is a simple matter to calculate the likelihood of the observation noise, $p(\mathrm{y}|\sigma_{\mathrm{y}}^2) = \mathrm{Norm}(\mathrm{y}; 0, \sigma_{\mathrm{y}}^2 + 1)$, and therefore this quantity could be optimised directly to find the maximum-likelihood estimate. However, an alternative approach is to use the EM algorithm. The EM alorithm begins by initialisating the observation noise. Next, in the E-Step, the approximating distribution q is updated to the posterior distribution over the latent variables given the current value of the parameters, that is,

$$q(\mathrm{x}) = p(\mathrm{x}|\mathrm{y}, \sigma_{\mathrm{y}}^2) = \mathrm{Norm}\left(\mathrm{x}; \frac{\mathrm{y}}{1+\sigma_{\mathrm{y}}^2}, \frac{\sigma_{\mathrm{y}}^2}{1+\sigma_{\mathrm{y}}^2}\right). \tag{1.14}$$

Then, in the M-Step, the observation noise is updated by maximising the free-energy with respect to the parameter, which has a closed-form solution, $\sigma_{\mathrm{y}}^2 = \mathrm{y}^2 - 2\mathrm{y}\langle \mathrm{x} \rangle_q + \langle \mathrm{x}^2 \rangle_q$. The E- and M-Step updates are then iterated, and this amounts to coordinate ascent of the free-energy (with respect to the distribution and then to the parameter) as illustrated in the upper left panel of figure 1.1A. Moreover, as the free-energy is equal to the log-likelihood after each E-Step has been performed (see figure 1.1B), the algorithm converges to a local optimum of the likelihood.

An alternative to exact ML learning is to use the vEM algorithm to return approximate ML estimates. This requires the q distribution to be restricted to a particular class, and as factored approximations are not an option for this one-dimensional model, a parametric restriction is considered. An instructive constraint is that the approximating distribution is a Gaussian with a flexible mean, but with a fixed variance. In the E-Step of vEM, the mean is set to minimise the KL divergence, which occurs when it is equal to the posterior mean. Therefore,

$$q_{\mu_q}(\mathrm{x}) = \mathrm{Norm}(\mathrm{x}; \mu_q, \sigma_q^2) \;\text{ where }\; \mu_q = \frac{\mathrm{y}}{1+\sigma_{\mathrm{y}}^2} \;\text{ and }\; \sigma_q^2 = \text{const.} \tag{1.15}$$

The M-Step of vEM is identical to EM, but the expectations are taken with respect to the new, approximate, distribution. As a result the free-energy is no longer pinned to the log-likelihood after an E-Step and therefore vEM is not guaranteed to converge to a local optimum of the likelihood. In fact, for the model considered here, the vEM estimate is biased away from the maximum in the likelihood, towards regions of parameter space where the variational bound is tightest (see figure 1.1). One of the main questions considered in this chapter is what extent such biases are a general feature of vEM.

1.2.2 Chapter Organisation

The motivating example in the previous section is a simple one, but it indicates that parameter estimates from EM and vEM can be quite different. However, it is unclear whether similar biases will arise for more realistic models, and in particular in those for time-series. Moreover, the example considered in the previous section involved estimating one parameter from one observation and a complete analysis should also compare EM and vEM on large data-sets. After all, it is well known that maximum-likelihood estimators can perform poorly when a large number of parameters have to be estimated from a small data-set, and so the discrepancy between EM and vEM noted above is not necessarily concerning. Of particular interest is the behaviour of vEM in the limit of infinite data. Maximum-likelihood estimators are often consistent, meaning that they converge to the true parameters in this limit. Do vEM estimators inherit this property? The motivating example indicates that a key determinant is the parameter dependence of the tightness of the free-energy bound, given by $\mathrm{KL}(q(\mathrm{x})|p(\mathrm{y}|\mathrm{x}, \theta))$, and whether this is significant

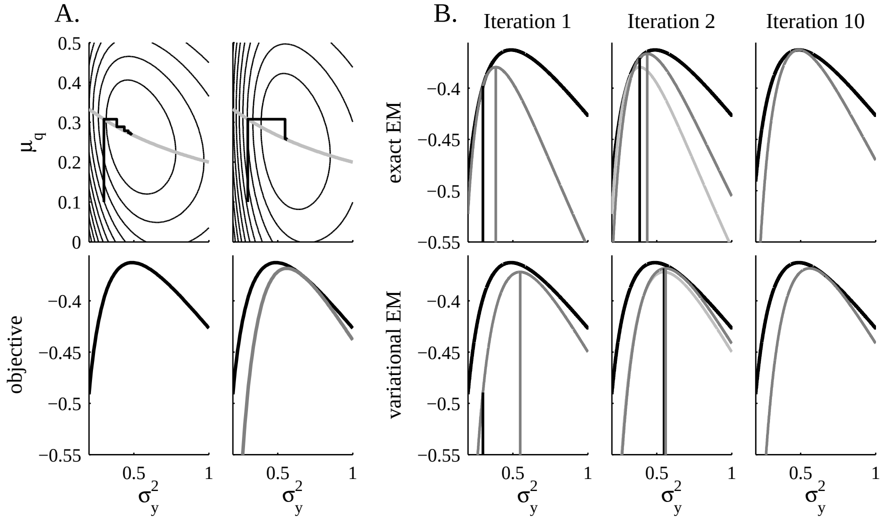

Figure 1.1: Schematics of EM and vEM using the model described in the text where the observation takes the value y = 0.4. **A. Top Left**: Thin black curves are the contours of the free-energy, $\mathcal{F}_{\text{EM}}(q, \sigma_{\text{y}}^2)$ for exact EM as a function of the observation noise (σ_{y}^2, abscissa) and the mean of the posterior distribution (μ_q, ordinate). The variance of the approximating distribution, σ_q^2 is set to the optimal value at each point. The thick grey line indicates the optimal choice for μ_q ie. the mean of the posterior distribution $p(\text{x}|\text{y}, \sigma_{\text{y}}^2)$. Ten updates using the EM algorithm are shown (thick black lines). Each update consists of an E-Step, which moves vertically to the optimal setting of μ_q (thick grey line), and an M-Step, which moves horizontally to the optimal setting of σ_{y}^2. By iterating these steps the algorithm converges via coordinate ascent to the optimum of the free-energy, which is also the optimum of the likelihood. **Bottom Left:** Log-likelihood of the observation noise. The value of the log-likelihood corresponds to the contour values along the thick grey line in the plot above. **Top Right**: Contours of the free-energy, $\mathcal{F}_{\text{vEM}}(q, \sigma_{\text{y}}^2)$, for vEM (black lines) in which the variance of the approximating distribution is fixed to the value $\sigma_q^2 = 0.4$. The position of the optimum has shifted to a larger value of the observation noise and the vEM algorithm converges onto this optimum (thick black lines). **Bottom Right:** The optimal free-energy (thick grey line) is a lower bound on the log-likelihood of the observation noise (thick black line). The value of the free-energy corresponds to the contour values along the thick grey line in the plot above. **B. Top Left:** Schematic showing the first M-Step for Exact EM. After an initial E-Step the free-energy, $\mathcal{F}_{\text{EM}}(q_1, \sigma_{\text{y}}^2)$, (thick grey line) is tight to the log-likelihood (thick black curved line) at the current value of the parameters (indicated by the vertical black line). In the M-Step q is fixed, and the optimal parameters are found (thick vertical grey line). This corresponds to the first horizontal line in the top left subplot of A. **Top Middle:** Schematic showing the second M-Step of exact EM. After the second E-Step, the updated free-energy, $\mathcal{F}_{\text{EM}}(q_2, \sigma_{\text{y}}^2)$, (thick, dark grey line) is tight to the log-likelihood (thick black line) at the current value of the parameters (indicated by the thick black vertical line). The old free-energy is shown for reference (thick, light grey line). The result of the second M-Step is indicated by the thick vertical grey line. **Top Right:** Schematic showing the free-energy after ten iterations, $\mathcal{F}_{\mathbf{EM}}(q_{10}, \sigma_{\text{y}}^2)$ (thick grey line). The optimum is clearly close to that of the log-likelihood (thick black line). **Bottom Left:** Schematic showing the first M-Step for vEM. As compared with the panel above, the free-energy, $\mathcal{F}_{\text{vEM}}(q(\mu_{q1}, \sigma_{q1}^2), \sigma_{\text{y}}^2)$ (tick grey line), is not tight to the log-likelihood (thick black line). **Bottom Middle:** Schematic showing the second M-Step for vEM. **Bottom Right:** Schematic showing the free-energy after ten iterations, $\mathcal{F}_{\mathbf{vEM}}(q(\mu_{q10}, \sigma_{q10}^2), \sigma_{\text{y}}^2)$ (thick grey line). The optimum clearly lies to the right of the optimum of the log-likelihood (thick black line). It is biased to where the variational approximation is tightest.

in comparison to the peak in the likelihood. This size of this contribution to the free-energy is studied in a simple setting in section 1.4. One intriguing possibility is that the best approximations for learning are not necessarily those that yield the tightest bounds, but rather those in which the tightness of the bounds depends least on the parameters. Evidence that such an effect exists is provided in section 1.4.4 and further investigation, in section 1.4.6, reveals that it is fairly common.

Before considering the biases in parameters learned using vEM, we first consider the E-Step of vEM in isolation. It is well known that variational approximations, like those derived in the vEM E-Step, tend to be compact (MacKay, 2003). In other words, the variational approximation has a tendency to have a smaller entropy than the true distribution. The evidence for this folk-theorem is reviewed in the next section with particular emphasis on the relevance to time-series modelling. Then, in section 1.3.3, we show a consequence of compactness in mean-field approximations is a complete failure to propagate uncertainty between time-steps, this makes the popular mean-field approximations most over-confident exactly when they are poorest.

Both the compactness and parameter learning biases are exemplified using very simple time-series models, although the conclusions are likely to apply more generally.

1.3 Compactness of variational approximations

In this section we consider approximating a known distribution, $p(\mathrm{x})$, with a simpler one, $q(\mathrm{x})$, by minimising the variational KL divergence, $\mathrm{KL}(q(\mathrm{x})||p(\mathrm{x}))$. This operation forms the E-Step of vEM (see equation 1.7) and so its behaviour has implications for how the full algorithm behaves. Before considering a number of instructive examples, it is immediately clear from the form of the variational KL that at any point where the true density is zero, the approximation must also be zero (otherwise the KL divergence will be infinity). A consequence of this fact is that when a distribution which has two modes that are separated by a region of zero density is approximated by a unimodal distribution, then the approximation will model just one of the modes, rather than averaging across them. This is one example of a general tendency for variational approximations to have a smaller entropy than the target distribution. This section explores this so-called compactness property of variational approximations, before considering the implications for time-series modelling.

1.3.1 Approximating mixtures of Gaussians with a single Gaussian

It has just been argued above that when the true distribution contains two modes which are separated by an intermediate region of zero density, the approximation will be compact. However, it is unclear what happens when the intermediate region does not dip to zero. In order to investigate this situation, consider approximating a one-dimensional mixture of Gaussians with a single Gaussian, where

$$p(\mathrm{x}|\theta) = \sum_{k=1}^{K} \pi_k \mathrm{Norm}(\mathrm{x}; \mu_k, \sigma_k^2), \quad \text{and} \quad q(\mathrm{x}) = \mathrm{Norm}(\mathrm{x}; \mu_q, \sigma_q^2). \tag{1.16}$$

In figure 1.2 a number of examples are shown for a range of different parameter choices for the mixture. As expected, for mixtures with two clearly defined modes (right-hand column of figure 1.2), the approximation matches the mode with the largest variance, rather than averaging across both of them (Bishop, 2006). In these cases the entropy of the approximation is less than that of the true distribution. However, for intermediate distributions, in which the modes are joined by a significant bridge of probability-density, the variational approximation does average across the modes and in some cases the entropy of the approximation is larger than the true distribution. The conclusion is that the compactness property is a useful guide to the behaviour of variational methods when applied to highly-multimodal distributions, but that there are examples when variational methods are not compact (as measured by their entropy relative to that of the true distribution). Variational approximations commonly used in clustering are an example of the former (Bishop, 2006), but variational approximations to independent component analysis can result in the latter (Turner et al., 2008).

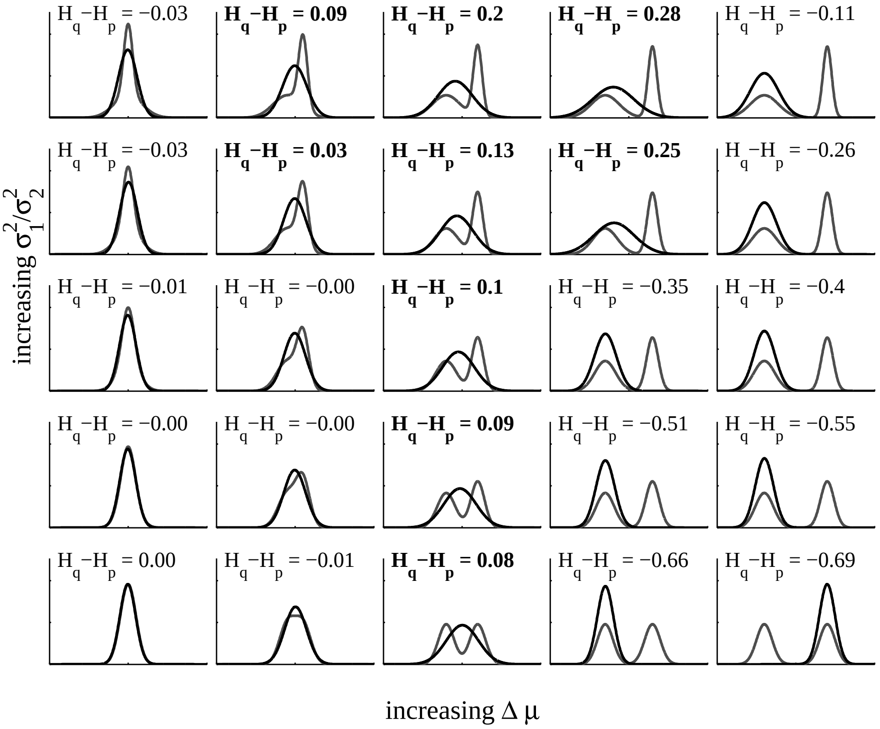

Figure 1.2: Each panel shows a variational approximation to a true distribution. The true distribution is a mixture of two Gaussians (grey line) and the approximating family is a Gaussian (black line). The parameters of the mixture were set so that each component has equal weight ($\pi_1 = \pi_2 = \frac{1}{2}$). The difference between the means of the mixture components increases from the left column of panels (where $\mu_1 - \mu_2 = 0$) to the right column of panels (where $\mu_1 - \mu_2 = 10$). The ratio of the variances increases from bottom row of panels (where $\sigma_1^2/\sigma_2^2 = 1$) to the top row of panels (where $\sigma_1^2/\sigma_2^2 = 10$). The smaller of the two variances is fixed, $\sigma_2^2 = 1$. The bottom left is therefore a mixture of two Gaussians with the same mean and variance and this is another Gaussian. The approximation is therefore exact and the entropy difference, shown at the top of each panel, is therefore zero. In general the entropy of the approximation can be less than (normal font) or greater than (bold font) the true entropy.

1.3.2 Approximating a correlated Gaussian with a factored Gaussian

The examples above indicate how compactness operates for univariate distributions, when the approximating distribution is restricted to a particular parametric form.

Next, we consider approximating a bivariate distribution using the mean-field approximation. The true distribution is a zero-mean, correlated Gaussian distribution with principal axes oriented in the directions $\mathbf{e}_1 = \frac{1}{\sqrt{2}}[1,1]^T$ and $\mathbf{e}_2 = \frac{1}{\sqrt{2}}[1,-1]^T$ with variances σ_1^2 and σ_2^2 respectively (after MacKay, 2003), so

$$p(\mathrm{x}_1, \mathrm{x}_2|\Sigma) = \mathrm{Norm}\left(\mathrm{x}_1, \mathrm{x}_2; 0, \Sigma\right), \quad \Sigma = \sigma_1^2 \mathbf{e}_1 \mathbf{e}_1^T + \sigma_2^2 \mathbf{e}_2 \mathbf{e}_1^T. \tag{1.17}$$

This correlated Gaussian is approximated in the mean-field approach by a separable distribution, $q(\mathrm{x}_1, \mathrm{x}_2) = q(\mathrm{x}_1)q(\mathrm{x}_2)$, and by considering the fixed-points of equation 1.12 the optimal updates are found to be,

$$q(\mathrm{x}_i) = \mathrm{Norm}\left(\mathrm{x}_i; 0, \frac{1}{2}\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right). \tag{1.18}$$

That is, the optimal factored approximating distribution is a spherical Gaussian that has a precision which is equal to the diagonal elements of the precision matrix of the original Gaussian (that is, $(\Sigma^{-1})_{i,i}$). This is an example of a more general result, that variational approximations between two Gaussians match precisions, which will be seen again later in the chapter. Consider now the behaviour of the variational approximation in the case where variance of the two components is very different (e.g. $\sigma_1^2 > \sigma_2^2$). The width of the approximating distribution becomes $\sigma_2^2/2$, and therefore independent of the longer length-scale. In this sense the approximation is becoming compact; it matches the smallest length scale in the posterior. In the next subsection, this result will be rediscovered from the contrasting perspective of mean-field inference in time-series models.

1.3.3 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 (Beal, 1998). 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 models (see Winn and Minka, 2007, for a related example).

We consider a time-series model with a single latent variable x_t at each time-step drawn from a first-order auto-regressive prior with coefficient λ and innovations variance σ^2,

$$p(\mathrm{x}_t|\mathrm{x}_{t-1}) = \mathrm{Norm}(\mathrm{x}_t; \lambda \mathrm{x}_{t-1}, \sigma^2). \tag{1.19}$$

The marginal mean of this distribution is zero and the marginal variance is $\sigma_\infty^2 = \frac{\sigma^2}{1-\lambda^2}$. As the time series of greatest interest tend to be those which exhibit strong temporal structure, we will study models in which with the autoregressive parameter λ is close to unity[3]. The observed variables y_t depend only on the latent variable at the corresponding time-steps. The precise form of $p(\mathrm{y}_t|\mathrm{x}_t)$ will not be important here.

If we choose a mean-field approximating distribution which is factored over time $q(\mathrm{x}_{1:T}) = \prod_{t=1}^T q(\mathrm{x}_t)$, the update for the latent variable at time t follows from

[3] In fact the effective time-scale of equation 1.19 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 in section 1.4.3.

equation 1.12,

$$q(\mathrm{x}_t) = \frac{1}{Z} p(\mathrm{y}_t|\mathrm{x}_t) \exp(\langle \log p(\mathrm{x}_t|\mathrm{x}_{t-1}) p(\mathrm{x}_{t+1}|x_t) \rangle_{q(\mathrm{x}_{t-1})q(\mathrm{x}_{t+1})}), \tag{1.20}$$

$$= \frac{1}{Z'} p(\mathrm{y}_t|\mathrm{x}_t) \mathrm{Norm}\left(\mathrm{x}_t; \frac{\lambda}{1+\lambda^2} \left(\langle \mathrm{x}_{t-1}\rangle + \langle \mathrm{x}_{t+1}\rangle\right), \frac{\sigma^2}{1+\lambda^2}\right) \tag{1.21}$$

$$= \frac{1}{Z'} p(\mathrm{y}_t|\mathrm{x}_t) q_{\mathrm{prior}}(\mathrm{x}_t). \tag{1.22}$$

That is, the variational update is formed by combining the likelihood with a variational prior-predictive $q_{\mathrm{prior}}(\mathrm{x}_t)$ 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. As such, *none* of the (potentially large) uncertainty in the value of the adjacent latent variables is propagated to $q(\mathrm{x}_t)$, and the width of the variational predictive is consequently narrower than the width of state-conditional distribution $p(\mathrm{x}_t|\mathrm{x}_{t-1})$ (compare to equation 1.19)[4].

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 (see 1.3). Thus, the mean-field approach is not all that different to the "zero-temperature EM" or MAP-based approach (in which the joint probability of observed data and latent variables is optimised alternately with respect to the latent variables—with no distributional information—and the parameters), 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 factors. In particular, in time-series with strong correlations between latents at adjacent times, the mean-field variational posterior becomes extremely concentrated, even though it is least accurate in this regime. An ideal distributional approximation would perhaps behave in the opposite fashion, returning larger uncertainty when it is likely to be more inaccurate.

1.4 Variational approximations are biased

In the last section we showed that variational approximations under-estimate the uncertainties in inference. We will now investigate how these inaccuracies 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".

What makes for a good variational approximation in this case? The instant reaction is that the free-energy should be as tight to the log-likelihood as possible. That is, the optimal KL divergence at each parameter setting,

$$\mathrm{KL}^*(\theta) = \underset{q(\mathrm{x})}{\arg\max}\, \mathrm{KL}(q(\mathrm{X})||p(\mathrm{X}|\mathrm{Y},\theta)), \tag{1.23}$$

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

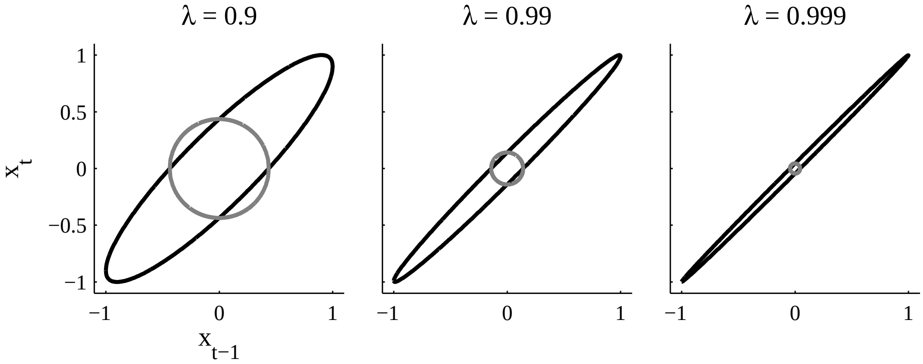

Figure 1.3: Compactness in mean-field approximations for time-series. The average true prior-predictive (black ellipses, showing the probability density contour at one standard deviation) is shown together with the mean-field approximations (grey circles, also showing the probability density contour at one standard deviation), for three settings of λ. The marginal variance of the true prior-predictive is 1. The marginal variance of the mean-field approximation is $(1-\lambda^2)/(1+\lambda^2)$ which is tiny for typical values of λ in time-series models. Notice that this example is equivalent to the previous example in section 1.3.2 involving a bivarate Gaussian when, $\sigma_1^2 = \frac{\sigma^2}{1-\lambda}$ and $\sigma_2^2 = \frac{\sigma^2}{1+\lambda}$.

should be as small as possible for all θ. However, the conclusion from the motivating example in section 1.2.1, is that from the perspective of learning it is more important to be *equally tight everywhere*. In other words it is more important for the KL-term to be as parameter-independent as possible: If $\mathrm{KL}^*(\theta)$ 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. This perspective explains a previous observation whereby variational Bayes typically prunes out too many components of a mixture model (MacKay, 2001).

We now illustrate this effect in a linear SSM and show that consequences can include mis-estimation of the time constant with which the latent variables evolve, under-estimation of the overlap of emission weights, and unwanted pruning of emission weights. Moreover, 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. We also show that the biases in parameter estimates increase considerably the more parameters are estimated.

1.4.1 Deriving the learning algorithms

In the following we first introduce an elementary SSM, for which we can find the exact log-likelihood, $\log p(\mathrm{y}|\theta)$. 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 (Neal and Hinton, 1998). 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, $\max_{q(\mathrm{x})} \mathcal{F}(\theta, q(\mathrm{x}))$

and $\max_{\mathrm{X}} \log p(\mathrm{Y}, \mathrm{X}|\theta)$ have closed form solutions, and do not require iterative updates to be computed as is usually the case. Thus, we can study the relationship between the peaks in the log-likelihood and the peaks in the free-energies and zero-temperature EM objective function, for any dataset. This is analogous to the lower right hand panel of figure 1.1A for the motivating example.

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

$$p(\mathrm{x}_{k,1}) = \mathrm{Norm}\left(\mathrm{x}_{k,1}; 0, \frac{\sigma_{\mathrm{x}}^2}{1-\lambda^2}\right), \tag{1.24}$$

$$p(\mathrm{x}_{k,2}|\mathrm{x}_{k,1}) = \mathrm{Norm}\left(\mathrm{x}_{k,2}; \lambda \mathrm{x}_{k,1}, \sigma_{\mathrm{x}}^2\right), \tag{1.25}$$

$$p(\mathrm{y}_{d,t}|\mathrm{x}_{1,t}, \mathrm{x}_{2,t}) = \mathrm{Norm}\left(\mathrm{y}_{d,t}; \sum_{k=1}^{2} w_{d,k}\mathrm{x}_{k,t}, \sigma_{\mathrm{y}}^2\right). \tag{1.26}$$

This defines a joint Gaussian over the observations and latent variables. From this we can compute the likelihood exactly by marginalising. Defining the vector of observations, $\mathbf{y} = [\mathrm{y}_{11}, \mathrm{y}_{21}, \mathrm{y}_{12}, \mathrm{y}_{22}]^T$, and the matrix $M_{d,d'} = \sum_{k=1}^{2} w_{d,k}w_{d',k}$, the likelihood is given by,

$$p(y_{1:2,1:2}|\theta) = \mathrm{Norm}(\mathbf{y}; 0, \sigma_{\mathrm{y}}), \quad \sigma_{\mathrm{y}} = I\sigma_{\mathrm{y}}^2 + \frac{\sigma_{\mathrm{x}}^2}{1-\lambda^2}\left[\begin{array}{cc} M & \lambda M \\ \lambda M & M \end{array}\right]. \tag{1.27}$$

The posterior distribution over the latent variables is also Gaussian, and is given by, $p(\mathbf{x}|\mathbf{y}) = \mathrm{Norm}(\mu_{\mathbf{x}|\mathbf{y}}, \Sigma_{\mathbf{x}|\mathbf{y}})$, where the vector of latent variables is $\mathbf{x} = [\mathrm{x}_{11}, \mathrm{x}_{21}, \mathrm{x}_{12}, \mathrm{x}_{22}]^T$. In order to ease notation, we define weight vectors and matrices:

$$\mathbf{w}_1 = \begin{bmatrix} w_{11} \\ w_{21} \end{bmatrix} = |\mathbf{w}_1| \begin{bmatrix} \cos(\phi_1) \\ \sin(\phi_1) \end{bmatrix}, \tag{1.28}$$

$$\mathbf{w}_2 = \begin{bmatrix} w_{12} \\ w_{22} \end{bmatrix} = |\mathbf{w}_2| \begin{bmatrix} \cos(\phi_2) \\ \sin(\phi_2) \end{bmatrix}, \quad \text{and } W = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}. \tag{1.29}$$

Then, the covariance and mean of the posterior distribution are given by

$$\Sigma_{\mathbf{x}|\mathbf{y}}^{-1} = \left[\begin{array}{cccc} \frac{|\mathbf{w}_1|^2}{\sigma_{\mathrm{y}}^2} + \frac{1}{\sigma_{\mathrm{x}}^2} & \frac{\mathbf{w}_1^T\mathbf{w}_2}{\sigma_{\mathrm{y}}^2} & -\frac{\lambda}{\sigma_{\mathrm{x}}^2} & 0 \\ \frac{\mathbf{w}_1^T\mathbf{w}_2}{\sigma_{\mathrm{y}}^2} & \frac{|\mathbf{w}_2|^2}{\sigma_{\mathrm{y}}^2} + \frac{1}{\sigma_{\mathrm{x}}^2} & 0 & -\frac{\lambda}{\sigma_{\mathrm{x}}^2} \\ -\frac{\lambda}{\sigma_{\mathrm{x}}^2} & 0 & \frac{|\mathbf{w}_1|^2}{\sigma_{\mathrm{y}}^2} + \frac{1}{\sigma_{\mathrm{x}}^2} & \frac{\mathbf{w}_1^T\mathbf{w}_2}{\sigma_{\mathrm{y}}^2} \\ 0 & -\frac{\lambda}{\sigma_{\mathrm{x}}^2} & \frac{\mathbf{w}_1^T\mathbf{w}_2}{\sigma_{\mathrm{y}}^2} & \frac{|\mathbf{w}_2|^2}{\sigma_{\mathrm{y}}^2} + \frac{1}{\sigma_{\mathrm{x}}^2} \end{array}\right], \tag{1.30}$$

$$\mu_{\mathbf{x}|\mathbf{y}} = \frac{1}{\sigma_{\mathrm{y}}^2}\Sigma_{\mathbf{x}|\mathbf{y}}\left[\begin{array}{cc} W & 0 \\ 0 & W \end{array}\right]\mathbf{y}. \tag{1.31}$$

The posterior is correlated through time because of the linear-Gaussian prior, and correlated across chains because of explaining away [5]. The correlations through time

[5] Explaining away is the name given to the phenomenon in probabilistic modelling where the observation of an effect of two possible independent causes, leads to (anti-)correlation in the posterior distribution over those two causal latent variables. Suppose that either latent may take on a value that could account for the observation. Then if one does so, it "explains away" the observed effect, and the observed data no longer constrains the other. Thus, conditioned on the observation, the distribution over each latent depends on the value of the other, even if there was no such dependence in the prior.

increase as the prior becomes "slower" ($|\lambda|$ increases) and less noisy (σ_{x}^2 decreases). The correlations across chains increase as the magnitude of the weights increase ($|\mathbf{w}_d|^2$), and the angle between the weights ($\phi_1 - \phi_2$) or 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_{MF}), factorisation over chains but not time (q_{FC}), and factorisation over time but not chains (q_{FT}), as shown in the following table:

	factored over time	unfactored over time
chains factored	$q_{\mathrm{MF}}(\mathbf{x})$	$q_{\mathrm{FC}}(\mathbf{x})$
chains unfactored	$q_{\mathrm{FT}}(\mathbf{x})$	$p(\mathbf{x}\vert\mathbf{y}) = q(\mathbf{x})$

The factorisations are therefore,

$$q_{\mathrm{MF}}(\mathbf{x}) = q_{\mathrm{MF}}^{(1)}(x_{11}) q_{\mathrm{MF}}^{(2)}(\mathrm{x}_{12}) q_{\mathrm{MF}}^{(3)}(\mathrm{x}_{21}) q_{\mathrm{MF}}^{(4)}(\mathrm{x}_{22}), \tag{1.32}$$

$$q_{\mathrm{FC}}(\mathbf{x}) = q_{\mathrm{FC}}^{(1)}(\mathrm{x}_{11}, \mathrm{x}_{12}) q_{\mathrm{FC}}^{(2)}(\mathrm{x}_{21}, \mathrm{x}_{22}), \tag{1.33}$$

$$q_{\mathrm{FT}}(\mathbf{x}) = q_{\mathrm{FT}}^{(1)}(\mathrm{x}_{11}, \mathrm{x}_{21}) q_{\mathrm{FT}}^{(2)}(\mathrm{x}_{12}, \mathrm{x}_{22}). \tag{1.34}$$

The optimal E-Step updates for these three distributions can be found by minimising the variational KL. Each factor is found to be Gaussian, with a mean and precision that match the corresponding elements in $\mu_{\mathbf{x}|\mathbf{y}}$ and $\Sigma_{\mathbf{x}|\mathbf{y}}^{-1}$. The fourth and final approximation is zero-temperature EM (q_{MAP}), for which the E-Step is given by the MAP estimate for the latent variables for the current parameter setting. 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 and the uncertainty preserving variational approximations enables the KL divergences to be calculated analytically:

$$\mathrm{KL}_i \left(\prod_{a=1}^{A} q_i^{(a)}(\mathbf{x}_a) || p(\mathbf{x}|\mathbf{y}) \right) = \frac{1}{2} \log \frac{\prod_{a=1}^{A} \det \left(\Sigma_i^{(a)} \right)}{\det \left(\Sigma_{\mathbf{x}|\mathbf{y}} \right)}. \tag{1.35}$$

That is, the KL divergence is the log-ratio of the volume of the approximation (as measured by the matrix determinants) to the volume of the true posterior. It should be noted that the whole point of variational methods is that this quantity is usually intractable to compute, and it is only because the example is very simple that it is possible here. Using this expression we find,

$$\mathrm{KL}_{\mathrm{MF}}^{*} = \frac{1}{2} \log \left(\sigma_{\mathrm{y}}^2 + |\mathbf{w}_1|^2 \sigma_{\mathrm{x}}^2 \right)^2 \left(\sigma_{\mathrm{y}}^2 + |\mathbf{w}_2|^2 \sigma_{\mathrm{x}}^2 \right)^2 / \gamma \tag{1.36}$$

$$\mathrm{KL}_{\mathrm{FC}}^{*} = \frac{1}{2} \log \left(\left(\sigma_{\mathrm{y}}^2 + |\mathbf{w}_1|^2 \sigma_{\mathrm{x}}^2 \right)^2 - \lambda^2 \sigma_{\mathrm{y}}^4 \right) \left(\left(\sigma_{\mathrm{y}}^2 + |\mathbf{w}_2|^2 \sigma_{\mathrm{x}}^2 \right)^2 - \lambda^2 \sigma_{\mathrm{y}}^4 \right) / \gamma \tag{1.37}$$

$$\mathrm{KL}_{\mathrm{FT}}^{*} = \frac{1}{2} \log \left(\sigma_{\mathrm{x}}^4 |\mathbf{w}_1|^2 |\mathbf{w}_2|^2 \sin^2(\phi_1 - \phi_2) + \left(|\mathbf{w}_1|^2 + |\mathbf{w}_2|^2 \right) \sigma_{\mathrm{x}}^2 \sigma_{\mathrm{y}}^2 + \sigma_{\mathrm{y}}^4 \right)^2 / \gamma \tag{1.38}$$

where

$$\begin{aligned} \gamma = & \left(\left(|\mathbf{w}_1|^2 + |\mathbf{w}_2|^2 \right) \sigma_{\mathrm{x}}^2 \sigma_{\mathrm{y}}^2 + \sigma_{\mathrm{x}}^4 |\mathbf{w}_1|^2 |\mathbf{w}_2|^2 \sin^2(\phi_1 - \phi_2) + (1 + \lambda^2) \sigma_{\mathrm{y}}^4 \right)^2 \\ & - \left(\lambda \sigma_{\mathrm{y}}^2 \sigma_{\mathrm{x}}^2 \left(|\mathbf{w}_1|^2 + |\mathbf{w}_2|^2 \right) + 2 \lambda \sigma_{\mathrm{y}}^4 \right)^2. \end{aligned} \tag{1.39}$$

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. To ease notation, in what follows $\mathrm{KL}_i = \mathrm{KL}_i^*$.

1.4.2 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, $\mathrm{KL}_{\mathrm{MF}} > \mathrm{KL}_{\mathrm{FC}}, \mathrm{KL}_{\mathrm{FT}} > 0$, which is bourne out by the expressions. Furthermore, q_{FT} becomes looser than q_{FC} when temporal correlations dominate over the correlations between chains. For instance, if the weights have identical magnitude, $|\mathbf{w}_1| = |\mathbf{w}_2| = |\mathbf{w}|$, then $\mathrm{KL}_{\mathrm{FT}} > \mathrm{KL}_{\mathrm{FC}}$ when explaining away (EA) becomes more important than temporal correlation (TC) in the posterior,

$$\frac{\mathrm{EA}}{\mathrm{TC}} < 1, \quad \text{where } \mathrm{EA} = \frac{|\cos(\phi_1 - \phi_2)||\mathbf{w}|^2}{\sigma_{\mathrm{y}}^2} \quad \text{and} \quad \mathrm{TC} = \frac{|\lambda|}{\sigma_{\mathrm{x}}^2}. \tag{1.40}$$

Moreover, q_{FC} is equivalent to the mean field approximation, $\mathrm{KL}_{\mathrm{MF}} = \mathrm{KL}_{\mathrm{FC}}$, when there are no temporal correlations, $\lambda = 0$ or $\sigma_{\mathrm{x}}^2 = \infty$, and in this case the true posterior matches q_{FT}, $\mathrm{KL}_{\mathrm{FT}} = 0$. Similarly, q_{FT} is equivalent to the mean-field approximation when the observation noise is infinity $\sigma_{\mathrm{y}}^2 = \infty$, and here q_{FC} is exact ($\mathrm{KL}_{\mathrm{FC}} = 0$). Finally, it is noted that as q_{FT} is the only one which captures cross-chain correlations due to explaining away, it is the only one which is dependent on the relative angle between the weights.

Having verified that the expressions for the KL divergences appear reasonable, we can now consider how the maxima in the log-likelihood relate to the maxima in the free-energies. Unfortunately, there is no closed form solution for the location of these maxima, but in the simple examples which follow, the free-energies and likelihoods can be visualised. In general, we will be concerned with the consistency of the variational estimators, which means the behaviour when we have a large number of observations from the same time series. In this case the average likelihood becomes,

$$\lim_{N\to\infty} \frac{1}{N} \log p(\mathbf{y}_{1:N}|\Sigma_{\mathbf{y}}) = -\frac{1}{2}\log\det\Sigma_{\mathbf{y}} - \frac{1}{2}\Sigma_{\mathbf{y}}^{-1} \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n\mathbf{y}_n^T \tag{1.41}$$

$$= -\frac{1}{2}\log\det\Sigma_{\mathbf{y}} - \frac{1}{2}\Sigma_{\mathbf{y}}^{-1}\left\langle \mathbf{y}\mathbf{y}^T\right\rangle. \tag{1.42}$$

When the data are drawn from the forward model, $\left\langle \mathbf{y}\mathbf{y}^T\right\rangle$ can be computed analytically. In all cases the ML estimators are found to be consistent, and therefore equal to the true parameters in the limit of infinite data.

Although the model is simple, it has seven parameters and this means there are a great number of possible learning scenarios ranging from learning one parameter with the others fixed, to learning all parameters at once. In the following we highlight several illustrative examples in order to elucidate the general properties of the variational approach. First we consider learning a single parameter (the dynamical parameter, the observation noise, the innovations noise, the orientation of one of the weights, and the magnitude of one of the weights) with the other parameters set to their true values. 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 log-likelihood and free-energies are easy to visualise; some typical

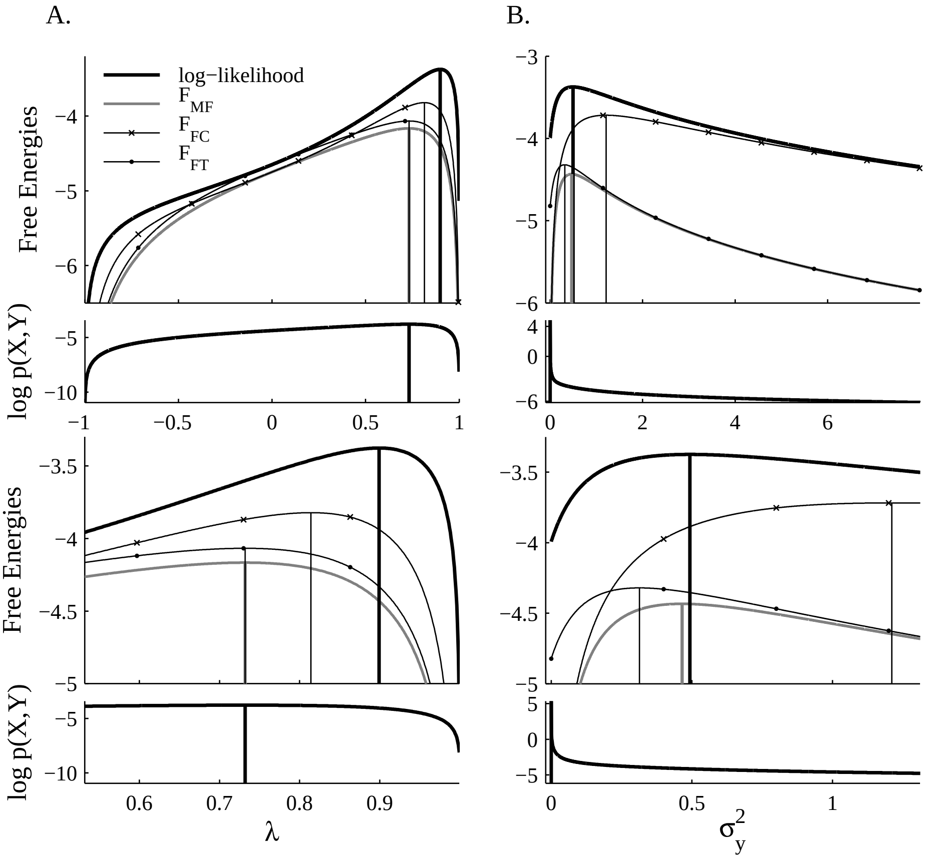

Figure 1.4: Biases in the free-energies for learning the dynamical parameter, λ, and the observation noise, σ_{y}^2, of a simple linear dynamical system. The true parameters were $\lambda = 0.9$, $\sigma_{\mathrm{x}}^2 = 1 - \lambda^2 = 0.19$, $\mathbf{w}_1^T = [1, 0]$,$\mathbf{w}_2^T = [1, 0]$, and $\sigma_{\mathrm{y}}^2 = 0.43$. In both columns A and B one parameter is learned and the others are set to their true values. A. Shows the results of learning λ, and B. learning σ_{y}^2. Large panels show the log-likelihood (thick black line) and the free-energies of the uncertainty preserving methods ($\mathcal{F}_{\mathrm{MF}}$ by a thick grey line, $\mathcal{F}_{\mathrm{FC}}$ by the crosses, and $\mathcal{F}_{\mathrm{FT}}$ by the circles). Small panels show the zero-temperature EM approach (q_{MAP}). The maxima of these functions are indicated by the vertical lines. The maximum of the log-likelihood lies at the true value of the parameters. The bottom two panels show a zoomed in region of the top two panels.

examples are shown in figure 1.4 and figure 1.5. We then consider how the bias changes as a function of the true 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 parameters. Next we will study the bias when learning the dynamic parameter and the observation noise simultaneously, as this provides a typical example of how the variational approach performs when multiple parameters are learned. The conclusion is that the biases become significantly larger as more parameters are estimated.

1.4.3 Learning the dynamical parameter, λ

We begin by considering learning the dynamical parameter λ, with the other parameters fixed to their true values. In order to ensure the effects of explaining away are properly considered the weights are set to be identical, with unit magnitude ($\mathbf{w}_1 = \mathbf{w}_2$ and $|\mathbf{w}_k|^2 = 1$).

As the magnitude of the dynamical parameter increases, so does the correlation in the posterior between successive latent variables in the same chain ($\mathrm{x}_{k,1}$ and $\mathrm{x}_{k,2}$). This means that q_{FT}, which factorises over time, results in a looser variational bound as the magnitude of λ increases ($\mathrm{KL}_{\mathrm{FT}}$ increases, equation 1.38). Furthermore, as the correlation between latents in the same chain increases, ($\mathrm{x}_{k,1}$ and $\mathrm{x}_{k,2}$), so does the correlation between x_{11} and x_{22} (because explaining away is propagated through time by the dynamics). This means, somewhat surprisingly, that q_{FC} which does not factorise over time, but over chains, also becomes looser as the magnitude of λ increases. That is, $\mathrm{KL}_{\mathrm{FC}}$ also increases with the magnitude of λ. In both cases, this λ-dependence in the tightness of the bound means that the corresponding variational free-energies peak at lower values of λ than the likelihood, and therefore both approximations yield under-estimates (see Wang and Titterington 2004 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, $\mathrm{KL}_{\mathrm{MF}}$ and $\mathrm{KL}_{\mathrm{FT}}$ are equivalent. Consequently the mean-field approximation, q_{MF}, and q_{FT}, which factors over time, recover identical values for the dynamical parameter, even though the former is looser. Curiously, the solution from zero-temperature EM (q_{MAP}) is also identical to those solutions. One of the conclusions to draw from this is that most severe approximation need not necessarily yield the most biased parameter estimates.

1.4.4 Learning the observation noise, σ_{y}^2, and the dynamical noise, σ_{x}^2

Next we consider learning the observation noise σ_{y}^2, with the other parameters fixed to their true values. Once again, in order to ensure the effects of explaining away are properly considered we consider identical, unit magnitude weight vectors ($\mathbf{w}_1 = \mathbf{w}_2$ and $|\mathbf{w}_k|^2 = 1$).

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 q_{FC}, which factors over chains, becomes worse as σ_{y}^2 decreases, and therefore $\mathrm{KL}_{\mathrm{FC}}$ is an increasing function of σ_{y}^2. On the other hand, as the observation process becomes less noisy the hidden states are more precisely determined by local information, and so correlations between them in the prior become less important. Thus, q_{FT}, which factorises over time but not over chains, becomes tighter as σ_{y}^2 decreases i.e. $\mathrm{KL}_{\mathrm{FT}}$ is a decreasing function of σ_{y}^2. We have now indicated that $\mathrm{KL}_{\mathrm{FC}}$ and $\mathrm{KL}_{\mathrm{FT}}$ have opposite dependencies on σ_{y}^2. As the mean-field approximation shares both of these effects its maximum lies somewhere between the two, depending on the settings of the parameters. This means that whilst q_{FT} under-estimates the observation noise, and q_{FC} 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. In the next section we will characterise the parameter regime over which this occurs.

The final approximation scheme, zero-temperature EM, behaves catastrophically when it is used to learn the observation noise, σ_{y}^2. This is caused by a narrow spike in the likelihood-surface at $\sigma_{\mathrm{y}}^2 = 0$. At this point the latent variables arrange themselves to explain the data perfectly, and so there is no likelihood penalty (of the sort $-\frac{1}{2\sigma_{\mathrm{y}}^2}(\mathrm{y}_{1,t} - \mathrm{x}_{1,t} - \mathrm{x}_{2,t})^2$). In turn, this means the noise variance can be shrunk to zero which maximises the remaining terms ($\propto -\log \sigma_{\mathrm{y}}^2$). 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[6]: there is an infinitesimal region of the likelihood-posterior surface with an infinite peak. By integrating over the latent variables, even if only approximately in a variational method for example, such peaks are discounted, as they are associated with negligible probabilitiy mass and so make only a small contribution to the integral. Thus, 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 may still provide better parameter estimates than the catastrophic zero-temperature EM approach.

Finally we note that learning the dynamical noise σ_{x}^2 with the other parameters fixed at their true values results in a very similar situation: q_{FC} under-estimates σ_{x}^2, and q_{FT} 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. The fact that learning σ_{y}^2 and σ_{x}^2 results in similar effects indicates that the conclusions drawn from these examples are quite general.

1.4.5 Learning the magnitude and direction of one emission weight

Finally we consider learning the emission weights. In order to explicate the various factors at work it is useful to separately consider learning the orientation of the weight vector and its magnitude. Consider first learning the orientation of one of the weights whilst its magnitude, and the value of the other parameters in the model, are known and fixed to their true values (shown in figure 1.5). The relative orientation of the pair of weights is the critical quantity, because this determines the amount of explaining away. If the weights are orthogonal ($\phi_1 - \phi_2 = \pi(n + 1/2)$), there is no explaining away ($\langle \mathrm{x}_{1t}\mathrm{x}_{2t}\rangle_{p(\mathbf{x}|\mathbf{y})} = 0$), and so q_{FC} is exact and q_{MF} and q_{FT} are equivalent. In contrast, if the weights are parallel ($\phi_1 - \phi_2 = n\pi$), explaining away is maximised and q_{MF} and q_{FC} are at their loosest because they do not model the dependencies between the chains. q_{FT} is also at its loosest in this region (because it does not capture the 'diagonal' correlations $\langle \mathrm{x}_{11}\mathrm{x}_{22}\rangle_{p(\mathbf{x}|\mathbf{y})}$ and $\langle \mathrm{x}_{21}\mathrm{x}_{12}\rangle_{p(\mathbf{x}|\mathbf{y})}$ which are strongest here). The result is that all the approximations are biased toward settings of the weights which are more orthogonal than the true setting. The bias in q_{MF}, q_{FC} and q_{MAP} are equal and can be substantial (see figure 1.5 and figure 1.6). The bias in q_{FT} is somewhat less as it captures the correlations induced by explaining away. Finally, we consider learning the magnitude of the second weight when all other parameters set to their true values (this includes the direction of the second weight). For low magnitudes, q_{FC} is tightest as the temporal correlations dominate over explaining away, but for high magnitudes the situation is reversed and q_{FT} is tightest. Consequently, q_{FC} under-estimates the magnitudes (often severely thereby pruning the weight entirely, see figure 1.5B), whilst q_{FT} over-estimates the magnitudes. As the mean-field approximation suffers from both effects, its estimates lie between the two and can therefore be less biased. Once again the MAP solution suffers from an over-fitting problem, where the estimated weight magnitudes blow up to infinity.

1.4.6 Characterising the space of solutions

In the previous section we found examples where the mean-field approximation was the most unbiased (see figure 1.4B. and 1.5B.). How typical is this scenario? To answer this question we look at the extent of the bias in parameter values returned by the four approximate learning schemes, over a wide range of different data sets

[6] This is the SSM analogue to Mackay's (2003) so-called KABOOM! problem in soft K-means.

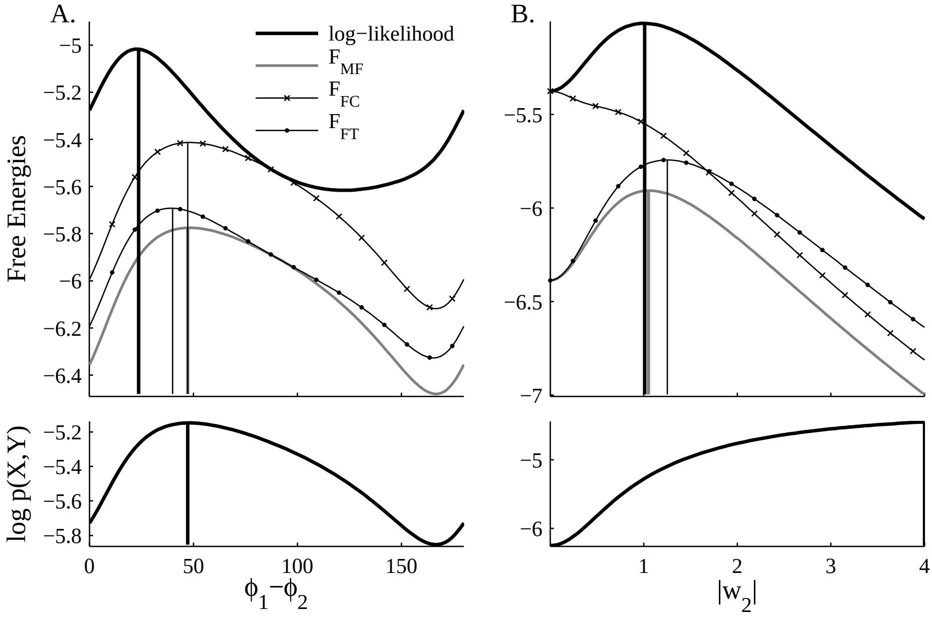

Figure 1.5: Biases in the free-energies for learning the weights of a simple linear dynamical system. The true parameters are $\lambda = 0.9$, $\sigma_{\text{x}}^2 = 1 - \lambda^2 = 0.19$, $\mathbf{w}_1^T = [1, 0]$, $\mathbf{w}_2^T = [\cos(\pi/8), \sin(\pi/8)]$ and $\sigma_{\text{y}}^2 = 0.3$. In both columns, A and B, one parameter is learned and the others are set to their true values. A. Learning the orientation (ϕ_2) of the second weight, $\mathbf{w}_2^T = [\cos(\phi_2), \sin(\phi_2)]$. B. Learning the magnitude of the second weight, $\mathbf{w}_2^T = |\mathbf{w}_2|[\cos(\pi/8), \sin(\pi/8)]$. Large panels show the log-likelihood (thick black line) and the free-energies of the uncertainty preserving methods ($\mathcal{F}_{\text{MF}}$ by a thick grey line, $\mathcal{F}_{\text{FC}}$ by the crosses, and $\mathcal{F}_{\text{FT}}$ by the circles). Small panels show the zero-temperature EM approach (q_{MAP}). The maxima of these functions are indicated by the vertical lines. The maximum of the log-likelihood lies at the true value of the parameters.

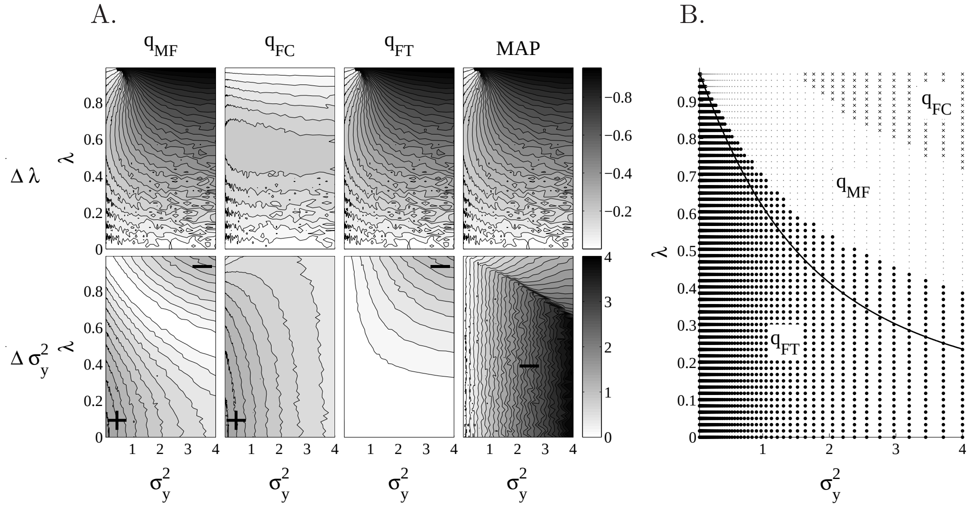

Figure 1.6: A. Biases for infering a single parameter as a function of σ^2_{y} and λ. Lighter colours indicate a bias of smaller magnitude. The bias is defined as $\Delta\Theta = \Theta_{INF} - \Theta_{ML}$ so that over-estimation results in a positive bias. For all points, $\sigma^2_{\text{x}} = 1 - \lambda^2$. The columns correspond to the four approximations. Top Row: Bias in estimating λ. All the schemes return underestimates and so the biases are always negative. Bottom Row: Bias in estimating σ^2_{y}. The sign of the bias is indicated by the '+' and '-' symbols. B. The best approximation for finding σ^2_{y} indicated by marker-type (q_{MF} grey filled-circles, q_{FC} black crosses and q_{FT} black open-circles). The black solid line is $r = \sigma^2_{\text{x}}/|\lambda|\sigma^2_{\text{y}} = 1$ and below it q_{FT} is tightest, and above it q_{FC} is tightest.

with different true parameter values. As the likelihood or free-energy surfaces may be multimodal—and we are not interested here in failures of learning due to local optima—the optimal parameters were found using three different optimisation schemes: grid-based search; direct gradient ascent on the free-energy; and coordinate ascent of the free-energy (or vEM). For the examples of this section, all three methods returned identical results up to experimental error.

One typical example—the bias in inferring λ for many different maximum-likelihood settings of σ^2_{y} and λ—appears in figure 1.6A. In each case σ^2_{x} was set to its true value, $1 - \lambda^2$. The parameter λ is under-estimated in all cases, often substantially (e.g. for q_{MF}, q_{FT} and q_{MAP}, at high σ^2_{y} and λ values, the bias is almost one). The bias from using q_{FC} is always smaller than that from using the others, and thus in this case it is to be preferred everywhere. However, this simple situation where one of the approximation schemes is universally superior does not generalise when learning other parameters. For example, the bias for inferring σ^2_{y} is shown in figure 1.6B. As noted in the previous section, q_{FC} over-estimates the observation noise, whilst q_{FT} and q_{MAP} under-estimate it. The mean-field approximation combines the behaviours of q_{FC} and q_{FT} and therefore under-estimates in regions where λ and σ^2_{y} 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 of all the approximations. This is shown in figure 1.6C which indicates the best approximation to use for inferring the observation noise at different parts of the space. The figure shows that the mean-field solution is to be preferred over a fairly large part of the space.

Next we consider biases in estimating the weight vectors (see figure 1.7). When learning the vector orientations, q_{MF}, q_{FC} and q_{MAP} turn out to exhibit identical biases. (Indeed, it is generally true that if the MAP solution does not suffer from over-fitting, then it is equal to the mean-field approximation in these Gaussian models.) These approximations do not model explaining away and so they are most biased in regions where the true weights are approximately parallel ($\phi_1 \approx \phi_2$). On the other hand, q_{FT} does capture inter-chain correlations, and so is superior for any setting of the true parameters [7]. When learning the weight vector magnitudes, q_{FT} is superior in regions where explaining away is large compared to the temporal correlations, whilst q_{FC} is superior in regions where temporal correlations dominate over explaining away. However, there is a large intermediate region where the mean-field approximation is the least biased. Once again, the tightness of the free-energy approximations is a poor indicator of which is the least biased.

The main conclusions from this section are that the biases in variational methods are often severe. The examples above indicate that factorisations across time can ignore strong temporal correlations in the data, and factorisations across chains can erroniously prune out emission weights. Furthermore, which is the best approximation depends not only on which parameter has to be learned, but also on the true value of parameters. Suprisingly, mean-field approximations are often superior to structured methods when a single parameter is estimated.

1.4.7 Simultaneous learning 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 true value (see figure 1.8). As before, three methods are used to find the optimal parameter settings (gridding, gradient ascent and vEM). In this case, a small minority of the objective functions are multi-modal, and then the agreement between the methods depends on the initialisation. 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 figure 1.8A. is that the biases are often very large (even in regimes where the structural approximations are at their tightest). In principle, if the mapping between the inferred paramters and true parameters were known it might be possible to correct for the biases in the variational estimates. However, multiple different settings of the true parameters result in the same inferred parameters and so it is impossible to correct the variational estimates in this way.

Figure 1.8B. 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.

1.4.8 Discussion of the scope of the results

The examples considered in this chapter were chosen to be simple so that exact results could be computed and visualised. It is necessary, then, to consider how the effects described here generalise to longer time-series ($T > 2$) with more hidden variables ($K > 2$). Unfortunately, it is generally not tractable to analyse these, more complex, scenarios. However, increasing the length of the time-series and the

[7] It is common practice to use zero-temperature EM (q_{MAP}) to learn weights in sparse-coding models and then to make a detailed statistical comparison of the learned weights to biological analogues derived from experiments in visual cortex. The result here—that zero-temperature EM recovers weights which are significantly more orthogonal than the true weights—raises concerns that this practice will be seriously affected by biases in the learned weights.

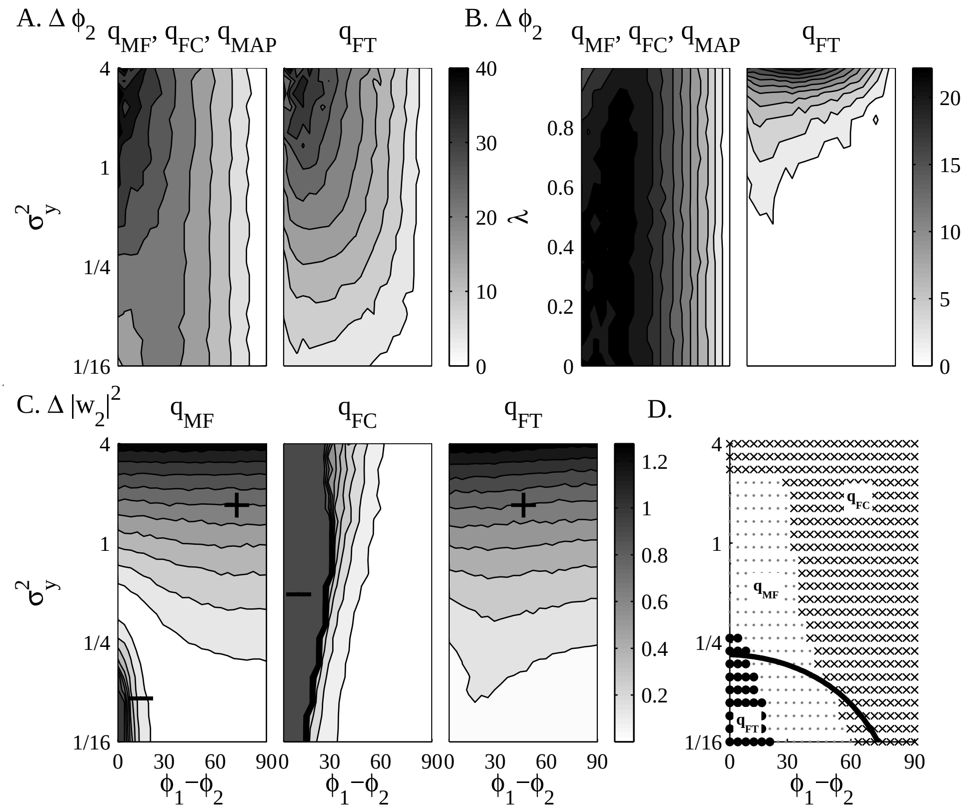

Figure 1.7: Parameter dependence of the biases in learning the weights. A. Biases in learning the relative orientation ($\Delta\phi$) of the weights as a function of the true relative orientation ($\phi_1 - \phi_2$) and the observation noise σ^2_{y}. The magnitude of the weights is unity, $|\mathbf{w}_k| = 1$ and the dynamical parameters are set to $\lambda = 0.9$ and $\sigma^2_{\mathrm{x}} = 1-\lambda^2$. All of the approximations over-estimate the angular separation of the weights, but q_{FT} is less or equally biased everywhere. B. Biases in learning the relative orientation of the weights as a function of the true orientation ($\phi_1-\phi_2$) and the dynamical parameter, λ. The observation noise is fixed to $\sigma^2_{\mathrm{y}} = 0.1$ and the state-noise to $\sigma^2_{\mathrm{x}} = 1-\lambda^2$. Again, q_{FT} less or equally biased everywhere. C. Biases in learning the magnitude of the second weight as a function of the true relative orientation ($\phi_1 - \phi_2$) and the observation noise. The other parameters are set to, $\lambda = 0.9$ and $\sigma^2_{\mathrm{x}} = 1-\lambda^2$. The MAP estimate q_{MAP} returns an infinite value for the weights everywhere and is therefore not shown. D. The least biased approximation for finding the magnitude of the weight (indicated by marker-type; q_{MF} grey filled-circles, q_{FC} black crosses and q_{FT} black open-circles) as a function of the relative orientation of the weights and the observation noise. Above the solid line, q_{FC} is the tighter approximation and below it q_{FT} is the tighter approximation.

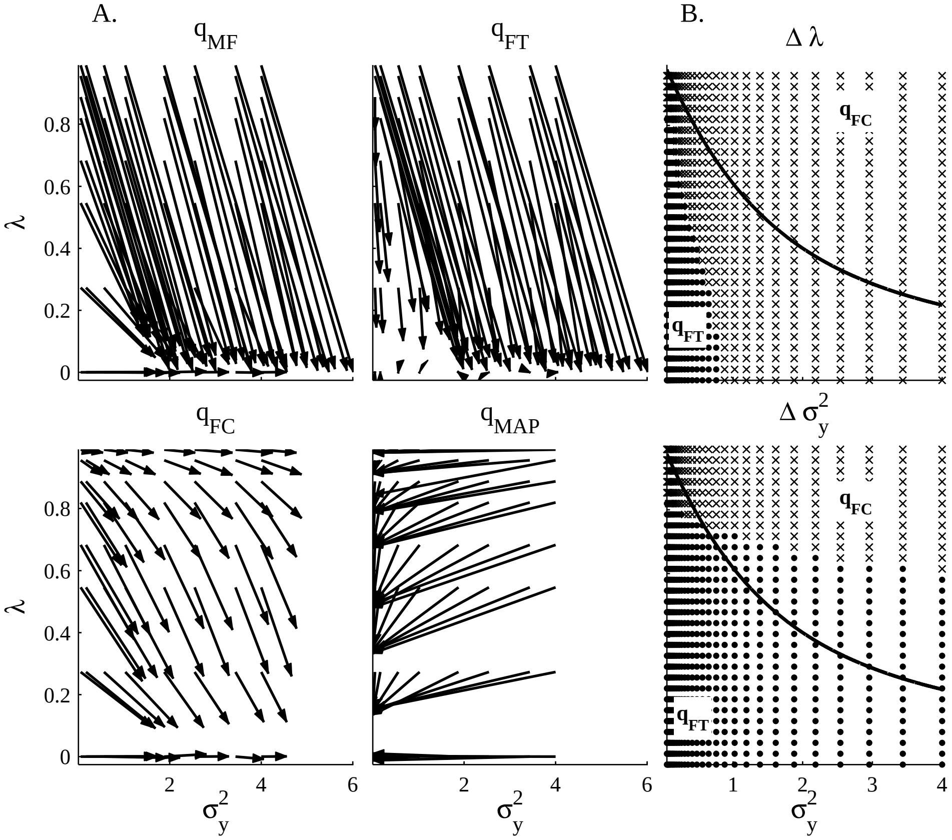

Figure 1.8: Simultaneous inference of λ and σ_{y}^2 with biases shown as a function of the true settings of the parameters. A. For each approximation a number of simulations are run and each is represented by an arrow. The arrow begins at the true setting of the parameters and the tip ends at the inferred value. Ideally the arrows would be very short, but in fact they are often very large. B. The best uncertainty preserving approximation ($\{q_{\mathrm{MF}}, q_{\mathrm{FC}}, q_{\mathrm{FT}}\}$) for finding λ (Top) and σ_{y}^2 (Bottom) indicated by marker-type (q_{MF} is never superior, q_{FC} black crosses, and q_{FT} black open-circles). The black solid line is $r = \sigma_{\mathrm{x}}^2/|\lambda|\sigma_{\mathrm{y}}^2 = 1$ and below it q_{FT} is tightest, and above it q_{FC} is tightest.

number of latent chains results in a posterior distribution that has a richer correlational structure. That is, the posterior covariance matrix has a greater number of off-diagonal terms. The variational approximations considered here would therefore ignore larger parts of this structure, and so one might expect the KL terms, and the associated biases, to be correspondingly larger.

There are many ways of assessing the performance of vEM. This chapter has focussed exclusively on the consistency of the methods and the biases in learned parameters. However, another relevant set of criteria come from tasks which require prediction of some kind, for instance, to fill-in missing data or to denoise. How does vEM fair in this new scenario? In order to answer this question it is necessary to specify the task more accuarately. Consider then, a task in which the first stage involves learning the model parameters from a training set, and the second involves filling in a section of missing-data in a test set using the mean of the approximate posterior. Given the same set of parameters, all four appoximations will make identical predications for the missing section (the mean of the true posterior). The differences between the four approximations are therefore entirely dependent on the quality of the parameters learned during the first stage of the experiments. As the task requires accurate learning of both the temporal dynamics (to predict the missing latent variables), and the emission weights (to predict the missing data from the missing latent variables), all of the approximation schemes will perform poorly compared to the optimal prediction.

1.5 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 tightness 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, the 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 the 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 (e.g. mixtures of variational approximations Jaakkola and Jordan, 1998).

Acknowledgements

We thank David Mackay for inspiration. This research was supported by the Gatsby Charitable Foundation.

Bibliography

Beal, M. J. (1998). *Variational Algorithms for approximate Bayesian Inference*. PhD thesis, University College London.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Dempster, A. P. (1977). Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

Hathaway, R. (1986). Another interpretation of the EM algorithm for mixture distributions. *Statistics and Probability Letters*, 4:53–56.

Jaakkola, T. and Jordan, M. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37.

Jaakkola, T. S. and Jordan, M. I. (1998). Improving the mean field approximation via the use of mixture distributions.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233.

MacKay, D. J. C. (2001). A problem with variational free energy minimization.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press. available from http://www.inference.phy.cam.ac.uk/mackay/itila/.

Neal, R. and Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers.

Turner, R. E., Berkes, P., Sahani, M., and MacKay, D. J. C. (2008). Counterexamples to variational free-energy compactness folk theorems. Technical report, University College London.

Wang, B. and Titterington, D. M. (2004). Lack of consistency of mean field and variational bayes approximations for state space models. *Neural Processing Letters*, 20(3):151–170.

Winn, J. and Minka, T. (2007). Expectation propagation and variational message passsing: a comparison with infer.net. Neural Information Processing Systems Workshop: Inference in continuous and hybrid models.