Bayesian Inference for Nonlinear Dynamical Systems

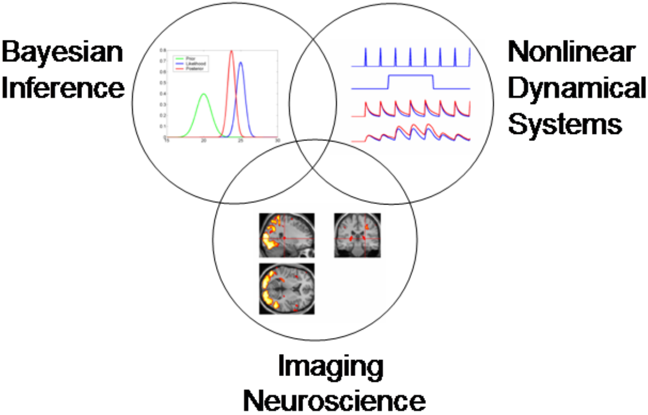
Will Penny

MPI Workshop on Networks in the Human Brain, Leipzig, 8th March 2011

Acknowledgements

Klaas Stephan, Alex Leff, Tom Schofield, Jean Daunizeau, Karl Friston, Maria Joao.

Context


Bayesian Inference
Nonlinear Dynamical Systems
Imaging Neuroscience

Model

We consider Bayesian estimation of nonlinear models of the form

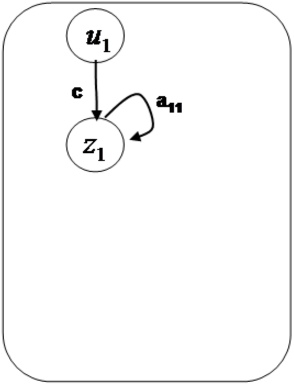
$$y = g(\theta, m) + e$$

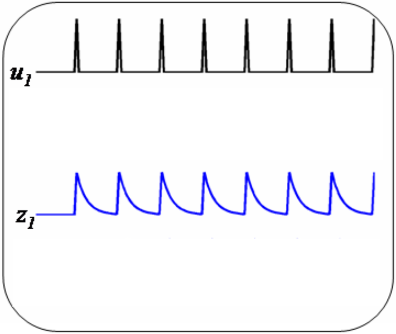
where $g(\theta)$ is some nonlinear function, and e is Gaussian noise.

As an example we consider $g(\theta, m)$ to be the prediction from a nonlinear differential equation model, such as a Dynamic Causal Model (DCM) for fMRI (Friston et al, 2003).

Single Region

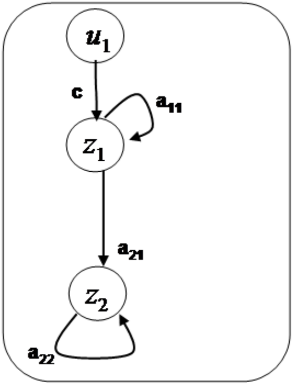
$$\dot{z}_1 = a_{11} z_1 + c u_1$$

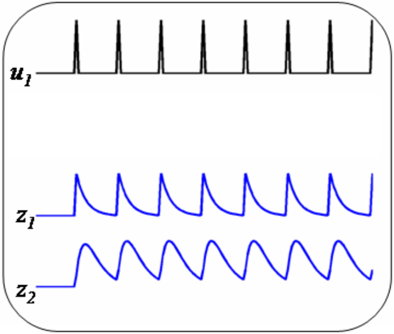




Two Regions

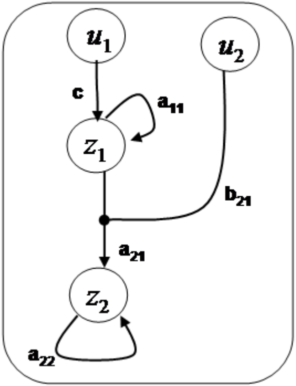
$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} c \\ 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

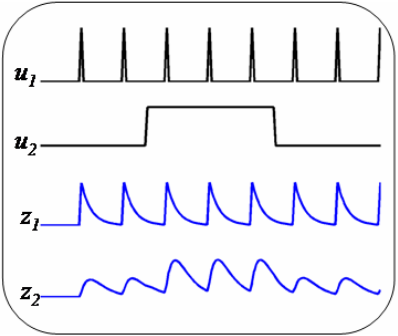




Modulations

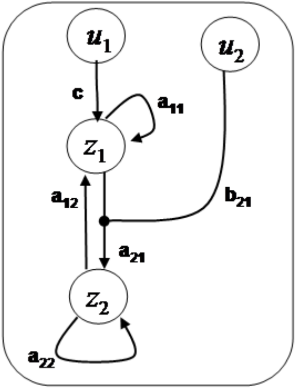
$$\begin{bmatrix}\dot{z}_1\\ \dot{z}_2\end{bmatrix} = \begin{bmatrix}a_{11} & 0\\ a_{21} & a_{22}\end{bmatrix}\begin{bmatrix}z_1\\ z_2\end{bmatrix} + u_2\begin{bmatrix}0 & 0\\ b_{21} & 0\end{bmatrix}\begin{bmatrix}z_1\\ z_2\end{bmatrix} + \begin{bmatrix}c & 0\end{bmatrix}\begin{bmatrix}u_1\\ u_2\end{bmatrix}$$

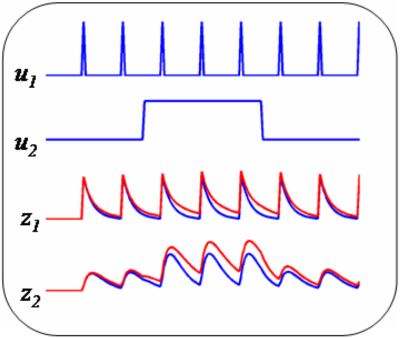




Recurrent Connections

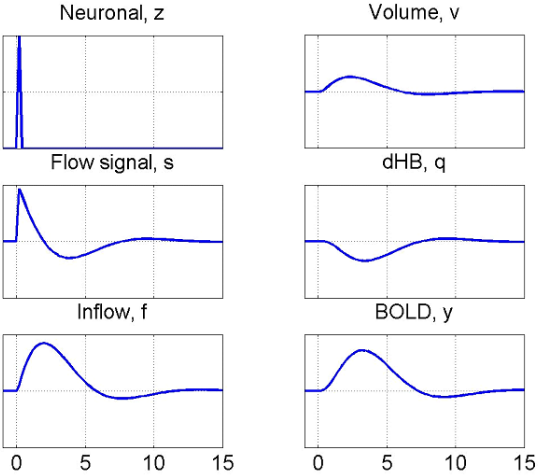
$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + u_2 \begin{bmatrix} 0 & 0 \\ b_{21} & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} c \\ 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$





Hemodynamics

Hemodynamic variables $x = [s, f, v, q]$.



Buxton et al (2004).

Integrating dynamics

Integrating neurodynamic

$$\dot{z} = \left(A + \sum_i u_i B_i \right) z + Cu$$

and hemodynamic

$$\dot{x} = h(x, z, w)$$

equations gives predictions of BOLD data y

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

where θ are all parameters and m indexes model structure.

Likelihood and Prior

We consider Bayesian estimation of nonlinear models of the form

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

The likelihood of the data is therefore

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

We allow Gaussian priors over model parameters

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

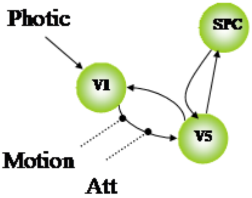
where the prior mean and covariance are assumed known.

Posteriors

Because we have a nonlinear model there is no simple formula for the posterior density. We therefore have to resort to approximate inference methods such as variational inference or sampling methods.

Here, we assumed the posterior to have a Gaussian form

$$q(\theta|y,m) = \mathrm{N}(\theta; m_\theta, S_\theta)$$



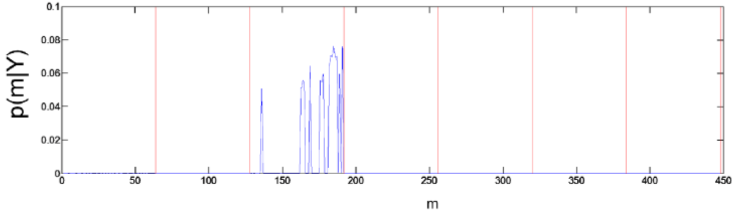
Its parameters are set using the Variational Laplace (VL) algorithm (Friston et al. 2007). This allows for inferences to be made about model parameters.

DCM for fMRI good for event-related designs.

Bayes rule for model inference

The posterior model probability is given by Bayes rule

$$p(m|y) = \frac{p(y|m)p(m)}{p(y)}$$



where $p(y|m)$ is the model evidence.

Bayesian model comparison can of course be applied to all statistical models eg. multivariate autoregressive models used for multivariate Granger causality (Penny et al. 2002).

Model Evidence

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

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

It can however be approximated using a number of methods

- Akaike's Information Criterion
- Bayesian Information Criterion
- Variational Free Energy
- Prior Arithmetic Mean
- Posterior Harmonic Mean
- Thermodynamic Integration

Free Energy

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

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

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

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

Free Energy

This can be rearranged as

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

where

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

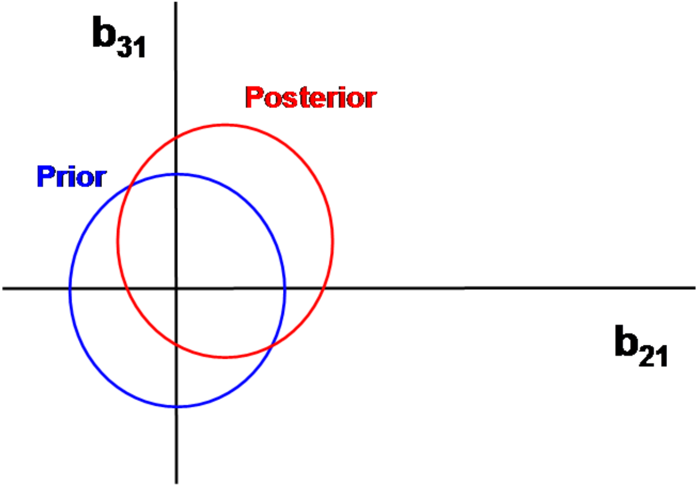
$$\begin{aligned} Complexity(m) &= KL[q(\theta|Y)||p(\theta)] \\ &= \frac{1}{2} e_\theta^T C_\theta^{-1} e_\theta + \frac{1}{2} \log \frac{|C_\theta|}{|S_\theta|} \end{aligned}$$

with prediction errors

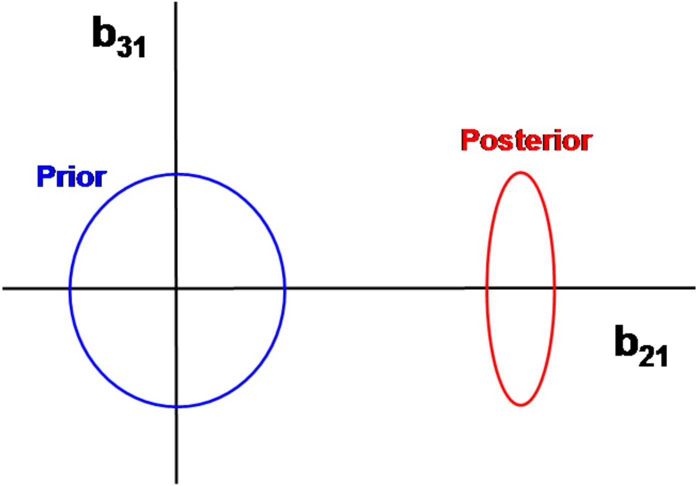
$$\begin{aligned} e_y &= y - g(m_\theta) \\ e_\theta &= m_\theta - \mu_\theta \end{aligned}$$

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

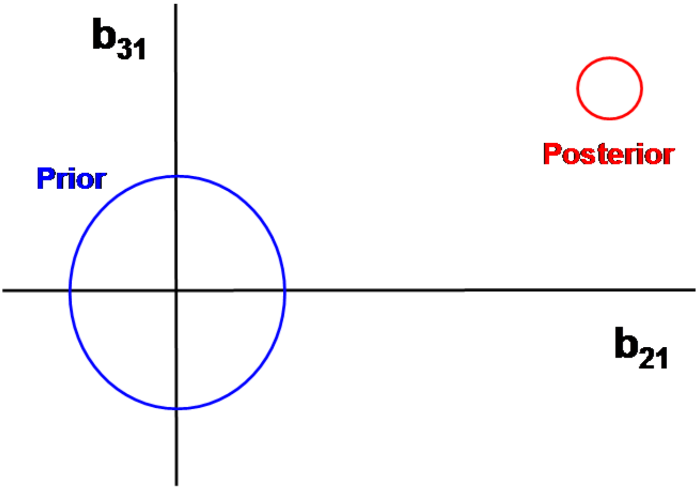
Small KL



Medium KL



Big KL



AIC and BIC

A simple approximation to the log model evidence is given by the Bayesian Information Criterion

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

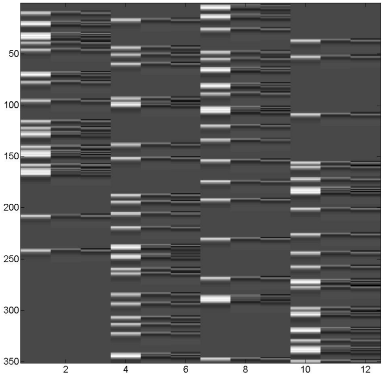
where $\hat{\theta}$ are the estimated parameters and hyperparameters, p is the number of parameters, and N_y is the number of data points. The BIC is a special case of the Free Energy approximation that drops all terms that do not scale with the number of data points

An alternative approximation is Akaike's Information Criterion (or 'An Information Criterion')

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

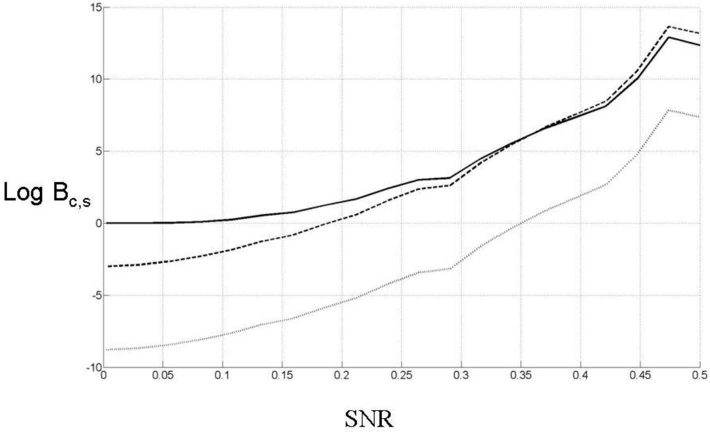
Synthetic fMRI example

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.



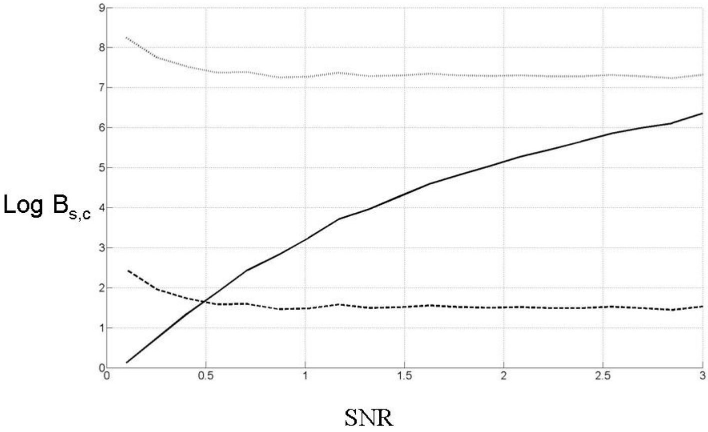
True Model: Complex GLM

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).



fMRI study of auditory word processing

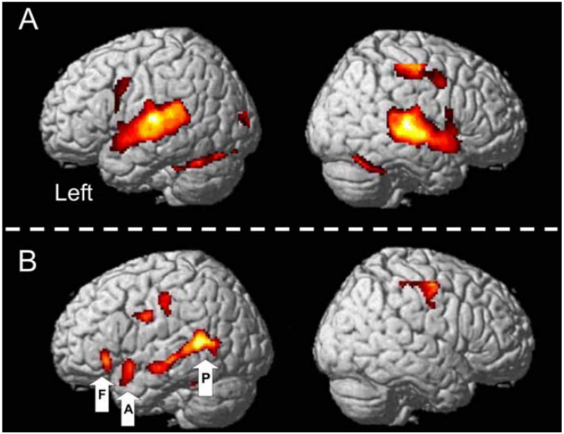
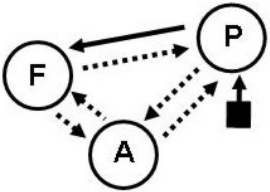
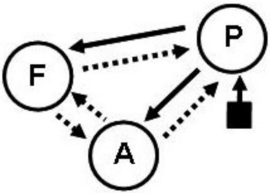


Figure 1. Results from the two SPM analyses. *A*, Main effects of all auditory stimuli; *B*, main effect of intelligible − unintelligible stimuli (intelligibility contrast). Results from analysis *A* show bilateral activation of Heschel's gyrus, planum temporal, and STG as well as cerebellar, visual, and right motor areas associated with making a decision on the gender of the speaker of the auditory stimuli. Results from analysis *B* show areas activated by intelligible auditory stimuli and include the length of the STS and part of the IFG on the left (POrb). Data from three of these areas (VOIs) were entered into the DCM analysis: A, aSTS; F, IFG (POrb); P, pSTS. *A* is thresholded at a voxel level of $p < 0.05$ (corrected for the search volume) and at a cluster level of 100 contiguous voxels; *B* is thresholded at a voxel level of $p = 0.01$ (uncorrected) and a cluster level of 50 contiguous voxels.

Synthetic data

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.

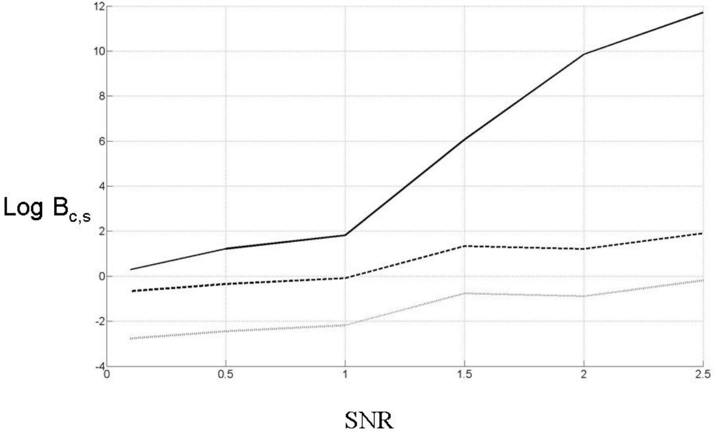




Use empirical regressors (i) auditory input and (ii) intelligibility (speech versus reversed speech)

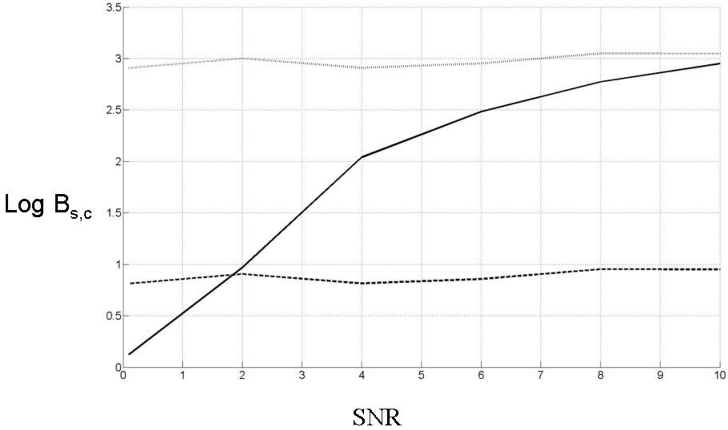
True Model: Complex DCM

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).



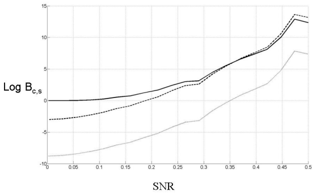
True Model: Simple DCM

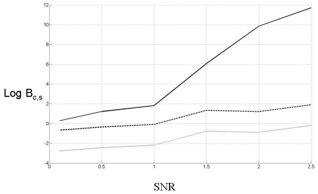
Log Bayes factor of simple versus complex model, Log $B_{s,c}$, versus the signal to noise ratio, SNR, when true model is the simple DCM for F (solid), AIC (dashed) and BIC (dotted).



A surprise

For generating data from the simpler models the results are the same for GLMs and DCMs. But for generating data from complex models they are not (left: GLM, right:DCM)



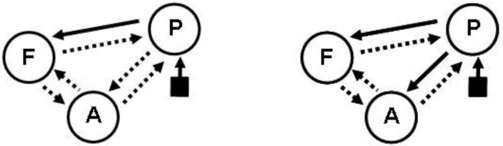


What is going on ?

Complexity

By decomposing the Free Energy difference into contributions from different parameters, we found that this ability was mainly due to penalising the simple model for having a very large, and a-priori unlikely, intrinsic connection from brain region F to A.

$$\frac{1}{2} e_\theta^T C_\theta^{-1} e_\theta = \frac{1}{\sigma_a} \sum_i a_i^2 + \ldots$$



Because AIC and BIC use the same complexity penalty for every parameter (regardless of its magnitude) they lack this sensitivity.

Model Evidence

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

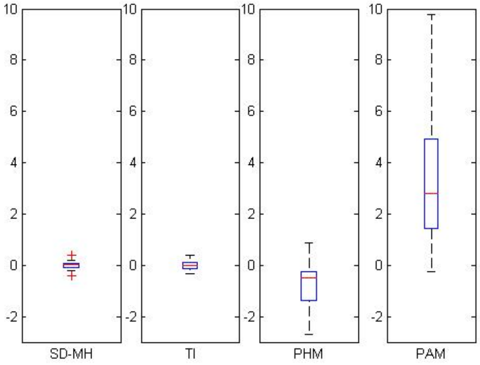
$$p(y|m) = \int p(y|\theta, m)p(\theta|m)d\theta.$$

It can however be approximated using a number of methods

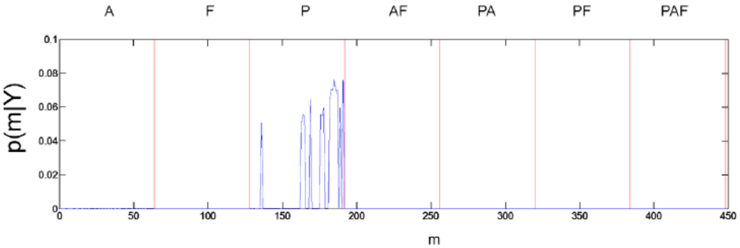
- Akaike's Information Criterion
- Bayesian Information Criterion
- Variational Free Energy
- Prior Arithmetic Mean
- Posterior Harmonic Mean
- Thermodynamic Integration

Sample-based methods

For GLMs the free energy defaults to the exact model evidence. Bayes factors are therefore exact. The boxplots show estimated minus true logBF for each sample-based approach.



Comparing large numbers of models

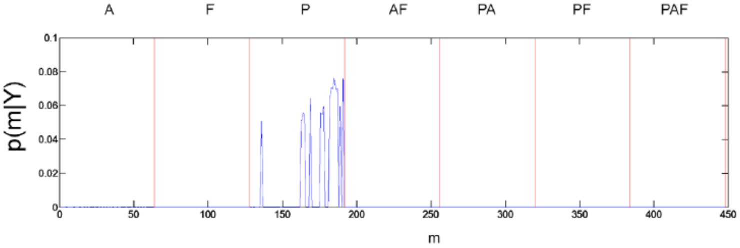


Bayes rule for families

$$p(f|y) = \frac{p(y|f)p(f)}{p(y)}$$

For comparing model families having unequal numbers of models, the model level prior $p(m)$ can be adjusted to make $p(f)$ uniform (Penny et al. 2010).

Comparing model families



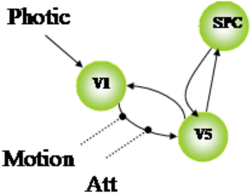
The same model space can be partitioned into different ways, like a factorial design, and inferences can be made about a particular factor by collapsing over others eg linear vs nonlinear, recurrent versus feedforward.

Bayesian Model Averaging

Integrating out model uncertainty

$$p(\theta|Y) = \sum_m p(\theta|Y,m)p(m|Y)$$

Make inferences about parameters eg between subjects

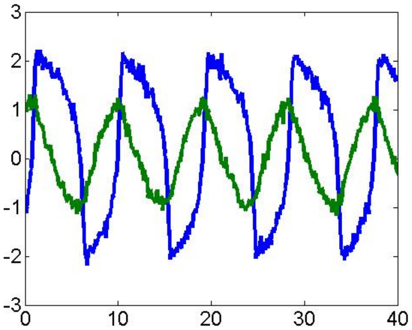


and groups.

Intrinsic Dynamics

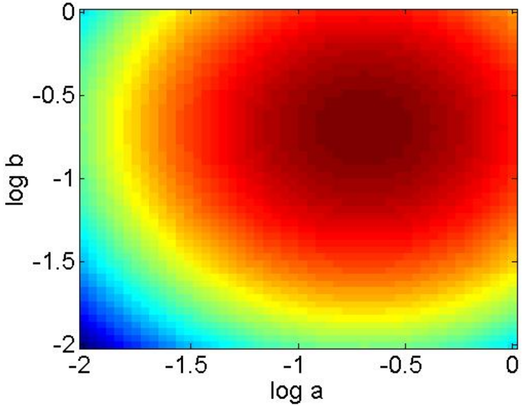
Nonlinear oscillator with $a = 0.2$, $b = 0.2$, $c = 3$.

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



Priors

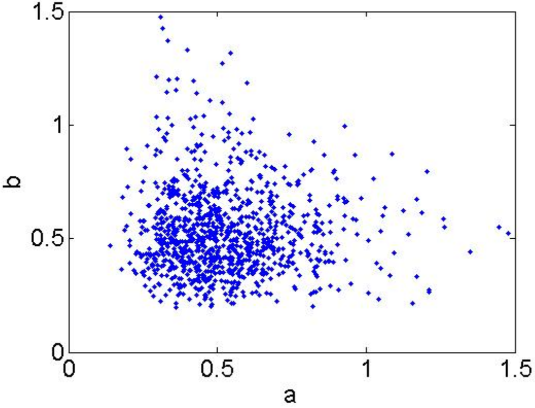
A plot of $\log p(\theta)$



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

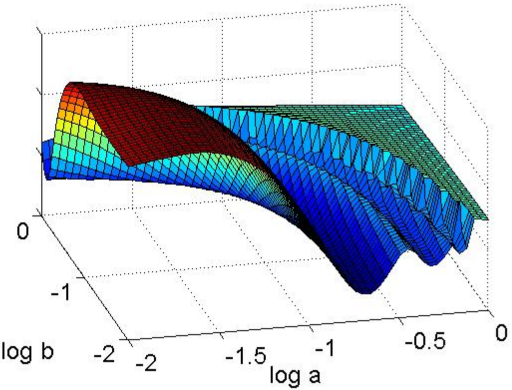
Priors

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



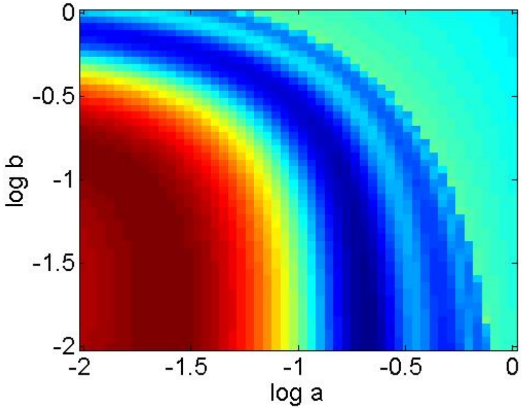
Posterior

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



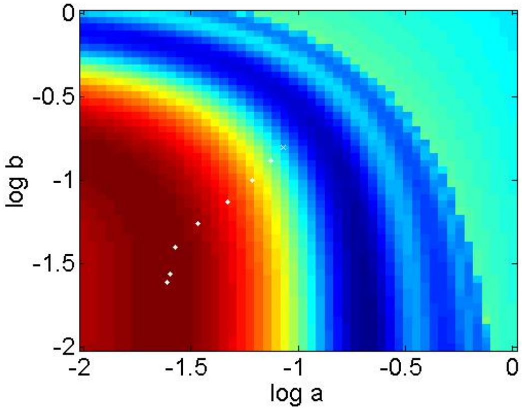
Posterior

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



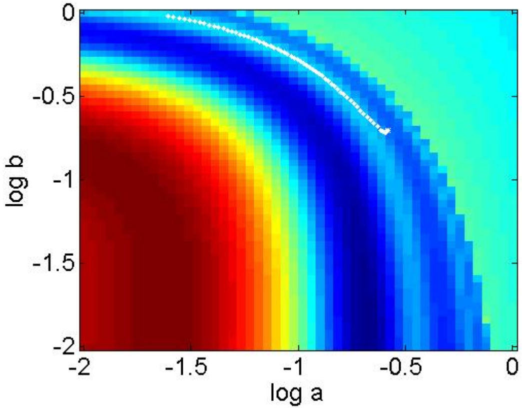
VL optimisation I

Global maxima



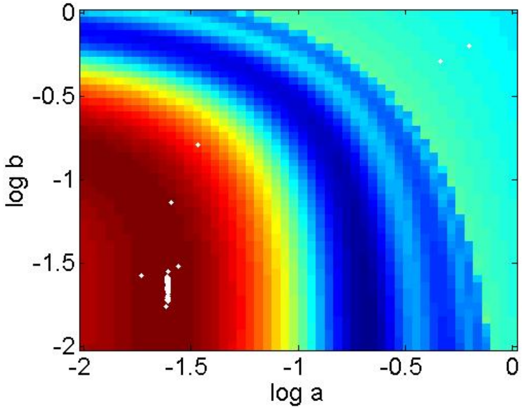
VL optimisation II

Local maxima



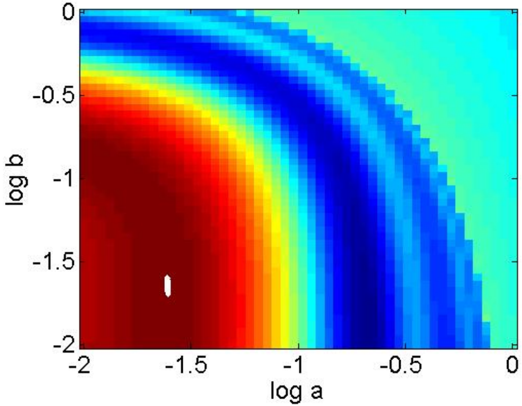
MH - Scaling

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



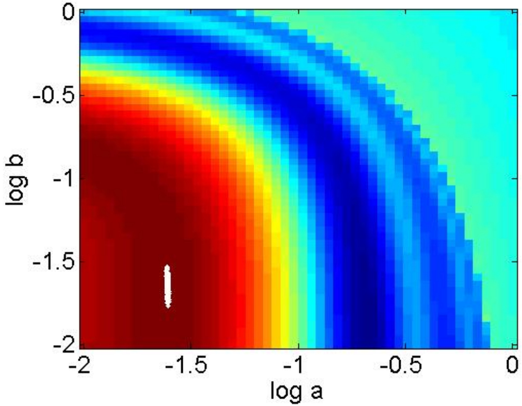
MH - Tuning

1000 samples



MH - Sampling

2000 samples



References

Friston et al (2003) Dynamic Causal Modelling. Neuroimage 19, 1273-1302.

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

Penny et al. (2002) Bayesian multivariate autoregressive models with structured priors. IEE Prov Vis Im Sig 149, 33-41.

Penny et al. (2004) Comparing Dynamic Causal Models. Neuroimage 22, 1157-1172.

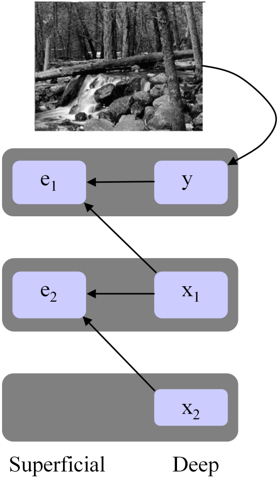
Penny et al. (2010) Comparing Families of Dynamic Causal Models. PLoS CB 6, e1000709.

Hierarchical Predictive Coding

Mumford (1991) ”I put forward a hypothesis on the role of the reciprocal, topographic pathways between two cortical areas, one often a ‘higher’ area dealing with more abstract information about the world, the other ’lower’ dealing with more concrete data. The higher area attempts to fit its abstractions to the data it receives from lower areas by sending back to them from its deep pyramidal cells a template reconstruction best fitting the lower level view. The lower area attempts to reconcile the reconstruction of its view that it receives from higher areas with what it knows, sending back from its superficial pyramidal cells the features in its data which are not predicted by the higher area. The whole calculation is done with all areas working simultaneously, but with order imposed by synchronous activity in the various top-down, bottom-up loops”

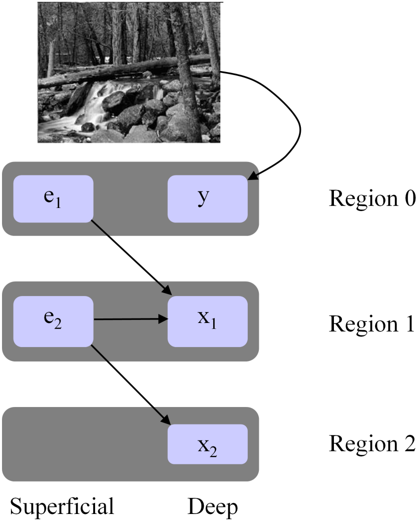
Predictive Coding

Top-down connections (from deep layers) embody an (Empirical Bayesian) generative model.



Predictive Coding

Bottom-up connections (from superficial layers) send prediction errors.



Fries,Weeden,Turner.

But see Lohmann/Logothetis for hidden node problem ?

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.

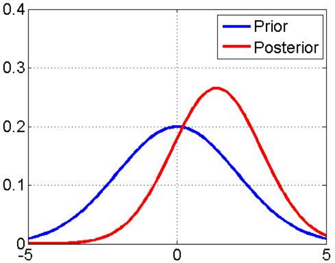
Savage-Dickey

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

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

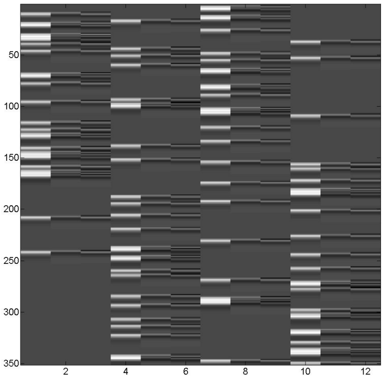
where

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

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

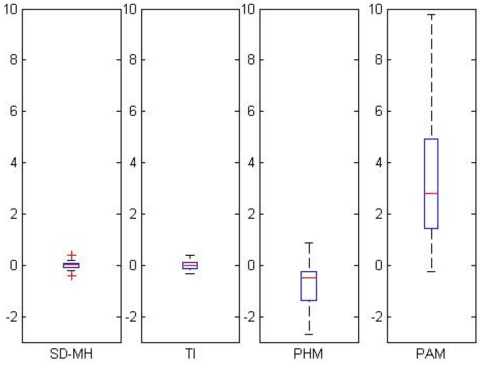
Synthetic fMRI example

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. The boxplots show estimated minus true logBF for each approach.



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

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

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

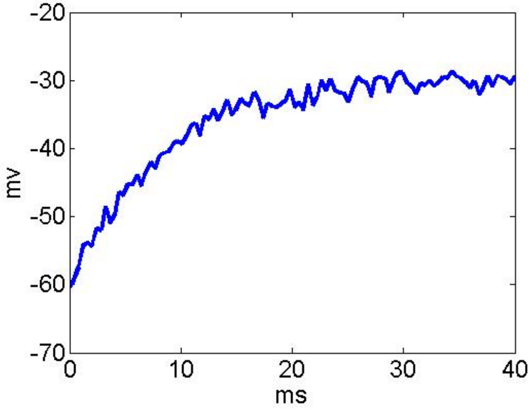
$$
\begin{aligned}
j_\theta(i) &= \frac{dI(\theta)}{d\theta(i)} \\
H_\theta(i,j) &= \frac{d^2 I(\theta)}{d\theta(i) d\theta(j)}
\end{aligned}
$$

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

$$
m_\theta(new) = m_\theta(old) - H_\theta^{-1} j_\theta
$$

Likelihood

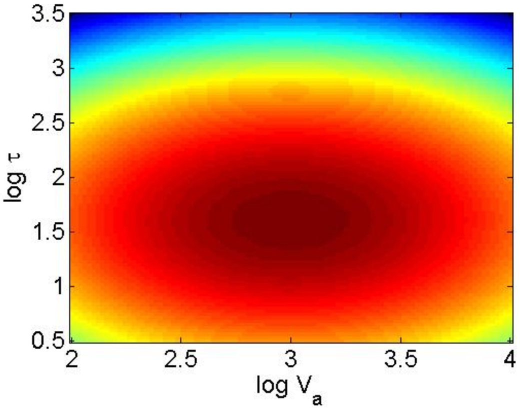
$$y(t) = -60 + V_a[1 - \exp(-t/\tau)] + e(t)$$



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

Prior Landscape

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



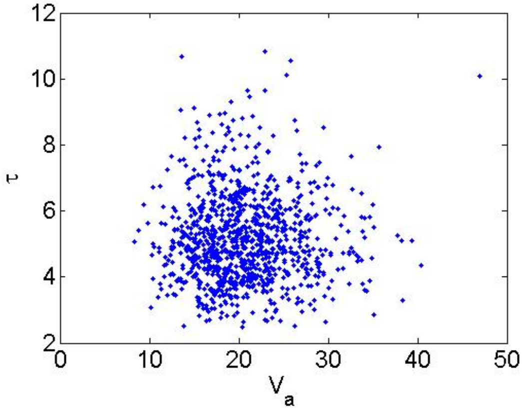
$$\mu_\theta = [3, 1.6]^T, C_\theta = diag([1/16, 1/16]);$$

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

Samples from Prior

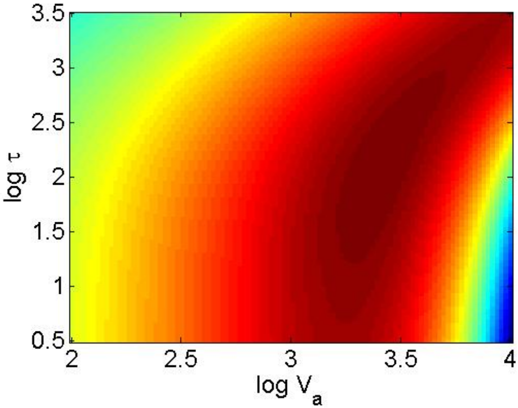
The true model parameters are unlikely apriori

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



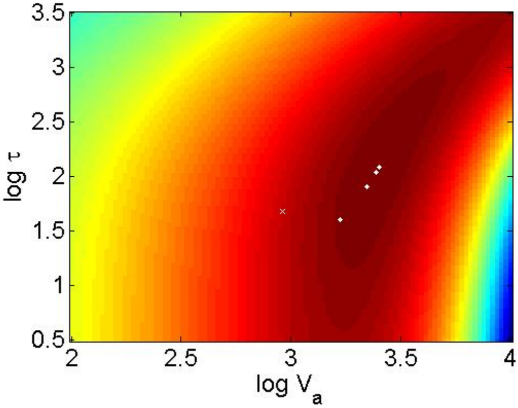
Posterior Landscape

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



VL optimisation

Path of 6 VL iterations (x marks start)



Modulations

Stephan et al (2010): "...rapid changes of connection strength can result either from membrane excitability changes, synaptic plasticity, or a combination of both. For example, postsynaptic responses of ionotropic glutamatergic receptors are modulated by metabotropic receptors (Coutinho and Knopfel, 2002) and by receptors of various neuromodulatory transmitters (McCormick and Williamson, 1989). Alternatively, various forms of short-term synaptic plasticity can lead to fast changes in synaptic strength, e.g. synaptic depression and facilitation (Zucker and Regehr, 2002), NMDA- and dopamine-dependent phosphorylation of AMPA receptors (Chao et al., 2002;Wang et al., 2005), or dendritic spine motility (Holtmaat and Svoboda, 2009). All of these changes in synaptic strength can unfold within milliseconds to seconds."