Bayesian Inference for DCMs

Will Penny

23rd 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

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

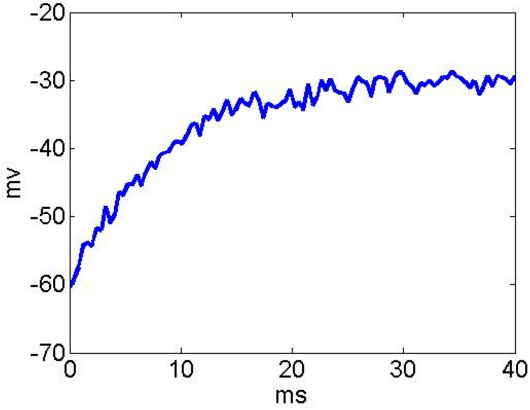
This last expression implements a 'temporal regularisation' with parameter v. In the limit $v \rightarrow \infty$ the update reduces to

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

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

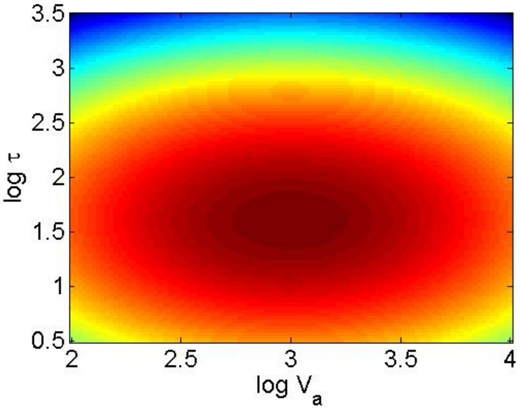
$$y(t) = -60 + V_a[1 - \exp(-t/\tau)] + e(t)$$



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

Prior Landscape

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



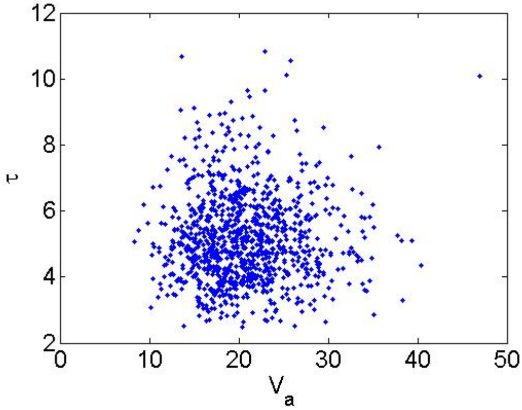
$$\mu_\theta = [3, 1.6]^T, C_\theta = diag([1/16, 1/16]);$$

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

Samples from Prior

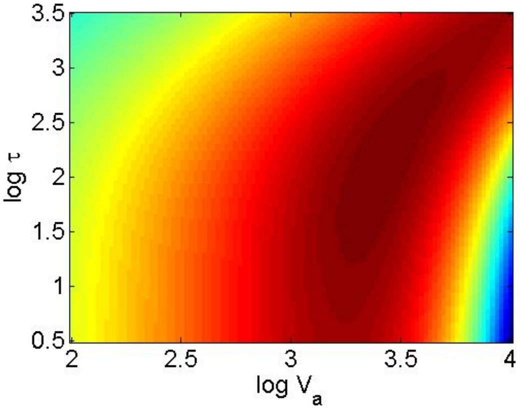
The true model parameters are unlikely apriori

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



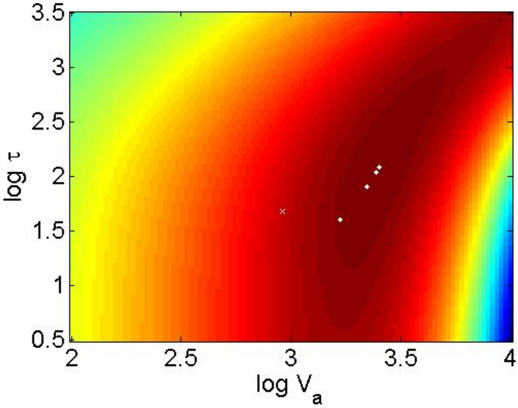
Posterior Landscape

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



VL optimisation

Path of 6 VL iterations (x marks start)



Metropolis-Hastings

MH creates as series of random points $(\theta(1), \theta(2), ...)$ whose distribution converges to the target distribution of interest. For us, this is the posterior density $p(\theta|y)$. Each sequence can be considered a random walk whose stationary distribution is $p(\theta|y)$.

MH makes use of a proposal density $q(\theta'; \theta)$ which is dependent on the current state vector θ. For symmetric q (such as a Gaussian) samples from the posterior density can be generated as follows.

MH update

First, start at some point $\theta(0)$ in parameter space. Then generate a proposal θ' using the density q. This proposal is then accepted according to the standard Metropolis-Hastings procedure.

That is, with probability $\min(1, r)$ where

$$r = \frac{p(y|\theta')p(\theta')}{p(y|\theta)p(\theta)}$$

If the step is accepted we set $\theta(n+1) = \theta'$. If it is rejected we set $\theta(n+1) = \theta(n)$.

Adaptive proposal density

We use a zero mean Gaussian proposal density with covariance C_s. This covariance is initialised to

$$C_s = \sigma C_\theta \tag{4}$$

where C_θ is the prior covariance and $\sigma = 1$.

We then use a three stage procedure comprising (i) scaling, (ii) tuning and (iii) sampling steps in which the scaling and tuning stages are used to optimize the proposal covariance C_s.

The first two stages are regarded as a burn-in phase and samples from this period are later discarded. At the end of this C_s is fixed and sampling proper begins.

Scaling

The proposal covariance is given by

$$C_s = \sigma C_\theta \tag{5}$$

In the scaling step σ is adjusted as follows.

If the acceptance rate, as measured over the last $n_s = 100$ proposals, is less than 20 per cent then σ is halved.

If the acceptance rate is greater than 40 per cent σ is doubled.

Otherwise, σ remains unchanged.

Tuning

The tuning step makes use of adaptive estimation of a covariance matrix C_{tune} based on a Robbins-Monro update.

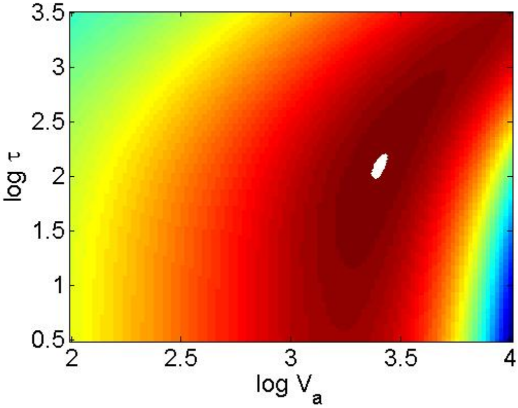
At the beginning of the tuning stage we set $C_{tune} = C_s$. We then update according to

$$\begin{aligned} \mu_t &= \mu_{t-1} + \frac{1}{n_t}(x_t - \mu_t) \\ \Delta C_{tune} &= \frac{1}{n_t}[(x_t - \mu_t)(x_t - \mu_t)^T - C_{tune}(t-1)] \end{aligned} \tag{6}$$

where n_t is the number of elapsed iterations in the tuning period. At the end of tuning set $C_s = C_{tune}$.

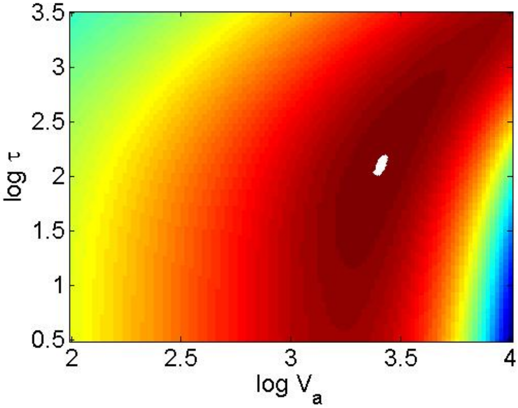
MH Samples

64,000 samples from MH posterior



VL Samples

64,000 samples from VL posterior

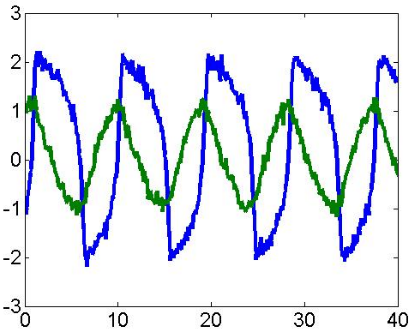


Likelihood

Nonlinear oscillator with $a = 0.2$, $b = 0.2$, $c = 3$.

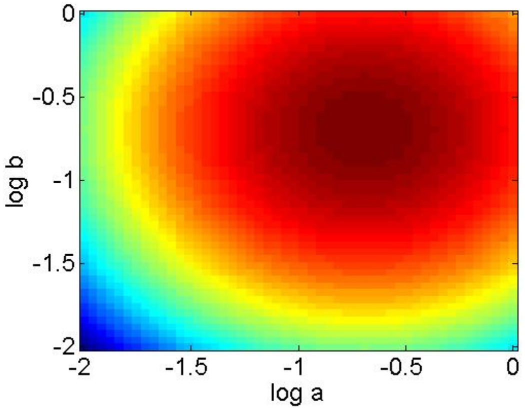
$$\begin{aligned} \dot{v} &= c[v - \frac{1}{3}v^3 + r] \\ \dot{r} &= -\frac{1}{c}[v - a + br] \end{aligned} \tag{7}$$

We have noise level $\exp(\lambda) = 10$.



Priors

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

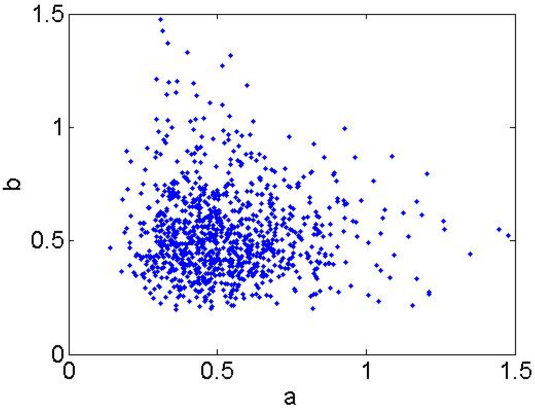


$$\mu_\theta = [-0.69, -0.69]^T, C_\theta = diag([1/8, 1/8]);$$

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

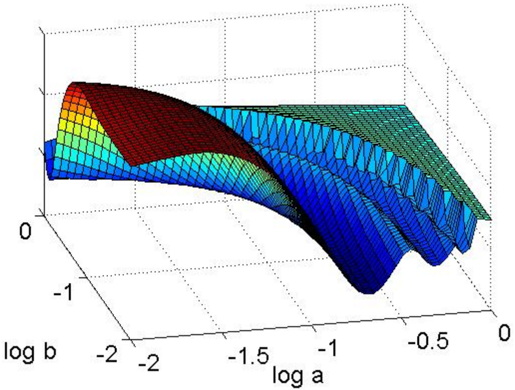
Priors

True value $a = 0.2$, $b = 0.2$ is apriori unlikely



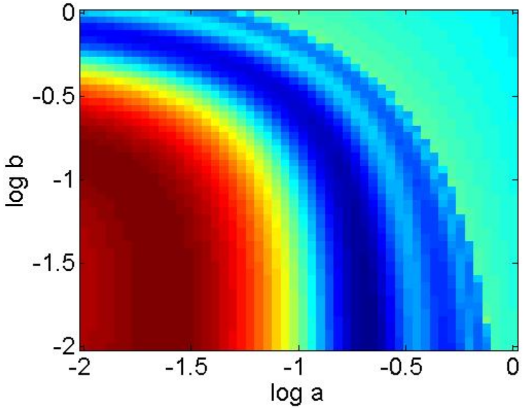
Posterior

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



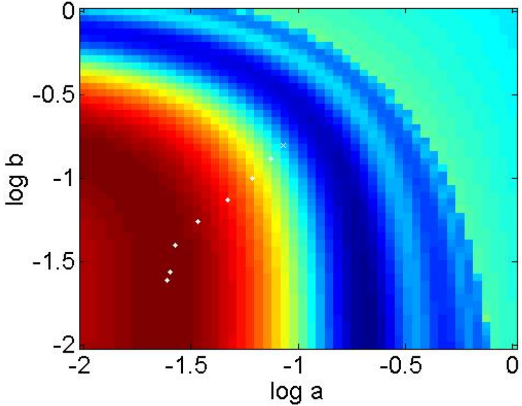
Posterior

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



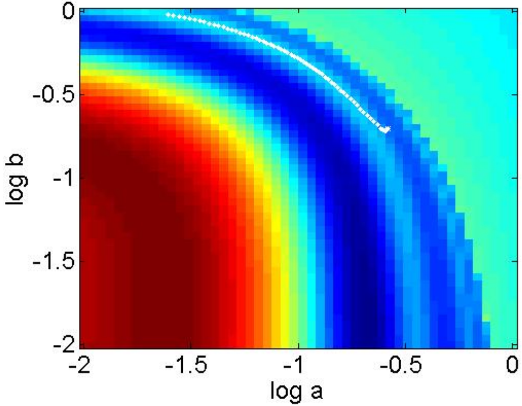
VL optimisation I

Global maxima



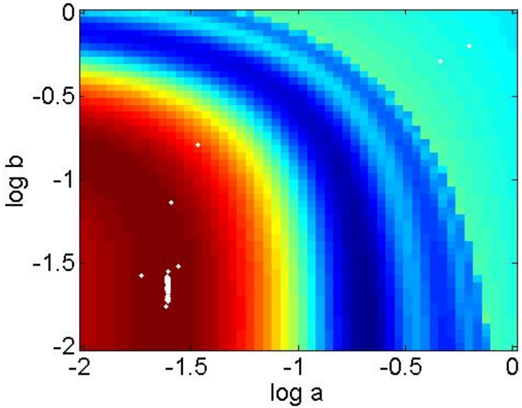
VL optimisation II

Local maxima



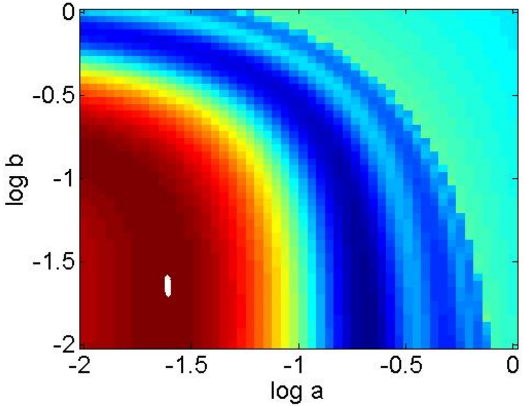
MH - Scaling

Init: $[-0.2, -0.2]$. Then 1000 samples



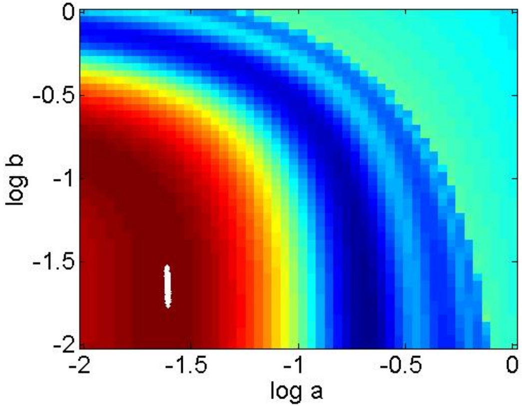
MH - Tuning

1000 samples



MH - Sampling

2000 samples



Pat Benatar Interlude

Happy Birthday Jean !

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

$$
\begin{aligned}
F &= -\frac{1}{2} e_y^T C_y^{-1} e_y - \frac{1}{2}\log|C_y| - \frac{N_y}{2}\log 2\pi \\
&= -\frac{1}{2} e_\theta^T C_\theta^{-1} e_\theta - \frac{1}{2}\log\frac{|C_\theta|}{|S_\theta|} \\
&= -\frac{1}{2} e_\lambda^T C_\lambda^{-1} e_\lambda - \frac{1}{2}\log\frac{|C_\lambda|}{|S_\lambda|}
\end{aligned}
\tag{8}
$$

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

$$
\begin{aligned}
e_y &= y - g(m_\theta) \\
e_\theta &= m_\theta - \mu_\theta \\
e_\lambda &= m_\lambda - \mu_\lambda
\end{aligned}
\tag{9}
$$

Prior Arithmetic Mean

The simplest approximation to the model evidence

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

is the Prior Arithmetic Mean

$$p_{PAM}(y|m) = \frac{1}{S}\sum_{s=1}^{S} p(y|\theta_s, m)$$

where the samples θ_s are drawn from the prior density.

A problem with this estimate is that most samples from the prior will have low likelihood. A large number of samples will therefore be required to ensure that high likelihood regions of parameter space will be included in the average.

Posterior Harmonic Mean

A second option is the Posterior Harmonic Mean

$$p_{PHM}(y|m) = \left[\frac{1}{S} \sum_{s=1}^{S} \frac{1}{p(y|\theta_s, m)} \right]^{-1}$$

where samples are drawn from the posterior (eg. through MH sampling).

A problem with the PHM is that the largest contributions come from low likelihood samples which results in a high-variance estimator.

Both PAM and PHM can be motivated from the perspective of importance sampling.

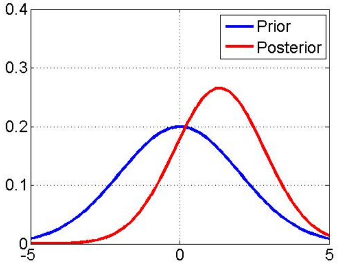
Savage-Dickey

For models 1 and 2 having common parameters θ_1 and model 2 having additional parameters θ_2, then if

$$p(\theta_1|m_2) = p(\theta_1|m_1)$$

the Bayes factor is given by

$$B_{12} = \frac{p(\theta_2 = 0|y, m_2)}{p(\theta_2 = 0|m_2)}$$



Here $B_{12} = 0.9$.

Thermodynamic Integration

We define inverse 'temperatures' β_k such that

$$0 = \beta_0 < \beta_1 < .. < \beta_{k-1} < \beta_K = 1$$

For example

$$\beta_k = \left(\frac{k}{K}\right)^5$$

We also define

$$f_k(\theta) = p(y|\theta, m)^{\beta_k} p(\theta|m)$$

Sample from kth chain using MH with prob

$$r = \frac{f_k(\theta'_k)}{f_k(\theta_k)}$$

Thermodynamic Integration

We can define the normalising constants

$$z_k = \int f_k(\theta) d\theta$$

where $z_0 = 1$ and $z_K = p(y|m)$. Now

$$\log p(y|m) = \log z_K - \log z_0$$

We can write this as

$$\log p(y|m) = \int_0^1 \frac{d \log z(\beta)}{d\beta} d\beta$$

Thermodynamic Integration

The log evidence can therefore be approximated as

$$\log p_{TI}(y|m) = \sum_{k=1}^{K-1} (\beta_{k+1} - \beta_k) \left(\frac{E_{k+1} + E_k}{2} \right)$$

where

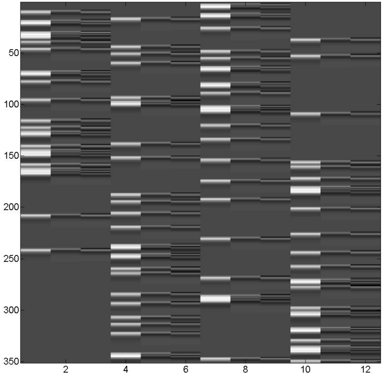
$$E_k = \frac{1}{N_k} \sum_{s=1}^{N_k} \log p(y|\theta_{ks})$$

where θ_{ks} is the sth sample from the kth chain.

Will Penny

Synthetic fMRI example

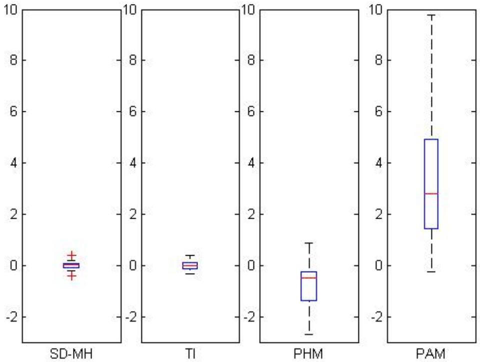
Design matrix from Henson et al. Regression coefficients from responsive voxel in occipital cortex. Data was generated from a 12-regressor model with SNR=0.2. We then fitted 12-regressor and 9-regressor models. This was repeated 25 times.



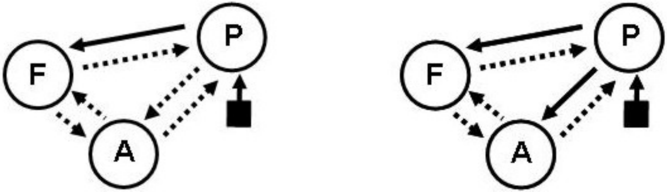
Log Bayes factors

For these linear Gaussian models the free energy defaults to the exact model evidence. Bayes factors are therefore exact. This also holds for Savage-Dickey. The average true logBF was 3.45 in favour of the 12-regressor model.

The boxplots show estimated minus true logBF for each approach:

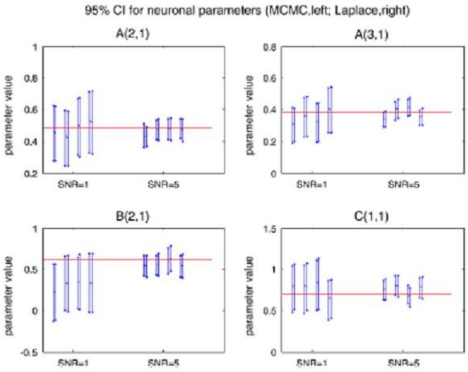


Auditory DCMs



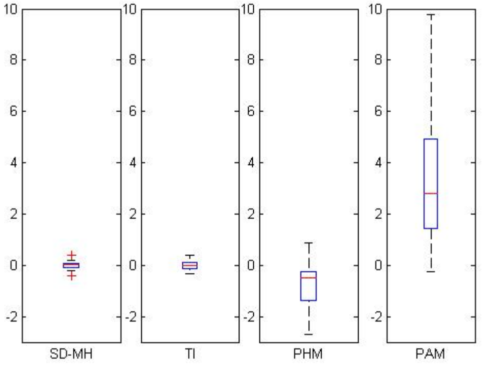
DCM for fMRI

Chumbley et al (2007) 175,000 burn samples +...



What Bayes factor results might look like !

Estimated logBF - F_{diff} for each approach:



Conclusions

VL is good. Care with $p(\theta), p(\lambda)$.

Bottleneck for sampling methods is speed of function evaluation. For DCMs we can generate 2-20 samples/second per core. This equates to 7200 to 72000 samples per hour.

May be worth looking at Metropolis Adjusted Langevin Algorithms (MALA) - basically a stochastic VL with MH step.