Approximate Inference

Will Penny

31st March 2011

Information

Shannon (1948) asked how much information is received when we observe a specific value of the variable x ?

If an unlikely event occurs then one would expect the information to be greater. So information must be inversely proportional to $p(x)$, and monotonic.

Shannon also wanted a definition of information such that if x and y are independent then the total information would sum

$$h(x_i, y_j) = h(x_i) + h(y_j)$$

Given that we know that in this case

$$p(x_i, y_j) = p(x_i)p(y_j)$$

then we must have

$$h(x_i) = \log \frac{1}{p(x_i)}$$

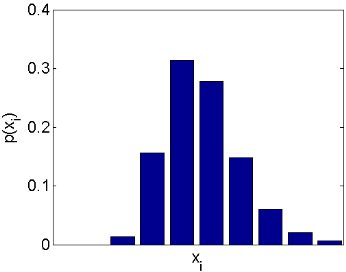
This is the self-information or surprise.

Entropy

The entropy of a random variable is the average surprise. For discrete variables

$$H(x) = \sum_i p(x_i) \log \frac{1}{p(x_i)}$$

The uniform distribution has maximum entropy.

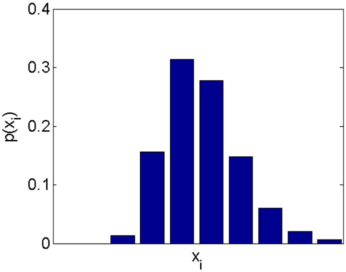


A single peak has minimum entropy. We define

$$0 \log 0 = 0$$

If we take logs to the base 2, entropy is measured in bits.

Source Coding Theorem



Assigning code-words of length $h(x_i)$ to each symbol x_i results in the maximum rate of information transfer in a noiseless channel. This is the Source Coding Theorem (Shannon, 1948).

$$h(x_i) = \log \frac{1}{p(x_i)}$$

If channel is noisy, see Noisy Channel Coding Theorem (Mackay, 2003)

Prefix Codes

No code-word is a prefix of another. Use number of bits $b(x_i) = ceil(h(x_i))$. We have

$$
\begin{aligned}
h(x_i) &= \log_2 \frac{1}{p(x_i)} \\
b(x_i) &= \log_2 \frac{1}{q(x_i)}
\end{aligned}
$$

Hence, each code-word has equivalent

$$q(x_i) = 2^{-b(x_i)}$$

i	$p(x_i)$	$h(x_i)$	$b(x_i)$	$q(x_i)$	CodeWord
1	0.016	5.97	6	0.016	101110
2	0.189	2.43	3	0.125	100
3	0.371	1.43	2	0.250	00
4	0.265	1.92	2	0.250	01
5	0.115	3.12	4	0.063	1010
6	0.035	4.83	5	0.031	10110
7	0.010	6.67	7	0.008	1011110
8	0.003	8.53	9	0.002	101111100

Relative Entropy

Average length of code word

$$B(x) = \sum_i p(x_i) b(x_i)$$
$$= \sum_i p(x_i) \log \frac{1}{q(x_i)} = 2.65 bits$$

Entropy

$$H(x) = \sum_i p(x_i) h(x_i)$$
$$= \sum_i p(x_i) \log \frac{1}{p(x_i)} = 2.20 bits$$

Difference is relative entropy

$$KL(p||q) = B(x) - H(x)$$
$$= \sum_i p(x_i) \log \frac{p(x_i)}{q(x_i)}$$
$$= 0.45 bits$$

Continuous Variables

For continuous variables the (differential) entropy is

$$H(x) = \int p(x) \log \frac{1}{p(x)} dx$$

Out of all distributions with mean m and standard deviation σ the Gaussian distribution has the maximum entropy. This is

$$H(x) = \frac{1}{2}(1 + \log 2\pi) + \frac{1}{2} \log \sigma^2$$

Relative Entropy

We can write the Kullback-Liebler (KL) divergence

$$KL[q||p] = \int q(x) \log \frac{q(x)}{p(x)} dx$$

as a difference in entropies

$$KL(q||p) = \int q(x) \log \frac{1}{p(x)} dx - \int q(x) \log \frac{1}{q(x)} dx$$

This is the average surprise assuming information is encoded under $p(x)$ minus the average surprise under $q(x)$. Its the extra number of bits/nats required to transmit messages.

Univariate Gaussians

For Gaussians

$$\begin{aligned} p(x) &= N(x; \mu_p, \sigma_p^2) \\ q(x) &= N(x; \mu_q, \sigma_q^2) \end{aligned}$$

we have

$$KL(q||p) = \frac{(\mu_q - \mu_p)^2}{2\sigma_p^2} + \frac{1}{2} \log \left(\frac{\sigma_p^2}{\sigma_q^2} \right) + \frac{\sigma_q^2}{2\sigma_p^2} - \frac{1}{2}$$

Multivariate Gaussians

For Gaussians

$$
\begin{aligned}
p(x) &= \mathrm{N}(x; \mu_p, C_p) \\
q(x) &= \mathrm{N}(x; \mu_q, C_q)
\end{aligned}
$$

we have

$$KL(q||p) = \frac{1}{2} e^T C_p^{-1} e + \frac{1}{2} \log \frac{|C_p|}{|C_q|} + \frac{1}{2} \mathrm{Tr}\left(C_p^{-1} C_q \right) - \frac{d}{2}$$

where $d = dim(x)$ and

$$e = \mu_q - \mu_p$$

Asymmetry

For densities $q(x)$ and $p(x)$ the Relative Entropy or Kullback-Liebler (KL) divergence from q to p is

$$KL[q||p] = \int q(x) \log \frac{q(x)}{p(x)} dx$$

The KL-divergence satisfies Gibbs' inequality

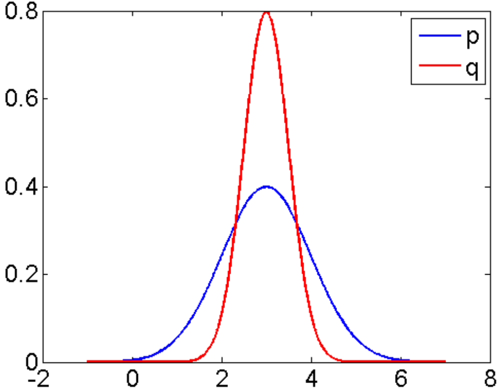
$$KL[q||p] \geq 0$$

with equality only if $q = p$. In general $KL[q||p] \neq KL[p||q]$, so KL is not a distance measure.

Different Variance - Asymmetry

$$KL[q||p] = \int q(x) \log \frac{q(x)}{p(x)} dx$$

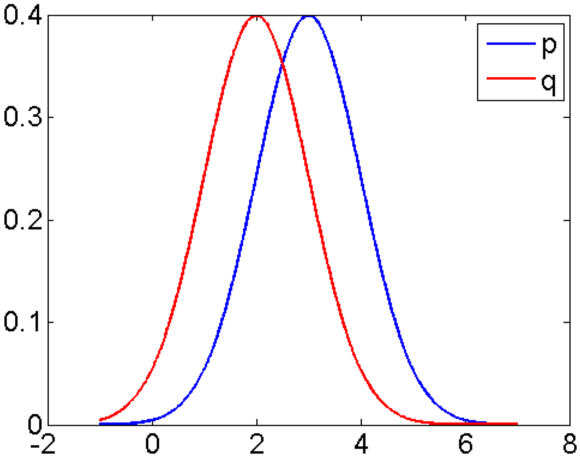
If $\sigma_q \neq \sigma_p$ then $KL(q||p) \neq KL(p||q)$



Here $KL(q||p) = 0.32$ but $KL(p||q) = 0.81$.

Same Variance - Symmetry

If $\sigma_q = \sigma_p$ then $KL(q||p) = KL(p||q)$ eg. distributions that just have a different mean

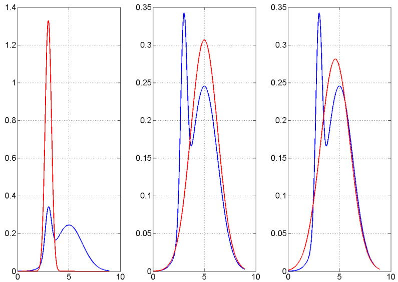


Here $KL(q||p) = KL(p||q) = 0.12$.

Approximating multimodal with unimodal

We approximate the density *p* (blue), which is a Gaussian mixture, with a Gaussian density *q* (red).

	Left Mode	Right Mode	Moment Matched
KL(q,p)	1.17	0.09	0.07
KL(p,q)	23.2	0.12	0.07

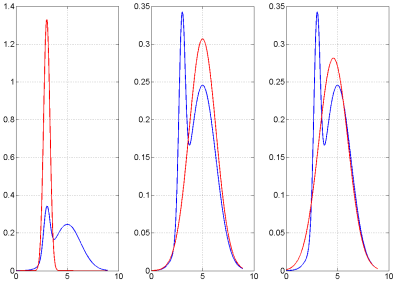


Minimising either KL produces the moment-matched solution.

Approximate Bayesian Inference

True posterior *p* (blue), approximate posterior *q* (red).
Gaussian approx at mode is a Laplace approximation.

	Left Mode	Right Mode	Moment Matched
KL(q,p)	1.17	0.09	0.07
KL(p,q)	23.2	0.12	0.07

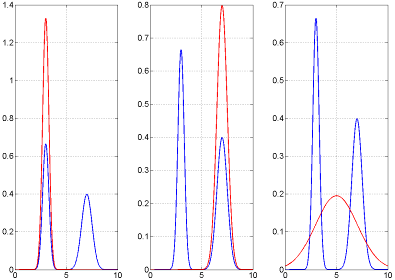


Minimising either KL produces the moment-matched solution.

Distant Modes

We approximate the density p (blue), which is a Gaussian mixture, with a Gaussian density q (red).

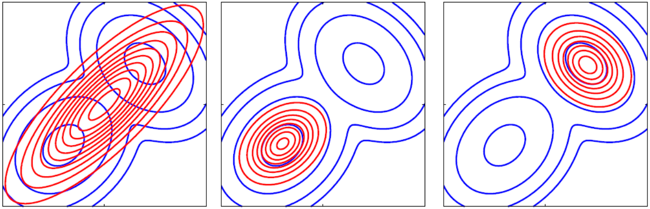
	Left Mode	Right Mode	Moment Matched
KL(q,p)	0.69	0.69	3.45
KL(p,q)	43.9	15.4	0.97



Minimising $KL(q||p)$ produces mode-seeking. Minimising $KL(p||q)$ produces moment-matching.

Multiple dimensions

In higher dimensional spaces, unless modes are very close, minimising $KL(p||q)$ produces moment-matching (a) and minimising $KL(q||p)$ produces mode-seeking (b and c).



(a) (b) (c)

Minimising $KL(q||p)$ therefore seems desirable, but how do we do it if we don't know p ?

Variational Free Energy

Given a probabilistic model of some data, the log of the evidence can be written as

$$
\begin{aligned}
\log p(Y) &= \int q(\theta) \log p(Y) d\theta \\
&= \int q(\theta) \log \frac{p(Y,\theta)}{p(\theta|Y)} d\theta \\
&= \int q(\theta) \log \left[\frac{p(Y,\theta) q(\theta)}{q(\theta) p(\theta|Y)} \right] d\theta \\
&= \int q(\theta) \log \left[\frac{p(Y,\theta)}{q(\theta)} \right] d\theta \\
&+ \int q(\theta) \log \left[\frac{q(\theta)}{p(\theta|Y)} \right]
\end{aligned}
$$

where $q(\theta)$ is the approximate posterior. Hence

$$\log p(Y) = F + KL(q(\theta)||p(\theta|Y))$$

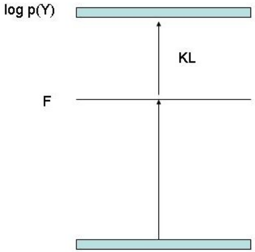
Variational Free Energy

We have

$$F = \int q(\theta) \log \frac{p(Y,\theta)}{q(\theta)} d\theta$$

which in statistical physics is known as the *negative* variational free energy.

Variational Free Energy



$$\log p(Y) = F + KL[q(\theta)||p(\theta|Y)]$$

Because KL is always positive, due to the Gibbs inequality, F provides a lower bound on the model evidence. Moreover, because KL is zero when two densities are the same, F will become equal to the model evidence when $q(\theta)$ is equal to the true posterior. For this reason $q(\theta)$ can be viewed as an *approximate posterior*.

Factorised Approximations

To obtain a practical learning algorithm we must also ensure that the integrals in F are tractable. One generic procedure for attaining this goal is to assume that the approximating density factorizes over groups of parameters. In physics, this is known as the mean field approximation. Thus, we consider:

$$q(\theta) = \prod_i q(\theta_i)$$

where θ_i is the ith group of parameters. We can also write this as

$$q(\theta) = q(\theta_i) q(\theta_{\backslash i})$$

where $\theta_{\backslash i}$ denotes all parameters *not* in the ith group.

Variational Energy

The distributions $q(\theta_i)$ which maximise F can then be derived as follows.

$$
\begin{aligned}
F &= \int q(\theta) \log \left[\frac{p(Y,\theta)}{q(\theta)} \right] d\theta \\
&= \int \int q(\theta_i) q(\theta_{\backslash i}) \log \left[\frac{p(Y,\theta)}{q(\theta_i) q(\theta_{\backslash i})} \right] d\theta_{\backslash i} d\theta_i \\
&= \int q(\theta_i) \left[\int q(\theta_{\backslash i}) \log p(Y,\theta) d\theta_{\backslash i} \right] d\theta_i - \int q(\theta_i) \log q(\theta_i) d\theta_i + C \\
&= \int q(\theta_i) I(\theta_i) d\theta_i - \int q(\theta_i) \log q(\theta_i) d\theta_i + C
\end{aligned}
$$

where the constant C contains terms not dependent on $q(\theta_i)$ and

$$
I(\theta_i) = \int q(\theta_{\backslash i}) \log p(Y,\theta) d\theta_{\backslash i}
$$

This quantity is known as the variational energy for the ith partition.

Approximate Posteriors

Writing $I(\theta_i) = \log \exp I(\theta_i)$ gives

$$F = \int q(\theta_i) \log \left[\frac{\exp(I(\theta_i))}{q(\theta_i)} \right] d\theta_i + C$$
$$= KL\left[q(\theta_i) || \exp(I(\theta_i))\right] + C$$

This is minimised when

$$q(\theta_i) = \frac{\exp[I(\theta_i)]}{Z}$$

where Z is the normalisation factor needed to make $q(\theta_i)$ a valid probability distribution.

Free-form versus Fixed-form approximations (Beal, 2003).

Mean Field Models

For mean field approaches

$$q(\theta_i) = f(m_i, S_i)$$

where moments of densities are functions of each other

$$\begin{aligned} m_i &= g_1(m_j, S_j) \\ S_i &= g_2(m_j, S_j) \end{aligned}$$

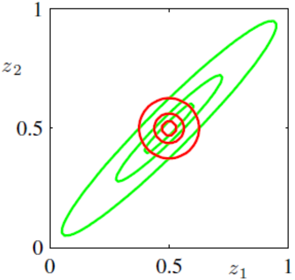
Neural populations interact with each other via sufficient statistics (Deco et al. 2008). For example, cells in one population are only affected by average firing rate in other populations (the mean field, m_j). Or additionally, by synchronisation level of other populations (S_j).

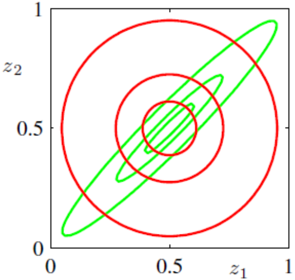
Factorised Approximations

For

$$q(z) = q(z_1)q(z_2)$$

minimising $KL(q,p)$ where p is green and q is red produces left plot, where minimising $KL(p,q)$ produces right plot.





Hence minimising variational free energy tends to produce approximations on left rather than right. That is, uncertainty is underestimated. See Minka (2005) for other divergences.

Nonlinear Regression

We consider the framework implemented in the SPM function *spm-nlsi-GN.m*. It implements Bayesian estimation of nonlinear models of the form

$$y = g(w) + e$$

where $g(w)$ is some nonlinear function of parameters w, and e is zero mean additive Gaussian noise with covariance C_y. The likelihood of the data is therefore

$$p(y|w, \lambda) = \mathrm{N}(y; g(w), C_y)$$

The error *precision* matrix is assumed to decompose linearly

$$C_y^{-1} = \sum_i \exp(\lambda_i) Q_i$$

where Q_i are known precision basis functions and λ are hyperparameters eg $Q = I$, noise precision $s = \exp(\lambda)$.

Priors

We allow Gaussian priors over model parameters

$$p(w) = \mathrm{N}(w; \mu_w, C_w)$$

where the prior mean and covariance are assumed known.

The hyperparameters are constrained by the prior

$$p(\lambda) = \mathrm{N}(\lambda; \mu_\lambda, C_\lambda)$$

This is not Empirical Bayes.

VL Posteriors

The Variational Laplace (VL) algorithm, implemented in *spm-nlsi-GN.m*, assumes an approximate posterior density of the following factorised form

$$
\begin{aligned}
q(w,\lambda|y) &= q(w|y)q(\lambda|y) \\
q(w|y) &= \mathrm{N}(w; m_w, S_w) \\
q(\lambda|y) &= \mathrm{N}(\lambda; m_\lambda, S_\lambda)
\end{aligned}
$$

This is a fixed-form variational method.

Energies

The above distributions allow one to write down an expression for the joint log likelihood of the data, parameters and hyperparameters

$$L(w,\lambda) = \log[p(y|w,\lambda)p(w)p(\lambda)]$$

The negative of this is known as the Gibbs Energy. Here it splits into three terms

$$\begin{aligned} L(w,\lambda) &= \log p(y|w,\lambda) \\ &+ \log p(w) \\ &+ \log p(\lambda) \end{aligned}$$

Joint Log Likelihood

The joint log likelihood is composed of sum squared precision weighted prediction errors and entropy terms

$$
\begin{aligned}
L &= -\frac{1}{2} e_y^T C_y^{-1} e_y - \frac{1}{2} \log |C_y| - \frac{N_y}{2} \log 2\pi \\
&- \frac{1}{2} e_w^T C_w^{-1} e_w - \frac{1}{2} \log |C_w| - \frac{N_w}{2} \log 2\pi \\
&- \frac{1}{2} e_\lambda^T C_\lambda^{-1} e_\lambda - \frac{1}{2} \log |C_\lambda| - \frac{N_\lambda}{2} \log 2\pi
\end{aligned}
$$

where prediction errors are the difference between what is expected and what is observed

$$
\begin{aligned}
e_y &= y - g(m_\theta) \\
e_w &= m_w - \mu_w \\
e_\lambda &= m_\lambda - \mu_\lambda
\end{aligned}
$$

Variational Energies

The approximate posteriors are estimated by minimising the Kullback-Liebler (KL) divergence between the true posterior and these approximate posteriors. This is implemented by maximising the following (negative) variational energies

$$
\begin{aligned}
I(w) &= \int L(w,\lambda) q(\lambda) \\
I(\lambda) &= \int L(w,\lambda) q(w)
\end{aligned}
$$

Gradient Ascent

This maximisation is effected by first computing the gradient and curvature of the variational energies at the current parameter estimate, $m_w(old)$. For example, for the parameters we have

$$
\begin{aligned}
j_w(i) &= \frac{dI(w)}{dw(i)} \\
H_w(i,j) &= \frac{d^2 I(w)}{dw(i)dw(j)}
\end{aligned}
$$

where i and j index the ith and jth parameters, j_w is the gradient vector and H_w is the curvature matrix. The estimate for the posterior mean is then given by

$$
m_w(new) = m_w(old) + \Delta m_w
$$

Adaptive Step Size

The change is given by

$$\Delta m_w = [\exp(vH_w) - I]\, H_w^{-1} j_w$$

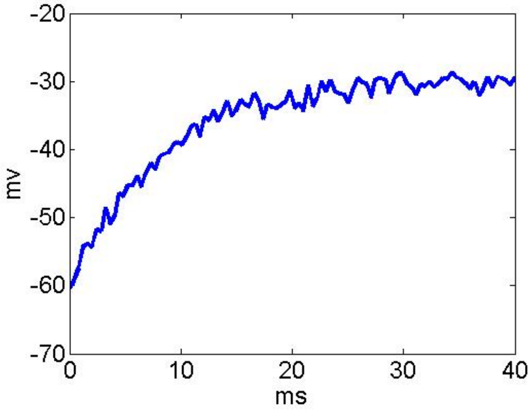
This last expression implements a 'temporal regularisation' with parameter v (Friston et al. 2007). In the limit $v \rightarrow \infty$ the update reduces to

$$\Delta m_w = -H_w^{-1} j_w$$

which is equivalent to a Newton update. This implements a step in the direction of the gradient with a step size given by the inverse curvature. Big steps are taken in regions where the gradient changes slowly (low curvature).

Approach to Limit

$$y(t) = -60 + V_a[1 - \exp(-t/\tau)] + e(t)$$



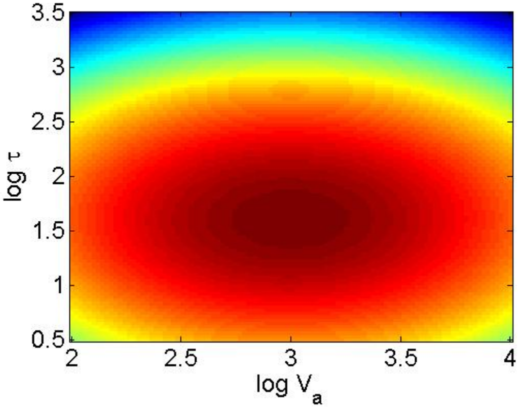
$$V_a = 30, \tau = 8$$

Noise precision

$$s = \exp(\lambda) = 1$$

Prior Landscape

A plot of $\log p(w)$ where $w = [\log \tau, \log V_a]$

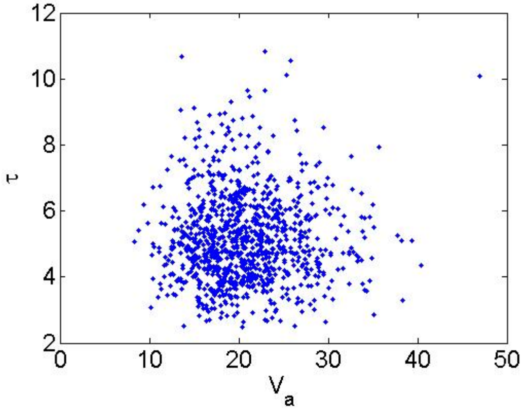


$$\mu_w = [3, 1.6]^T, C_w = diag([1/16, 1/16]);$$

Samples from Prior

The true model parameters are unlikely apriori

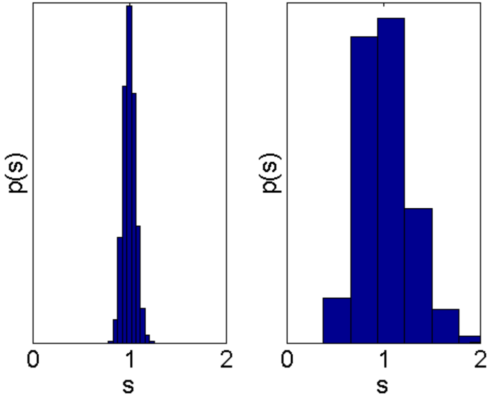
$$V_a = 30, \tau = 8$$



Prior Noise Precision

$Q = I$. Noise precision $s = \exp(\lambda)$ with

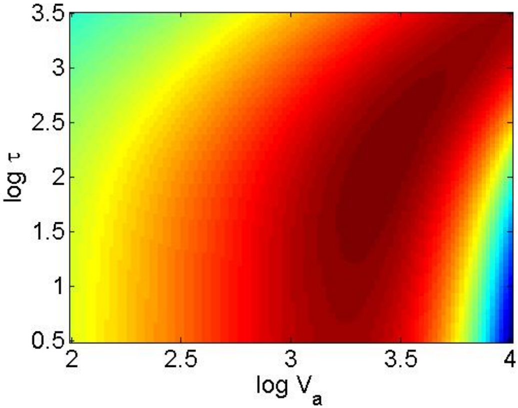
$$p(\lambda) = N(\lambda; \mu_\lambda, C_\lambda)$$



with $\mu_\lambda = 0$. We used $C_\lambda = 1/16$ (left) and $C_\lambda = 1/4$ (right). True noise precision, $s = 1$.

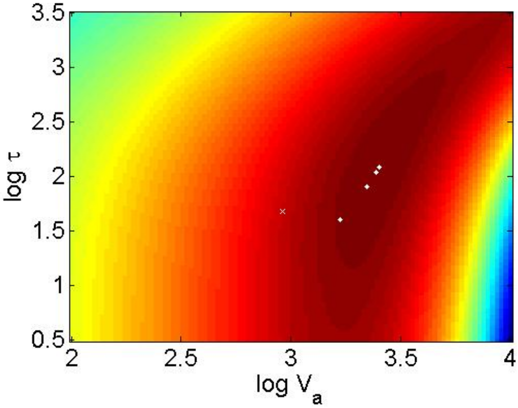
Posterior Landscape

A plot of $\log[p(y|w)p(w)]$



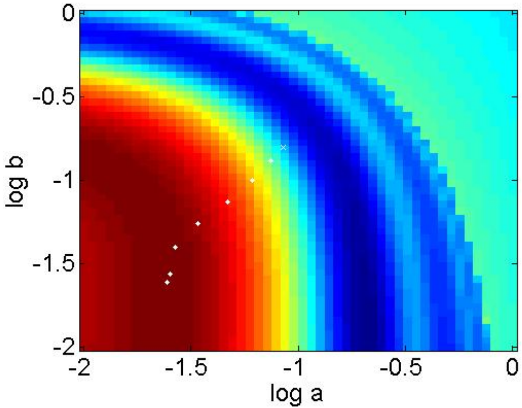
VL optimisation

Path of 6 VL iterations (x marks start)



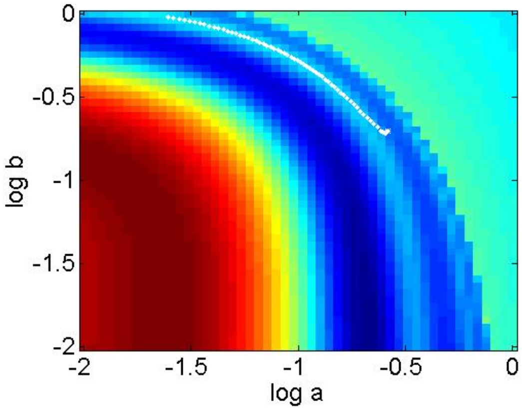
VL optimisation I

Global maxima



VL optimisation II

Local maxima



References

M. Beal (2003) PhD Thesis. Gatsby Computational Neuroscience Unit, UCL.

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

G. Deco et al. (2008) The Dynamic Brain: From Spiking Neurons to Neural Masses and Cortical Fields. PLoS CB 4(8), e1000092.

D. Mackay (2003) Information Theory, Inference and Learning Algorithms, Cambridge.

K. Friston et al. (2007) Variational Free Energy and the Laplace Approximation. Neuroimage 34(1), 220-234.

T. Minka et al. (2005) Divergence Measures and Message Passing. Microsoft Research Cambridge.

W. Penny (2006) Variational Bayes. In SPM Book, Elsevier.