Bayesian Inference for Nonlinear Models

Will Penny

26th November 2010

Likelihood

We consider Bayesian estimation of nonlinear models of the form

$$y = g(\theta, m) + e$$

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

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

The error covariances are assumed to decompose into terms of the form

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

where Q_i are known precision basis functions and λ are hyperparameters.

Priors

We allow Gaussian priors over model parameters

$$p(\theta|m) = \mathrm{N}(\theta; \mu_\theta, C_\theta)$$

where the prior mean and covariance are assumed known.

The hyperparameters are constrained by the prior

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

VL Posteriors

The Variational Laplace (VL) algorithm assumes an approximate posterior density of the following factorised form

$$
\begin{aligned}
q(\theta, \lambda | y, m) &= q(\theta | y, m) q(\lambda | y, m) \qquad (1) \\
q(\theta | y, m) &= \mathrm{N}(\theta; m_\theta, S_\theta) \\
q(\lambda | y, m) &= \mathrm{N}(\lambda; m_\lambda, S_\lambda)
\end{aligned}
$$

Energies

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

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

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 variational energies

$$\begin{aligned} I(\theta) &= \int L(\theta, \lambda) q(\lambda) \\ I(\lambda) &= \int L(\theta, \lambda) q(\theta) \end{aligned} \tag{2}$$

Gradient Ascent

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

$$\begin{aligned} j_\theta(i) &= \frac{dI(\theta)}{d\theta(i)} \\ H_\theta(i,j) &= \frac{d^2 I(\theta)}{d\theta(i) d\theta(j)} \end{aligned} \tag{3}$$

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

$$m_\theta(new) = m_\theta(old) + \Delta m_\theta$$

Adaptive Step Size

The change is given by

$$\Delta m_{\theta} = \left[\exp(vH_{\theta}) - I\right] H_{\theta}^{-1} j_{\theta}$$

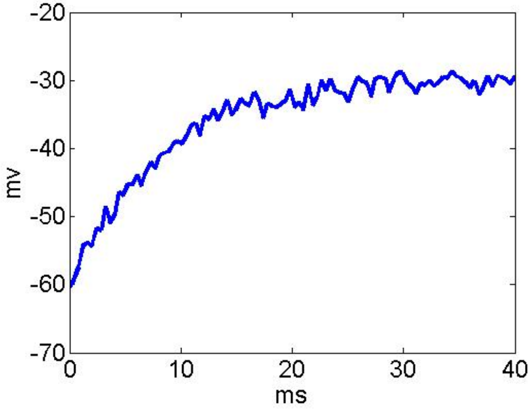
This last expression implements a 'temporal regularisation' with parameter v. In the limit $v \rightarrow \infty$ the update reduces to

$$\Delta m_{\theta} = -H_{\theta}^{-1} j_{\theta}$$

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).

Likelihood

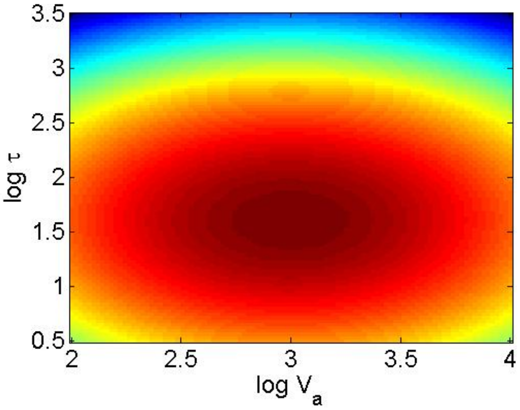
$$y(t) = -60 + V_a[1 - \exp(-t/\tau)] + e(t)$$



$$V_a = 30, \tau = 8, \exp(\lambda) = 1$$

Prior Landscape

A plot of $\log p(\theta)$



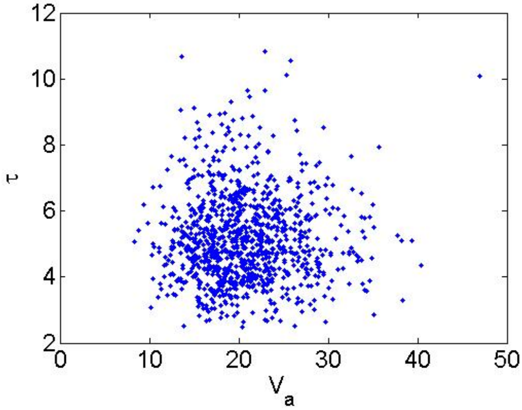
$$\mu_\theta = [3, 1.6]^T, C_\theta = diag([1/16, 1/16]);$$

$$\mu_\lambda = 0, C_\lambda = 1/16$$

Samples from Prior

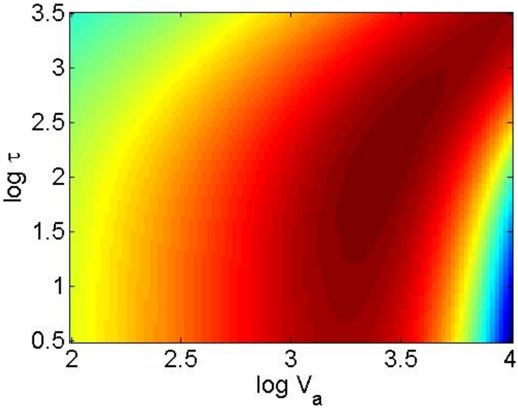
The true model parameters are unlikely apriori

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



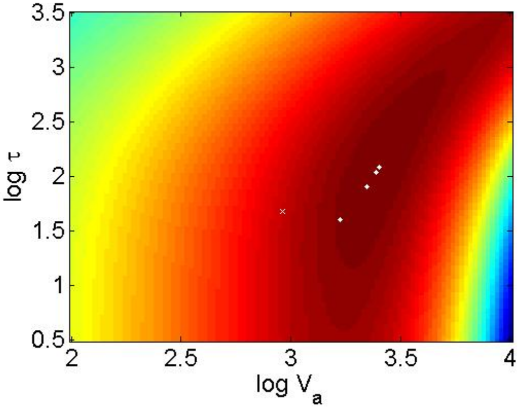
Posterior Landscape

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



VL optimisation

Path of 6 VL iterations (x marks start)



Model Evidence

The model evidence is not straightforward to compute, since this computation involves integrating out the dependence on model parameters

$$p(y|m) = \int p(y|\theta, m)p(\theta|m)d\theta.$$

Once computed two models can be compared via the Bayes factor

$$B_{12} = \frac{p(y|m_1)}{p(y|m_2)}$$

Free Energy

The free energy is composed of sum squared precision weighted prediction errors and Occam factors

$$
\begin{aligned}
F &= -\frac{1}{2} e_y^T C_y^{-1} e_y - \frac{1}{2} \log |C_y| - \frac{N_y}{2} \log 2\pi \\
&- \frac{1}{2} e_\theta^T C_\theta^{-1} e_\theta - \frac{1}{2} \log \frac{|C_\theta|}{|S_\theta|} \\
&- \frac{1}{2} e_\lambda^T C_\lambda^{-1} e_\lambda - \frac{1}{2} \log \frac{|C_\lambda|}{|S_\lambda|}
\end{aligned} \tag{4}
$$

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

$$
\begin{aligned}
e_y &= y - g(m_\theta) \\
e_\theta &= m_\theta - \mu_\theta \\
e_\lambda &= m_\lambda - \mu_\lambda
\end{aligned} \tag{5}
$$

Free Energy

This can be rearranged as

$$F(m) = Accuracy(m) - Complexity(m)$$

where

$$Accuracy(m) = -\frac{1}{2} e_y^T C_y^{-1} e_y - \frac{1}{2} \log |C_y| - \frac{N_y}{2} \log 2\pi$$

$$\begin{aligned} Complexity(m) &= \frac{1}{2} e_\theta^T C_\theta^{-1} e_\theta + \frac{1}{2} \log \frac{|C_\theta|}{|S_\theta|} \\ &+ \frac{1}{2} e_\lambda^T C_\lambda^{-1} e_\lambda + \frac{1}{2} \log \frac{|C_\lambda|}{|S_\lambda|} \end{aligned} \tag{6}$$

Model complexity will tend to increase with the number of parameters because distances tend to be larger in higher dimensional spaces.

AIC and BIC

A simple approximation to the log model evidence is given by the Bayesian Information Criterion [**?**]

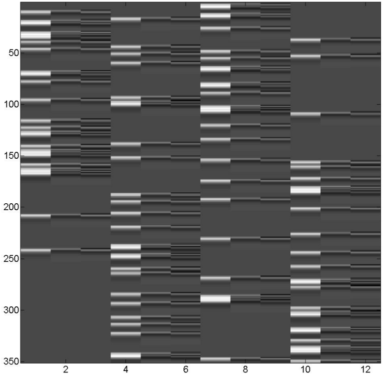
$$BIC = \log p(y|\hat{\theta}, \hat{\lambda}, m) - \frac{p}{2} \log N_y$$

where $\hat{\theta}$,$\hat{lambda}$, are the estimated parameters and hyperparameters, p is the number of parameters, and N_y is the number of data points. The BIC is a special case of the Free Energy approximation that drops all terms that do not scale with the number of data points
An alternative approximation is Akaike's Information Criterion (or 'An Information Criterion')

$$AIC = \log p(y|\hat{\theta}, \hat{\lambda}, m) - p$$

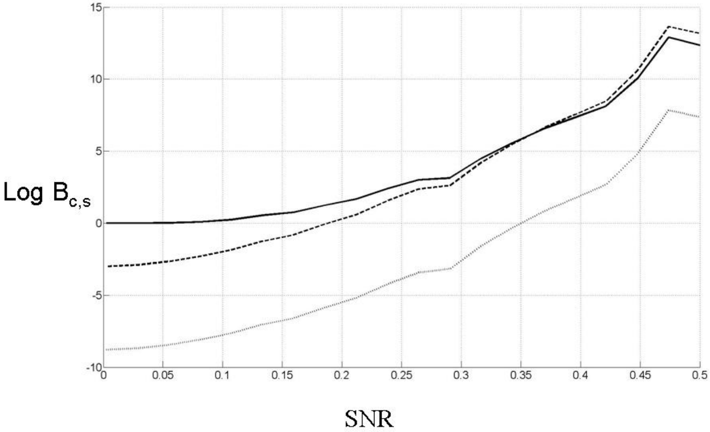
Synthetic fMRI example

Design matrix from Henson et al. Regression coefficients from responsive voxel in occipital cortex. Data was generated from a 12-regressor model with SNR=0.2. We then fitted 12-regressor and 9-regressor models. This was repeated 25 times.



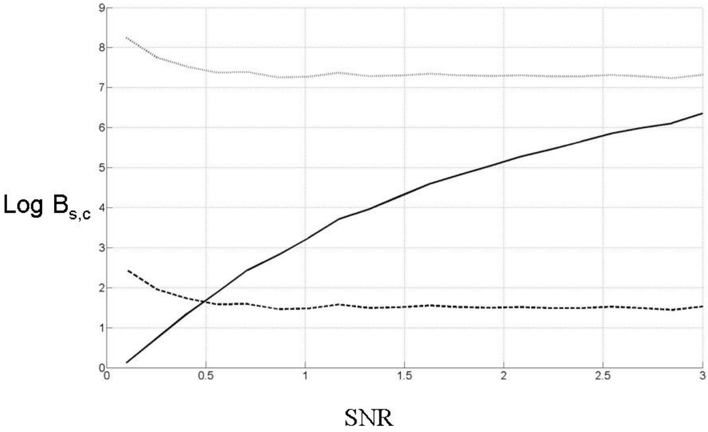
True Model: Complex GLM

Log Bayes factor of complex versus simple model, Log $B_{c,s}$, versus the signal to noise ratio, SNR, when true model is the
complex GLM for F (solid), AIC (dashed) and BIC (dotted).



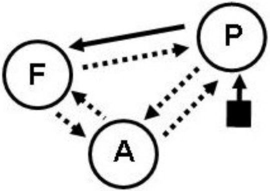
True Model: Simple GLM

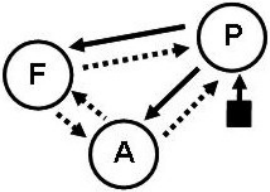
Log Bayes factor of simple versus complex model, Log $B_{s,c}$, versus the signal to noise ratio, SNR, when true model is the simple GLM for F (solid), AIC (dashed) and BIC (dotted).



DCM for fMRI

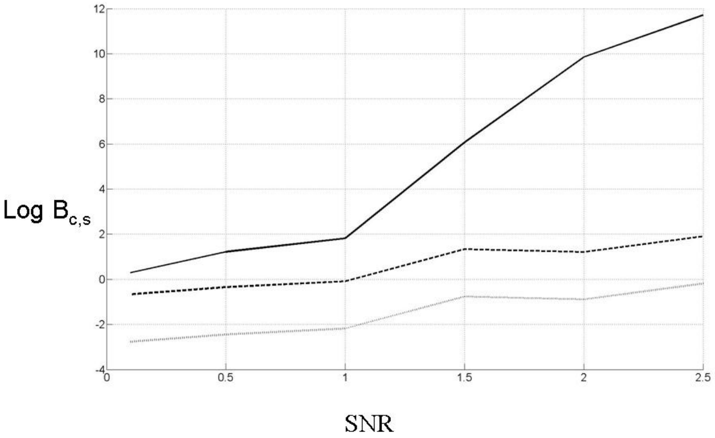
A simple (left) and complex (right) DCM. The complex DCM is identical to the simple DCM except for having an additional modulatory forward connection from region P to region A.





True Model: Complex DCM

Log Bayes factor of complex versus simple model, Log $B_{c,s}$, versus the signal to noise ratio, SNR, when true model is the
complex DCM for F (solid), AIC (dashed) and BIC (dotted).



True Model: Simple DCM

Log Bayes factor of simple versus complex model, Log $B_{s,c}$, versus the signal to noise ratio, SNR, when true model is the simple DCM for F (solid), AIC (dashed) and BIC (dotted).