Local Minima, Symmetry-breaking, and Model Pruning in Variational Free Energy Minimization

David J.C. MacKay
University of Cambridge
Cavendish Laboratory
Madingley Road
Cambridge CB3 0HE
mackay@mrao.cam.ac.uk

June 27, 2001

Abstract

Approximate inference by variational free energy minimization (also known as variational Bayes, or ensemble learning) has maximum likelihood and maximum *a posteriori* methods as special cases, so we might hope that it can only work better than these standard methods. However, cases have been found in which degrees of freedom are 'pruned', perhaps inappropriately. This paper investigates this phenomenon in a toy example.

Approximate inference by variational free energy minimization (also known as variational Bayes, or ensemble learning, or learning with noisy weights – see (MacKay 1995) for a review) has maximum likelihood and maximum *a posteriori* methods as special cases, so we might hope that it can only work better than these standard methods. However, cases have been found in which degrees of freedom are 'pruned', perhaps inappropriately. This paper investigates this phenomenon in a toy example.

Motivations for VFE: want to incorporate uncertainty about parameters into the model-fitting process. Also worried about the electric monastry – location in parameter space where the likelihood diverges.

Uncertainty is greatest (and singularities in the likelihood more prominent) when there is little data, so VFE is of most interest for small N.

Problem observed by Zoubin Ghahramani (studying ensemble learning for HMMs (MacKay 1997)): extra degrees of freedom are not used. The model self-prunes. Annoying because we don't want the pruned model, we want the model we believe in – with lots of parameters, and big error bars on them! Parameter pruning is bad news because we would like predictions to take into accoutn uncertainty.

Comment on spontaneous pruning: is sometimes viewed as a convenient automatic Occam's razor, but does it behave correctly? Occam effect should be very weak for small N.

1 Example

Consider fitting a mixture of K Gaussians to a one-dimensional data set $D = \{x^{(n)}\}_{n=1}^{N}$. The standard parameters of the model are the means $\{\mu_k\}_{k=1}^{K}$, the standard deviations $\{\sigma_k\}_{k=1}^{K}$, and the mixing coefficients $\{\pi_k\}_{k=1}^{K}$; the latent variables are the class labels $\{c_n\}_{n=1}^{N}$.

For such a model, if we assume a separable approximating distribution $Q(\{\mu\})Q(\{\sigma\})Q(\pi)Q(\{c\})$ an ensemble learning algorithm equivalent to the E–M algorithm (soft K-means clustering) is easily derived. For each mean, for example, the variational approximation $Q(\mu_k)$ is a normal distribution with mean m_k and variance s_k^2. The two steps of the iterative algorithm are

Update $Q(\{c\})$ – an assignment step in which the 'responsibilities' $q_k^{(n)}$ of clusters k for points n are computed;

Update $Q(\{\mu\})$**,** $Q(\{\sigma\})$**, and** $Q(\pi)$ – an update step in which, for example, $Q(\mu_k)$ is updated to match the posterior distribution of μ_k given the data weighted by the responsibilities $q_k^{(n)}$.

The differences of this algorithm from the maximum likelihood algorithm are few in number. Are the differences healthy?

For simplicity, let's assume

1. all the standard deviations are fixed to $\sigma = 1$,
2. all the mixing coefficients are fixed to $1/K$, and
3. the number of components K is 2.

The prior on the means μ_k is zero mean and has standard deviation σ_0. We'll further assume that the data points have empirical mean zero and are sufficiently closely clumped that the update algorithm has no incentive to move the pseudo-posterior's means m_1 and m_2 from $m_1 = 0$, $m_2 = 0$.

1.1 One point

Simplest example of all: one data point only at $x = 0$. Bear in mind that symmetry breaking and pruning are the same thing.

What is the evidence? In the case of a single data point, the predictive distributions are identical, so $P(D|\mathcal{H}_2) = P(D|\mathcal{H}_1)$, and we do not expect any evidence in favour of either model. Indeed the evidence difference must be nil. But what do we find? Figure 1 shows the free energy as a function of the parameter q (top) and the intersections of the function q

$$q = \frac{1}{1 + \exp\left(\frac{1}{2}(s_1^2 - s_2^2)\right)}, \quad \text{where } \frac{1}{s_1^2} = \frac{1}{\sigma_0^2} + \frac{Nq}{\sigma^2} \text{ and } \frac{1}{s_2^2} = \frac{1}{\sigma_0^2} + \frac{N(1-q)}{\sigma^2} \tag{1}$$

which characterize extrema.

Why the symmetry-breaking? I think in this example an interpretation of it is that it represents a good guess that the *fluctuations* might lead all the points to actually come from one class only. For small N this is quite probable! For larger N it is improbable and we are obliged to accept the sensible conclusion that both means are zero.

2 More general picture

Let N data points be distributed in the ratio 1:3 between two points $x_a = -1$ and $x_b = 3$. The responsibilities of the two clusters for these all data points at x_a are $r_a, (1 - r_a)$; similarly, the responsibilities of the two clusters for these all data points at x_a are $r_b, (1 - r_b)$; we can plot the free energy as a function of r_a and r_b, assuming all the other disttributions Q are optimized. See figure 7.

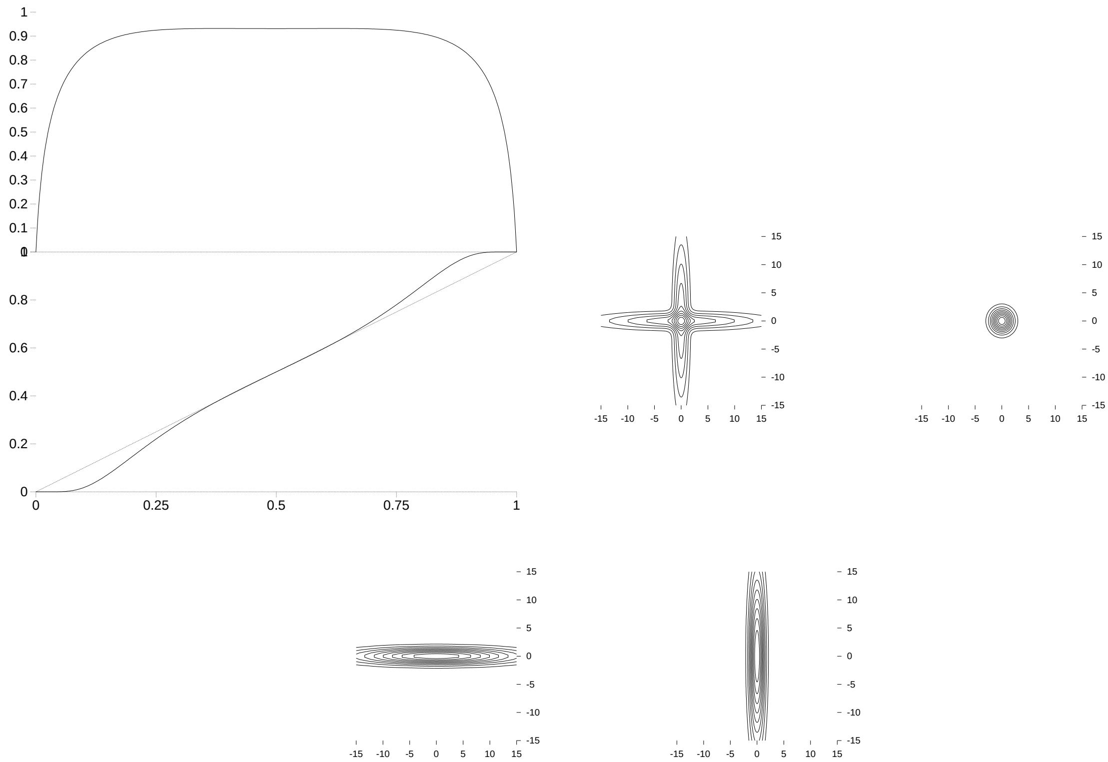

Figure 1: 1 data point only; Prior sd = 10; true posterior and 3 approximating distributions.

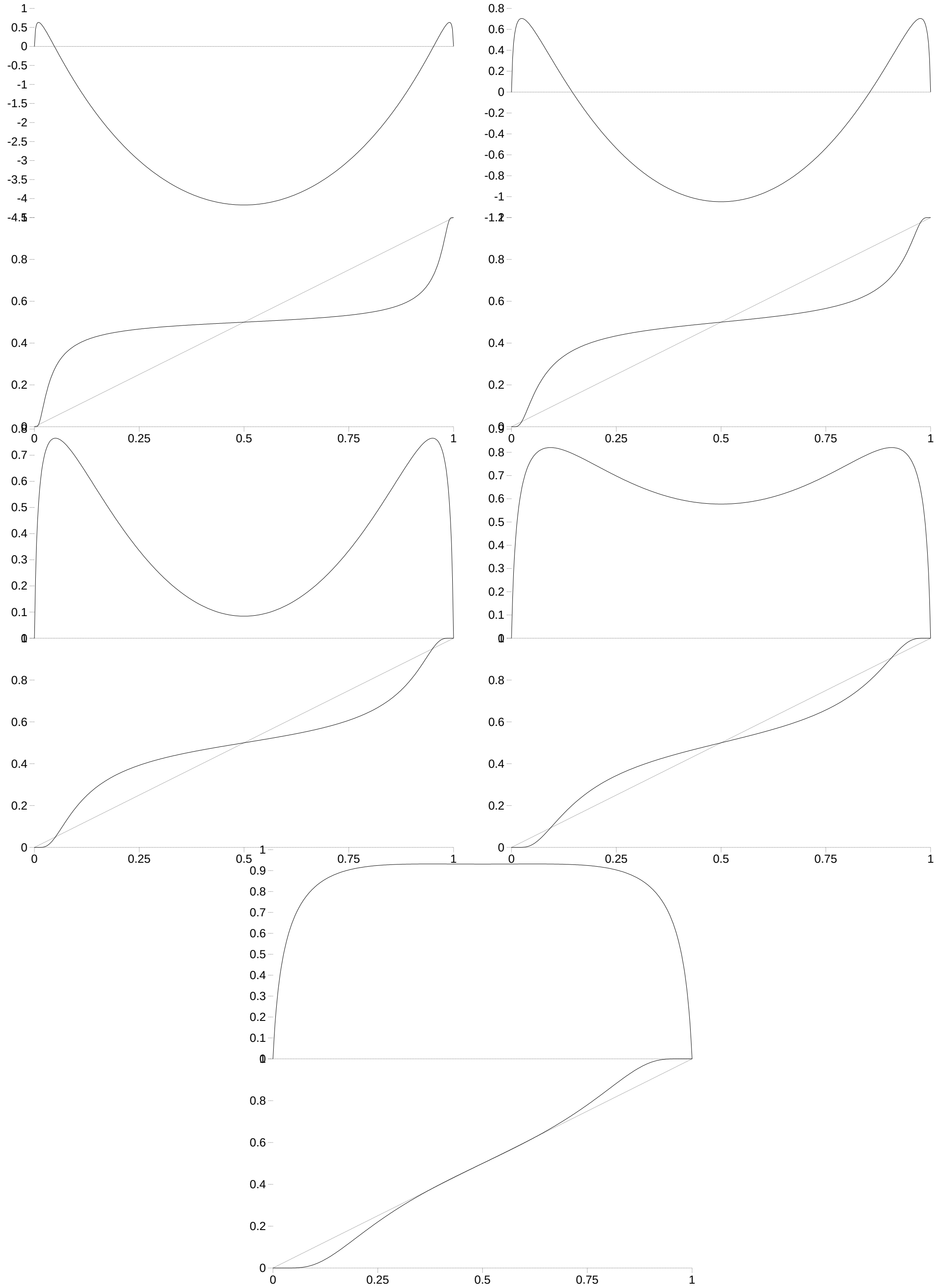

Figure 2: 10,5,3,2,1 Prior sd = 10

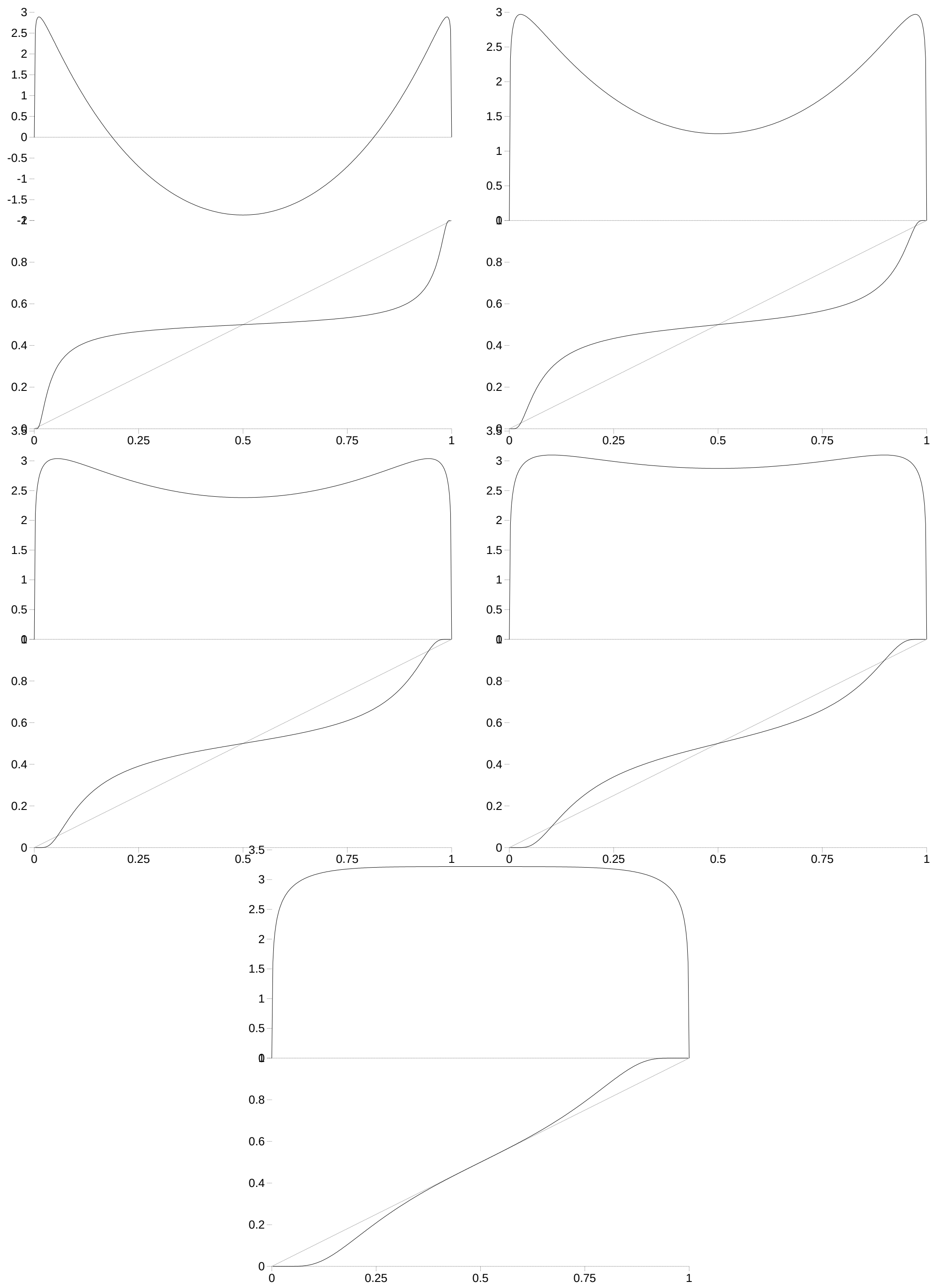

Figure 3: 10,5,3,2,1 Prior sd = 100

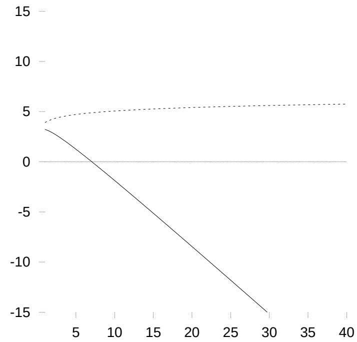

Figure 4: This shows that the scaling of the evidence difference is not one bit emulated by the behaviour of the variational free energy difference between the symmetric and symmetry-broken states.
Upper line shows the correct asymptotic scaling of $\log \frac{P(D|\mathcal{H}_1)}{P(D|\mathcal{H}_2)}$ with N; the advantage of $\mathcal{H}_1$ over $\mathcal{H}_2$ *grows* as $\log N$ (the correct value at $N = 1$ is zero). The evidence favours the *simpler* model more as N increases. The lower line shows the variational free energy difference. It starts above zero and *decreases* with N, becoming negative (*i.e.*, favouring the more complex, unpruned, model) when N is *bigger*.

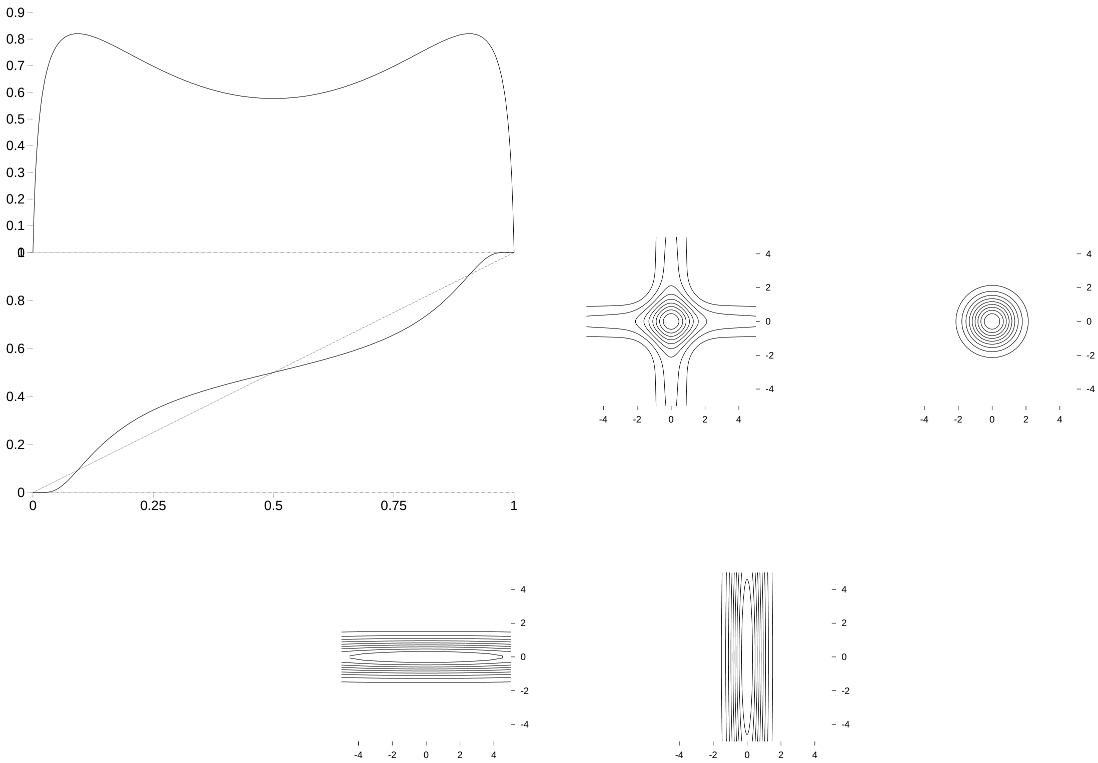

Figure 5: two data points - true posterior and 3 approximating distributions.

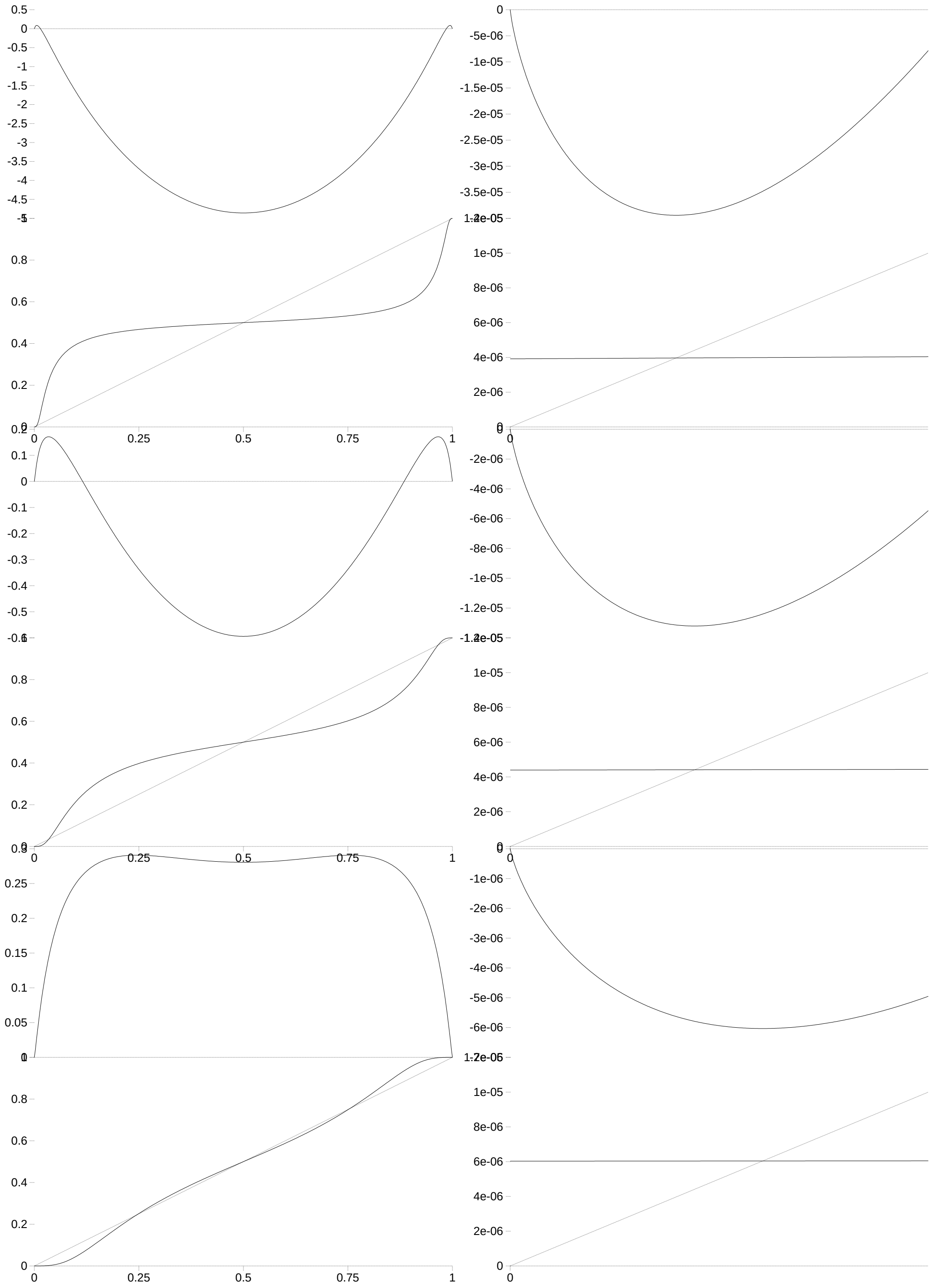

Figure 6: Strictly speaking, the minima at $q = \pm 1$ are not exactly at 0 and 1, they are at very slightly non-extreme locations. The broader the prior sd, the closer to the extremes. These pictures show (in the right hand blowups) the xrange from 0 to 1e-5. 10,3,1 Prior sd = 5

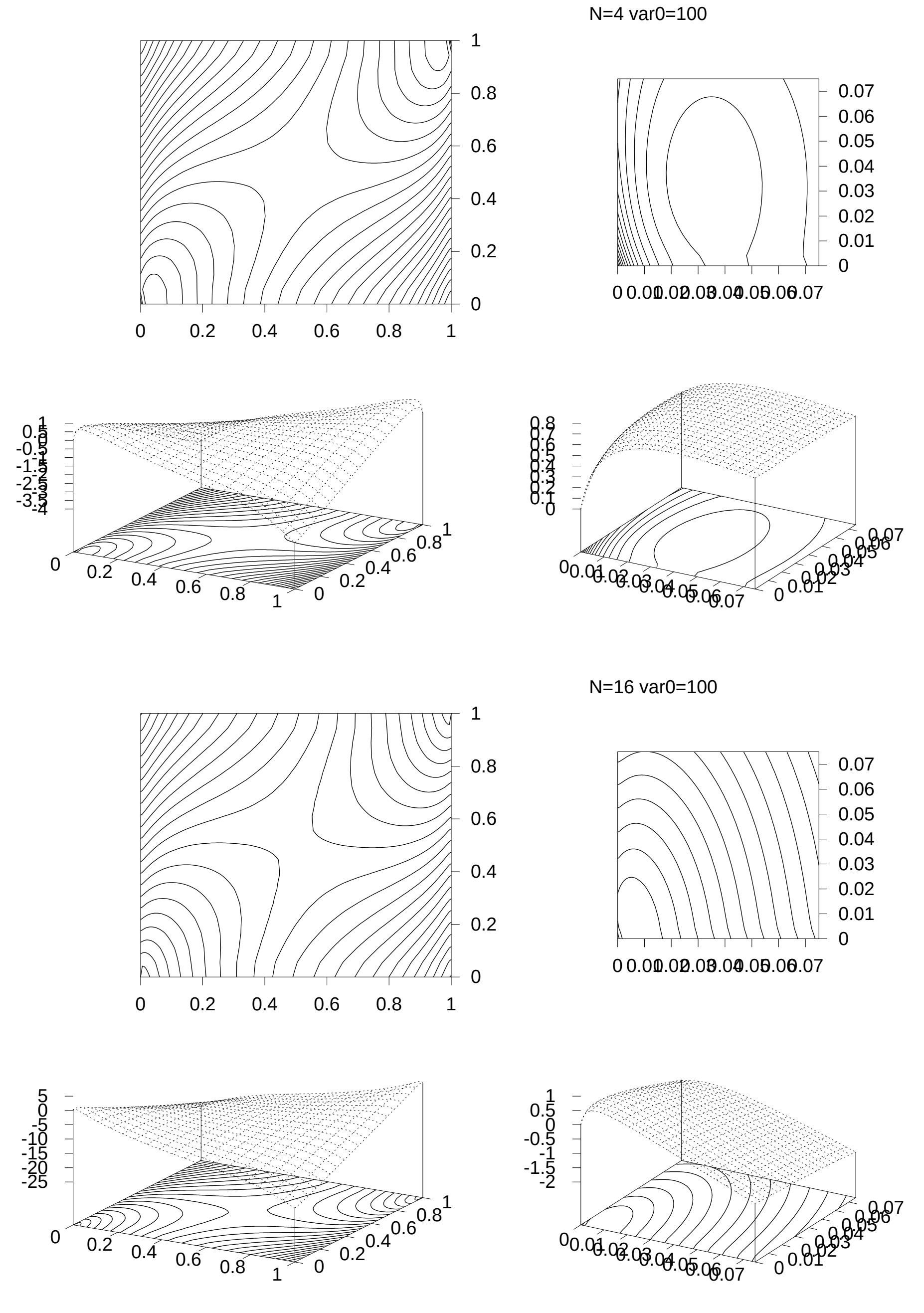

Figure 7: Multiple data points at two different locations. $N = 4$ and $N = 16$, and $\sigma_0 = 10$.

Acknowledgements

Thanks to Zoubin Gharamani and John Winn for helpful discussions.

References

MacKay, D. J. C. (1995) Developments in probabilistic modelling with neural networks – ensemble learning. In *Neural Networks: Artificial Intelligence and Industrial Applications. Proceedings of the 3rd Annual Symposium on Neural Networks, Nijmegen, Netherlands, 14-15 September 1995*, pp. 191–198, Berlin. Springer.

MacKay, D. J. C., (1997) Ensemble learning for hidden Markov models. `http://wol.ra.phy.cam.ac.uk/mackay/abstracts/ensemblePaper.html`.