Annealed Importance Sampling for Neural Mass Models

Will Penny, UCL.

WTCN, UCL, October 2015.

Generative Model

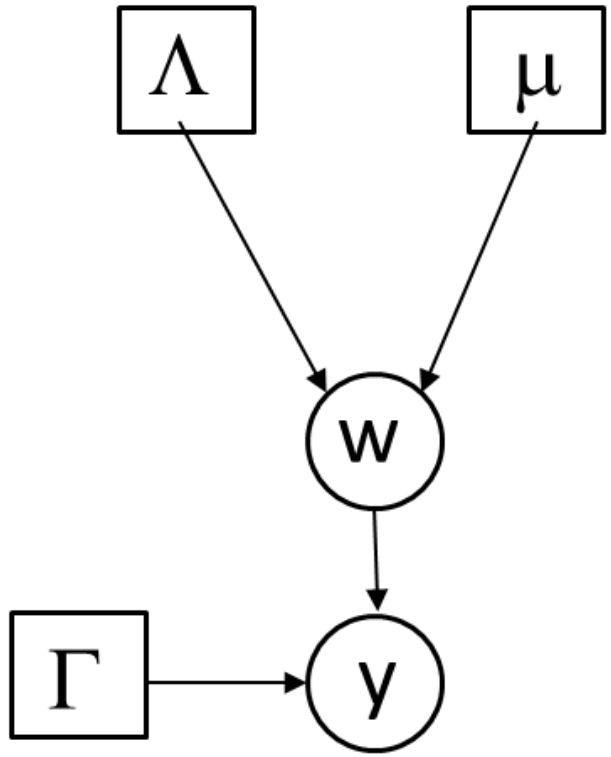
Behavioural or imaging data y.

Likelihood $p(y|w, \Gamma)$.

We have a Gaussian prior over model parameters

$$p(w|\mu, \Lambda) = \mathcal{N}(w; \mu, \Lambda)$$

Assume observation noise precision, Γ, prior mean μ and precision Λ are known.



Estimation and Inference

Variational Laplace (VL):

- Local optimisation based on gradients and curvatures
- Posterior assumed Gaussian
- Provides model evidence estimate
- Very fast

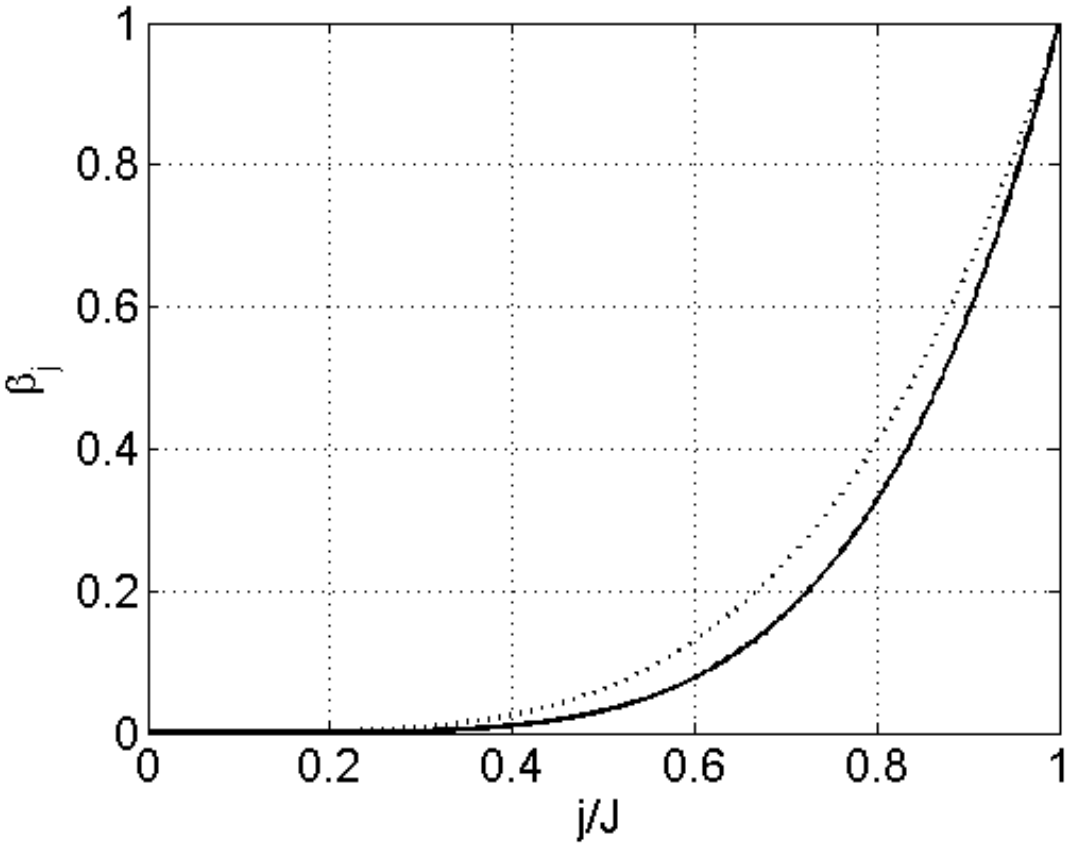
Annealed Importance Sampling (AIS):

- Avoid Local Maxima
- Posterior not Assumed Gaussian
- Provides model evidence estimate
- Use Langevin Monte Carlo (grad and curve used) for proposals
- Test parametric assumptions
- Slow

Work with Biswa Sengupta @ FIL.

Annealed Importance Sampling

Inverse temperatures β_j with $j = 0..J$, $\beta_0 = 0$ and $\beta_J = 1$.
Geometric schedule $\beta_j = (j/J)^5$ (solid), $\beta_j = (j/J)^4$ (dotted).



For the jth temperature the algorithm produces a sample from

$$f_j(w) = p(y|w)^{\beta_j} p(w)$$

Annealed Importance Sampling

An independent sample $w^{(i)}$ from the posterior density is produced by generating a sequence of points $w_1, w_2, \ldots w_J$ as follows

- Generate w_1 from $p(w)$
- Generate w_2 from w_1 using $T_1(w_2|w_1)$
- ...
- Generate w_j from w_{j-1} using $T_{j-1}(w_j|w_{j-1})$
- ...
- Generate w_J from w_{J-1} using $T_{J-1}(w_J|w_{J-1})$

and then let $w^{(i)} = w_J$. We refer to the process of producing a single independent sample as a 'trajectory'.

We are using Langevin Monte Carlo for the T_j's.

Langevin Monte Carlo

Given log joint and its gradient as a function of w

$$\begin{aligned} f_j(w) &= p(y|w,\Gamma)^{\beta_j} p(w|\mu,\Lambda) \\ L_j(w) &= \beta_j \log p(y|w,\Gamma) + \log p(w|\mu,\Lambda) \\ g_j(w) &= \frac{dL_j(w)}{dw} \end{aligned}$$

the LMC Proposal is drawn as

$$\begin{aligned} w_j^* &\sim p(w_j^*|w_{j-1}) \\ p(w_j^*|w_{j-1}) &= \mathcal{N}(w_j^*; m_j, C_j) \\ m_j &= w_{j-1} + \frac{1}{2} C_j g_j(w_{j-1}) \\ C_j &= h^2 \left(\Lambda + \beta_j S^T \Gamma S\right)^{-1} \end{aligned}$$

where S is a sensitivity matrix

$$S(i,k) = \frac{dy(i)}{dw_s(k)}$$

Langevin Monte Carlo

The proposal is accepted using the standard Metropolis-Hastings probability

$$a = \frac{f_j(w_j^*)}{f_j(w_{j-1})} \frac{p(w_{j-1}|w_j^*)}{p(w_j^*|w_{j-1})}$$

The proposal is always accepted if $a > 1$.

If the step is accepted we set $w_j = w_j^*$. If it is rejected we set $w_j = w_{j-1}$.

The second term above ensures reversibility, and in principle that we visit all of parameter space in proportion to its (posterior) probability.

Annealed Importance Sampling

The above process is repeated $i = 1..I$ times to produce I independent samples from the posterior density.

Because the samples are produced independently, without interaction among trajectories, the AIS algorithm is amenable to 'embarrassing parallelization'

We need not concern ourself with within-trajectory correlation (as e.g. Hamiltonian Monte Carlo does) as we're only taking one sample from each

Effectively, AIS is a multistart algorithm, that has a principled way of combining information from multiple starts/trajectories

Annealed Importance Sampling

Each sample is also accompanied by an importance weight

$$v^{(i)} = \frac{f_1(w_1)}{f_0(w_1)} \frac{f_2(w_2)}{f_1(w_2)} \frac{f_3(w_3)}{f_2(w_3)} \cdots \frac{f_J(w_J)}{f_{J-1}(w_J)}$$

which can be evaluated as

$$\log v^{(i)} = \sum_{j=1}^{J} (\beta_j - \beta_{j-1}) \log p(y|w_j)$$

The importance weights, or average of them, provide an approximation to the model evidence.

Annealed Importance Sampling

We define the normalising constant at each temperature as

$$\begin{aligned} Z_j &= \int f_j(w) dw \\ &= \int p(y|w,m)^{\beta_j} p(w|m) dw \end{aligned}$$

We then have

$$\begin{aligned} Z_1 &= \int p(w|m) dw = 1 \\ Z_J &= \int p(y|w,m) p(w|m) dw \\ &= p(y|m) \end{aligned}$$

Annealed Importance Sampling

Therefore

$$\begin{aligned} p(y) &= \frac{Z_J}{Z_1} \\ &= \frac{Z_2}{Z_1}\frac{Z_3}{Z_2}\cdots\frac{Z_J}{Z_{J-1}} \\ &= \prod_{j=1}^{J-1} r_j \end{aligned}$$

where $r_j = Z_{j+1}/Z_j$. We can then write

$$\begin{aligned} r_j &= \frac{1}{Z_j}\int f_{j+1}(w)dw \\ &= \int \frac{f_{j+1}(w)}{f_j(w)}\frac{f_j(w)}{Z_j}dw \\ &\approx \frac{1}{N}\sum_{n=1}^{N}\frac{f_{j+1}(w_n)}{f_j(w_n)} \end{aligned}$$

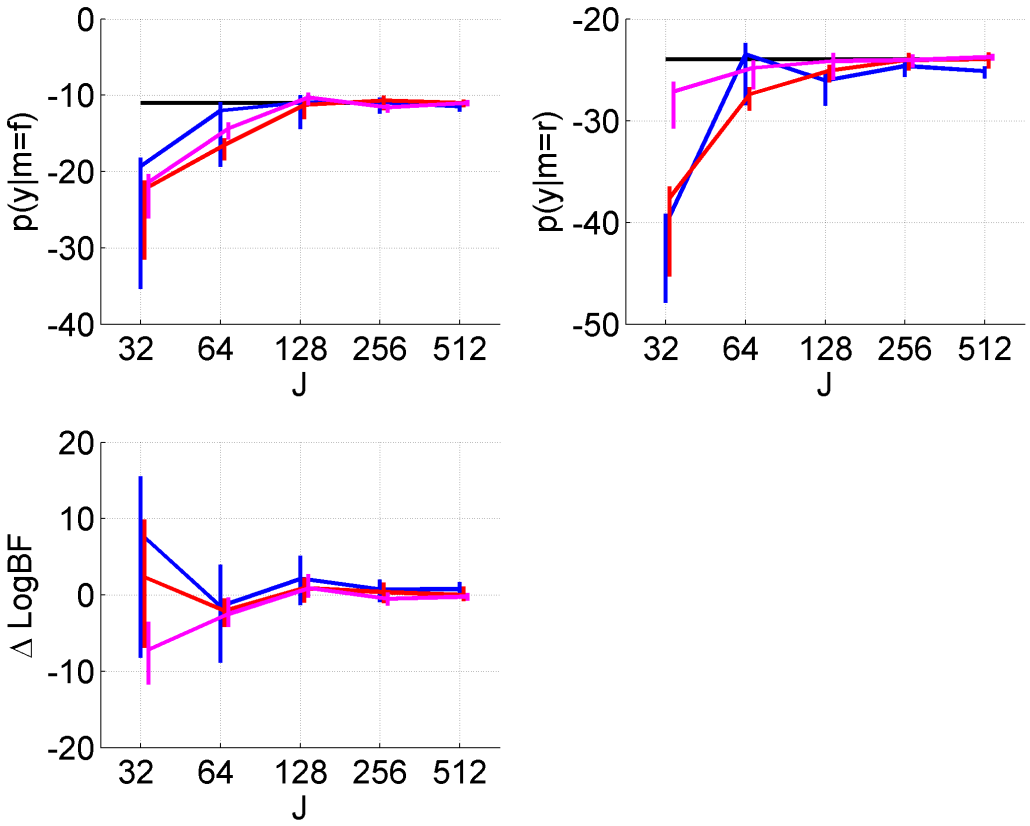
where the last line indicates a Monte-Carlo approximation of the integral with samples w_n drawn from the distribution at temperature β_j. This can in turn be written as

$$r_j = \frac{1}{N}\sum_{n=1}^{N} p(y|w_n, m)^{\beta_{j+1}-\beta_j}$$

For $N = 1$ this equals the importance weight.

Linear Regression

AIS approximations with $I = 16$ (blue), $I = 32$ (red) and $I = 64$ (magenta) trajectories.The black lines show the equivalent analytic quantities.



Vertical lines span the 5th and 95th percentiles from bootstrapping.

Linear Regression

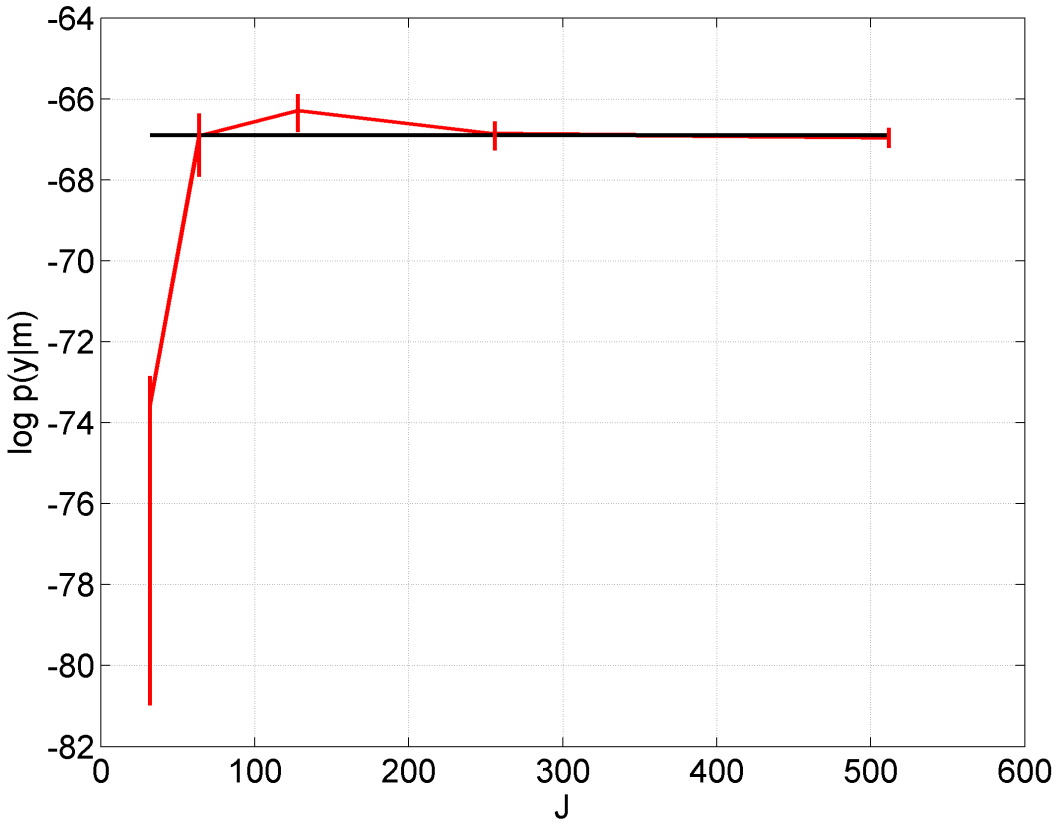
Using the 32 samples produced by AIS, we could not reject the hypothesis that the posterior is Gaussian using Royston's test for the full ($p = 0.67$) and reduced ($p = 0.68$) models.

Royston, J.P. (1992). Approximating the Shapiro-Wilk W-Test for non-normality. Statistics and Computing, 2:117-119.

As the samples from AIS are IID we can use this test without e.g. correcting for temporal autocorrelation (c.f. other MCMC schemes)

Approach to Limit

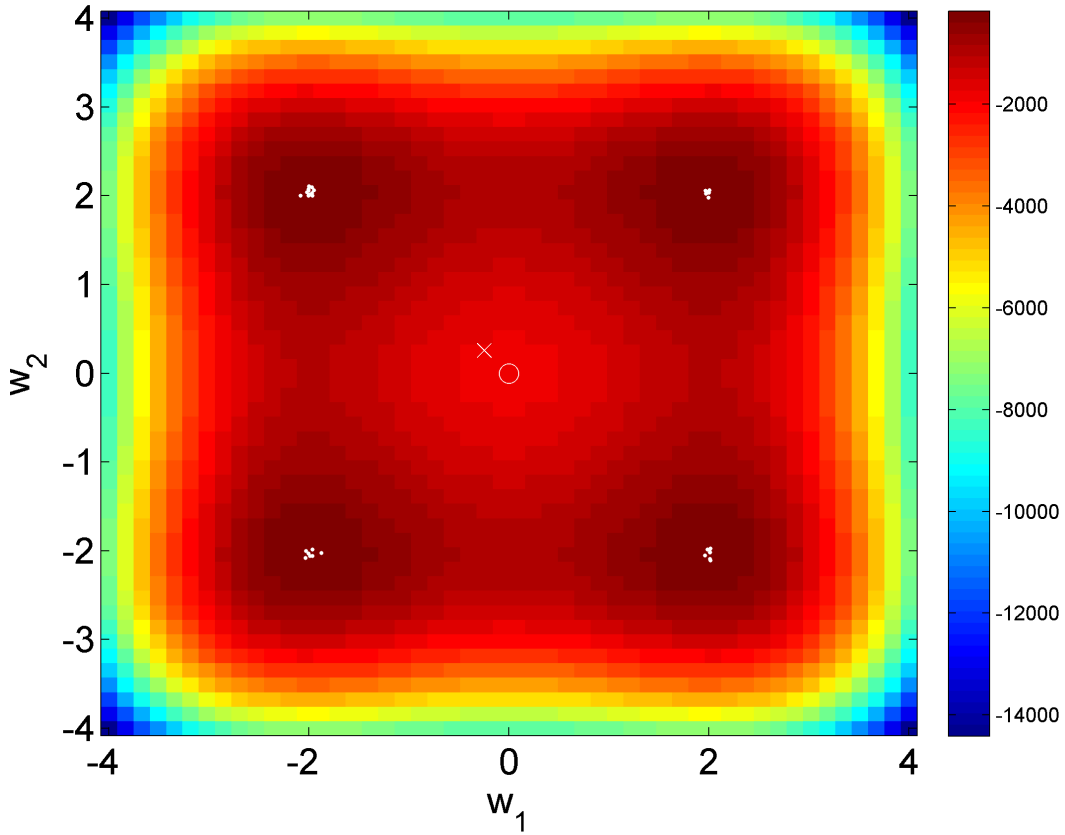
AIS with $J = 32$ trajectories.



Using the 32 samples produced by AIS, we could not reject the hypothesis that the posterior is Gaussian using Royston's test (p = 0.96).

Squared Parameters

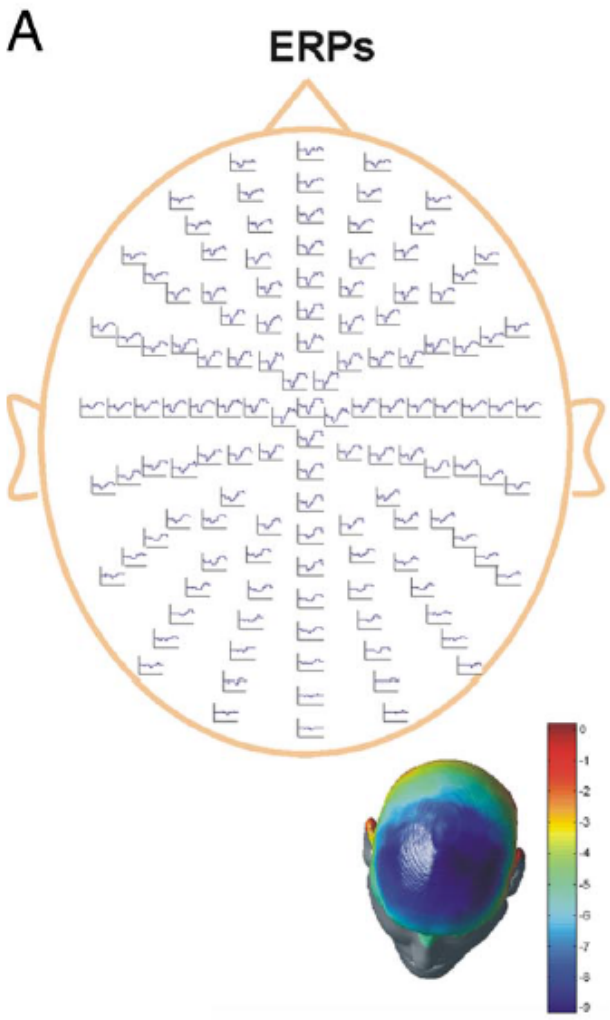
Linear regression but with squared parameters

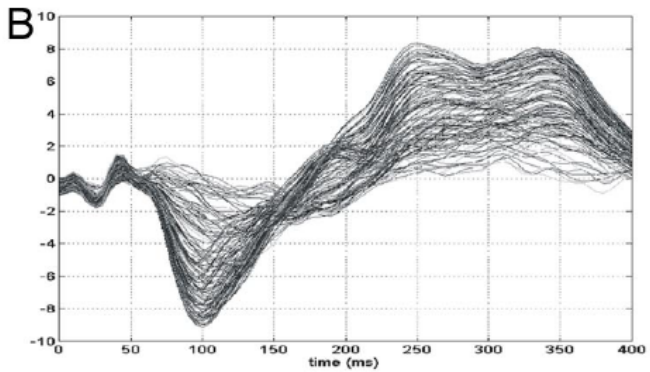


We can reject the hypothesis that the posterior is Gaussian using Royston's test ($p = 10^{-12}$).

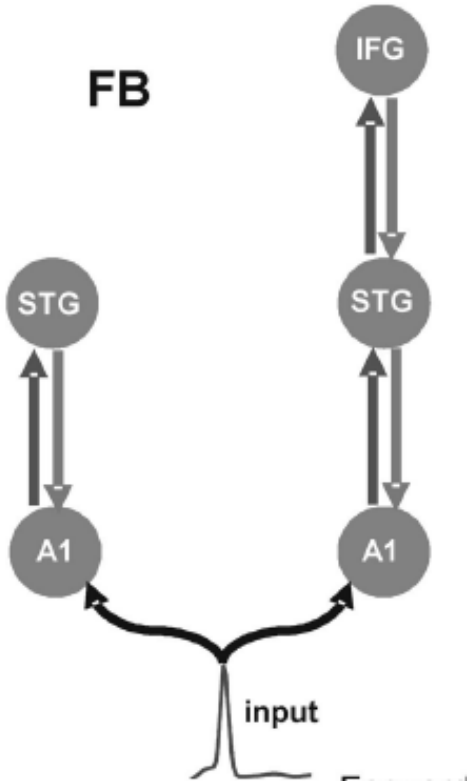
Brain Connectivity

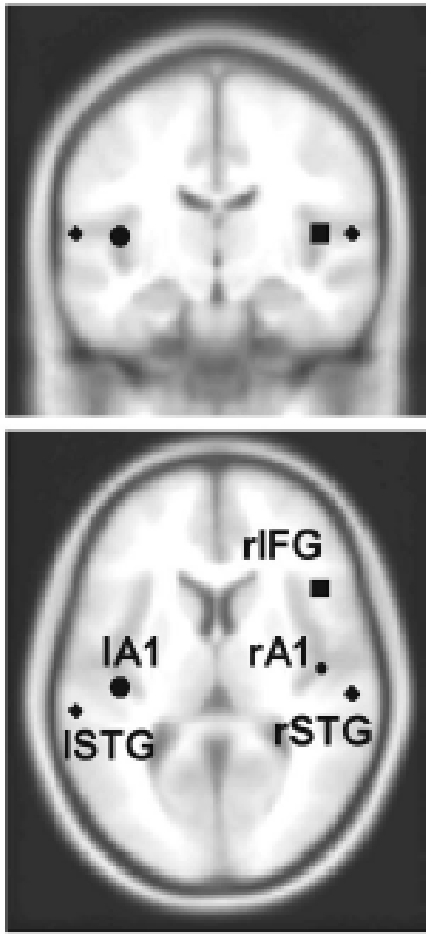
EEG data y for subject n.





Brain Connectivity Model

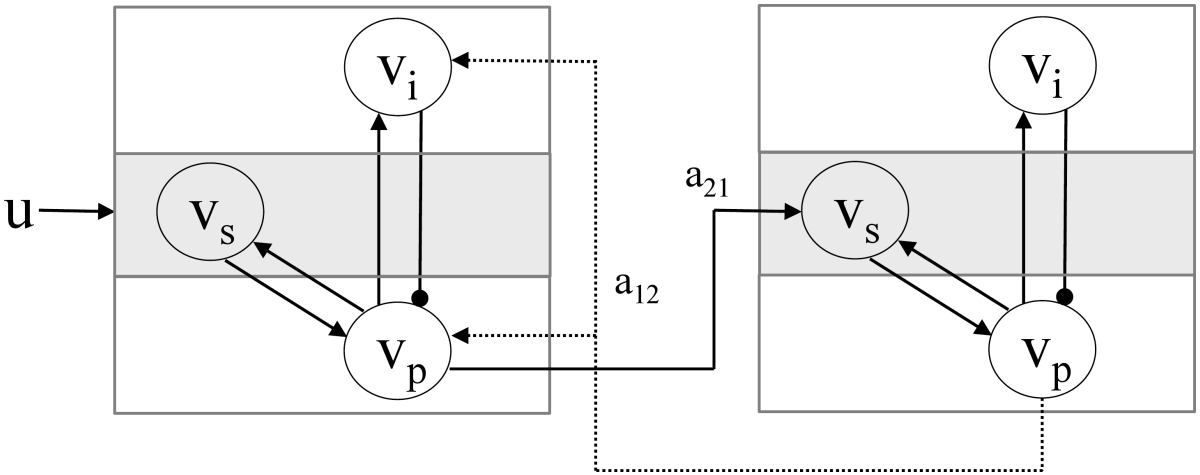




Garrido et al. Evoked brain responses are generated by feedback loops. *PNAS*, 2007.

Neural Masses

Neural mass models have been proposed as network models of cortical activity.

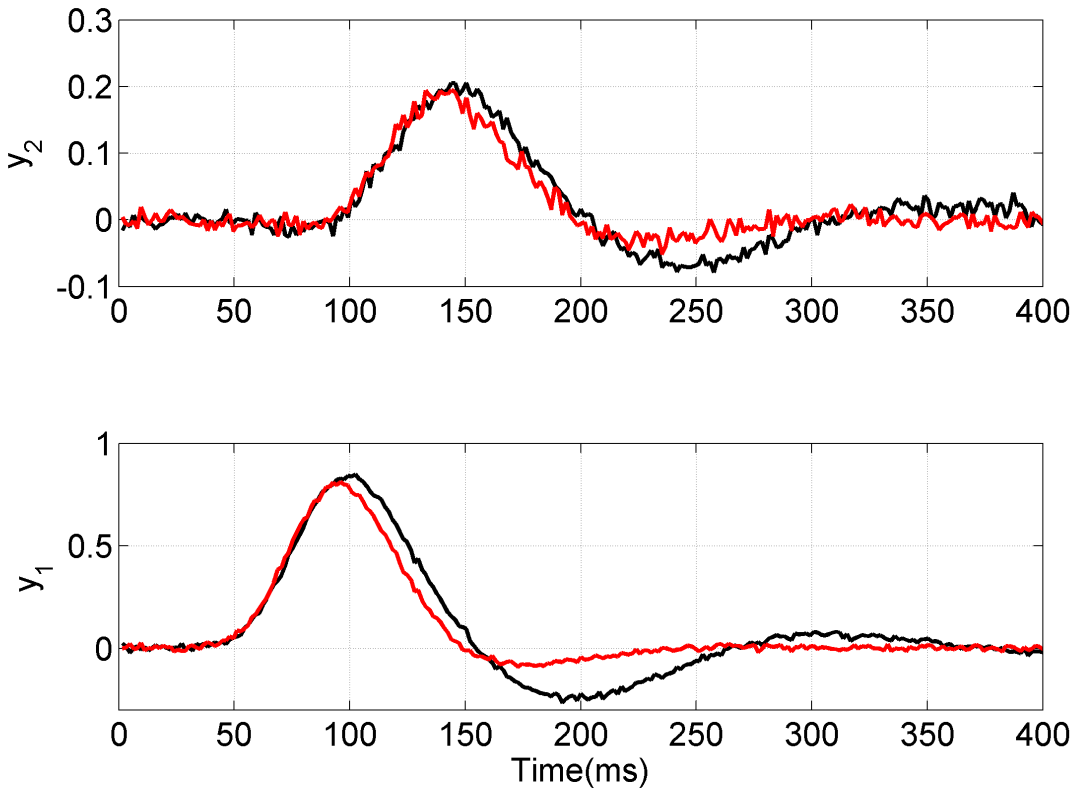


We estimate a 10-dimensional parameter vector w. These are between-region connex, a_{12}, a_{21}, between region delays δ_{12}, δ_{21}, within-region connex $\gamma_{1..4}$ and parameters of firing rate function r_1, r_2.

David et al, Dynamic Causal Models for Event-Related Potentials. *Neuroimage*, 2006.

Two region model

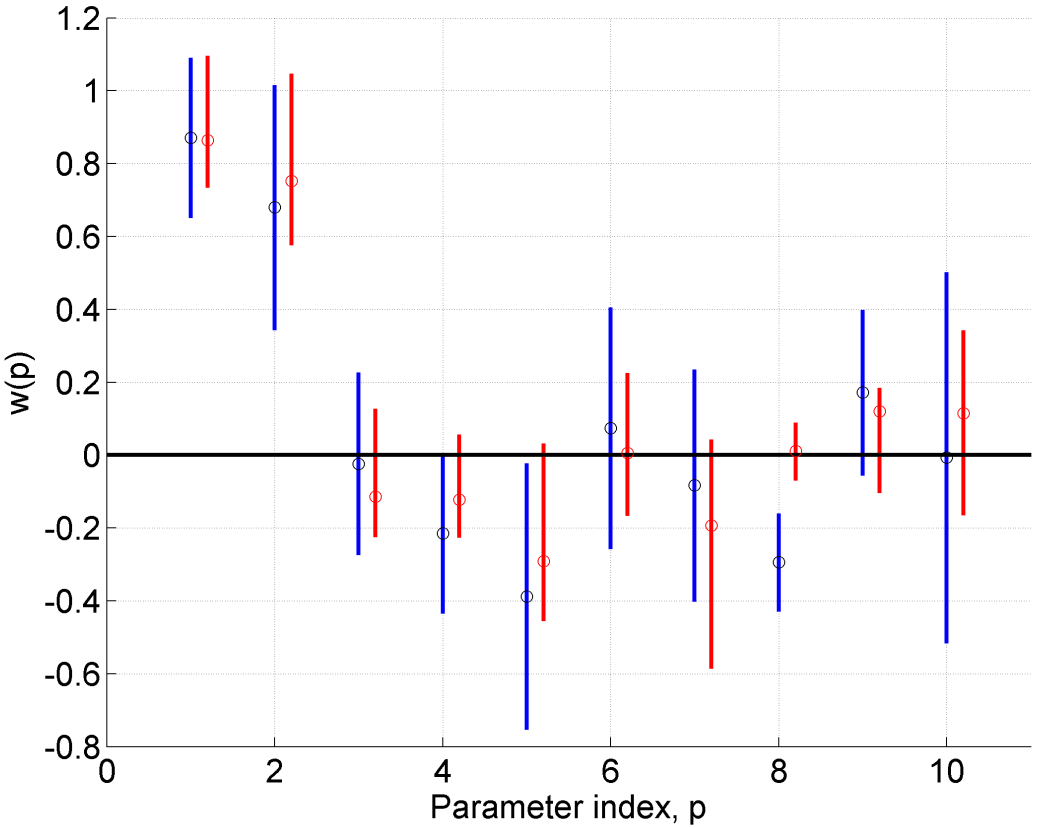
Impulse of activity, u, at $t = 0$ produces observed time series, y_1, being pyramidal cell activity in lower-level (sensory) region.



Observed time series, y_2, is pyramidal cell activity in higher-level region.

Parameter Estimates

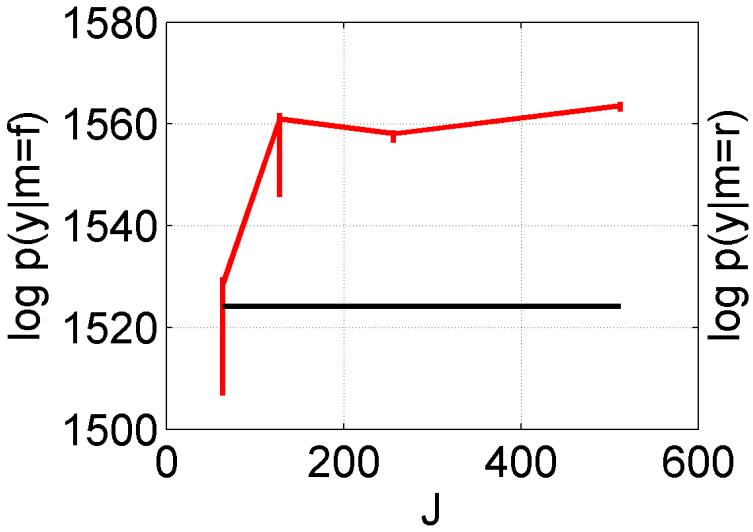
With 95% confidence intervals. AIS (red) VL (blue).

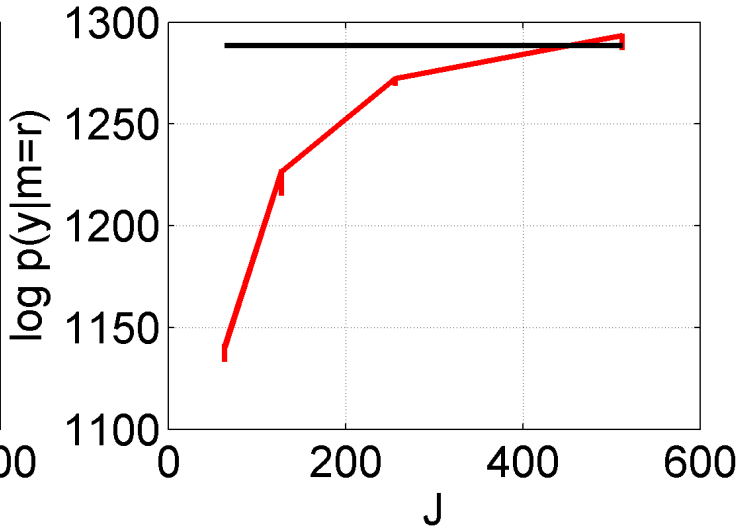


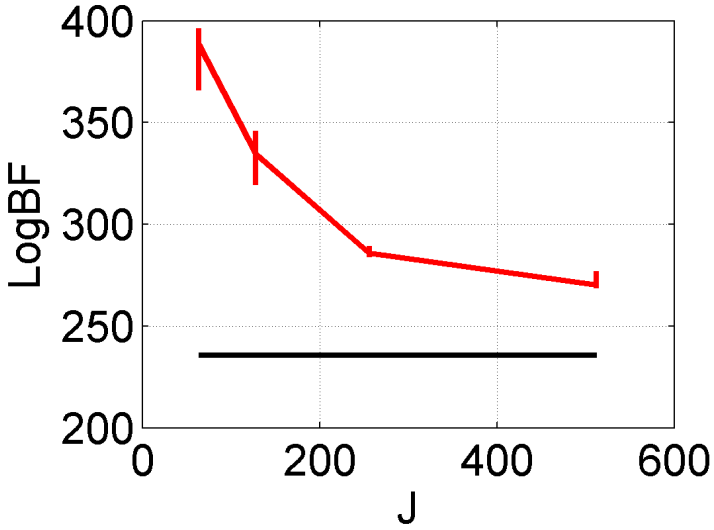
True parameters are all zero except first two.

Model Evidence

Vary resolution of annealing schedule







AIS (red), VL (black)

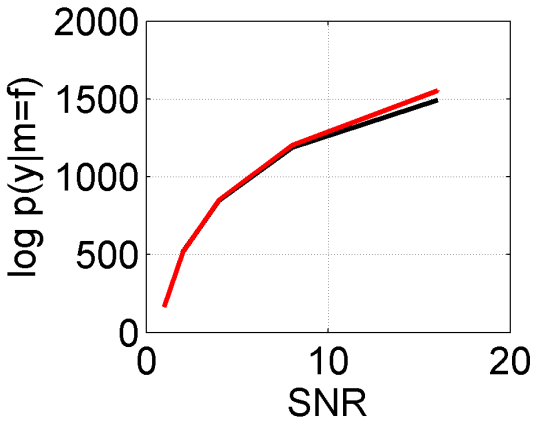
Evidence, Bayes Factors, Compute Time

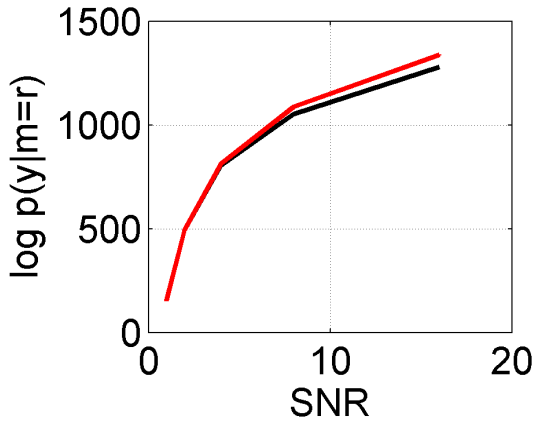
Model	Estimate		Time(s)	
	VL	AIS	VL	AIS
Linear, LogEv, Full	-11.02	-11.00	0.005	15.4
Linear, LogEv, Red	-23.97	-23.94	0.002	3.1
Linear, LogBF	12.95	12.94	-	-
Approach, LogEv, Full	-73.88	-73.77	0.58	19.4
Approach, LogEv, Red	-783.62	-783.61	0.02	2.9
Approach, LogBF	709.74	709.84	-	-
Neural Mass, LogEv, Full	1524.1	1563.6	22	5290
Neural Mass, LogEv, Red	1288.4	1293.4	24	4610
Neural Mass, LogBF	235.74	270.2	-	-

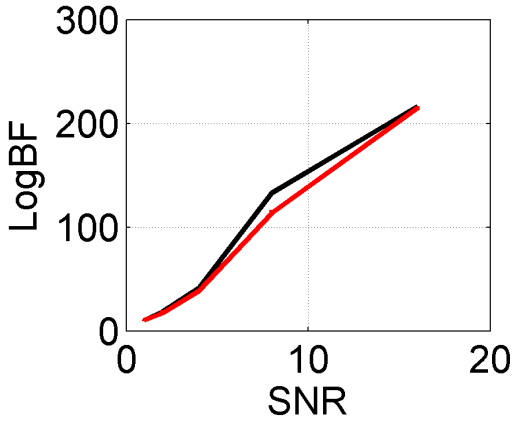
AIS estimates from $I = 32$ samples and $J = 512$ trajectories.
The linear model VL results are for analytic solution.

Effect of SNR

True model has full connectivity. AIS (red), VL (black).



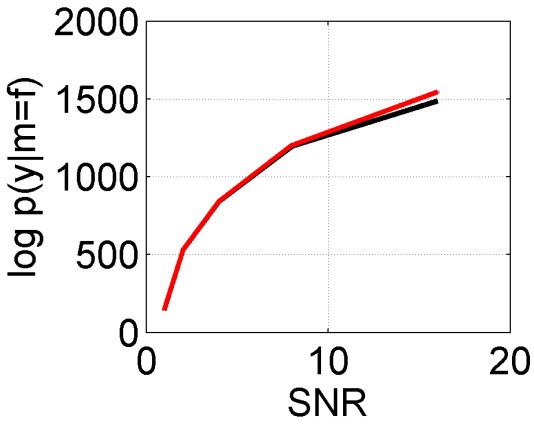


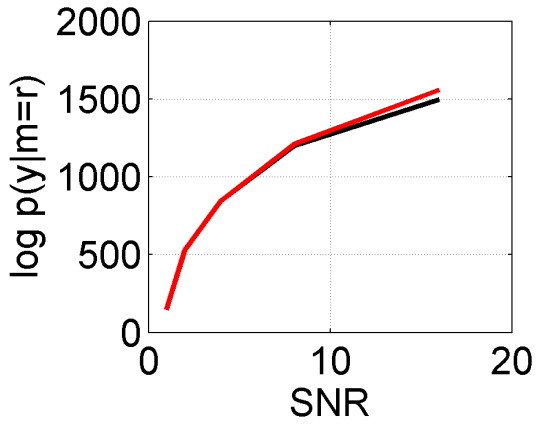


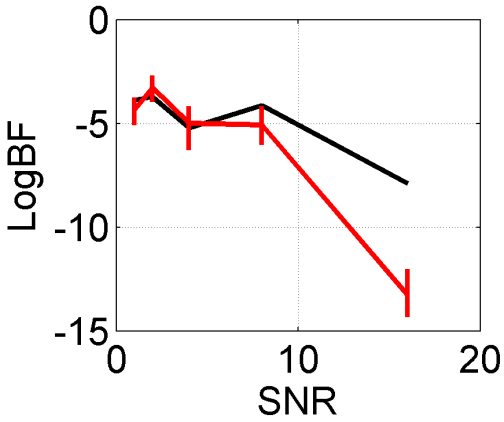
AIS and VL are always in agreement in favouring the true model.

Effect of SNR

True model has reduced connectivity. AIS (red), VL (black).







AIS and VL are always in agreement in favouring the true model.

Gaussianity

Table: *p-values from Royston's Gaussianity test applied to AIS samples from 'Full' NMM.*

SNR	Full	Reduced
1	0.02	0.18
2	0.40	0.10
4	0.80	0.54
8	0.33	0.51
16	0.40	0.17

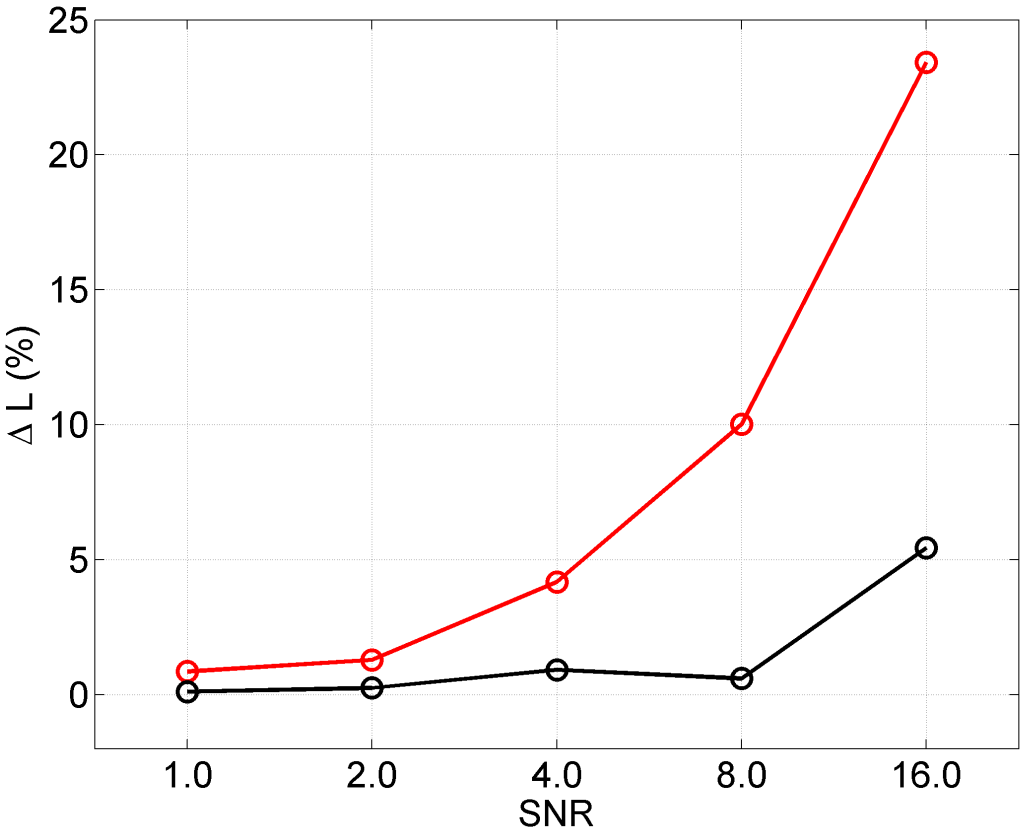
Gaussianity

Table: *p-values from Royston's Gaussianity test applied to AIS samples from 'Reduced' NMM.*

SNR	Full	Reduced
1	0.04	0.60
2	0.86	0.87
4	0.40	0.72
8	0.02	0.52
16	0.11	0.15

AIS versus Multistart VL

Baseline log joint, L, is from single default VL (start from prior mean). We are then plotting percentage improvement in this.



AIS (red) finds better parameters than best Multistart VL (black). They are matched for computer time.

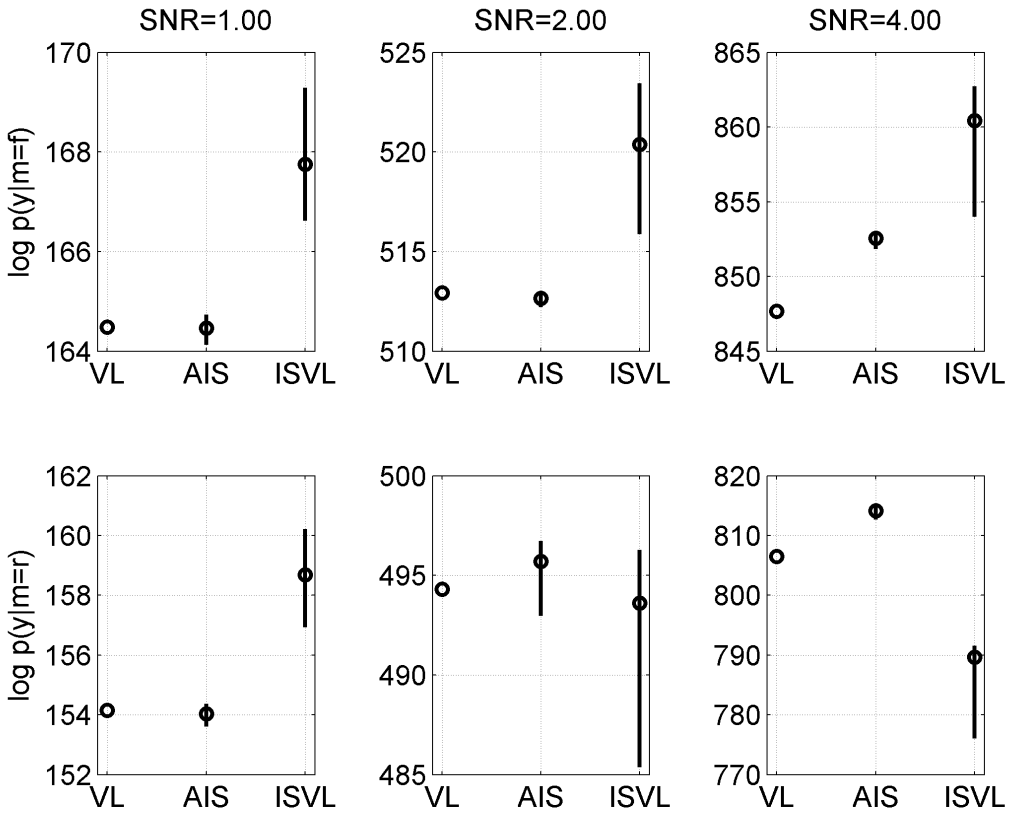
Summary

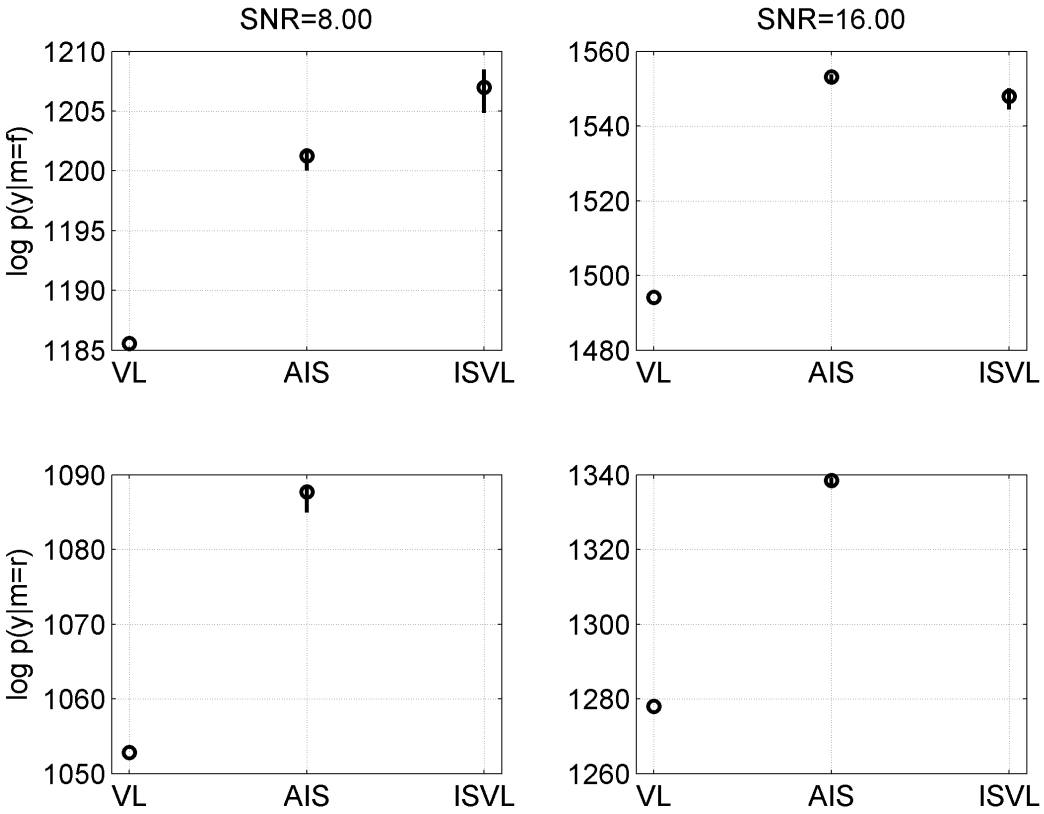
- LMC explores local parameter space using gradients and curvatures
- Embed this in AIS for global Bayesian optimisation
- Better parameter estimates than Multistart VL
- AIS provides an estimate of the model evidence
- PAM and PHM are special cases of AIS
- Test parametric assumptions of VL

But its slow (about 80 mins per Neural Mass Model)

- Anneal from posterior of full model to posterior of reduced (or other) model to compute Bayes Factor
- Automatically tune annealing schedules whilst preserving parallelisation

wellcome trust


SNR=1.00
SNR=2.00
SNR=4.00
log p(y|m=f)
log p(y|m=r)
VL
AIS
ISVL
170
168
166
164
525
520
515
510
865
860
855
850
845
162
160
158
156
154
152
500
495
490
485
820
810
800
790
780
770


SNR=8.00
SNR=16.00
log p(y|m=f)
log p(y|m=r)
1210
1205
1200
1195
1190
1185
1560
1540
1520
1500
1480
1090
1080
1070
1060
1050
1340
1320
1300
1280
1260
VL
AIS
ISVL

Reverse Annealing

By inverting the equation for the model evidence we have

$$\begin{aligned}\frac{1}{p(y|m)} &= \frac{Z_1}{Z_J} \\ &= \frac{Z_{J-1}}{Z_J} \cdots \frac{Z_2}{Z_3} \cdots \frac{Z_1}{Z_2} \\ &= \prod_{j=1}^{J-1} \frac{1}{r_j}\end{aligned}$$

Importance weights for reverse annealing are given by

$$v^{(i)} = \frac{f_{J-1}(w_{J-1})}{f_J(w_{J-1})} \cdots \frac{f_2(w_2)}{f_3(w_2)} \frac{f_1(w_1)}{f_2(w_1)}$$

and a series of samples $w_J, w_{J-1}, \ldots w_2, w_1$ are created by starting with w_J from forward annealing, and generating the others sequentially using LMC.

Reverse Annealing

For $J = 2$ temperatures $\beta_2 = 1$, $\beta_1 = 0$ we get

$$\frac{1}{p(y|m)} = \frac{1}{p(y|w,m)}$$

Averaging over multiple trajectories gives

$$\frac{1}{p(y|m)} = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{p(y|w_i,m)}$$

which shows that the PHM approximation to the model evidence is a special case of AIS with a reverse annealing schedule and only $J = 2$ temperatures.