PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Annealed Importance Sampling for Neural Mass Models

Will Penny*, **Biswa Sengupta**

Wellcome Trust Centre for Neuroimaging, University College London, London, United Kingdom

* w.penny@ucl.ac.uk

OPEN ACCESS

Citation: Penny W, Sengupta B (2016) Annealed Importance Sampling for Neural Mass Models. PLoS Comput Biol 12(3): e1004797. doi:10.1371/journal.pcbi.1004797

Editor: Jean Daunizeau, Brain and Spine Institute (ICM), FRANCE

Received: September 28, 2015

Accepted: February 5, 2016

Published: March 4, 2016

Data Availability Statement: The data used in this study are generated from mathematical models which are available from http://www.fil.ion.ucl.ac.uk/spm/download/spm12_updates/.

Funding: This work was supported by two grants from the Wellcome Trust (www.wellcome.ac.uk). WP was funded by grant number 091593/Z/10/Z, and BS was funded by grant number 088130/Z/09/Z. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Neural Mass Models provide a compact description of the dynamical activity of cell populations in neocortical regions. Moreover, models of regional activity can be connected together into networks, and inferences made about the strength of connections, using M/EEG data and Bayesian inference. To date, however, Bayesian methods have been largely restricted to the Variational Laplace (VL) algorithm which assumes that the posterior distribution is Gaussian and finds model parameters that are only locally optimal. This paper explores the use of Annealed Importance Sampling (AIS) to address these restrictions. We implement AIS using proposals derived from Langevin Monte Carlo (LMC) which uses local gradient and curvature information for efficient exploration of parameter space. In terms of the estimation of Bayes factors, VL and AIS agree about which model is best but report different degrees of belief. Additionally, AIS finds better model parameters and we find evidence of non-Gaussianity in their posterior distribution.

Author Summary

The activity of populations of neurons in the human brain can be described using a set of differential equations known as a neural mass model. These models can then be connected to describe activity in multiple brain regions and, by fitting them to human brain imaging data, statistical inferences can be made about changes in macroscopic connectivity among brain regions. For example, the strength of a connection from one region to another may be more strongly engaged in a particular patient population or during a specific cognitive task. Current statistical inference approaches use a Bayesian algorithm based on principles of local optimization and the assumption that uncertainty about model parameters (e.g. connectivity), having seen the data, follows a Gaussian distribution. This paper evaluates current methods against a global Bayesian optimization algorithm and finds that the two approaches (local/global) agree about which model is best, but finds that the global approach produces better parameter estimates.

Introduction

Dynamical systems models instantiated using differential equations are a mainstay of modern neuroscience and provide mathematical descriptions of neuronal activity over multiple spatial and temporal scales [1, 2]. In imaging neuroscience a widely adopted framework, called Dynamic Causal Modelling (DCM), has been developed for fitting such models to brain imaging data using a Bayesian approach [3]. This allows inferences to be made about changes in parameters (eg. effective connectivity) in the human brain using noninvasive imaging data. There is now a library of DCMs which differ according to their level of biological realism and the data features they explain. DCM can be applied to fMRI [3], EEG and MEG [4] and invasive electrophysiological data [5].

The Bayesian approach to model fitting in DCM is based on the Variational Laplace (VL) algorithm [6]. One of its core assumptions, the 'Laplace Assumption', is that the posterior distribution is Gaussian. This assumption is typically instantiated by finding the maximum posterior parameter vector, using numerical optimisation, and making a Taylor expansion around this value and retaining terms up to second order [7]. It has been found to be more robust than higher-order moment expansions on empirical data [8]. In VL, the posterior is assumed to factorise into a product of probability distributions, one over latent variables controlling noise variances and one over model parameters. Each distribution is multivariate Gaussian with mean and covariance that are iteratively updated to maximise an approximation to the model evidence [6].

The Laplace approximation is attractive because it provides a computationally simple method for both quantifying posterior uncertainty in model parameters and approximating the model evidence for Bayesian model comparison.

A theoretical motivation for the the Laplace approximation is that the posterior will tend to a Gaussian in the limit where the number of data points goes to infinity [9]. But as previously noted in the context of DCM [10], it is questionable as to whether posteriors are Gaussian for datasets that are encountered in practice which naturally have a finite number of data points. The VL algorithm has two potential weaknesses (i) as with any local optimisation method working in a non-convex domain [11] it may fall into a local maxima and (ii) the distribution around the maxima may be non-Gaussian.

In this paper we compare VL to Monte Carlo methods in the challenging context of identifying Neural Mass Models (NMMs) [12]. The advantage of Monte Carlo methods is that, provided the sampling process runs for a sufficiently long time, the samples converge in distribution to the exact posterior. This obviates the need for Gaussian assumptions but at the cost of potentially very long sampling times. To address these issues this paper uses the Annealed Importance Sampling (AIS) algorithm [13] with proposals made using a Langevin Monte Carlo (LMC) procedure [14]. The use of AIS has two benefits (i) it can accomodate multiple local maxima and (ii) it provides an estimate of the Bayesian model evidence. The use of LMC improves convergence properties because proposals are made using local gradient and curvature information [14, 15].

Previously, the Metropolis-Hastings (MH) algorithm has been used to validate VL in the context of DCM for fMRI [16]. Whilst these findings are largely consistent with the Laplace assumption this study is incomplete in a number of respects (i) only results from a single Markov chain were reported thus raising the possibility that a local maxima was found, (ii) no sample-based estimate of the model evidence was provided, and (iii) the neurodynamical models used in fMRI are based on linear dynamical systems, so this finding may not hold for the nonlinear dynamical models [17] underlying other DCMs such as those for M/EEG data.

This paper assesses how well the two Bayesian estimation algorithms (AIS-LMC and VL) perform inference for NMMs. These models have been chosen as they are highly nonlinear and

underlie the first proposed DCM for M/EEG data [17]. In order to validate our software implementation and fine tune parameters of the AIS algorithm, we additionally evaluate these algorithms in the simpler context of linear and nonlinear regression models.

Materials and Methods

In what follows $\mathcal{N}(x; m, \Lambda)$ denotes a multivariate Gaussian variable x with mean m and precision Λ. We consider Bayesian inference for data Y, or y, models with parameters w, priors $p(w)$ and likelihoods $p(Y|w)$ or $p(y|w)$. All models in this paper use Gaussian priors with mean μ and precision Λ. In the subsections that follow we describe the AIS algorithm and show how LMC can be used within it to provide proposals. We then describe the linear regression, nonlinear regression and neural mass models that we will use to test the inference methods. To provide a convenient reference for some of the underlying concepts we provide supplementary material on Importance Sampling S1 Text, Fisher Information S2 Text, Neural Mass Models S3 Text, Variational Laplace S4 Text and Chib's method for estimating model evidence S5 Text.

Annealed Importance Sampling

Annealed Importance Sampling (AIS) [13] provides samples from a posterior density using a sequence of densities at a series of monotonically increasing inverse temperatures β_j with $j = 0..J$, $\beta_0 = 0$ and $\beta_J = 1$. For the jth temperature the algorithm produces a sample from the unnormalised density

$$f_j(w) = p(y|w)^{\beta_j} p(w) \tag{1}$$

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'. The transition densities T_j can be chosen in any of the usual ways for constructing Markov chains [18] and may themselves involve several steps. The only requirement is that T_j is chosen to leave f_j as the invariant distribution. For example, for a simple density estimation problem, Neal [13] specified each T_j to be a sequence of Metropolis moves each defined using an isotropic Gaussian proposal with increasing width. For a linear regression problem with non-Gaussian priors he employed a Hamiltonian Monte-Carlo (HMC) approach [19]. In this paper we will use Langevin Monte Carlo (LMC), as recent work shows this to provide higher effective sample size per unit of computation time as compared to HMC [15].

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' [20]. Specifically, trajectories can be assigned to individual computer processors or processor cores thus greatly speeding up the implementation.

Each sample is also accompanied by an importance weight

$$v^{(i)} = \frac{f_1(w_1)}{f_0(w_1)}\frac{f_2(w_2)}{f_1(w_2)}\frac{f_3(w_3)}{f_2(w_3)}\cdots\frac{f_J(w_J)}{f_{J-1}(w_J)} \tag{2}$$

which can be evaluated as

$$\log v^{(i)} = \sum_{j=1}^{J}\left(\beta_j - \beta_{j-1}\right)\log p(y|w_j) \tag{3}$$

To avoid numerical overflow we first create adjusted weights u_i

$$\begin{aligned} v_{max} &= \max(\log v) \\ u_i &= \exp(\log v^{(i)} - v_{max}) \end{aligned} \tag{4}$$

and let $\bar{u}$ be the mean adjusted weight. The normalised importance weights are

$$q_i = \frac{u_i}{\sum_i u_i} \tag{5}$$

A derivation of the formula for the importance weights is provided in [13] and included in S1 Text. The variance of the importance weights is an indicator of the quality of the approximation to the posterior density [13].

Annealing schedule. An important choice in any AIS implementation is the annealing schedule, that is, how to space the β_j over the (0, 1) interval. Calderhead and Girolami [21] show that, for estimates of the model evidence for linear regression models, the annealing schedule that minimises the Monte Carlo variance has a power-law form. Following [21, 22] the applications in this paper use a 5th-order geometric annealing schedule

$$\beta_j = \left(\frac{j}{J}\right)^5 \tag{6}$$

Additionally, one must choose the number of trajectories, and number of temperatures per trajectory. In the original AIS paper [13] $I = 1000$ trajectories were used with either $J = 200$ or 1000 temperatures. The AIS algorithm has also been compared to a Variational Bayes (VB) approach for scoring graphical models [23]. This implementation used only $I = 5$ trajectories with $J = 16{,}384$ temperatures. Proposals were made using a standard MH step which is perhaps one reason for the very large number of temperatures required. Only with $J > 5000$ temperatures did the AIS model evidence estimate exceed that produced by VB (which provides a provable lower bound [7]). In an application of AIS to score differential equation models [24], $I = 10$ trajectories with $J = 40$ temperatures were used along a 4th order geometric schedule, with a transition kernel implemented using an MH step with 4000 samples at each temperature. Because LMC provides better proposals than MH we envisage that a finer grained schedule can be used at similar computational expense. This will be examined in the results section in the context of linear and nonlinear regression models.

Model evidence. The importance weight, or the average importance weight across multiple trajectories, provides an approximation to the model evidence $p(y|m)$ for model m, as shown below. This section uses the notation $p(y|w, m)$ and $p(w|m)$ to make it explicit that the likelihood and prior depend on model assumptions. We define the normalising constant at

each temperature as

$$\begin{aligned} Z_j &= \int f_j(w)dw \\ &= \int p(y|w,m)^{\beta_j} p(w|m)dw \end{aligned} \tag{7}$$

We then have

$$\begin{aligned} Z_0 &= \int p(w|m)dw = 1 \\ Z_J &= \int p(y|w,m)p(w|m)dw \\ &= p(y|m) \end{aligned} \tag{8}$$

Therefore

$$\begin{aligned} p(y) &= \frac{Z_J}{Z_0} \\ &= \frac{Z_1}{Z_0}\frac{Z_2}{Z_1}\frac{Z_3}{Z_2}\cdots\frac{Z_J}{Z_{J-1}} \\ &= \prod_{j=0}^{J-1} r_j \end{aligned} \tag{9}$$

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

$$\begin{aligned} r_j &= \frac{1}{Z_j}\int f_{j+1}(w)dw \\ &= \int \frac{f_{j+1}(w)}{f_j(w)}\frac{f_j(w)}{Z_j}dw \\ &\approx \frac{1}{N}\sum_{n=1}^{N}\frac{f_{j+1}(w_n)}{f_j(w_n)} \end{aligned} \tag{10}$$

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

$$r_j = \frac{1}{N}\sum_{n=1}^{N} p(y|w_n,m)^{\beta_{j+1}-\beta_j} \tag{11}$$

For $N = 1$ we can therefore see that $\log p(y)$ is equal to Eq 3. To avoid numerical overflow we compute the log evidence as

$$\log p(y|m)_{AIS} = v_{max} + \log \bar{u} \tag{12}$$

We can now see that estimation of the model evidence using the Prior Arithmetic Mean (PAM) (see S1 Text), in which the average likelihood is computed over samples drawn from the prior, is a special case of the AIS estimate with just two temperatures, $\beta_1 = 1$ and $\beta_0 = 0$. It is also possible to define a reverse annealing schedule in which the temperature is gradually increased and defines a path from the posterior to the prior [13]. Agreement between forward and reverse estimates of the model evidence can then be used to ensure one has a sufficiently fine-grained annealing schedule [23]. For reverse schedules the Posterior Harmonic Mean (PHM) emerges as a special case of AIS with two temperatures (see S1 Text). AIS therefore generalises both PAM and PHM. In high dimensional spaces PAM underestimates the model evidence because it doesn't sufficiently explore regions of high probability, whereas PHM

overestimates it because it doesn't sufficiently explore regions of low probability. These problems are ameliorated in AIS by the use of intermediate densities that form 'bridges' as described in a related method called bridge sampling [25].

In this paper our empirical results are based on forward annealing schedules only. Confidence intervals in model evidence estimates are provided using bootstrapping [26], by resampling the I estimates $N_{boot} = 1000$ times with replacement, computing the evidence for each, and finding the 5th and 95th percentiles. Thus, bootstrapping is implemented over trajectories.

Langevin Monte Carlo

In this paper the transition densities T_j in AIS are implemented using a Langevin Monte Carlo (LMC) sampler, which leads to proposals being accepted with high probability even for nonlinear and high dimensional inference problems, as it uses information about the gradient and curvature of the unnormalised density, f_j.

The use of LMC follows from the definition of the log joint and its gradient as a function of w

$$\begin{aligned} L(w) &= \log p(y|w) + \log p(w|\mu, \Lambda) \\ g(w) &= \frac{dL(w)}{dw} \end{aligned} \tag{13}$$

A proposal is drawn as

$$\begin{aligned} w_s^* &\sim p(w_s^*|w_s) \\ p(w_s^*|w_s) &= \mathcal{N}(w_s^*; m, C) \\ m &= w_s + \frac{1}{2} C g(w_s) \\ C &= h^2(\Lambda + F)^{-1} \end{aligned} \tag{14}$$

where Λ is the prior precision, w_s is the sth sample, and h is a step size parameter (fixed at 0.5 for all applications in this paper). The quantity F is the Fisher Information matrix (see S2 Text) and quantifies the precision of the parameters conferred by the data. This has analytic forms for many probabilistic models such as logistic regression [14] and is readily computed for differential equation models using an approach based on forward sensitivity analysis [27, 28].

The Metropolis-Hastings (MH) criterion is then applied to accept proposals with probability

$$r = \frac{p_w(w_s^*)}{p_w(w_s)} \frac{p(w_s|w_s^*)}{p(w_s^*|w_s)} \tag{15}$$

where $p_w(w_s) = \exp[L(w_s)]$. The proposal is always accepted if $r > 1$. We set $w_{s+1} = w_s^*$ if the sample is accepted and $w_{s+1} = w_s$ if it is rejected.

The above proposal (Eq 14) has the same functional form as the Simplified Manifold MALA algorithm as applied to ODEs [14, 27]. Here the 'manifold' is defined by C and m and its computation has been 'simplified' as the curvature has been assumed to be locally constant. For Gaussian likelihoods, this same local linearity assumption is also the basis of the Gauss-Newton optimization algorithm [29].

In the usual application of LMC [14, 15], Eqs 14 and 15, are repeatedly applied until one obtains samples from the posterior density. However, in this paper we use LMC to provide a single sample at each temperature in an AIS trajectory. Specifically, the transition kernel, $T_{j-1}(w_j|w_{j-1})$, starts at $w_s = w_{j-1}$ and produces $w_j = w_s^*$ using the modified log joint L_{j-1}. This

modification requires multiplication of the likelihood, gradient and Fisher information by β_{j-1}. The LMC updates are otherwise identical. Because LMC is used to produce only a single sample at each temperature the total number of LMC steps is equal to the number of temperatures.

We now briefly comment on the computational scalability of the combined AIS-LMC algorithm. Because AIS is based on importance sampling its accuracy is proportional to the number of annealing runs ("trajectories") [13]. As trajectories are independent, and can be assigned to cores on multiple core computer architectures, the accuracy will therefore scale with the number of cores (at almost no increase in computer time). For a fixed number of cores computer time scales linearly with the number of trajectories. The computational bottleneck within each AIS trajectory is the evaluation of the gradient of the log joint and the Fisher information, required for each LMC step. These quantities can be efficiently computed for ODE models using forward sensitivity or adjoint methods [27, 28]. The computation time of these methods scales linearly with the length of time series being modelled, and adjoint methods are typically more efficient than forward sensitivity methods if the number of parameters is much larger than the number of dynamical states.

Linear Regression

In multiple linear regression an $[N \times 1]$ data vector y is generated as

$$y = X\beta + e \tag{16}$$

where X is an $[N \times p]$ design matrix, β is a $[p \times 1]$ vector of regression coefficients, and e is an $[N \times 1]$ zero-mean IID Gaussian noise vector with entries having variance σ^2.

Nonlinear Regression

To provide a simple nonlinear model with multiple maxima, we consider a regression model where the parameters of interest are nonlinearly related to the regression coefficients

$$\begin{aligned} y &= \sum_i x_i \beta_i + e \\ \beta_i &= w_i^2 \end{aligned} \tag{17}$$

This model will have multiple maxima over the various combinations of positive and negative values of w_i.

We also consider an exponential approach-to-limit or 'approach' model where

$$y(t) = -60 + V_a[1 - \exp(-t/\tau)] + e(t) \tag{18}$$

with parameters $w_1 = \log \tau$ and $w_2 = \log V_a$. This models the ramping up of a voltage from −60 to $-60 + V_a$ with a time constant τ, and has the same mathematical form as Biochemical Oxygen Demand (BOD) models [30] previously used to evaluate Bayesian inference methods [31].

Neural Mass Models

Single region. In Neural Mass Models (NMMs) [17], postsynaptic potentials (PSPs) at excitatory synapses are related to firing rates via convolutions with synaptic kernels

$$v_{out}(t) = h_e(t) \otimes s(v_{in}) \tag{19}$$

where the population firing rate function

$$s(x) = \frac{1}{1 + \exp(-r_1(x - r_2))} - \frac{1}{1 + \exp(r_1 r_2)} \tag{20}$$

has parameters r_1 and r_2, and the synaptic kernel is given by an alpha function

$$h_e(t) = \frac{H_e}{\tau_e} t \exp(-t/\tau_e) \tag{21}$$

with magnitude H_e and time constant τ_e. Inhibitory synapses are similarly defined but with kernels $h_i(t)$ and parameters H_i, τ_i.

The activity of a single neocortical unit is then defined by the convolution equations

$$\begin{aligned} v_i &= \gamma_3 s(\tilde{v}_p) \otimes h_e \\ v_s &= \left[s(u) + \gamma_1 s(\tilde{v}_p) \right] \otimes h_e \\ v_{pe} &= \gamma_2 s(\tilde{v}_s) \otimes h_e \\ v_{pi} &= \gamma_4 s(\tilde{v}_i) \otimes h_i \\ v_p &= v_{pe} - v_{pi} \end{aligned} \tag{22}$$

where v_{pe} and v_{pi} are potentials at excitatory and inhibitory synapses in the pyramidal cell population, $\tilde{v}$ denotes the potential after a delay δ_{ii} due to signalling delays among the different populations within a single brain region. Following [17] a first order Taylor series approximation is used to capture these delays, $\tilde{v} = v - \delta_{ii}\dot{v}$. The connection strengths among neural populations are specified by the parameters $\gamma_{1..4}$. These within-region values are also referred to as the 'intrinsic connectivity'.

Each of the above convolution equations can be written as a second order differential equation, or two first order DEs, as shown in [12] (see also S3 Text). Thus a single cortical unit has $N_x = 9$ state variables. The input to the cortical region, u, is a surrogate for event-related subcortical brain activity and is specified by a Gaussian function peaking at 64ms post-stimulus with width 16ms.

Two region model. David et al. [17] describe how cortical units can be connected into hierarchical networks that follow known anatomical connectivity patterns [32]. A two region network with forward connection a_{21} (from region 1 to 2) and backward connection a_{12} is shown in Fig 1. The convolution equations for this network are given in S3 Text.

There are two between-region or 'extrinsic' connectivity parameters (a_{12}, a_{21}) and two extrinsic delay parameters (δ_{12} and δ_{21}). Additionally, we have four 'intrinsic' connectivity

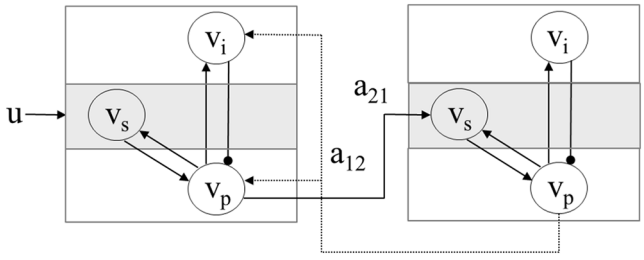

Fig 1. Neural mass model of two cortical regions in a hierarchical network. The first unit receives thalamic input u, and projects output $v_p(1)$ via a forward connection of strength a_{21} to region 2. The second unit produces output $v_p(2)$ and projects it via a backward connection of strength a_{12} to region 1.

doi:10.1371/journal.pcbi.1004797.g001

parameters ($\gamma_{1..4}$) and parameters of firing rate functions (r_1 and r_2) that are constrained to be identical in each region. This gives a total of $N_p = 10$ neurophysiological variables to estimate.

The intrinsic delay parameters (δ_{11}, δ_{22}—one for each region) are assumed known. The synaptic time constants (τ_e, τ_i) and synaptic response magnitudes (H_e, H_i) are fixed to be the same for all regions, and are also assumed known. This two region neural mass model has $N_x = 18$ state variables. The differential equations are integrated to produce time series of currents and potentials for each population in each cortical unit, at N_t time points. The resulting 'neuronal state matrix' X is of dimension $[N_x \times N_t]$. The generative model is then specified as

$$Y = L_2 X + e \tag{23}$$

where L_2 is a matrix that picks off the pyramidal cell activities in each of the regions, and e is zero mean Gaussian noise. For the simulations in this paper Y is therefore a $[2 \times N_t]$ data matrix containing the pyramidal cell activities of each of the brain regions. In applications to empirical M/EEG data [17, 33] an $[N_d \times N_x]$ lead field matrix L is used to model Event-Related Potentials (ERPs) at N_d sensors.

We assume that the noise variance on the sth output (where $s = 1..2$) is σ_s^2. The model likelihood is therefore

$$p(Y|w) = \prod_{t=1}^{T} \mathsf{N}(y_t; \hat{y}_t, C_e^{-1}) \tag{24}$$

where w are the parameters, $\hat{y}_t = L_2 x_t$ and $C_e = \mathsf{diag}(\sigma_s^2)$. The unknown neurophysiological variables are related to model parameters according to the transformations shown in S1 Table which enforce positivity and constrain parameters within a physiologically plausible range. The Gaussian prior over model parameters has zero mean μ, and Λ^{-1} is a diagonal matrix with entries of 0.16 for the first two parameters (a_{12} and a_{21}) and 0.0625 for the rest. The above choice of parameter transformation and prior are the same as that used in DCM for ERP [33].

Testing for Normality

As the VL algorithm assumes that the posterior distribution is Gaussian it will be interesting to see if this is indeed the case. We use Royston's test for multivariate normality [34] using a Matlab implementation by Trujillo-Ortiz et al [35]. This is a multivariate extension of the Shapiro-Wilks test and we apply it to Monte Carlo samples from the posterior densities produced by AIS. As these samples are independent there is no need for 'thinning' or assessments of Effective Sample Size [36].

Software

The algorithms on which this research is based have been implemented in Matlab in the 'Monte Carlo Inference (MCI)' toolbox and will be distributed as part of a forthcoming release of the Statistical Parametric Mapping (SPM) package. AIS and LMC, for example, are implemented in the spm_mci_ais.m and spm_mci_lgv.m functions available in the subdirectory /toolbox/mci/inference/.

Variational Laplace

The Variational Laplace (VL) algorithm is instantiated in the SPM software [33] (in the function spm_nlsi_GN.m) and described elsewhere [6, 37]. We also include a brief mathematical description in S4 Text. In VL, the posterior is assumed to factorise into a product of probability distributions, one over latent variables controlling noise variances and one over model

parameters. Each distribution is multivariate Gaussian with mean and covariance that are iteratively updated to maximise an approximation to the model evidence [6]. Importantly, the multivariate nature of each Gaussian allows parameter dependencies to be accommodated. This optimiser is the standard approach used for the majority of DCM applications in neuroimaging. Known noise variances (see below) are implemented for the VL algorithm by setting the prior over the log noise precision to have a mean corresponding to the true (known) value, and a variance of 10^{-8} (i.e. very tight).

By default, the implementation of VL in SPM initialises parameters at the prior mean. A simple way of potentially handling optimisation problems with multiple maxima, however, is to run the VL algorithm multiple times where each run is initialised using a different sample from the prior. We will refer to this procedure as Multistart VL.

Results

We present results on linear and nonlinear regression models to demonstrate the effect of the number of temperatures J and trajectories I in AIS. The algorithms were run on a high-end desktop computer (Hewlett Packard Z440) with 32G memory, 8 cores, and a 64-bit operating system. All the results are derived from synthetic data for which the ground truth parameters are known. Following other recent comparisons of inference algorithms for differential equation models [15, 21, 38, 39], our simulations assume that the noise variances are known for models with Gaussian likelihoods.

All AIS results were produced using a fifth order geometric annealing schedule and the posterior mean was computed using the mean over trajectories. The AIS implementation was parallelized using the Matlab Parallel Computing toolbox such that independent 'pool workers' (in this case cores) were assigned to different trajectories. The distribution of normalised importance weights, u_i, is characterised in two ways. Firstly, by the entropy. For I trajectories the maximum entropy is $\log_2 I$ e.g. 5 bits for $I = 32$. Secondly, by the number of significantly non-zero values, I_q, which we define as the number above 0.01.

Linear Regression

We first provide results on a multiple linear regression model, as there are analytic formulae for the posterior distribution and model evidence [7], and the Laplace approximation is exact. This comprised $p = 7$ regressors chosen from a discrete cosine basis set over $N = 20$ 'time points', with additive noise of standard deviation $\sigma = 0.2$. The prior variances, Λ_{pp}^{-1} were set to 10 for each regressor and the prior means, μ_p to zero. The regression coefficients were drawn from the prior.

The AIS algorithm was applied to this data using $J = 512$ temperatures and $I = 32$ independent samples. We fitted the true model (with 7 regressors) and a reduced model to the same data but this time using only the first 6 regressors.

Using the 32 samples produced by AIS, we could not reject the hypothesis that the posterior was Gaussian using Royston's test for the full ($p = 0.67$) and reduced ($p = 0.68$) models. This is of course to be expected as the posterior distribution is indeed Gaussian for linear regression models [7]. For the full model, the normalised importance weights had high entropy, $H = 4.07$, and many trajectories had significant weight, $I_q = 21$.

The AIS estimates of the log model evidences for the full, $\log p(y|m = f)$, and reduced models, $\log p(y|m = r)$ and the corresponding log Bayes factor, and computation times, are provided in Table 1. The estimates very closely match the analytic values. Note that the VL estimates correspond to the analytic values for the case of linear regression [7]. We then re-estimated the evidences using different numbers of AIS samples and temperatures, with results plotted in

Table 1. Evidence and Bayes Factor Approximations (Single Run).

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	-	-

These results are for a single run of each inference algorithm (AIS or VL). AIS estimates from $I = 32$ samples and $J = 512$ trajectories. The results for the linear model here* are for the analytic solution, which also corresponds to the VL solution.

doi:10.1371/journal.pcbi.1004797.t001

Fig 2. Theses results show good agreement with analytic values for $J = 128$ and above. The error bars on AIS model evidence estimates were computed using bootstrapping (over trajectories) as described in the section on 'Annealed Importance Sampling', in the subsection on 'Model Evidence'.

Nonlinear Regression

Multiple maxima. We now report results for the nonlinear regression model that was designed to have multiple maxima. This model has two independent variables, x_1 and x_2,

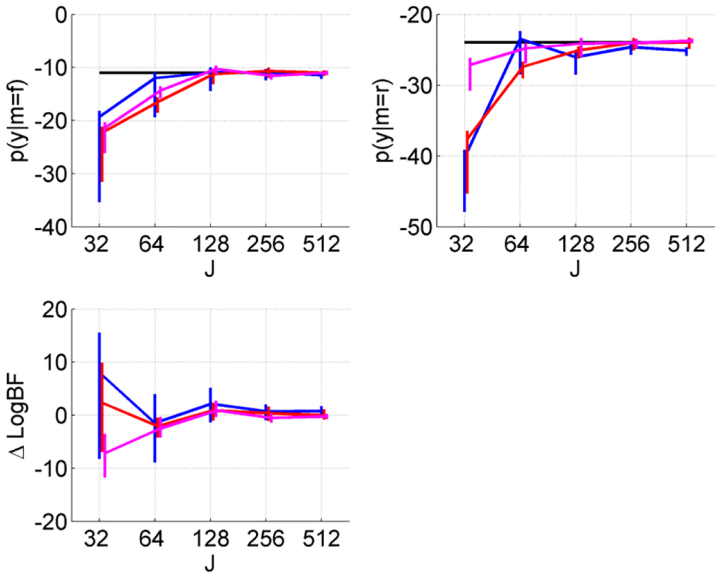

Fig 2. Linear regression. AIS approximations of log evidence for a 'full' model with 7 parameters (top left), and a 'reduced' model with 6 parameters (top right) as a function of number of temperatures J. These approximations use $I = 16$ (blue), $I = 32$ (red) and $I = 64$ (magenta) trajectories. The black lines show the equivalent analytic quantities. The bottom left plot shows the difference between the AIS estimated log Bayes factor and the true value. Vertical lines span the 5th and 95th percentiles from bootstrapping.

doi:10.1371/journal.pcbi.1004797.g002

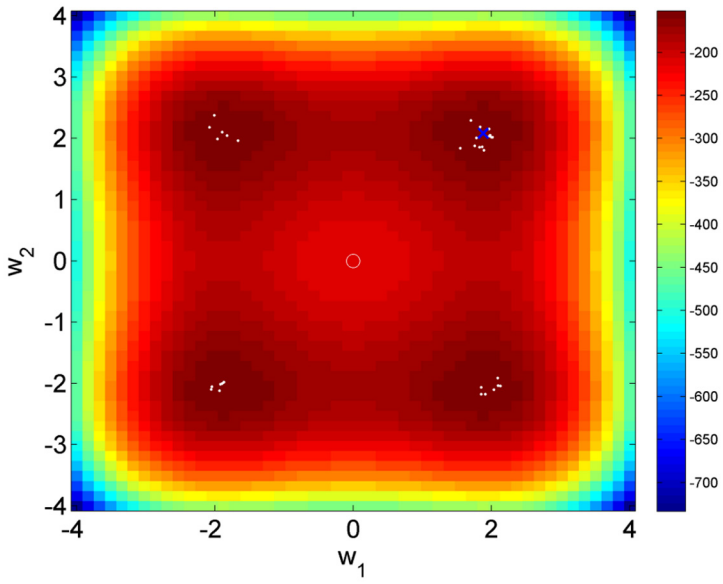

Fig 3. Log posterior of nonlinear regression model with multiple maxima. The true parameters are $w_1 = w_2 = 2$. The circle denotes the prior mean. Samples from the posterior density as computed using AIS are shown as white dots (in each of the four maxima), and the blue cross close to the true parameters denotes the VL posterior mean.

doi:10.1371/journal.pcbi.1004797.g003

corresponding to two components of a discrete cosine basis set (the first two from the linear regression problem described above). The priors were set to be the same as for the linear regression simulation but the observation noise was increased to $\sigma = 0.5$. AIS was run using the same parameters as before and Fig 3 shows samples from the posterior density which lie in all of the four posterior modes. The normalised importance weights had lower entropy than for the linear regression model, $H = 3.42$, and fewer trajectories with significant weight, $I_q = 16$. We can reject the hypothesis that the posterior is Gaussian using Royston's test ($p = 10^{-12}$). Thus AIS is able to accomodate multiple maxima as expected, and we correctly infer that the posterior is non-Gaussian. AIS is able to find the different maxima by virtue of employing multiple trajectories. Fig 3 also shows the posterior mean for VL. We also ran Multistart VL (see section on Variational Laplace) with 32 starts and, as expected, it was also able to identify each of the four maxima. The posterior distribution for this example is multimodal and is therefore not well represented by the posterior mean. The AIS samples do, however, collectively provide a good description of the posterior distribution.

Approach to limit. We now report results for the approach-to-limit model. Data were generated with parameters $V_a = 30$, $\tau = 8$ and Gaussian observation noise variance of unity. The prior has mean $\mu = [3, 1.6]^T$ and precision $\Lambda = \mathrm{diag}([16, 16])$. A 'reduced' model was defined as only having the V_a parameter, thus producing a constant prediction over the time interval.

The AIS algorithm was applied to this data using $J = 512$ temperatures and $I = 32$ independent samples. Using the 32 samples produced by AIS, we could not reject the hypothesis that the posterior was Gaussian using Royston's test ($p = 0.96$). The estimates of the model evidences and Bayes factors, shown in Table 1, agree very well with those from VL. The normalised importance weights had high entropy, $H = 4.27$, and many trajectories had significant weight, $I_q = 21$.

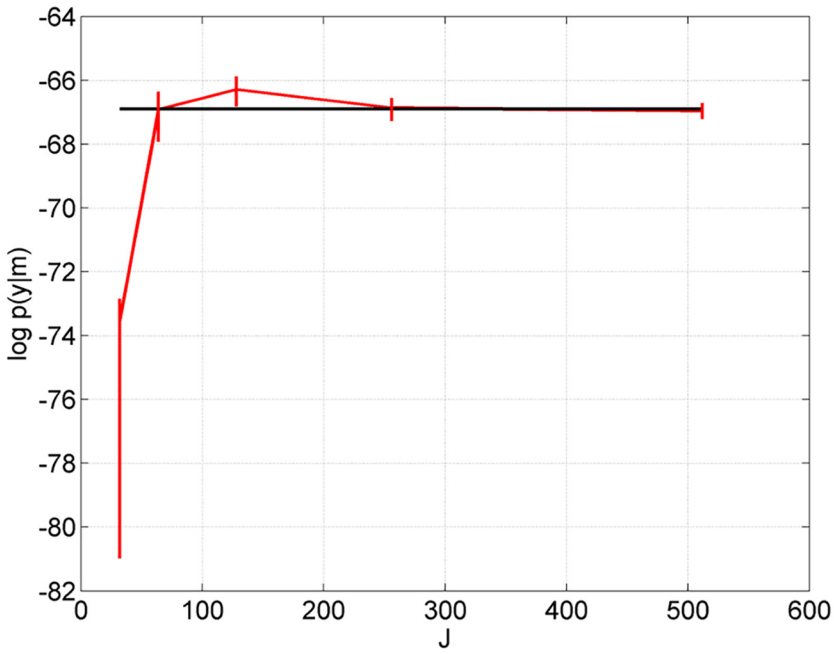

Fig 4. Approach-to-limit model. AIS approximation of log evidence (red line) as a function of number of temperatures *J*. Vertical lines span the 5th and 95th percentiles from bootstrapping. These approximations use *I* = 32 samples. The VL approximation is shown as the black line.

doi:10.1371/journal.pcbi.1004797.g004

Fig 4 shows the AIS approximation to the log model evidence, using $I = 32$ samples, as a function of the number of temperatures J. We see good agreement with VL for J larger than 128. These simulation results were based on a second data set from the approach model created by sampling parameters from the prior and producing time series as above (because this is a different data set the log evidence values are different to those in Table 1).

Neural Mass Models

To produce the following results the differential equations underlying the neural mass models (see S3 Text) were integrated using implicit backward-differentiation formulas (BDFs) and the resulting nonlinear equations solved using Newton's method as implemented in the CVODES software [40]. With a relative tolerance of 10^{-2} and an absolute tolerance 10^{-4} this algorithm took an average of 75ms (averaged over ten runs) to produce the time series for the two-region model. This was lower than the 229ms for Matlab's ODE15s integrator and the 90ms for SPM's (implemented in the function spm_int_L.m). Both VL and AIS model estimation approaches therefore used the CVODES implementation. For the LMC algorithm used in AIS, gradients were computed using a forward sensitivity method as implemented in CVODES. For VL, gradients and curvatures were computed using central differences as implemented in the SPM function spm_nlsi_GN.m.

The simulations that follow make use of the two-region neural mass model depicted in Fig 1 and described above. We generated data from a model with strong forward and backward connections. This is specified using the parameter values $w_1 = w_2 = 1$ which set the connections a_{21} and a_{12} according to S1 Table. The other parameters were set to zero. Data was then generated from the model as described above using zero mean additive Gaussian noise having standard deviation $\sigma_s = 0.01$. The resulting time series are shown in black in Fig 5. The priors over model

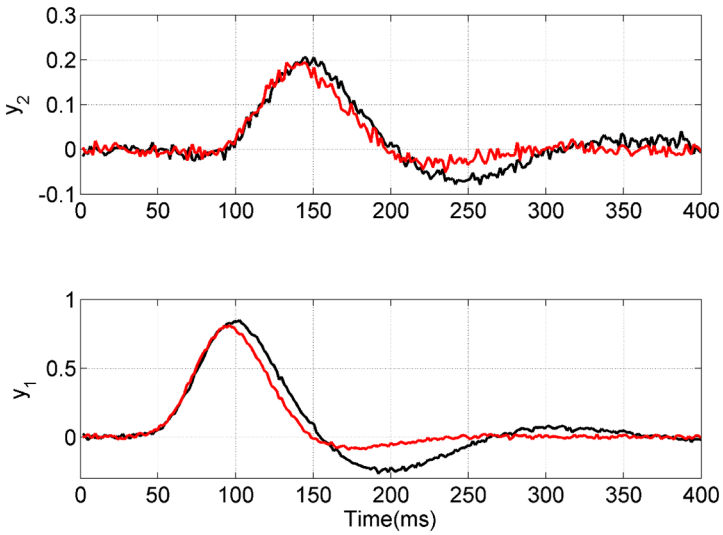

Fig 5. Time series from neural mass models. The bottom figure shows the pyramidal cell potential in region 1 for the full model (black) and reduced model (red). The top figure shows the same for the pyramidal cells in region 2. The reduced model is identical to the full model except that it does not have the backward connection from region 2 to 1. All time series contain additive Gaussian observation noise with standard deviation $\sigma_s = 0.01$.

doi:10.1371/journal.pcbi.1004797.g005

parameters for Bayesian model fitting are as described at the end of the above subsection 'Two-Region Model' in the section on 'Neural Mass Models'.

We then fitted two models to the data using AIS, a 'full' model, which has the same structure as the model from which the data were generated, and a 'reduced' model which did not have the backward connection. We used $I = 32$, $J = 512$ and model estimation took 5290s and 4610s for the full and reduced models. The estimated log model evidences were 1563.6 for the full model and 1293.4 for the reduced model, corresponding to a Log Bayes Factor of 270.2 in favour of the full model. Using the 32 samples produced by AIS, we could not reject the hypothesis that the posterior was Gaussian using Royston's test for the full ($p = 0.32$) and reduced ($p = 0.15$) models.

The AIS acceptance rates, a_j, averaged over the $I = 32$ trajectories, showed a gradual decrease with β_j. Averaging a_j over the high temperatures ($\beta_j < 0.5$) gave a value of $a_{high} = 0.43$ and over the low temperatures of $a_{low} = 0.19$. These acceptance rates show that the cost function is being sufficiently explored and are in line with other Bayesian annealing methods [38]. The normalised importance weights had lower entropy than for the previous models above, $H = 2.59$, and fewer trajectories with significant weight, $I_q = 12$.

We also fitted the full and reduced models using VL, which took 22s and 24s (using 19 and 22 VL iterations) respectively. The estimated log model evidences were 1524.1 for the full model and 1288.4 for the reduced model, corresponding to a Log Bayes Factor of 235.74 in favour of the full model. Thus, the VL and AIS estimates agree reasonably well for the reduced model (within 0.4 per cent) but not for the full model (within only 2.5 per cent). Which are we to believe?

As described in S1 Text, it is also possible to use the VL posterior as a proposal density to provide an importance sampling estimate of the model evidence, without using any annealing. We refer to this procedure as ISVL and used it to generate 1000 samples. ISVL is highly computationally efficient, requiring only 90s of compute time. The estimate of the log evidence was 1562.8 for the full model which agrees very well with the AIS estimate (within 0.05 per cent).

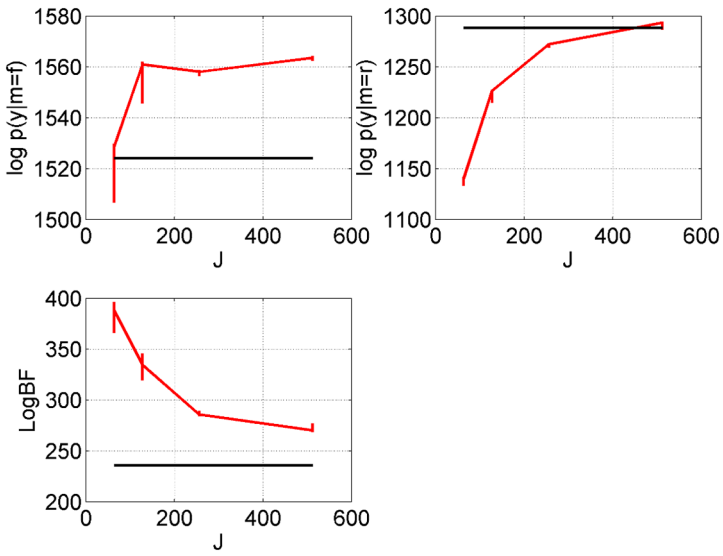

Fig 6. Two-region neural mass model: Temperature discretisation. The red lines indicate the AIS approximation of log evidence for full model, log $p(y|m = f)$ (top left), log evidence for reduced model, log $p(y|m = r)$ (top right), and log Bayes factor for full versus reduced, as a function of number of temperatures J. The vertical lines span the 5th to 95th percentiles from bootstrapping. These approximations use $I = 32$ samples. The black lines show the equivalent quantities for VL.

doi:10.1371/journal.pcbi.1004797.g006

Fig 6 plots the log evidences and log Bayes factor as a function of the number of temperatures J. These indicate that a fine-grained temperature resolution J is required to obtain good results. We also note that the log joint probability, L (see Eq 13), of the posterior mean AIS solution increases with J, with values of $L = 1583, 1584, 1588, 1589$ for $J = 64, 128, 256, 512$. The log joint probability of the true parameters is $L = 1589$, whereas the log joint of the VL posterior mean is only $L = 1157$.

Fig 7 plots the posterior densities from fitting the full model for VL and the AIS solution with $J = 512$. The estimates are generally in agreement but the AIS posterior means are closer to the true parameter values ($w_1 = w_2 = 1$, w_3 to w_{10} equal to 0) for eight out of ten parameters. This is reflected in the higher joint probability mentioned above. Given that we know the true parameters we can also compute the Root Mean Squared Error (RMSE) between true and posterior mean parameters. For VL this is 0.21 and for AIS it is 0.11.

Multiple runs. Perhaps it is not surprising that AIS has found a better solution given that it requires 240 times as much computer time (for the full model and with $J = 512$). We therefore compared AIS to a Multistart VL procedure (see above description in the section 'Variational Laplace') using 240 multi-starts, so as to equate computation time with AIS. The best solution had a log joint of $L = 1428$. The remaining solutions had a log joint of less than $L = 1275$, with 84% having $1130 \leq L \leq 1175$. Our initial solution (with $L = 1157$) is therefore fairly typical. On this evidence multi-start VL doesn't seem to be the best strategy.

Both AIS and VL will produce slightly different results over different runs of the algorithm (sampling trajectories for AIS, initialisations for VL). To quantify this variation we ran each algorithm twenty times and report the mean results and standard deviations for estimates of the log Bayes Factors, log joint density and RMSE in tables 2, 3 and 4.

The VL estimates of LogBF for the neural mass model have very low standard deviation, a point which we will comment on further in the next subsection. The corresponding AIS estimates have a standard deviation (or Monte Carlo error) of 5.27. However, this is a small

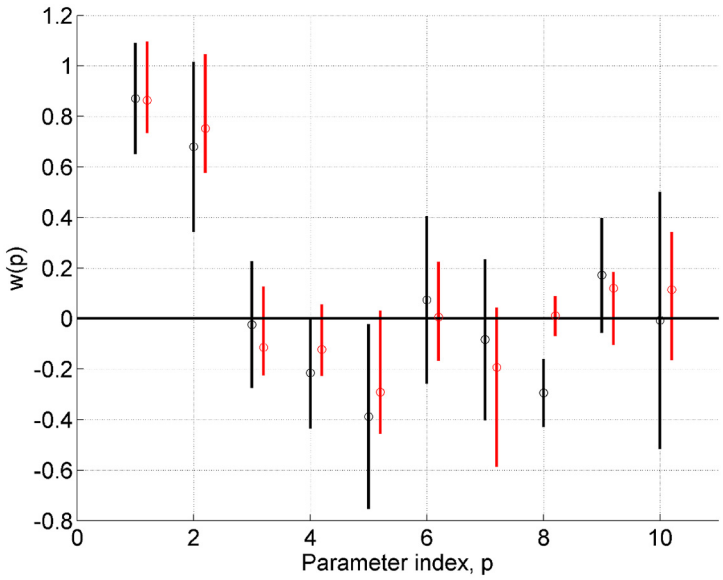

Fig 7. Two-region neural mass model: Posterior densities. Univariate posterior densities from a single model fit for AIS (red) and VL (black). The vertical lines span the 5th to 95th percentiles and the circles denote the posterior means. AIS provides better estimates for eight out of ten parameters.

doi:10.1371/journal.pcbi.1004797.g007

proportion of the absolute value of 271.86. For smaller Bayes factors we expect the AIS Monte Carlo error to be commensurately smaller [13]. This was confirmed by running AIS ten times on the same neural mass models, but with data (additive noise) chosen to produce a signal to noise ratio of unity (see next section). The mean Log Bayes Factor was 11.14 with a standard deviation of 0.66.

To provide an indication as to what level of performance other MCMC methods can provide, we also implemented an Adaptive Monte Carlo (AMC) approach which has been applied to related problems [39, 41]. Specifically we implemented "Algorithm 4" in [42] (which we refer to as AMC4) and collected 2000 samples. The proposal density was adapted for the first 600 samples, and these were then discarded as burn-in. The remaining 1400 samples provided the estimate of the posterior density and were used to compute the model evidence using the

Table 2. Evidence and Bayes Factor Approximations (Multiple Runs).

Model	VL	AIS	AMC4-PHM	AMC4-Chib
Linear, LogEv, Full	-11.02*	-11.07 (0.39)	-0.62 (3.49)	-11.31 (0.48)
Linear, LogEv, Red	-23.97*	-24.00 (0.31)	-14.09 (1.84)	-24.06 (0.24)
Linear, LogBF	12.95*	12.94 (0.49)	13.48 (3.49)	12.75 (0.56)
App, LogEv, Full	-60.85 (0.02)	-60.85 (0.27)	-57.42 (0.29)	-60.86 (0.04)
App, LogEv, Red	-662.67 (0.00)	-666.52 (0.13)	-664.36 (0.59)	-666.53 (0.02)
App, LogBF	605.68 (0.02)	605.67 (0.33)	606.94 (0.63)	605.67 (0.05)
NMM, LogEv, Full	1524.11 (0.00)	1563.12 (1.22)	1405.94 (130.88)	1476.05 (50.73)
NMM, LogEv, Red	1288.37 (0.00)	1291.26 (5.10)	63.81 (567.51)	451.93 (541.37)
NMM, LogBF	235.7 (0.01)	271.86 (5.27)	1342.13 (599.68)	1024.13 (559.24)

AIS estimates from I = 32 samples and J = 512 trajectories. The results for the linear model here* are for the analytic solution, which also corresponds to the VL solution. Entries shows the mean values from 20 runs of each algorithm with standard deviations shown in brackets. 'App' denotes the Approach model and 'NMM' the two-region neural mass model.

doi:10.1371/journal.pcbi.1004797.t002

Table 3. Log Joint Density of Posterior Mean (Multiple Runs).

Model	VL	AIS	AMC4
Linear, Full	-11.74*	-11.87 (0.06)	-12.07 (0.40)
Linear, Red	-24.67*	-24.78 (0.06)	-24.78 (0.10)
App, Full	-54.83 (0.02)	-54.97 (0.03)	-54.83 (0.00)
App, Red	-662.67 (0.00)	-662.70 (0.05)	-662.67 (0.00)
NMM, Full	1158.12 (13.57)	1588.43 (1.09)	1509 (51.27)
NMM, Red	-15911.88 (214.35)	1330.18 (5.50)	477.96 (541.59)

AIS estimates from I = 32 samples and J = 512 trajectories. The results for the linear model here* are for the analytic solution, which also corresponds to the VL solution. Entries shows the mean values from 20 runs of each algorithm with standard deviations shown in brackets. 'App' denotes the Approach model and 'NMM' the two-region neural mass model.

doi:10.1371/journal.pcbi.1004797.t003

Posterior Harmonic Mean (PHM) method (see equation 9 in S1 Text). As the PHM is known to overestimate the model evidence we also implemented Chib's method [43]. This uses samples from the posterior density and an additional set of samples produced by applying the proposal density to a chosen parameter vector (e.g. posterior mean). For completeness, this is described in S5 Text.

Whilst AMC4 worked well for the linear and approach models (with PHM overestimating the model evidence, as expected) it does not work so well for the neural mass model, as the model evidence approximations are highly variable.

Although AMC4 was run with a modest number of samples this was the same number as in [39], and results were not improved by running the algorithm for longer (we tried collecting 38,000 samples with 3,000 adaption/burn-in). Moreover, we implemented another AMC approach which had two separate phases of adaption (i) tuning of a global scaling parameter for 300 samples, to ensure acceptance rates of between 20 and 40 percent, (ii) tuning of proposal covariance for 300 samples using updates in [44]. Results were again not improved.

Effect of signal to noise ratio. The results presented so far have been found in a very high Signal to Noise (SNR) regime, using a very small value for the observation noise standard deviation (SD). Here the SNR is defined as the ratio of the observation noise SD to the signal SD in one of the brain regions (taken arbitrarily to be region 2). So far we have used SNR = 16.

Figs 8 and 9 show results for simulations in which the SNR was varied over a broad range. The results indicate that VL and AIS are generally in agreement, with monotonically increasing estimates of the log evidence as a function of SNR. For both VL and AIS, the log Bayes factors

Table 4. RMSE between Posterior Mean and True Parameters for Full Model (Multiple Runs).

Model	VL	AIS	AMC4
Linear	0.57 *	0.58 (0.04)	0.57 (0.05)
App	0.015 (0.004)	0.013 (0.008)	0.016 (0.003)
NMM	0.21 (0.00)	0.11 (0.01)	0.38 (0.13)

AIS estimates from I = 32 samples and J = 512 trajectories. The results for the linear model here* are for the analytic solution, which also corresponds to the VL solution. Entries shows the mean values from 20 runs of each algorithm with standard deviations shown in brackets. 'App' denotes the Approach model and 'NMM' the two-region neural mass model.

doi:10.1371/journal.pcbi.1004797.t004

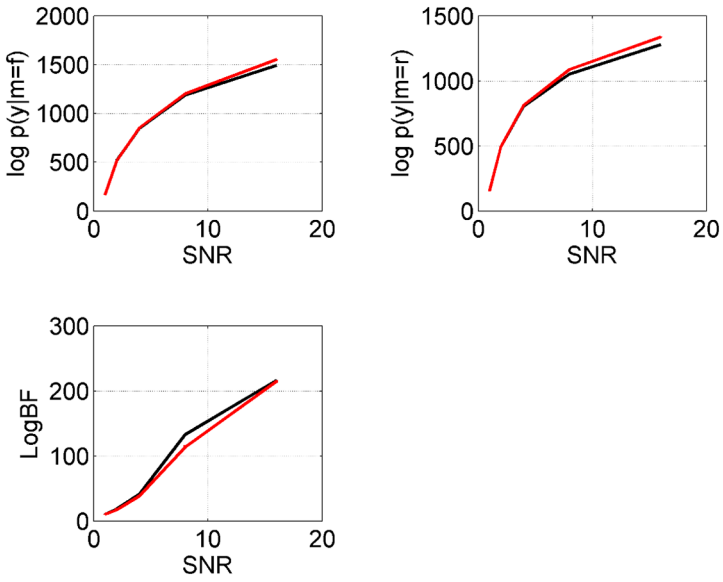

Fig 8. Data from 'full' neural mass model: Effect of SNR. Estimates of the log model evidence for full model, log $p(y|m = f)$, and reduced model, log $p(y|m = r)$, and Bayes factors for full versus reduced for VL (black) and AIS (red) over a range of SNRs.

doi:10.1371/journal.pcbi.1004797.g008

in favour of the full model are increasingly positive with data generated from the full model, and (generally) increasingly negative with data from the reduced model.

There are a number of discrepancies, however, with larger disagreements at high SNR. Overall, AIS tends to produce higher estimates of the log evidence. This is shown more clearly in S1 and S2 Figs for data generated from the full model. ISVL estimates of the model evidence (obtained using 10,000 samples) are also on the high side but have large error bars.

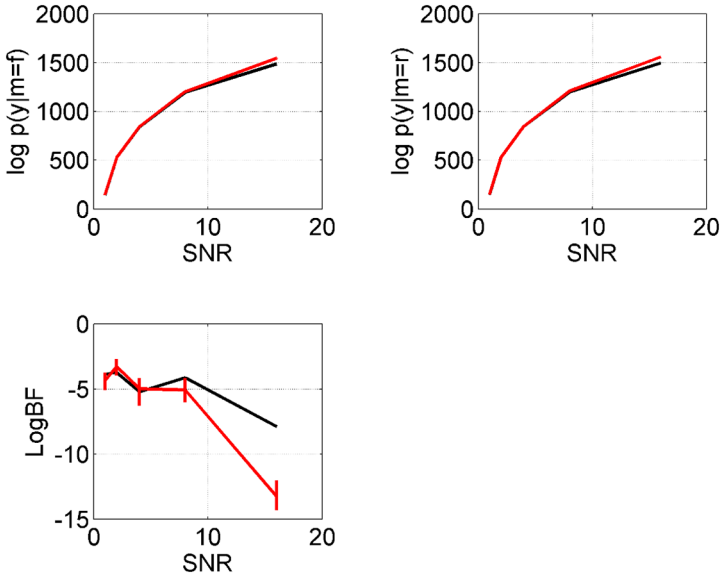

Fig 9. Data from 'reduced' neural mass model: Effect of SNR. Estimates of the log model evidence for full model, log $p(y|m = f)$, and reduced model, log $p(y|m = r)$, and Bayes factors for full versus reduced for VL (black) and AIS (red) over a range of SNRs.

doi:10.1371/journal.pcbi.1004797.g009

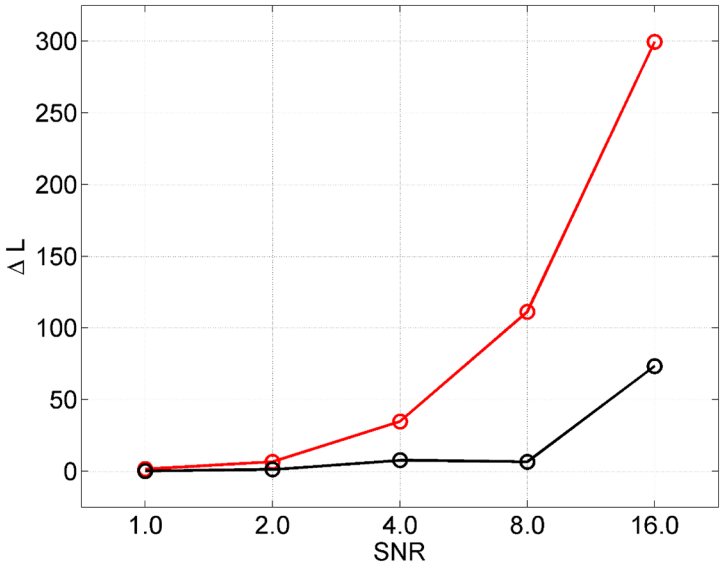

Fig 10. Data from 'full' neural mass model: AIS versus Multistart VL. The red curve plots the difference in log joint of the posterior mean from AIS versus that from VL, $L_{AIS} - L_{VL}$, and the black curve plots the difference in log joint for the best multistart VL versus VL, $L_{MVL} - L_{VL}$.

doi:10.1371/journal.pcbi.1004797.g010

Additionally, the ISVL estimates for the reduced model at high SNR were roughly 1000 or more less than for AIS/VL and had huge error bars, so were not plotted on the same figure. We therefore conclude that ISVL is unreliable.

At low SNRs the AIS acceptance rates, a_j, averaged over the $I = 32$ trajectories, were relatively constant over β_j whereas for high SNRs there was a gradual decrease with β_j. For example, at SNR = 2, $a_{high} = 0.53$ and $a_{low} = 0.48$ whereas at SNR = 8, $a_{high} = 0.49$ and $a_{low} = 0.34$.

As earlier, our AIS posterior means tend to have higher log joint probability, L, than those from VL. This is demonstrated in Fig 10 which plots the increase in L (over baseline VL) as a function of SNR. Here our baseline VL result uses the standard approach of initialisation with the prior mean. A multistart VL approach, however, can also produce better solutions. We used 240 multistarts as before. The maximum number of VL iterations over all multistarts was 42 (which did not exceed our maximal number of 128) and this individual model fit took 58s. Fig 10 plots the improvement offered by the best Multistart VL solution over the standard one showing, for example, an increase of $\Delta L = 73$ at the highest SNR. Overall, however, we find the improvement offered by AIS to be superior, with an increase of $\Delta L = 300$ at the highest SNR.

Perhaps surprisingly, there was hardly any improvement (or variation) in estimates of the VL model evidence or Bayes Factors over multistarts. As shown in S4 Fig, the variations in Log Bayes Factors are no greater than 0.1 (Bayes Factor = 1.1). Thus, for this neural mass model, VL model inferences show no meaningful variation over multistarts (according to [45] Bayes factors of less than 3 are 'barely worth a mention'). This is to be contrasted with large variations in the posterior mean over parameters (which led to the improvements in Multistart over baseline VL in Fig 10—see also S3 Fig). This result can perhaps be understood by noting that the model evidence approximation is the cost function that is optimised by VL, as contrasted to more standard Laplace approaches which find parameters that maximise the log joint.

In our initial (high SNR) comparison of AIS and VL estimates of model evidence the discrepancy was larger for the full than for the reduced model (see Fig 6). This did not translate, however, into an incorrect sign in the resulting log Bayes factor as the difference between full

and reduced model evidences dominated. A potential concern therefore is that when the model evidences of two models are more similar, errors in evidence estimates will result in errors in log Bayes factors that could produce radically different inferences. However, examples of when model evidences are more similar are provided in the low SNR regimes in Figs 8 and 9. Fig 8, for example, shows that even at the lowest SNR VL agrees with AIS in correctly favouring the more complex model. To examine this further we repeated the simulations at even lower SNR. The results in S5 Fig show, reassuringly, that as the SNR reduces to zero so does the log Bayes Factor (it does not become negative); VL Bayes Factors therefore do not show a bias towards simpler models. Additionally, the VL algorithm exhibits similar behaviour in the context of DCM for fMRI (see eg. Fig. 2 in [37]).

The p-values from Royston's tests for the various data sets are provided in S2 and S3 Tables. Results are provided for both 32 and 64 AIS trajectories. Most conservatively, considering the 40 multiple statistical comparisons a Bonferroni-corrected p-value of 1.25×10^{-3} or less would be seen as significant at the nominal 0.05 level. Given this threshold there is one significantly non-Gaussian posterior distribution. More descriptively, seven of the ten tests with 64 trajectories on data from the full model (two rightmost columns of S2 Table) have p-values of less than 0.05. We can therefore summarise these results by saying we have evidence for non-Gaussianity.

Discussion

Annealed Importance Sampling has a number of appealing properties. It can provide accurate estimates of the posterior parameter distribution and of the model evidence by avoiding local maxima and without making assumptions of Gaussianity. Samples from AIS converge in distribution to the true posterior density. Sub-optimal model evidence approximations [46] based on the Prior Arithmetic Mean (PAM) or Posterior Harmonic Mean (PHM) emerge as special cases of AIS with only two temperatures. Unlike Markov chain Monte Carlo methods, the samples produced are not serially correlated thus making any corrections involving effective sample size unnecessary.

We have described an implementation of AIS using a transition kernel based on an LMC sampler. The use of LMC here is critical as it allows proposals to be made based on local gradient and curvature information. Our empirical results show that the resulting proposals are accepted with probabilities in a desirable range (similar to the target of 20 to 40% in Zhou et al. [38]) even for nonlinear dynamical systems models at low temperature.

We have compared AIS to inferences based on the VL approximation in the context of neural mass models. In terms of the estimation of Bayes factors, the two methods agree as to which model is best but report different degrees of belief, especially at high signal to noise ratio. AIS tends to produce higher model evidence estimates both for optimal and suboptimal models. AIS finds better parameter estimates than does VL, as quantified by the joint log probability, especially in data regimes with high signal to noise ratio. A possible explanation as to the dependence on SNR could be that there are more or deeper local minima at high SNR. Moreover, a multistart VL procedure with computer time matched to AIS does not find better solutions. Additionally, we found evidence of non-Gaussianity in the AIS posteriors. Thus it appears that AIS is useful due to its ability to avoid local maxima, and its ability to characterise non-Gaussian parameter posteriors.

We have also used an Importance Sampling procedure to estimate the model evidence. This method, which we've referred to as ISVL, is highly computationally efficient as it uses the posterior from VL as a proposal density, but it proved unreliable. Similarly, other more standard approaches such as AMC worked well on linear and nonlinear regression problems but it was not possible to derive good AMC-based model evidence estimates for neural mass models.

In order to apply AIS one must decide upon an annealing schedule and in this paper we used a 5th-order geometric schedule, discretised using 512 temperatures and explored using 32 trajectories. This proved sufficient over a range of statistical models from linear and nonlinear regression to nonlinear differential equation models. Our empirical work has shown that the required number of temperatures and trajectories did not show a strong dependence on the number of model parameters or model nonlinearity. However, the need to specify the parametric form of the schedule, number of temperatures and trajectories is clearly a weakness of the AIS approach and is an area of ongoing research.

Previous work in this direction has focussed on the Sequential Monte Carlo (SMC) method which can be viewed as a generalisation of AIS. SMC represents probability densities using particles, as in the particle filter, but is applied at a sequence of temperatures rather than to a sequence of temporally ordered data. In particular Zhou et al. [38] have shown how SMC can be used for model comparison. Automatic annealing schedules can be derived by resampling at every temperature so as to maximise the effective sample size of the particle ensemble. An alternative approach grounded in statistical physics is based on the notion of contact flows and thermodynamic processes [47].

A potential drawback of SMC as compared to AIS, however, is that because particles interact during optimisation, SMC is not amenable to embarrasing parallelisation. Additionally, an application of SMC to nonlinear differential equations [38] used a similar number of temperatures as we do (500 as compared to our 512) but used many more trajectories (1000 as compared to our 32). This suggests that SMC may be more computationally demanding. Another development in this direction is Langevin Importance Sampling [48] which does not require specification of an annealing schedule as temperatures are sampled using Langevin dynamics. This flexibility again comes at the cost of interaction among trajectories (or particles) and therefore also compromises parallelisation.

Beal [23] has also suggested interesting ways of improving AIS. First, automatic annealing schedules could be produced by introducing finer graining of temperatures in regions of the path for which forward and reverse estimates are inconsistent. Second, Eq 11 suggest that better model evidence estimates could be produced by generating more samples at each temperature. This algorithim would then become more similar to thermodynamic integration [46] which, however, is naturally more computationally demanding than AIS [24].

Whilst our model fitting using AIS was parallelised over multiple cores, alternative efforts can be made to speed up implementation. For example, Wang et al. [49] have shown how the integration of neural mass models can be implemented on Graphical Processing Units (GPUs), resulting in a reduction of computing time by a factor of approximately seven. Additionally, Aponte et al. [50] have pursued a similar GPU approach for DCM for fMRI and shown how it can be used in the context of model evidence computation using thermodynamic integration. This GPU approach has been used to estimate parameters of DCM for fMRI models using an Adaptive Monte Carlo algorithm, again resulting in an order of magnitude reduction in computation time [41]. See also [51] for generic methods for parallelisation of single Markov chains.

Dynamical models have also been fitted to neuroimaging data using a range of global optimisation methods. For example, mean field models have been fitted to EEG using particle swarm optimisation [52] and stochastic nonlinear oscillator models have been fitted to EEG using a multi-start algorithm [53]. Additionally, DCMs have been fitted to fMRI data using a method that combines local search with Gaussian process approximation [41]. This method provides better parameter estimates than VL with only a modest increase in computational cost (much less than AIS). However, like the other global optimisation methods (see also [54]), it does not produce an estimate of the posterior distribution or model evidence.

This paper has compared the ability of VL and AIS to make inferences about two-region neural mass models based on simulated data. These simulations are a caricature of the DCM for ERP approach [17] as they are simplified in a number of respects (i) we have fixed parameters such as time delays between regions, synaptic time constants and synaptic response magnitudes, to known true values, (ii) we have not estimated observation noise, (iii) we have used only two brain regions whereas most practical applications use upwards of four [55–57], (iv) we have assumed that the electrical activities of brain regions are directly observed, rather than being filtered through a lead field matrix to produce observations in M/EEG sensor space, (v) we have used simulated rather than empirical M/EEG data. Further work will be needed to establish whether the findings from our caricature follow over to DCM for ERP.

This paper has used an independent model optimisation approach to compute Bayes factors, in which the evidence is computed separately for each model of interest. But in the context of AIS one can traverse a path from the posterior of one model to the posterior of another, with the resulting importance weights providing a direct approximation of the corresponding Bayes factor [13]. Direct computation of Bayes factors in this way is also possible in the context of SMC and a transdimensional AIS algorithm [58]. If one has a nested model, as in the empirical NMM examples in this paper in which the reduced model is nested within the full model model, Savage-Dickey approximations can also be used [59]. It would be interesting to compare Savage-Dickey against the direct path integral methods based on AIS.

This paper has explored one method for combining VL and sampling methods, ISVL, in which the VL posterior is used as a proposal density for importance sampling. However, this method did not provide good estimates of the model evidence. Other proposals for combining sampling with variational methods view the sequence of samples produced by a Markov chain as auxiliary variables in a variational inference problem [60]. An alternative approach, proposed in [13] would be to use AIS to traverse a path from the VL posterior to the true posterior at a series of intermediate temperatures, another interesting avenue for future work.

Supporting Information

S1 Text. Importance sampling.
(PDF)

S2 Text. Fisher information.
(PDF)

S3 Text. Neural mass models.
(PDF)

S4 Text. Variational laplace.
(PDF)

S5 Text. Chib's estimate of model evidence.
(PDF)

S1 Fig. Model evidence estimates for neural mass model: Low SNR. Estimates of the log model evidence for full model, $\log p(y|m = f)$, and reduced model, $\log p(y|m = r)$, at low SNR. Vertical lines indicate 95% confidence intervals.
(TIF)

S2 Fig. Model evidence estimates for neural mass model: High SNR. Estimates of the log model evidence for full model, $\log p(y|m = f)$, and reduced model, $\log p(y|m = r)$, at high SNR.

Vertical lines indicate 95% confidence intervals.
(TIF)

S3 Fig. VL estimates of log joint over multiple restarts. Estimates of the log joint for full model, log $p(y|m = f)$, over multiple restarts and range of SNRs. The baseline VL value (initialisation from prior mean) is shown in red.
(TIF)

S4 Fig. VL estimates of Log Bayes Factors over multiple restarts. Estimates of the Log Bayes Factor for full versus reduced models, over multiple restarts and range of SNRs. Data was generated from the full model. The baseline VL value (initialisation from prior mean) is shown in red.
(TIF)

S5 Fig. VL estimates of Log Bayes Factors at very low SNR. VL estimates of the Log Bayes Factor for full versus reduced models in very low SNR regime. Data was generated from the full model and the graph plots the mean and 95% confidence intervals computed over 5 data realisations at each SNR.
(TIF)

S1 Table. Neural mass model: Parameter transformations.
(PDF)

S2 Table. Neural mass model: Gaussianity tests on full models.
(PDF)

S3 Table. Neural mass model: Gaussianity tests on reduced models.
(PDF)

Author Contributions

Conceived and designed the experiments: WP. Performed the experiments: WP. Analyzed the data: WP. Contributed reagents/materials/analysis tools: WP BS. Wrote the paper: WP BS.

References

1. Dayan P, Abbott LF. Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. MIT Press; 2001.
2. Deco G, Jirsa V, Robinson P, Breakspear M, Friston K. The dynamic brain: from spiking neurons to neural masses and cortical fields. PLoS Comput Biol. 2008; 4(8):e1000092. doi: 10.1371/journal.pcbi.1000092 PMID: 18769680
3. Friston K, Harrison L, Penny W. Dynamic Causal Modelling. Neuroimage. 2003; 19(4):1273–1302. doi: 10.1016/S1053-8119(03)00202-7 PMID: 12948688
4. Daunizeau J, Kiebel SJ, Friston KJ. Dynamic causal modelling of distributed electromagnetic responses. Neuroimage. 2009; 47(2):590–601. doi: 10.1016/j.neuroimage.2009.04.062 PMID: 19398015
5. Moran R, Jung F, Kumagai T, Endepols H, Graf R, Dolan R, et al. Dynamic causal models and physiological inference: a validation study using isoflurane anaesthesia in rodents. PLoS One. 2011; 6(8): e22790. doi: 10.1371/journal.pone.0022790 PMID: 21829652
6. Friston K, Mattout J, Trujillo-Barreto N, Ashburner J, Penny W. Variational free energy and the Laplace approximation. Neuroimage. 2007; 34(1):220–234. doi: 10.1016/j.neuroimage.2006.08.035 PMID: 17055746
7. Bishop CM. Pattern Recognition and Machine Learning. New York: Springer; 2006.
8. Tierney L, Kadane J. Accurate Approximations for Posterior Moments and Marginal Densities. Journal of the American Statistical Association. 1986; 81:82–86. doi: 10.1080/01621459.1986.10478240

9. Walker A. On the asymptotic behaviour of posterior distributions. Journal of the Royal Statistical Society. 1969; 31.
10. Daunizeau J, David O, Stephan K. Dynamic causal modelling: A critical review of the biophysical and statistical foundations. Neuroimage. 2011; 58:312–322. doi: 10.1016/j.neuroimage.2009.11.062 PMID: 19961941
11. Nocedal J, Wright S. Numerical Optimization. Springer; 1999.
12. Grimbert F, Faugeras O. Bifurcation analysis of Jansen's neural mass model. Neural Comput. 2006; 18 (12):3052–68. doi: 10.1162/neco.2006.18.12.3052 PMID: 17052158
13. Neal RM. Annealed Importance Sampling. Statistics and Computing. 2001; 11:125–139. doi: 10.1023/A:1008923215028
14. Girolami M, Calderhead B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. Journal of the Royal Statistical Society Series B. 2011 03; 73(2):123–214. doi: 10.1111/j.1467-9868.2010.00765.x
15. Sengupta B, Friston K, Penny W. Gradient-based MCMC samplers for Dynamic Causal Modelling. Neuroimage, 125:1107–1118. doi: 10.1016/j.neuroimage.2015.07.043 PMID: 26213349
16. Chumbley J, Friston K, Fearn T, Kiebel S. A Metropolis-Hastings algorithm for dynamic causal models. Neuroimage. 2007; 38(3):478–87. doi: 10.1016/j.neuroimage.2007.07.028 PMID: 17884582
17. David O, Kiebel S, Harrison L, Mattout J, Kilner J, Friston K. Dynamic causal modeling of evoked responses in EEG and MEG. Neuroimage. 2006 May; 30(4):1255–1272. doi: 10.1016/j.neuroimage.2005.10.045 PMID: 16473023
18. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis. Chapman and Hall, Boca Raton; 1995.
19. Neal R. MCMC using Hamiltonian Dynamics. In: Brooks S Gelman GJ A, Meng X, editors. Handbook of Markov Chain Monte Carlo. CRC Press; 2011.
20. Foster I. Designing and building parallel programs. Addison Wesley; 1995.
21. Calderhead B, Girolami M. Estimating Bayes factors via thermodynamic integration and population MCMC. Computational Statistics & Data Analysis. 2009; 53(12):4028–4045. doi: 10.1016/j.csda.2009.07.025
22. Friel N, Pettitt A. Marginal likelihood estimation via power posteriors. Journal of the Royal Statistical Society: Series B. 2008; 70(3):589–607. doi: 10.1111/j.1467-9868.2007.00650.x
23. Beal M. Variational Algorithms for Approximate Bayesian Inference. Gatsby Computational Neuroscience Unit, University College London; 2003.
24. Vyshemirsky V, Girolami M. Bayesian ranking of biochemical system models. Bioinformatics. 2008; 24 (6):833–9. doi: 10.1093/bioinformatics/btm607 PMID: 18057018
25. Gelman A, Meng X. Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling. Statistical Science. 1998; 13(2):163–185.
26. Efron B, Tibshirani RJ. An introduction to the bootstrap. Chapman and Hall; 1993.
27. Calderhead B, Girolami M. Statistical analysis of nonlinear dynamical systems using differential geometric sampling methods. Interface Focus. 2011; 1:821–235. doi: 10.1098/rsfs.2011.0051 PMID: 23226584
28. Sengupta B, Friston K, Penny W. Efficient Gradient Computation for Dynamical Models. Neuroimage. 2014; 98:521–527. doi: 10.1016/j.neuroimage.2014.04.040 PMID: 24769182
29. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes in C (Second Edition). Cambridge: Cambridge; 1992.
30. Bates D, Watts D. Nonlinear Regression Analysis and its Applications. Wiley; 1988.
31. DiCiccio T, Kass R, Raftery A, Wasserman L. Computing Bayes Factors by Combining Simulation and Asymptotic Approximations. Journal of the American Statistical Association. 1997; 92(439):903–915. doi: 10.1080/01621459.1997.10474045
32. Felleman DJ, Essen DCV. Distributed hierarchical processing in the primate cerebral cortex. Cerebral Cortex. 1991; 1:1–47. PMID: 1822724
33. Litvak V, Mattout J, Kiebel S, Phillips C, Henson R, Kilner, et al. EEG and MEG data analysis in SPM8. Comput Intell Neurosci. 2011; 2011:852961. doi: 10.1155/2011/852961 PMID: 21437221
34. Royston J. Approximating the Shapiro-Wilk W-Test for non-normality. Statistics and Computing. 1992; 2:117–119. doi: 10.1007/BF01891203
35. Trujillo-Ortiz A, Walls RH, Barba-Rojo K, Cupul-Magana L. Roystest:Royston's Multivariate Normality Test. A MATLAB file. 2007;.

36. Geyer C. Practical Markov Chain Monte Carlo. Statistical Science. 1992; 7. doi: 10.1214/ss/1177011147

37. Penny WD. Comparing Dynamic Causal Models using AIC, BIC and Free Energy. Neuroimage. 2011; 59(1):319–330. doi: 10.1016/j.neuroimage.2011.07.039 PMID: 21864690

38. Zhou Y, Johansen M, Aston J. Towards automatic model comparison: an adaptive sequential Monte Carlo approach. ArCHIve. 2013;p. 1–33.

39. Sengupta B, Friston K, Penny W. Gradient-free MCMC methods for Dynamic Causal Modelling. Neuroimage; 112:375–81. doi: 10.1016/j.neuroimage.2015.03.008 PMID: 25776212

40. Hindmarsh A, Serban R. User Documentation for CVODES, and ODE Solver with Sensitivity Analysis Capabilities. Centre for Applied Scientific Computing, Lawrence Livermore National Laboratory; 2002.

41. Lomakina E, Paliwal S, Diaconescu A, Brodersen K, Aponte E, Buhmann J, et al. Inversion of Hierarchical Bayesian models using Gaussian processes. Neuroimage. 2015; 118:133–145. doi: 10.1016/j.neuroimage.2015.05.084 PMID: 26048619

42. Andrieu C, Thoms J. A tutorial on adaptive MCMC. Statistics and Computing. 2008; 18:343–373. doi: 10.1007/s11222-008-9110-y

43. Chib S, Jeliazkov I. Marginal Likelihood from the Metropolis-Hastings Output. Journal of the American Statistical Association. 2001; 96(453):270–281.

44. Haario H, Saksman E, Tamminen J. An adaptive Metropolis Algorithm. Bernoulli. 2001; 7(2):223–242. doi: 10.2307/3318737

45. Kass RE, Raftery AE. Bayes factors. Journal of the American Statistical Association. 1995; 90:773–795. doi: 10.1080/01621459.1995.10476572

46. Lartillot N, Philippe H. Computing Bayes factors using thermodynamic integration. Systematic Biology. 2006; 55(2):195–207. doi: 10.1080/10635150500433722 PMID: 16522570

47. Betancourt M. Thermodynamic Monte Carlo. Department of Statistics, University of Warwick; 2015.

48. Ma J, Peng J, Wang S, Xu J. Estimating the partition function of graphical models using Langevin Importance Sampling. In: 16th International Conference on Artifical Intelligence and Statistics (AISTATS); 2013.

49. Wang W, Hsieh I, Chen C. Accelerating computation of DCM for ERP in MATLAB by external function calls to the GPU. PLoS ONE. 2013; 8(6).

50. Aponte E, Raman S, Sengupta B, Penny W, Stephan K, Heinzle J. MPDCM: A toolbox for massively parallel dynamic causal modelling. Journal of Neuroscience Methods, 257:7–16. doi: 10.1016/j.jneumeth.2015.09.009 PMID: 26384541

51. Calderhead B. A general construction for parallelizing Metropolis-Hastings algorithms. Proceedings of the National Academcy of Sciences. 2014; 111(49):17408–17413. doi: 10.1073/pnas.1408184111

52. Bojak I, Liley DTJ. Modeling the effects of anesthesia on the electroencephalogram. Phys Rev E. 2005; 71(4):041902. doi: 10.1103/PhysRevE.71.041902

53. P Ghorbanian SR, Ashrafiuon H. Stochastic non-linear oscillator models of EEG: the Alzheimer's disease case. Frontiers in Computational Neuroscience. 2015; 9. doi: 10.3389/fncom.2015.00048

54. Mesejo P, Saillet S, David O, Benar C, Warnking J, Forbes F. Estimating biophysical parameters from BOLD signals through evolutionary-based optimization. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI); 2015.

55. Garrido M, Kilner J, Kiebel S, Friston K. Evoked brain responses are generated by feedback loops. Proc Natl Acad Sci U S A. 2007 Dec; 104(52):20961–20966. doi: 10.1073/pnas.0706274105 PMID: 18087046

56. Boly M, Garrido M, Gosseries O, Bruno M, Boveroux P, Schnakers C, et al. Preserved feedforward but impaired top-down processes in the vegetative state. Science. 2011 May; 332(6031):858–862. doi: 10.1126/science.1202043 PMID: 21566197

57. Schofield T, Penny W, Stephan K, Crinion J, Thompson A, Price C, et al. Changes in Auditory Feedback Connections Determine the Severity of Speech Processing Deficits after Stroke. J Neurosci. 2012 Mar; 32(12):4260–4270. doi: 10.1523/JNEUROSCI.4670-11.2012 PMID: 22442088

58. Karagiannis G, Andrieu C. Annealed Importance Sampling Reversible Jump MCMC Algorithms. Journal of Computational and Graphical Statistics. 2013; 22. doi: 10.1080/10618600.2013.805651

59. Friston K, Penny W. Post-hoc Bayesian model selection. Neuroimage. 2011 Jun; 56(4):2089–2099. doi: 10.1016/j.neuroimage.2011.03.062 PMID: 21459150

60. Salimans T, Welling M. Markov Chain Monte Carlo and Variational Inference: Bridging the Gap. In: International Confernece on Machine Learning; 2014.