Technical Report for Biases in Variational Learning

Richard E. Turner, Pietro Berkes, and Maneesh Sahani
Gatsby Computational Neuroscience Unit
17 Alexandra House, Queen Square, London, WC1N 3AR, London

Abstract

This is the technical report for the paper "Biases in Variational Learning". It fills in some of the technical steps and contributes a couple of different examples. It should *not* be read in isolation as it does not include all of the examples in the paper and there is little in the way of explanation especially of the high level points.

1 Factored Gaussian Variational Approximations

We will often find ourselves minimizing the KL between a factored Gaussian and a correlated Gaussian. That is,

$$\mathrm{KL}(q(\mathbf{x})|p(\mathbf{x})) = -\frac{1}{2}\sum_{k=1}^{K}\log\sigma_k^2 + \frac{1}{2}\log\det\Sigma^{-1} + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma_x\right), \tag{1}$$

$$\Sigma_x = \left\langle(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right\rangle_{q(\mathbf{x})}. \tag{2}$$

Optimising this for the means μ_k and then the variances σ_k^2 gives us the following optimal settings,

$$\mu_k = \boldsymbol{\mu}_k, \quad \sigma_k^2 = \frac{1}{\Sigma^2_{kk}}. \tag{3}$$

That is, the mean and precisions (diagonal of the inverse covariance matrix) of the approximating distribution match the mean and precisions of the true distribution.

2 Examples of Compactness

2.1 Compactness in continuous models

As a warm-up, we consider a simple illustration of the utility of the result from the last section. (This is a variant on Mackay, 2003 page 434). Consider a zero-mean two-dimensional Gaussian distribution with axes oriented in the directions $\mathbf{e}_1 = [1, 1]$ and $\mathbf{e}_2 = [1, -1]$ with variances σ_1^2 and σ_2^2 . That is,

$$\Sigma = \frac{1}{2}\sigma_1^2\mathbf{e}_1\mathbf{e}_1^T + \frac{1}{2}\sigma_2^2, \mathbf{e}_2\mathbf{e}_1^T. \tag{4}$$

Inverting this matrix, the variational updates are,

$$q(x_i) = \mathrm{Norm}\left(0, \frac{1}{2}\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right). \tag{5}$$

The approximating distribution is spherical. If the variance of the two components is very different, say $\sigma_1^2 >> \sigma_2^2$ then the width of the approximating distribution becomes, σ_2^2, and therefore independent of the longer length-scale. In this sense the approximation is becoming compact, matching the smallest length scale structure in the posterior.

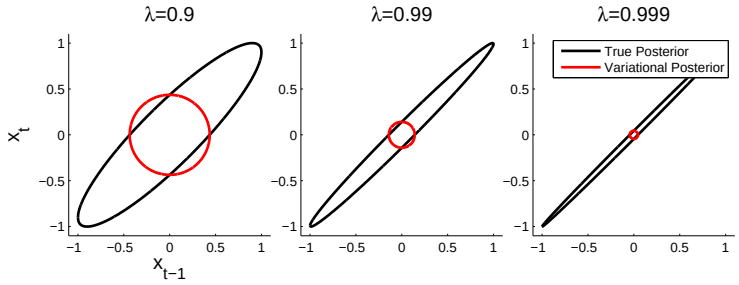

Figure 1: Temporally factored variational approximation for a linear dynamical system. As the slowness of the distribution increases, the variational approximation narrows, but the marginal variance is fixed. Notice the variational ellipse intersects the posterior ellipse on the axes indicating the variational distribution has the width of the conditional distribution.

One common situation where this effect is important arises when we factorise over strongly coupled variables and this is fairly routine for linear dynamical systems. By way of example, imagine we have the following AR(1) distribution,

$$p(x_t|x_{t-1}) = \mathrm{Norm}(\lambda x_{t-1}, \sigma^2). \tag{6}$$

The marginal mean of this system is zero and the marginal variance,

$$\sigma_\infty^2 = \frac{\sigma^2}{1-\lambda^2}. \tag{7}$$

Consider then the joint distribution,

$$p(x_{t-1}, x_t) = p(x_t|x_{t-1})p(x_{t-1}), \tag{8}$$

$$p(x_{t-1}) = \mathrm{Norm}(0, \sigma_\infty^2). \tag{9}$$

This is a Gaussian distribution with zero mean and covariance,

$$\Sigma^{-1} = \frac{1}{\sigma^2}\begin{bmatrix} 1 & -\lambda \\ -\lambda & 1 \end{bmatrix}, \quad \Sigma = \frac{\sigma^2}{1-\lambda^2}\begin{bmatrix} 1 & \lambda \\ \lambda & 1 \end{bmatrix}. \tag{10}$$

Let's approximate this using a factored variational approximation and use the results above to calculate the variational updates. Clealy this will cause problems as the latent variables are strongly correlated across time. In fact the example is exactly identical to the previous one, but $\sigma_1^2 = \frac{\sigma^2}{1-\lambda}$ and $\sigma_2^2 = \frac{\sigma^2}{1+\lambda}$.

So we can use the result directly and the variational approximation is,

$$q(x_t) = \mathrm{Norm}(0, \sigma^2), \tag{11}$$

$$q(x_{t-1}) = \mathrm{Norm}(0, \sigma^2). \tag{12}$$

That is, the distribution over x_t is exactly the same as if we had *known* the latent variable the previous time step to be zero $x_{t-1} = 0$. *None* of the uncertainty in the value of the previous latent has been folded in.

In fact the ratio of the variance of the variational approximation to the true variance is $\frac{\sigma_{\text{var}}^2}{\sigma_{\text{true}}^2} = 1 - \lambda^2$. Typically for LDS where the dynamics are slow, i.e. $\lambda = 1 - \Delta$ and Δ is a small quantity (e.g 10^{-1}-10^{-3}). The ratio of the variance of the approximating distribution to the truth is $\approx 2\Delta$ (meaning that the approximation is for example, 20% to 0.2% of the true variance).

2.2 Compactness in Discrete Models

Imagine a binary Markov Model which is symmetric, with slow dynamics,

$$p(x_1) = \frac{1}{2}\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \tag{13}$$

$$p(x_2|x_1) = \begin{bmatrix} 1-\Delta & \Delta \\ \Delta & 1-\Delta \end{bmatrix}. \tag{14}$$

So the system tends to stay in the same state (probability $1-\Delta$) rather than switching (probability Δ). The joint is simply,

$$p(x_1, x_2) = \frac{1}{2}\begin{bmatrix} 1-\Delta & \Delta \\ \Delta & 1-\Delta \end{bmatrix}. \tag{15}$$

We make a fully factored approximation to this distribution, $p(x_1, x_2) = q(x_1)q(x_2)$. This approximation is bound to fail catastrophically as we are approximating an asymmetric distribution with a symmetric one. As we shall see, the resulting distribution is compact. Defining $q(x_i = 1) = \rho_i$, the KL is,

$$\begin{aligned}&\mathrm{KL}(q(x_1, x_2)||p(x_1, x_2)) = \\ &- \rho_1\rho_2 \log \frac{1}{2}(1-\Delta) - (1-\rho_1)\rho_2 \log \frac{1}{2}\Delta - \rho_1(1-\rho_2)\log\frac{1}{2}\Delta - (1-\rho_1)(1-\rho_2)\log\frac{1}{2}(1-\Delta) \\ &+ \rho_1 \log \rho_1 + (1-\rho_1)\log(1-\rho_1) + \rho_2 \log \rho_2 + (1-\rho_2)\log(1-\rho_2).\end{aligned} \tag{16}$$

The variational distribution is,

$$\rho_1 = \frac{1}{1 + \left(\frac{\Delta}{1-\Delta}\right)^{2\rho_1 - 1}}, \quad \rho_1 = \rho_2. \tag{17}$$

So if there are strong temporal correlations so that $\Delta \to 0$ then there is symmetry breaking and $\rho_1 \to 1$ or $\rho_1 \to 0$ so the approximate joint is either

$$q(x_1, x_2) \to \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad \text{or} \quad q(x_1, x_2) \to \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{18}$$

Both of which are compact versions of the posterior, and do not fold in the uncertainty in x_1: just as was the case for the previous example with a linear dynamical system.

3 Biases in Variational Parameter Learning

3.1 Biases for Gaussian Linear Dynamical Systems

Here we provide some more details on the derivation of the bias results for the LDS. The LDS under consideration is,

$$p(x_{k,1}) = \mathrm{Norm}\left(0, \frac{\sigma_x^2}{1-\lambda^2}\right), \tag{19}$$

$$p(x_{k,2}) = \mathrm{Norm}\left(\lambda x_{k,1}, \sigma_x^2\right), \tag{20}$$

$$p(y_t | x_{1,t}, x_{2,t}) = \mathrm{Norm}(x_{1t} + x_{2t}, \sigma_y^2). \tag{21}$$

The joint Gaussian over the observations and latent variables is therefore Gaussian. The likelihood of the parameters is formed by marginalising, using $y_t = x_{1t} + x_{2t} + \sigma_y \epsilon_t$, $\langle x_{kt}^2 \rangle = \frac{\sigma_x^2}{1-\lambda^2}$ and $\langle x_{k1}x_{k2}\rangle = \frac{\lambda\sigma_x^2}{1-\lambda^2}$,

$$p(y_1, y_2|\theta) = \mathrm{Norm}(0, \Sigma_Y), \tag{22}$$

$$\Sigma_Y = I\sigma_y^2 + 2\frac{\sigma_x^2}{1-\lambda^2}\begin{bmatrix} 1 & \lambda \\ \lambda & 1 \end{bmatrix}. \tag{23}$$

This means the likelihood of N observations of the time series is,

$$\log p(Y|\theta) = -\frac{N}{2}\left(\log\det 2\pi\Sigma_y + \mathrm{tr}\left(\Sigma_y^{-1}\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{y}_n\boldsymbol{y}_n^T\right)\right) \tag{24}$$

Next we form the posterior distribution over the latent variables is also Gaussian. It can be derived by first finding the joint,

$$\begin{aligned}p(Y, X|\theta) = \frac{1}{Z}\exp\Bigg(&-\frac{1}{2}\left(\frac{1}{\sigma_y^2} + \frac{1}{\sigma_x^2}\right)\left(x_{11}^2 + x_{21}^2 + x_{12}^2 + x_{22}^2\right) - \frac{1}{2\sigma_y^2}(y_1^2 + y_2^2) \\ &+ \frac{\lambda}{\sigma_x^2}\left(x_{11}x_{12} + x_{21}x_{22}\right) + \frac{1}{\sigma_y^2}\left(y_1\left(x_{11} + x_{21}\right) + y_2\left(x_{12} + x_{22}\right) - x_{11}x_{21} - x_{12}x_{22}\right)\Bigg)\end{aligned} \tag{25}$$

The posterior is simply derived from the joint by picking off the functional dependence on the X, which is Gaussian. A symbolic package can be used to invert the corresponding inverse-covariance matrix and find the mean. Defining $\mathbf{x} = [x_{11}, x_{21}, x_{12}, x_{22}]^T$, it is given by,

$$p(\mathbf{x}|\mathbf{y}) = \mathrm{Norm}(\mu_{\mathbf{x}|\mathbf{y}}, \Sigma_{\mathbf{x}|\mathbf{y}}). \tag{26}$$

Where the inverse-covariance and mean are given by,

$$\Sigma^{-1}_{\mathbf{x}|\mathbf{y}} = \begin{bmatrix} \frac{1}{\sigma_y^2}+\frac{1}{\sigma_x^2} & \frac{1}{\sigma_y^2} & -\frac{\lambda}{\sigma_x^2} & 0 \\ \frac{1}{\sigma_y^2} & \frac{1}{\sigma_y^2}+\frac{1}{\sigma_x^2} & 0 & -\frac{\lambda}{\sigma_x^2} \\ -\frac{\lambda}{\sigma_x^2} & 0 & \frac{1}{\sigma_y^2}+\frac{1}{\sigma_x^2} & \frac{1}{\sigma_y^2} \\ 0 & -\frac{\lambda}{\sigma_x^2} & \frac{1}{\sigma_y^2} & \frac{1}{\sigma_y^2}+\frac{1}{\sigma_x^2} \end{bmatrix}, \quad \mu_{\mathbf{x}|\mathbf{y}} = \frac{1}{\sigma_y^2}\Sigma_{\mathbf{x}|\mathbf{y}} \begin{bmatrix} y_1 \\ y_1 \\ y_2 \\ y_2 \end{bmatrix}. \tag{27}$$

(28)

We now consider three different variational approximations. The fully factored mean-field approximation, fatorisation over time, and factorisation over chains.

	unfactored over chains	factored over chains
unfactored over time	$p(\mathbf{x}\|\mathbf{y}) = q(x_{11}, x_{12}, x_{21}, x_{22})$	$q_2(\mathbf{x}) = q_{21}(x_{11}, x_{12})q_{22}(x_{21}, x_{22})$
factored over time	$q_3(\mathbf{x}) = q_{31}(x_{11}, x_{21})q_{32}(x_{12}, x_{22})$	$q_1(\mathbf{x}) = q_{11}(x_1)q_{12}(x_2)q_{13}(x_3)q_{14}(x_4)$

The optimal updates for these distributions can be found by minimising the variational KL (Eq. **??**). The solution is that each factor is Gaussians with a mean and precision that matches the corresponding elements in $\mu_{\mathbf{x}|\mathbf{y}}$ and $\Sigma^{-1}_{\mathbf{x}|\mathbf{y}}$ (see section 1). Furthermore, as the approximations and posterior distributions are all Gaussians, it is possible to compute the KL divergences between them using the relationship in Eq. 2. As the approximating distributions have been found by minimizing the KL, things simplify to leave,

$$\mathrm{KL}_i\left(\prod_{a=1}^{A} q_{ia}(\mathbf{x}_a)||p(\mathbf{x}|\mathbf{y})\right) = \frac{1}{2}\log\det\Sigma_{\mathbf{x}|\mathbf{y}} - \frac{1}{2}\sum_a \log\det\Sigma_{ia}. \tag{29}$$

Which is the log of the ratio between the volume of the true posterior and the volume of the approximation. As the KL-divergence is positive, the volume of the approximation is always smaller than the volume of the true posterior, which is our friend, the compactness property. Using this expression we find,

$$\mathrm{KL}_1 = \frac{1}{2}\log\frac{\left(\sigma_y^2+\sigma_x^2\right)^4}{\sigma_y^4\gamma} \tag{30}$$

$$\mathrm{KL}_2 = \frac{1}{2}\log\frac{\left(\left(\sigma_y^2+\sigma_x^2\right)^2-\lambda^2\sigma_y^4\right)^2}{\sigma_y^4\gamma} \tag{31}$$

$$\mathrm{KL}_3 = \frac{1}{2}\log\frac{\left(\sigma_y^2+2\sigma_x^2\right)^2}{\gamma} \tag{32}$$

where

$$\gamma = (1-\lambda^2)\left((2\sigma_x^2+\sigma_y^2)^2-\lambda\sigma_y^4\right) \tag{33}$$

These results are easily checked using a symbolic maths package.

For completeness we check everything using EM and the E-Step updates have essentially been given (matching means and correponding precisions). The M-Step updates are,

$$\lambda = -\frac{1}{a}+\mathrm{sign}(a)\sqrt{1+\frac{1}{a^2}}, \;\; a = \frac{1}{N\sigma_x^2}\sum_{n=1}^{N}\left(\langle x_{11}^{(n)}x_{12}^{(n)}\rangle + \langle x_{21}^{(n)}x_{22}^{(n)}\rangle\right) \tag{34}$$

$$\sigma_x^2 = \frac{1}{4N}\sum_{n=1}^{N}\Bigg(\langle\left(x_{11}^{(n)}\right)^2\rangle + \langle\left(x_{21}^{(n)}\right)^2\rangle + \langle\left(x_{12}^{(n)}\right)^2\rangle + \langle\left(x_{22}^{(n)}\right)^2\rangle - 2\lambda\left(\langle x_{11}^{(n)}x_{12}^{(n)}\rangle + \langle x_{21}^{(n)}x_{22}^{(n)}\rangle\right)\Bigg) \tag{35}$$

$$\sigma_y^2 = \frac{1}{2N}\sum_{n=1}^{N}\Bigg(\langle\left(x_{11}^{(n)}\right)^2\rangle + \langle\left(x_{21}^{(n)}\right)^2\rangle + \langle\left(x_{12}^{(n)}\right)^2\rangle + \langle\left(x_{22}^{(n)}\right)^2\rangle + \left(y_1^{(n)}\right)^2 + \left(y_2^{(n)}\right)^2 - 2\left(y_1\left(\langle x_{11}^{(n)}\rangle + \langle x_{21}^{(n)}\rangle\right) + y_2\left(\langle x_{12}^{(n)}\rangle + \langle x_{22}^{(n)}\rangle\right) + \langle x_{11}x_{21}\rangle + \langle x_{12}x_{22}\rangle\right)\Bigg) \quad (36)$$

Finally we derive the MAP estimate of the latent variables. As the log-joint (Eq. 25) is quadratic in the latent variables, the derivatives with respect to the latent variables yield the following linear updates,

$$\begin{bmatrix} x_{11}^{(i+1)} \\ x_{21}^{(i+1)} \\ x_{12}^{(i+1)} \\ x_{22}^{(i+1)} \end{bmatrix} = \frac{1}{\sigma_x^2+\sigma_y^2}\begin{bmatrix} 0 & -\sigma_x^2 & \lambda\sigma_y^2 & 0 \\ -\sigma_x^2 & 0 & 0 & \lambda\sigma_y^2 \\ \lambda\sigma_y^2 & 0 & 0 & -\sigma_x^2 \\ 0 & \lambda\sigma_y^2 & -\sigma_x^2 & 0 \end{bmatrix}\begin{bmatrix} x_{11}^{(i)} \\ x_{21}^{(i)} \\ x_{12}^{(i)} \\ x_{22}^{(i)} \end{bmatrix} + \frac{\sigma_x^2}{\sigma_x^2+\sigma_y^2}\begin{bmatrix} y_1 \\ y_1 \\ y_2 \\ y_2 \end{bmatrix} \quad (37)$$

Defining $\mathbf{x}^{i+1} = W\mathbf{x}^i + \gamma$, the optimal value for the latents is given by the eigenvector, with eigenvalue one, of the matrix,

$$M = \begin{bmatrix} W & \gamma \\ \mathbf{0^T} & 1 \end{bmatrix} \quad (38)$$

This can be computed using a sybolic maths package, and the solution is,

$$e_{\lambda=1} = \frac{\sigma_x^2}{Z}\begin{bmatrix} (2\sigma_x^2+\sigma_y^2)y_1 + \lambda\sigma_y^2 y_2 \\ (2\sigma_x^2+\sigma_y^2)y_1 + \lambda\sigma_y^2 y_2 \\ (2\sigma_x^2+\sigma_y^2)y_2 + \lambda\sigma_y^2 y_1 \\ (2\sigma_x^2+\sigma_y^2)y_2 + \lambda\sigma_y^2 y_1 \end{bmatrix} \quad (39)$$

Where $Z = 4\sigma_x^4 - \lambda^2\sigma_y^4 + 4\sigma_x^2\sigma_y^2 + \sigma_y^4$. Notice this has the symmetry properties we would expect: $x_{11} = x_{21}$ and $x_{k1}(y_1, y_2) = x_{k2}(y_2, y_1)$. Unsurprisingly this is identical to the variational updates for the means as the posterior is Gaussian and, as such, the mode and the mean are in the same place. These results can be substituted into the log-joint enabling us to evaluate the MAP-objective purely as a function of the parameters.

3.2 Biases for learning weights in factor analysis (or factorising over time in LDSSMs)

The results from the previous section seem to indicate that factorising over chains is a better approach than factorising over time in LDSSMs. This is because the usual regime in which to use time-series models is where the observation noise (σ_y^2) is large and the dynamics are strong ($\sigma_x^2 \approx 0$ and $|\lambda| \approx 1$). In this regime q_2 did rather better than q_3 as the dynamical information is critical.

However, the above example had fixed generative weights and one-dimensional observations ($y_t = x_{1t} + x_{2t}$). We now consider what happens when you factorise over chains in the case where you are learning the weights. As explaining away is neglected when using an approximation which is factored across chains, and as a proper treatment of explaining away is required to estimate the weights correctly, we might expect to see a severe bias.

Although we could consider a time-series model, it is simpler to consider a factor analysis (FA) model. The results will clearly generalise to time-series: The chain-factored approximation for linear Gaussian state-space models is very similar to a mean-field approximation for factor analysis from the perspective of learning the weights. In particular, we consider a FA model with two latent variables. In order to disentangle a bias in estimating magnitude and a bias in direction, we parameterise the weights in terms of these two quantities,

$$p(x_k) = \mathrm{Norm}(0,1), \quad p(\boldsymbol{y}|\boldsymbol{x}, W) = \mathrm{Norm}(\boldsymbol{w}_1 x_1 + \boldsymbol{w}_2 x_2, \sigma_y^2), \quad (40)$$

$$\boldsymbol{w}_k = w_k[\cos(\theta_k), \sin(\theta_k)]^T. \quad (41)$$

We will now consider maximum-likelihood, variational mean-field, and zero-temperature EM learning algorithms for the directions (θ_1 and θ_2) and magnitudes (w_1 and w_2). The likelihood is Gaussian

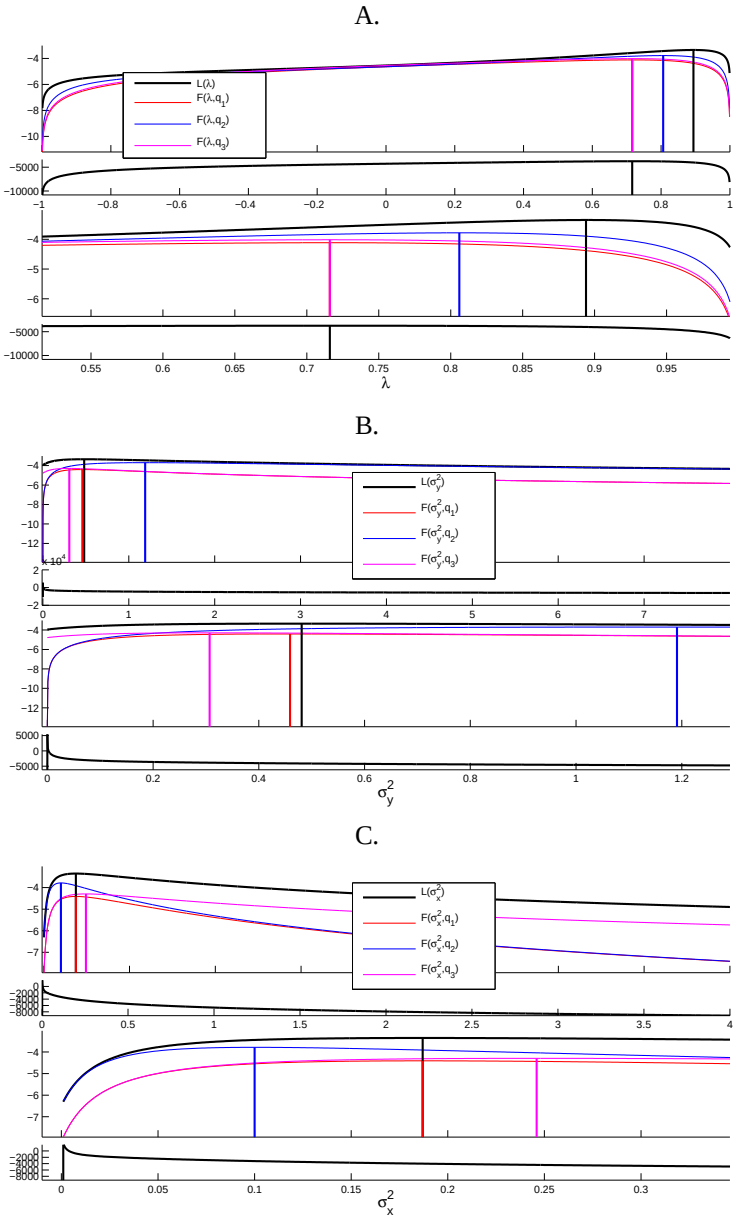

Figure 2: Biases in the Free-energies for a simple linear dynamical system. True/ML parameters are $\lambda = 0.9$, $\sigma_x^2 = 1 - \lambda^2 = 0.19$, and $\sigma_y^2 = 0.43$. In each case one parameter is learned and the others are set to their true/ML values. A. learning λ, B. learning σ_y^2, C. learning σ_x^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.

and given by,

$$p(\boldsymbol{y}) = \text{Norm}(\mathbf{0}, \Sigma_{\boldsymbol{y}}), \tag{42}$$

$$\Sigma_{\boldsymbol{y}} = \begin{bmatrix} w_1^2\cos^2(\theta_1) + w_2^2\cos^2(\theta_2) + \sigma_y^2 & w_1^2\sin(\theta_1)\cos(\theta_1) + w_2^2\sin(\theta_2)\cos(\theta_2) \\ w_1^2\sin(\theta_1)\cos(\theta_1) + w_2^2\sin(\theta_2)\cos(\theta_2) & w_1^2\sin^2(\theta_1) + w_2^2\sin^2(\theta_2) + \sigma_y^2 \end{bmatrix}. \tag{43}$$

The posterior is also Gaussian,

$$p(\boldsymbol{x}|\boldsymbol{y}) = \text{Norm}(\mu_{\boldsymbol{x}|\boldsymbol{y}}, \Sigma_{\boldsymbol{x}|\boldsymbol{y}}), \tag{44}$$

$$\mu_{\boldsymbol{x}|\boldsymbol{y}} = \frac{1}{\sigma_y^2}\Sigma_{\boldsymbol{x}|\boldsymbol{y}} \begin{bmatrix} y_1 w_1 \cos(\theta_1) + y_2 w_1 \sin(\theta_1) \\ y_1 w_2 \cos(\theta_2) + y_2 w_2 \sin(\theta_2) \end{bmatrix}, \tag{45}$$

$$\Sigma^{-1}_{\boldsymbol{x}|\boldsymbol{y}} = \frac{1}{\sigma_y^2} \begin{bmatrix} w_1^2 + \sigma_y^2 & w_1 w_2 \cos(\theta_1 - \theta_2) \\ w_1 w_2 \cos(\theta_1 - \theta_2) & w_2^2 + \sigma_y^2 \end{bmatrix}. \tag{46}$$

As expected the covariance of the posterior is only affected by changing the relative angle **between** the two weights, and not on the absolute angle. This is because explaining away gets larger when the weights point in the same direction.

The chain-factored variational approximation is $q(\boldsymbol{x}) = q(x_1)q(x_2)$ and we have seen that the optimal updates for the factors are Gaussians which match the means and precisions of the posterior. That is,

$$q(x_1) = \text{Norm}\left(\mu_{\boldsymbol{x}|\boldsymbol{y}}(1), \frac{\sigma_y^2}{\sigma_y^2 + w_1^2}\right), \quad q(x_1) = \text{Norm}\left(\mu_{\boldsymbol{x}|\boldsymbol{y}}(2), \frac{\sigma_y^2}{\sigma_y^2 + w_2^2}\right). \tag{47}$$

Using the results from the previous section the KL divergence between the posterior and this distribution is the log ratio of the determinant of the posterior to the determinant of the approximation,

$$\text{KL}(q(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{y})) = \log\left(\frac{(\sigma_y^2 + w_1^2)(\sigma_y^2 + w_2^2)}{(\sigma_y^2 + w_1^2)(\sigma_y^2 + w_2^2) - w_1^2 w_2^2 \cos^2(\theta_1 - \theta_2)}\right) \tag{48}$$

Dependence on the directions

Imagine for a moment that that we know the magnitude of the weights is unity ($w_1 = w_2 = 1$) and that the observation noise is very small, $\sigma_y^2 \to 0$. The posterior is highly correlated in this case and the variational approximation will be at its poorest. However, the compactness property means it becomes ulta-confident and the variance of the variational approximation becomes tiny (the uncertainty in x_k is σ_y^2/w_k^2). More interestingly, the KL-divergence becomes $\text{KL} \to \log\left(\frac{1}{1-cos^2(\theta_1-\theta_2)}\right)$. This is zero when the difference in angle between the weights is $\pi/2$ degrees. This makes sense as there is no explaining away in this case and so the mean-field approximation is exact. However, if the weights point in the same direction (and the difference in angle is zero) the KL is infinite. Again this makes sense as explaining away is greatest here and the mean-field approximation tends to a delta function. The conclusion is that **the weights derived from variational learning will tend to be more orthogonal than the true weights**. Furthermore, although the KL increases with decreasing observation noise (σ_y^2), the bias turns out to be independent of σ_y^2 and this because the likelihood becomes more peaked as the observation noise decreases, exactly cancelling the KL contribution (!!! I haven't proved this, but it is empirically true over a very wide range of values !!!).

Dependence on the directions

Imagine now we know that we know $w_1 = 1$, and the goal is to infer w_2. Moreover, let the directions of the weights also be known and the difference in angle is $\theta_1 - \theta_2 = \pi/4$. The KL-divergence becomes $\text{KL} \to \log\left(\frac{(\sigma_y^2+1)(\sigma_y^2+w_2^2)}{(\sigma_y^2+1)(\sigma_y^2+w_2^2)-w_2^2/2}\right)$. This is zero when the magnitude of the second weight is zero ($w_2 = 0$), and tends to $\log\frac{1+\sigma_y^2}{1/2+\sigma_y^2}$ as the magnitude increases to infinity $w_2 \to \infty$. This results in a small bias in inferring the magnitude of the weight.

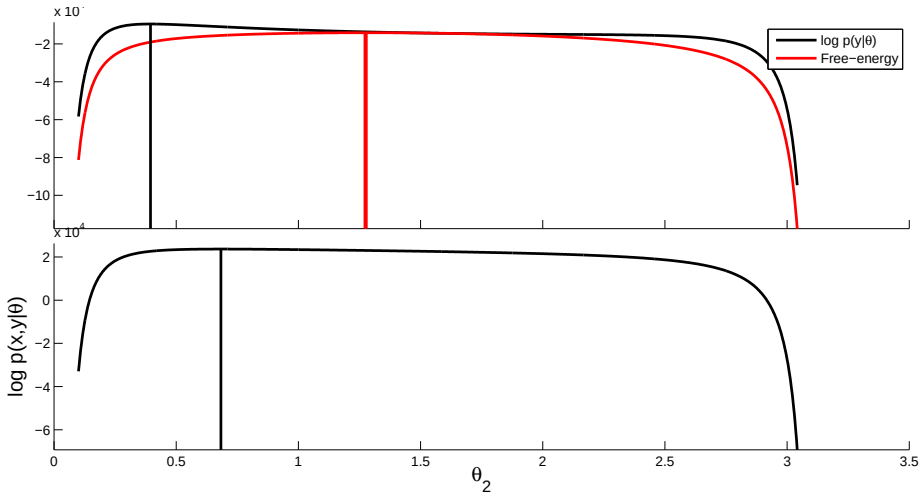

Figure 3: Top: Likelihood (black line) and Free-energy (red line) as a function of θ_2 when θ_1 is set to the true/ML value of $\theta_1 = 0$. The Free-energy peaks at a value closer to $\pi/2$ than the true peak in the likelihood (which is at $\theta = \pi/8 = 22.5$ degrees). Bottom: the log-joint or Zero-temperature EM objective function is significantly biased, though less so than the variational method.

We also consider the zero-temperature EM objective function which is the average log-joint evaluated at the MAP setting of the latents.

$$\log p(\boldsymbol{x}, \boldsymbol{y}) = -\log(2\pi) - \log(2\pi\sigma_y^2) - \frac{1}{2}\Bigg(x_1^2 + x_2^2 + \frac{1}{\sigma_y^2}\left(y_1 - w_1\cos(\theta_1)x_1 - w_2\cos(\theta_2)x_2\right)^2 + \frac{1}{\sigma_y^2}\left(y_2 - w_1\sin(\theta_1)x_1 - w_2\sin(\theta_2)x_2\right)^2 \Bigg) \tag{49}$$

The MAP and the mean are the same for a Gaussian and so the MAP setting for the latents is identical to the posterior mean.

Results: Learning the directions
We consider learning θ_2. For convenience θ_1 remains fixed at its true/ML value of 0. The results show,

1. The variational method returns an estimate of θ_2 which is biased toward the direction orthogonal to θ_1 ($\frac{1}{2}\pi$) (see Fig. 3 and Fig. 4). This bias is greatest when the true values of the angles are close together $\theta_1 = \theta_2$ and smallest when they are orthogonal $\theta_2 - \theta_1 = \frac{1}{2}\pi$.
2. The bias in the variational method is not affected by the observation noise. This is because the sharpening of the likelihood which occurs as the observation noise decreases cancels the increase in the KL term.
3. The zero-temperature EM method also returns an estimate of the weights which is often significantly biased toward the orthogonal direction (see Fig. 3). However, this bias is smaller than that for the variational approach. Moreover, the bias has a different trend being smallest when the weights are orthogonal *or* when they are very similar and greatest at intermediate regions.
4. The bias in the variational method is affected by the observation noise (see Fig. 5)

Results: Learning the magnitudes
We consider learning w_2. For convenience w_1 remains fixed at its ML/true value of 1.

1. The variational method returns an under-estimate of w_2 as it is always biased toward the zero (e.g. Fig. 6.

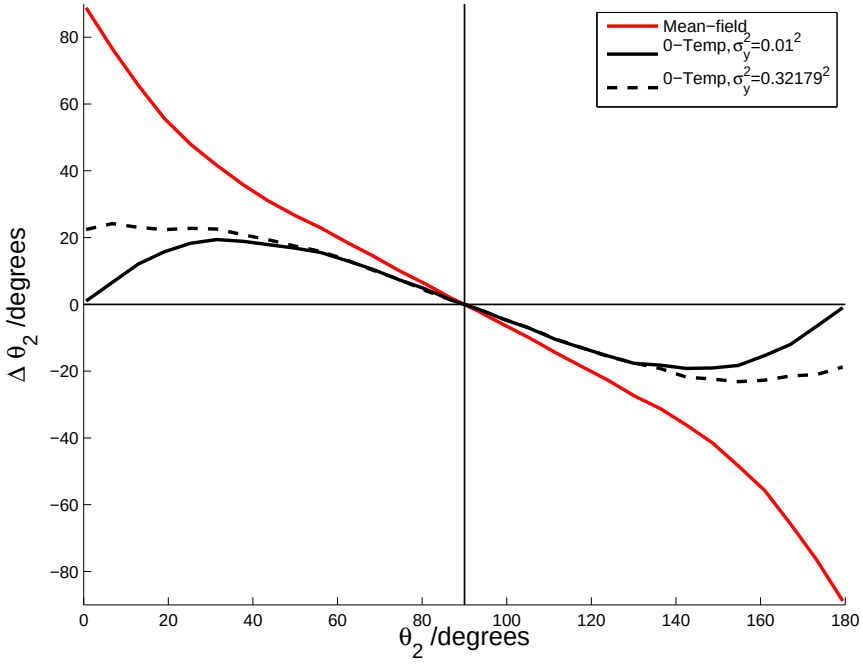

Figure 4: Bias in inferring θ_2 as a function of the true/ML value when $\theta_1 = 0$. Red line: The mean-field bias is smallest when the weights are orthogonal ($\Delta\theta = 0$) and greatest when they are (anti-) parallel ($\Delta\theta = 90$). Black lines: The zero-temperature EM bias is smaller than that for mean-field and is largest when the weights partially overlap. The bias grows with increasing noise level.

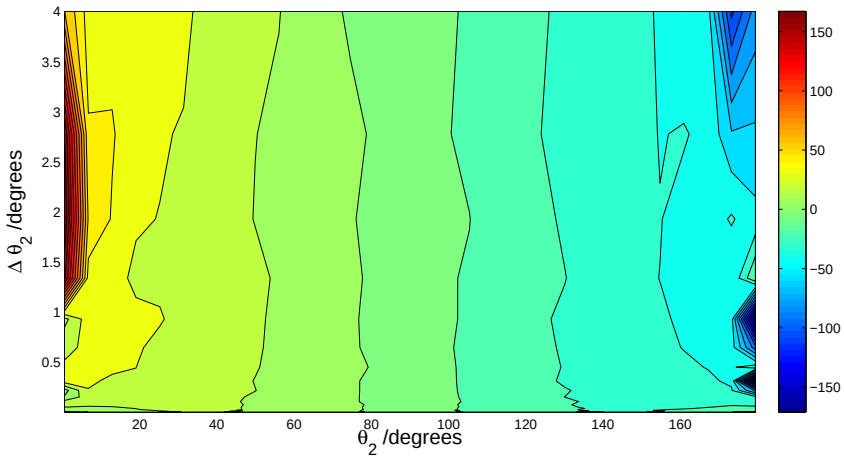

Figure 5: Bias in inferring θ_2 as a function of the observation noise σ_y^2 and θ_2 for zero-temperature EM. The bias is often very large and it depends on the observation noise (in a complex manner), unlike the case with the variational method.

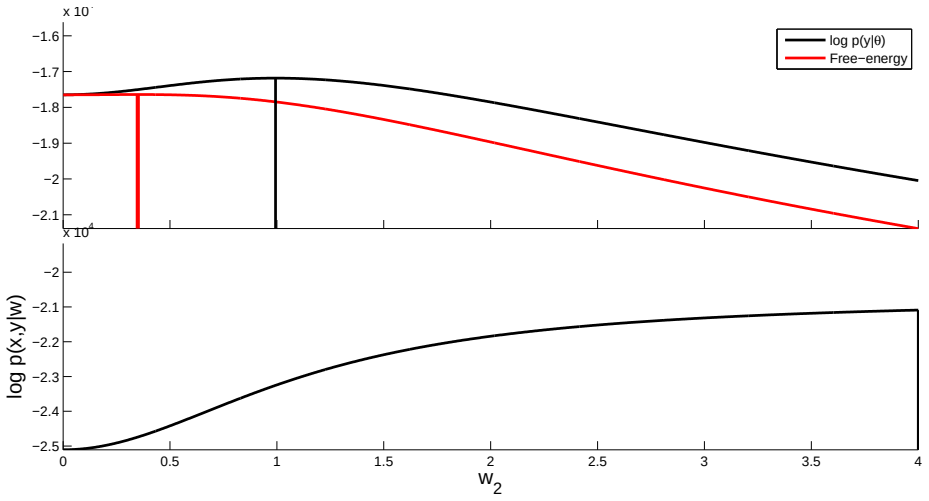

Figure 6: Top: Likelihood (black line) and Free-energy (red line) as a function of w_2 when w_1 is set to the true/ML value of $w_1 = 1$, $\Delta\theta = \pi/4 = 45$ degrees). The Free-energy peaks at a value closer to 0 than the true peak in the likelihood. Bottom: the log-joint or Zero-temperature EM objective function suffers from an over-fitting problem and returns an infinite value for the weights.

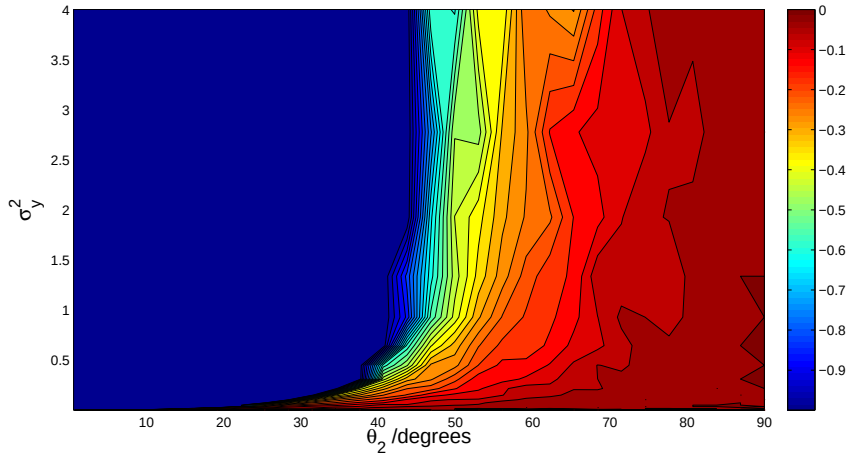

Figure 7: Top: Likelihood (black line) and Free-energy (red line) as a function of w_2 when w_1 is set to the true/ML value of $w_1 = 1$, $\Delta\theta = \pi/4 = 45$ degrees). The Free-energy peaks at a value closer to 0 than the true peak in the likelihood. Bottom: the log-joint or Zero-temperature EM objective function suffers from an over-fitting problem and returns an infinite value for the weights.

2. This bias is greatest when correlations in the posterior are largest and therefore when the observation noise is small and the angle between the weights small (see Fig. 7.
3. The zero-temperature EM method returns an infinite value for the weights (and an infinitesimal value for the latent variables).

Conclusions

1. Chain-factored approximations in time-series will have a bias toward finding weights which are more orthogonal and smaller in magnitude than the true/ML weights. This can cause 1. weights repelling one another 2. pruning of weights.
2. Zero-temperature EM also results in a similar bias for direction. As this is the method of choice for learning ICA models, this might warrant further investigation. Zero-temperature

EM cannot recover the magnitudes correctly, and this has to be handled by a hack (like renormalising the weights).

3. It is a general feature that the directions and magnitudes of the weights in a model will effect the dependencies in the posterior. So, although mean-field approximations are not the sort of approximation that are used to learn e.g. ICA models, other approaches (e.g. that used in the following section) do impose some type of factorisation, and it is likely that the weights will be able to arrange themselves away from the ML solution in order to reduce dependencies in the posterior to some extent (e.g. In the example that follows the factorisation ends up making the binary latent variables independent in the approximate posterior, and so if this method was extended to parameter learning the weights would be biased toward becoming orthogonal).

3.3 Biases for Independent Component Analysis

In order to look at the success of variational approximations when explaining away causes the posterior to be multi-modal we studied a very simple ICA model with two latent variables and one dimensional observations. The distributions over the latent variables are a mixture of two, zero mean Gaussians. The mixing proportions are $1/2$, but the variances differ,

$$p(s_k) = \frac{1}{2}, \tag{50}$$

$$p(x_k|s_k) = \mathrm{Norm}\left(0, (1-s_k)\sigma^2 + s_k(2-\sigma^2)\right), \tag{51}$$

$$p(y|x_1, x_2) = \mathrm{Norm}(x_1 + x_2, \sigma_y^2). \tag{52}$$

The variance of the latent variables is fixed to be unity as,

$$\langle x_1^2 \rangle - \langle x_1 \rangle^2 = \langle x_1^2 \rangle = \frac{1}{2}\sigma^2 + \frac{1}{2}(2-\sigma^2) = 1. \tag{53}$$

The kurtosis can be varied between zero ($\sigma = 1$) and three ($\sigma = 0$) as,

$$K(x) = \frac{\left\langle (x - \langle x \rangle)^4 \right\rangle}{\left(\langle x^2 \rangle - \langle x \rangle^2\right)^2} - 3 = \langle x^4 \rangle - 3 \tag{54}$$

$$\frac{3}{2}\left(\sigma^4 + (2-\sigma^2)^2\right) - 3 = 3(1-\sigma^2)^2. \tag{55}$$

The goal is to produce a one dimensional plot of the likelihood and free-energy as a function of this kurtosis (equivalently σ^2).

The key quantity in deriving the exact EM updates and the variational updates (and therefore the free-energies and likelihood) is the log-joint,

$$\begin{aligned}\log p(\boldsymbol{x}, \boldsymbol{s}, y) = &-\log 4 - \frac{3}{2}\log 2\pi - \frac{1}{2}\log\sigma_y^2 - \frac{1}{2}(s_1+s_2)\log(2-\sigma^2) - \frac{1}{2}(2-s_1-s_2)\log\sigma^2 \\ &- \frac{1}{2}x_1^2\left(2\gamma s_1 + \frac{1}{\sigma^2} + \frac{1}{\sigma_y^2}\right) - \frac{1}{2}x_2^2\left(2\gamma s_2 + \frac{1}{\sigma^2} + \frac{1}{\sigma_y^2}\right) \\ &- \frac{x_1 x_2}{\sigma_y^2} + \frac{y}{\sigma_y^2}(x_1 + x_2) - \frac{1}{2\sigma_y^2}y^2 \end{aligned} \tag{56}$$

where $\gamma = \frac{\sigma^2 - 1}{\sigma^2(2-\sigma^2)}$. The posterior distribution is formed from renormalising the exponential of this quantity and that is a mixture of Gaussians,

$$p(\boldsymbol{x}, \boldsymbol{s}|y) = \pi(\boldsymbol{s})\mathrm{Norm}_{\boldsymbol{x}}(\mu(\boldsymbol{s}), \Sigma(\boldsymbol{s})) \tag{57}$$

The covariance and mean of the mixture components are,

$$\Sigma(\boldsymbol{s})^{-1} = \begin{bmatrix} 2s_1\gamma + \frac{1}{\sigma_y^2} + \frac{1}{\sigma^2} & \frac{1}{\sigma_y^2} \\ \frac{1}{\sigma_y^2} & 2s_2\gamma + \frac{1}{\sigma_y^2} + \frac{1}{\sigma^2} \end{bmatrix}, \qquad \mu(\boldsymbol{s}) = \frac{y}{\sigma_y^2}\Sigma(\boldsymbol{s}) \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{58}$$

The mixing proportions are a little more complicated, and include a term which is the likelihood of the parameters,

$$\pi(\boldsymbol{s}) = \frac{1}{p(y)}\tilde{\pi}(\boldsymbol{s})\frac{1}{4\sqrt{2\pi\sigma_y^2}}\exp\left(-\frac{1}{2\sigma_y^2}y^2\right) \tag{59}$$

$$\tilde{\pi}(\boldsymbol{s}) = \left(2-\sigma^2\right)^{-\frac{1}{2}(s_1+s_2)}\left(\sigma^2\right)^{-\frac{1}{2}(2-s_1-s_2)}\left(\det\left(2\pi\Sigma_{\boldsymbol{s}}\right)\right)^{1/2}\exp\left(\frac{1}{2}\mu(\boldsymbol{s})^T\Sigma_{\boldsymbol{s}}^{-1}\mu(\boldsymbol{s})\right) \tag{60}$$

$$p(y) = \frac{1}{4\sqrt{2\pi\sigma_y^2}}\exp\left(-\frac{1}{2\sigma_y^2}y^2\right)\sum_{s_1,s_2}\tilde{\pi}(\boldsymbol{s}) \tag{61}$$

The next task is to compute the variational updates. We use an approach which ignores dependencies between the binary and Gaussian latent variables, but which captures dependencies between latent variables of the same type, that is $q(\boldsymbol{s},\boldsymbol{x}) = q(\boldsymbol{s})q(\boldsymbol{x})$. The updates are easily computed from the log-joint,

$$q(\boldsymbol{x}) = \frac{1}{Z_x}\exp\left\langle\log p(\boldsymbol{x},\boldsymbol{s},y)\right\rangle_{q(\boldsymbol{s})} = \mathrm{Norm}(\mu_q,\Sigma_q) \tag{62}$$

$$\Sigma_q^{-1} = \begin{bmatrix} 2\langle s_1\rangle\gamma + \frac{1}{\sigma_y^2} + \frac{1}{\sigma^2} & \frac{1}{\sigma_y^2} \\ \frac{1}{\sigma_y^2} & 2\langle s_2\rangle\gamma + \frac{1}{\sigma_y^2} + \frac{1}{\sigma^2} \end{bmatrix}, \qquad \mu_q = \frac{y}{\sigma_y^2}\Sigma_q\begin{bmatrix}1\\1\end{bmatrix} \tag{63}$$

$$q(\boldsymbol{s}) = \frac{1}{Z_S}\exp\left\langle\log p(\boldsymbol{x},\boldsymbol{s},y)\right\rangle_{q(\boldsymbol{x})} \tag{64}$$

$$q(s_1,s_2) = \frac{1}{Z_S}\begin{bmatrix} \frac{1}{\sigma^2} & \frac{1}{\sqrt{\sigma^2(2-\sigma^2)}}\exp\left(-\gamma\langle x_2^2\rangle\right) \\ \frac{1}{\sqrt{\sigma^2(2-\sigma^2)}}\exp\left(-\gamma\langle x_1^2\rangle\right) & \frac{1}{2-\sigma^2}\exp\left(-\gamma\left(\langle x_1^2\rangle + \langle x_2^2\rangle\right)\right) \end{bmatrix} \tag{65}$$

The optimal distribution over $\boldsymbol{s}$ is therefore factored, $q(\boldsymbol{s}) = q(s_1)q(s_2)$.

Ideally we would like to analytically find the fixed-point reached by iterating these two updates. However, this seems impossible. Instead we have to satisfy ourselves with picking a kurtosis, iterating these two updates until convergence and then computing the free-energy. We can check the maxima located in this way by running variational EM.

All that remains to be computed is the free energy which is the average log-joint plus the entropy of the approximating distribution,

$$F(q(\boldsymbol{x},\boldsymbol{s}),\theta) = \left\langle\log p(\boldsymbol{x},\boldsymbol{s},\theta)\right\rangle + \mathrm{H}(q(\boldsymbol{x})) + \mathrm{H}(q(\boldsymbol{s})) \tag{66}$$

The average log-joint is,

$$\begin{aligned}\log p(\boldsymbol{x},\boldsymbol{s},y) = &-\log 4 - \frac{3}{2}\log 2\pi - \frac{1}{2}\log\sigma_y^2 - \frac{1}{2}(\langle s_1\rangle + \langle s_2\rangle)\log(2-\sigma^2) - \frac{1}{2}(2-\langle s_1\rangle - \langle s_2\rangle)\log\sigma^2 \\ &- \frac{1}{2}\left\langle x_1^2\right\rangle\left(2\gamma s_1 + \frac{1}{\sigma^2} + \frac{1}{\sigma_y^2}\right) - \frac{1}{2}\left\langle x_2^2\right\rangle\left(2\gamma\langle s_2\rangle + \frac{1}{\sigma^2} + \frac{1}{\sigma_y^2}\right) \\ &- \frac{\langle x_1x_2\rangle}{\sigma_y^2} + \frac{y}{\sigma_y^2}(\langle x_1\rangle + \langle x_2\rangle) - \frac{1}{2\sigma_y^2}y^2\end{aligned} \tag{67}$$

The entropy of the Gaussian distribution over $\boldsymbol{x}$ is,

$$\mathrm{H}(q(\boldsymbol{x})) = \frac{1}{2}\left(\log\det 2\pi\Sigma_q + 2\right) \tag{68}$$

The entropy of the distribution over $\boldsymbol{s}$ is,

$$\mathrm{H}(q(\boldsymbol{s})) = -\sum_{s_1=0}^{1}\sum_{s_2=0}^{1}q(s_1,s_2)\log q(s_1,s_2) \tag{69}$$

If the sufficient statistics always converged on a symmetric solution significant simplification could be achieved. Unfortunately, small minority of solutions break this symmetry and so this is not possible.

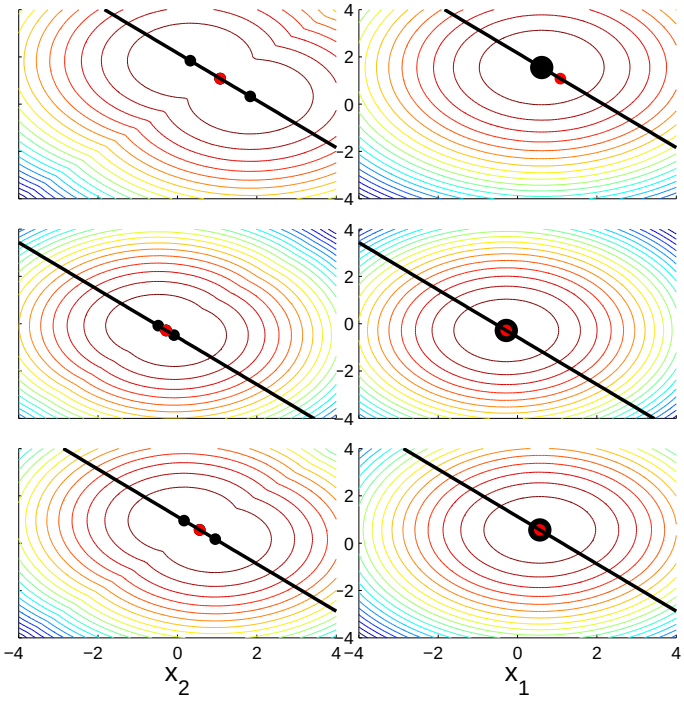

Figure 8: The posterior in over-complete ICA models is complex and multi-modal (left). The variational posterior is simple, unimodal and compact (right).

Finally the M-Step updates are made used a gradient based optimisation scheme (conjugate gradients) and are given by,

$$2\frac{d\langle \log p(y, \boldsymbol{s}, \boldsymbol{x})\rangle}{d\sigma^2} = -\frac{2}{\sigma^2} + (\langle s_1\rangle + \langle s_1\rangle)\frac{2}{\sigma^2(2-\sigma^2)} - \langle x_1^2\rangle\left(2\langle s_1\rangle\frac{d\gamma}{d\sigma^2} - \frac{1}{\sigma^4}\right) - \langle x_2^2\rangle\left(2\langle s_2\rangle\frac{d\gamma}{d\sigma^2} - \frac{1}{\sigma^4}\right) \quad (70)$$

$$\frac{d\gamma}{d\sigma^2} = \frac{1}{(2-\sigma^2)^2} + \frac{1}{\sigma^4} \quad (71)$$

Acknowledgments

We thank David Mackay for inspiration. This work has been supported by the Gatsby Charitable Foundation.

References

[1] Mackay, D. (2003) Information Theory, Inference and Learning Algorithms *Cambridge University Press*.

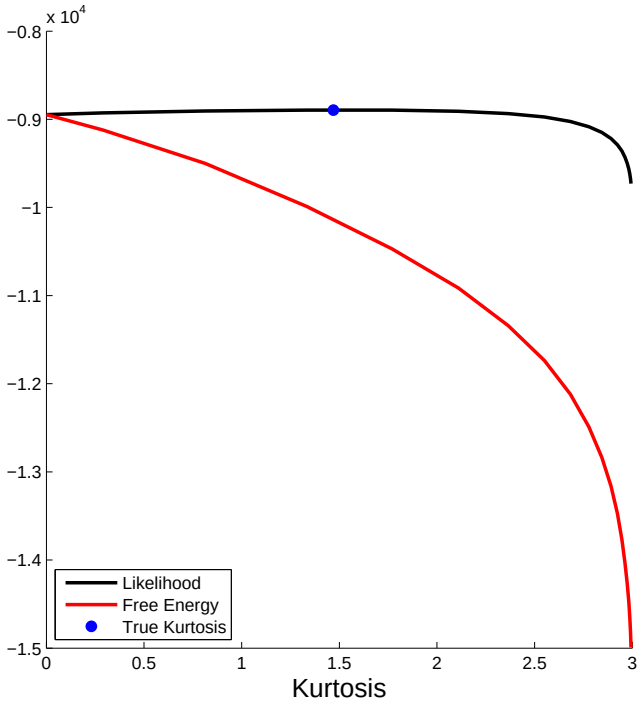

Figure 9: The likelihood is peaked at the true sparseness, but the free energy is biased with a peak at a sparseness of zero.