ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 38 (2007) 478–487

Technical Note

A Metropolis–Hastings algorithm for dynamic causal models

Justin R. Chumbley,[a,*] Karl J. Friston,[a] Tom Fearn,[b] and Stefan J. Kiebel[a]

[a] *Wellcome Centre for Neuroimaging, Institute of Neurology, UCL, 12 Queen Square, London, WC1N 3BG, UK*
[b] *Department of Statistical Science, UCL, Gower Street, London, WC1E 6BT, UK*

Received 24 November 2006; revised 6 July 2007; accepted 16 July 2007
Available online 7 August 2007

Dynamic causal modelling (DCM) is a modelling framework used to describe causal interactions in dynamical systems. It was developed to infer the causal architecture of networks of neuronal populations in the brain [Friston, K.J., Harrison, L, Penny, W., 2003. Dynamic causal modelling. NeuroImage. Aug; 19 (4): 1273–302]. In current formulations of DCM, the mean structure of the likelihood is a nonlinear and numerical function of the parameters, which precludes exact or analytic Bayesian inversion. To date, approximations to the posterior depend on the assumption of normality (i.e., the Laplace assumption). In particular, two arguments have been used to motivate normality of the prior and posterior distributions. First, Gaussian priors on the parameters are specified carefully to ensure that activity in the dynamic system of neuronal populations converges to a steady state (i.e., the dynamic system is dissipative). Secondly, normality of the posterior is an approximation based on general asymptotic results, regarding the form of the posterior under infinite data [Friston, K.J., Harrison, L, Penny, W., 2003. Dynamic causal modelling. NeuroImage. Aug; 19 (4): 1273–302]. Here, we provide a critique of these assumptions and evaluate them numerically. We use a Bayesian inversion scheme (the Metropolis–Hastings algorithm) that eschews both assumptions. This affords an independent route to the posterior and an external means to assess the performance of conventional schemes for DCM. It also allows us to assess the sensitivity of the posterior to different priors. First, we retain the conventional priors and compare the ensuing approximate posterior (Laplace) to the exact posterior (MCMC). Our analyses show that the Laplace approximation is appropriate for practical purposes. In a second, independent set of analyses, we compare the exact posterior under conventional priors with an exact posterior under newly defined uninformative priors. Reassuringly, we observe that the posterior is, for all practical purposes, insensitive of the choice of prior.

Introduction

Neuroscientists and psychologists try to understand human cognitive processes in terms of their mechanistic implementation in the brain. The prevailing view is that complex acts of perception, inference and learning are made possible by functional integration among basic computational components. Furthermore, these modular computations are performed in spatially distinct areas of the cortex, and integration proceeds by means of long distance cortico-cortical connections. At a micro-mechanistic level, experimental and theoretical neuroscientists have attempted to approach this question via invasive empirical techniques in experimental animals. Under most circumstances, such direct access in neurologically normal humans is clearly impractical.

This challenge has generated the development of a number of technologies aimed at acquiring information about brain function distally: EEG, MEG, fMRI, etc. Typically information from these techniques is inadequate to probe micro-computational structure (being of poor resolution in some dimension of importance). For this reason, emphasis is confined to functional integration at a relatively macroscopic level.[1] For any given cognitive process, an important question concerns which brain areas are active and what is the network architecture that explains (rather than simply predicts[2]) this activity in terms of other areas.

One recent approach (Friston et al., 2003) has been to describe the temporal interactions between neuronal populations in the network as a deterministic dynamical system (a system of differential equations). Critically, such 'dynamic causal models' (DCM) are parameterized in terms of unobserved connectivity among neuronal states rather than some observed surrogate (e.g., hemodynamics as measured with fMRI). Bayesian inversion is used to access a data-dependent probability distribution, over the unobserved parameters (the posterior distribution). More precisely,

* Corresponding author. Fax: +44 207 813 1445.
E-mail address: j.chumbley@fil.ion.ucl.ac.uk (J.R. Chumbley).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2007.07.028

[1] Happily, this level of analysis is also more accessible to cognitive psychologists and their corpus of theory.

[2] By convention, the term "effective connectivity" is used to distinguish an explicit causal connectivity from "functional" connectivity, which is identified by some data-led statistical process.

the core [3] of the model is a set of differential equations. Here, we demonstrate a new and independent means of accessing the posterior distribution of the model parameters in these equations. This approach has the dual benefit of furnishing an exact sample from the true posterior while requiring weaker prior assumptions. In this analysis, of one key model architecture (i.e., Mechelli et al., 2003), we show that our exact posterior is very close to that found under previous approximate methods (Friston et al., 2003). Importantly, we also ask, for the first time how prior distributional assumptions influence the posterior.

This paper comprises three sections. In the first, we review the nature and form of dynamic causal models and their inversion under the Laplace assumption. In the second section, we introduce a sampling scheme that provides samples from the posterior, while accommodating priors with bounded support. In the final section, we apply both inversion schemes to an exemplar DCM and evaluate the difference in the ensuing posterior densities. This section concludes by looking at the changes in the posterior induced by changing the priors. We conclude with a discussion of the implications for inference with DCM and the influence of priors.

Dynamic causal modelling

The basic idea behind dynamic causal modelling of [neuronal] systems is to construct a reasonably realistic model of interacting [neuronal] systems or nodes. This model is then supplemented with a forward model of how [neuronal] states are transformed into measured responses. This enables the parameters of the model (i.e., effective connectivity) to be estimated from observed data. These supplementary models may be forward models of electromagnetic measurements or hemodynamic models of fMRI measurements. In this work, we will focus on fMRI. Responses are evoked by known deterministic inputs that embody designed changes in stimulation or context. This is accomplished by using a dynamic input-state-output model with multiple inputs and outputs. The inputs correspond to conventional stimulus functions that encode experimental manipulations. The state variables cover both the neuronal states and other neurophysiological or biophysical variables needed to form the outputs. The outputs are measured electromagnetic or hemodynamic responses over the brain regions considered.

Neuronal dynamics

The equations in a DCM describe neuronal dynamics by a multivariate differential equation. Typically, for fMRI data, this is bilinear in the states and inputs,

$$\dot{z}=\left(A+\sum_{j=1}^{M} u_t B^j\right) z_t + C u_t$$

where t indexes continuous time and the dot notation denotes a time derivative. The neuronal activity z_t is an $L\times 1$ vector comprising activity in each of the L regions and the input u_t is an $M\times 1$ vector comprising the scalar inputs. These experimentally determined inputs drive the system and are described as "exogenous" (the spatial location of these driving variables is not represented in the model[4]). They are weighted by the elements of C. In contrast, the parameters in the matrix A (those not combined with the inputs) describe input-independent or "regional" connectivity among the states, z. Finally, the parameters in the matrix B may be interpreted as modulatory, interaction or bilinear parameters because u_t and z_t combine in a multiplicative (i.e., nonlinear) manner (Friston et al., 2003).[5] A crucial aspect of the parameters relates to the conditions for stability (dissipation) in the equations of a nonlinear dynamic system. In real brains, it is not possible for neuronal networks to diverge exponentially to infinite values. This implies that the real component of the eigenvalues of the regional coupling matrix A must be negative. Specifically, when and only when the largest real eigenvalue (Lyapunov exponent) is negative, the stable mode is a point attractor.

In short, the bilinear approximation reduces the parameters to three sets that control three distinct things. First, the direct or exogenous influence of inputs on brain states in any particular area. Second, the regional coupling of one area to another and finally, changes in this coupling that are induced by input. Although, in some instances, the relative strengths of coupling may be of interest, most DCMs focus on the changes in coupling encoded by the bilinear parameters.

Hemodynamics

In the fMRI community, an empirically grounded, biophysical argument has been made for a "forward" model linking the neuronal states above to observable blood oxygen level-dependent (BOLD) signals. For the precise form of this model see Friston et al. (2003). It is defined over four hemodynamic states and characterizes the hemodynamic response function (HRF) to neuronal input. The hemodynamic model entails five extra biophysical parameters for each region that pertain to the induced vasodilatory signal (rate of signal decay, rate of flow-dependent elimination, hemodynamic transit time, Grubb's exponent and resting oxygen extraction fraction; see Friston et al. (2003). We refer to parameters in this part of the model as HRF-parameters. In this paper, we adopt the notation 'HRF{i}(j)' to describe the jth HRF parameter for the ith neuronal area (i.e., state), 'A(k,j)' for the coupling from j to k, etc. Each state in the model also has a parameter quantifying the mean of the observed time series; as a variable of no interest, this is treated as a "confound parameter".

Bayesian inversion

The essence of the Bayesian approach to DCM estimation is to describe uncertainty about the unobservable parameters (rather than simply in the observable BOLD dynamics) in the language of probabilities. More specifically, inference proceeds by the following trio:

(1) Characterize a joint distribution over the variables (both observable BOLD dynamics and unobservable HRF- and neuronal parameters)

$$p(\theta, y) = p(y|\theta)p(\theta),$$

[3] Core here is intended to imply the mean structure of the observation likelihood.

[4] These inputs variables may equally be bottom-up (sensory) or top-down (executive) in psychological terms.

[5] Neuronal network dynamics may be cast as a nonstationary linear system that changes according to u_t (Penny et al., 2004a,b). Because u_t is known, parameter estimation is tractable.

where θ is a vector containing all parameters in the neuronal and hemodynamic model and y are possible BOLD data values. Note that this decomposes into a specification of the conditional distribution of observables, $p(y|\theta)$ and unconditional prior information on the parameters, $p(\theta)$.

(2) Make BOLD observations y, under some experimental context.

(3) Conditional on the observations, compute an updated, data-dependent distribution over the unobservable parameters, using the familiar Bayes' theorem.

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

This furnishes the conditional density on the parameters and corresponds to model inversion. Conventionally, these models are inverted using expectation maximization as described in Friston et al. (2003). This scheme is formally identical to a variational inversion under a fixed-form (i.e., Gaussian or Laplace) approximation to the posterior density over the signal parameters with the noise parameters (i.e., noise variances) fixed. In fact, when the covariances are parameterized in terms of precision parameters, expectation maximization becomes identical to variational learning (see Friston et al., 2007). We will use expectation maximization below to compare the ensuing posteriors to those obtained with sampling methods that do not impose a fixed-form on the posterior.

Gaussian assumptions

Conventional DCM inversion schemes represent all of the distributions above as (either exactly or approximately) multivariate normal/Gaussian (MVN). We remind the reader that a MVN density over a p-vector θ, is defined by

$$p(\theta) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)\right)\theta\in\Re^p$$

where $\mu\in\Re^p$ is a fixed joint mean and Σ is a symmetric, positive definite matrix containing the variance–covariance elements of θ. Analytic properties of the joint and marginal Gaussian distribution are well characterized and constitute clear testable criteria against which to assess any observed data distribution. For example, it is a characteristic of the MVN that dependence between the constituent random variables must be described by their linear correlation. Similarly for example, for any legitimate μ and Σ, it can be shown analytically (e.g., Bera et al., 1984) that the marginal distributions are normal and the standardized third central moments is zero.[6] In what follows, we recapitulate the rationale for the conventional MVM prior and posterior distributions.

Gaussian approximation to posterior

The conventional normal (Laplace) approximation to the posterior may be justified by appeal to general limit theorems (see Gelman et al., 2004). Under certain conditions,[7] these theorems show that for asymptotic data y, generated from any distribution $p(y|\theta)$, governed by parameters θ with prior distribution $p(\theta)$, the posterior distribution over the unobserved parameters, $p(\theta|y)$ is MVN and converges in probability to θ_{true}, the true parameters of the system. With large but not infinite samples, the MVN (Laplace) approximation may still be good, though this is not quantified in general (Gelman et al., 2004). In the first part of our enquiry, we ask whether, for typical n and a typical DCM model, the MVN is indeed a good approximation.

This is a particularly acute and general question for models of dynamic systems that are causal in a control theory sense. This is because the likelihood is a function of the response of a dynamical system to exogenous inputs. The response is a generalized convolution of the inputs by generalized [Volterra] kernels. The key point here is that the kernels are generally nonlinear functions of the systems parameters; for example, in DCM the kernels are matrix exponential of the underlying connectivity matrix, which plays the role of a Jacobian. This is important because it means the likelihood is nonlinear in the model parameters. In other words, the posterior must be non-Gaussian, even with Gaussian priors. The question is how non-Gaussian? Note that this issue pertains to any generative model of dynamic systems that is parameterized in terms some underlying state equations (e.g., the bilinear form above).

Gaussian priors

It is useful to think of prior knowledge of the DCM in terms of two distinct and exhaustive parameter[8] classes: neuronal and nonneuronal. Prior knowledge of the nonneuronal parameters exists from previous biophysical empirical work (Friston, 2002) and is well summarized by parameters of a normal distribution. In contrast, for any novel application of DCM, priors on the neuronal parameters may not be available from previous work. Instead, prior knowledge is in the form of rather general theoretical constraints, i.e., the eigenvalues of the connectivity matrices. In fMRI, the conventional joint prior distribution over the elements of A is specified in such a way as to ensure A's eigenvalues are negative with a fixed and known probability (Friston et al., 2003). In brief, the argument assumes independent and identical, normal distributions over the off-diagonal connection parameters of A. This assumption is used to establish sufficient conditions (in particular, conditions on the variance of these distributions), that determine the probability of positive real eigenvalues of A, see Friston et al. (2003).

[6] In general, the third moment is a measure of asymmetry in a distribution and is zero for exactly symmetrical distributions (approximately zero in the sample). A right or left skews imply positive and negative third central sample moments respectively (cubed positive deviations exceed the cubed negative or vice versa).

[7] The validity of the approach relies on certain other assumptions, which we do not assess in this work. (1) The approximation fails if the model is under-identified or nonidentified. (A model is said to be under-identified given observation vector y, if the likelihood is constant for some range of the parameter vector θ; Gelman et al., 2004.) In other words, different θ are observationally equivalent; there is more than one "most likely" parameter vector. In such a case, there is no single point to which the posterior will asymptotically converge. Bayesian analysis of a nonidentified model is always possible if a proper prior on all the parameters is specified, but at the cost that the marginal prior and posterior distributions are identical (the data do not "inform" the posterior; Poirier, 1998). (2) To attain a unique mode, the likelihood must be bounded. Because, we have no analytic form for the likelihood it is difficult to assess this possibility for a given data set, y. (3) The true solution must have nonzero probability in the prior.

[8] Parameters determining the mean structure of the likelihood (as opposed to the exponential family dispersion parameter[s]).

This approach, of shrinking the individual neuronal connectivity parameters probabilistically, satisfies the eigenvalue constraint on dissipation at the cost of assuming a Gaussian density with a carefully specified covariance matrix. Ideally, we would impose hard constraints on the prior that categorically, rather than stochastically, satisfy the eigenvalue constraint, thereby allowing us freedom to implement our true a priori beliefs on the remaining support of connectivity parameters (which may be uninformative or anatomically informed).[9] To make things clear, consider the 2×2 case for a coupling matrix

$$A = \begin{pmatrix} -1 & y \\ x & -1 \end{pmatrix}$$

then the eigenvalues are $v = -1 \pm \sqrt{xy}$. So the eigenvalues are both real and negative if $0<xy<1$. This region lies in the positive and negative quadrants of the x–y plane between the axes and the rectangular hyperbola $xy=1$. This is the region we would like to support our priors (but no simple conventional prior has this bounded support) (Fig. 1).

Ideally, we would like to use a nonnegative prior density only over this admissible domain of the neuronal coupling parameters. In what follows, we will define an uninformative prior over this admissible domain and compare the ensuing posterior with the posterior under conventional informative (shrinkage) priors.

Methods

The Metropolis–Hastings algorithm and DCM

Here use a Metropolis–Hastings (MH) algorithm[10] (Hastings, 1970) with a rejection step to attain a sample from the true posterior (for a fixed prior) without knowing the normalizing constant or analytic form for the posterior. Rejection of inadmissible proposals ensures that sampling only occurs on a truncated space and allows us to use priors on bounded supports, of the sort required by the stability constraint of the previous section.[11] In brief, MH involves the construction of a Markov chain (a sequence of random variables with Markov dependence) whose equilibrium distribution is the desired posterior distribution. At equilibrium, a sample from the chain is a sample from the posterior. Note that the posterior distribution, reconstructed in this way will not be constrained to the normal family, thereby evading a key limitation of conventional fixed-form inversion schemes. We describe the general form of the algorithm in Appendix A (for further details on the general MH algorithm, see Hastings, 1970). As an illustrative example, we apply the algorithm to a DCM that has been used in the peer-reviewed literature.

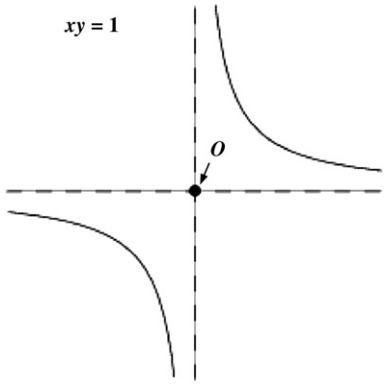

Fig. 1. Stability regions in parameter space.

The DCM

We simulated data from a tri-state DCM architecture (for generality, we refer to the states or regions as "1", "2", "3"). With regard to regional connectivity, "1" feeds into two successive states (see Fig. 2). With regard to modulatory influences, the 1–2 connection is subject to context-dependent modulation. This architecture was used in Mechelli et al. (2003) to assess the basis for computational specialization in the cortex. In total, this model has 23 unknown parameters: a global scale parameter, an input parameter, three coupling parameters, five HRF parameters for each region and three regional confounds. Posterior means of all parameters were obtained from a typical single-subject DCM analysis of this study and were used to generate synthetic data.

The synthetic data, used for our study, were obtained by integrating the system using the posterior parameter means of the real data. Our data sets comprised three-variate time series with 128 scans, generated in response to 107 events with jittered onsets and a TR of 3.0 s. We added Gaussian noise to these time series. For the first phase of simulations, we used two different signal-to-noise ratios (SNR); one and five. The low value of one simulates data that have been obtained without any spatial averaging during fMRI data preprocessing (i.e., smoothing) or DCM feature selection (i.e., eigenvariate extraction). For DCM, this low SNR represents the worst case encountered in practice. The high value of five simulates the upper bound of the SNR that one can expect in a DCM study. Having established algorithmic robustness at these extremes we proceed, in the second phase of simulations with a more typical SNR of 2.5. For details on generating synthetic data see spm_dcm_create.m (http://www.fil.ion.ucl.ac.uk/spm).

Convergence and preprocessing

Convergence of the MH Monte Carlo–Markov chain (MCMC) chains was assessed by two means. Firstly, we assessed whether, for any given chain, the marginal sample trace for all model parameters was stable. Secondly, we assessed whether multiple chains, starting

[9] It is helpful to consider this in terms of Bayesian ideas of hierarchical structure. To begin, note that the systemic constraint on the connectivity parameters restricts the support of possible real eigenvalues to the negative part of the real number line. Consider placing a uniform distribution on and only on this support. Conditional on this constraint, a subordinate joint distribution over the connectivity parameters may then be defined. We will show that a uniform prior on the eigenvalues has the effect of truncating inadmissible parts of the joint distribution on the subordinate parameters. This conceptualization differs from the standard implementation of hierarchical models in which the supraordinate parameter always figures explicitly in the functional form of the next level. In contrast, here the systemic constraint influences the posterior via modifications to its support. We do not formalize this approach in this work.

[10] Because the core of the likelihood is a numerical function of the parameters, the Gibbs sampling algorithm (which requires an analytic form in the likelihood) is not an option.

[11] We do not assess formally whether imposition of the hard constraint effects the equilibrium distribution. A similar potential limitation is described in Appendix A and relates to whether a jumping proposal distribution affects the sampling effect equilibrium.

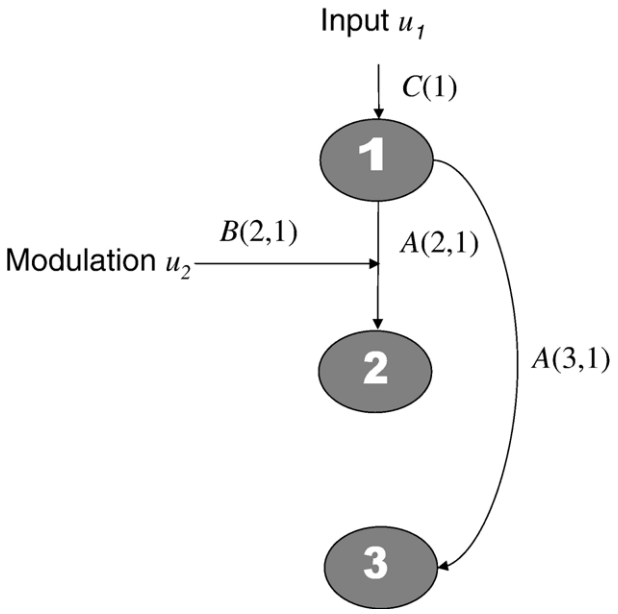

Fig. 2. Network architecture (and the four coupling parameters); each vertex or region in the graph has a corresponding time series. These series form the basis of inference, as described in the main text.

at different locations, converged to a single distribution.[12] These locations were over-dispersed with respect to the posterior density obtained with the Laplace approximation. For the model in question none of the diagnostic criteria refuted convergence. A sequence of $c=750{,}000$ preconvergence iterations (a "burn-in") was removed from the MH chain before further analysis. Because the sample comes from a Markov chain, the posterior realizations are not independent of one another. This dependence does not bias point estimates of the posterior, nor does it confound the graphical summaries we present of the posterior. However, for inference on the posterior we subsample from the chain to achieve approximate independence. In particular, a standard first-order linear autoregression indicated that discarding intervals of $b=20$ was more than sufficient to assume approximate independence among the remaining observations.

Following this preliminary treatment of the data, our analysis can be divided into two parts. First, because of its relevance to conventional practice, we use MH to reconstruct the true posterior given the conventional specification of the MVN prior (prior covariances were fixed as described in Appendix B; see Friston et al. (2003) for details). Second, we assess sensitivity of the posterior to key changes in the prior; the prior variance of the coupling parameters in A and B. As indicated above, this crucial question can now be addressed because the rejection/truncation step in the algorithm ensures that the system is dissipative and enables us to choose any prior on the remaining support. In this paper, we focus on what happens when the prior becomes uninformative.

In both analyses, we examine whether any differences in the posteriors influence the summaries used by experimentalists, e.g., the probability of the parameter over positive support or 95% Bayesian confidence intervals. In brief, we find that under the conventional priors (Friston et al., 2003), while deviations from posterior normality occur, they are subtle and do not affect practical inferences. We also show that posterior summaries are largely insensitive to the qualitative choice of prior distribution.

Results

MH vs. Laplace approximate posterior: conventional informative priors

We start with a comparison of the true (Metropolis–Hastings) and approximate (Laplace) posteriors by looking at the inferences based on univariate and bivariate marginal densities. To ensure stability of the results, we repeated the each simulation four times and report the results for all replicates separately.

Univariate marginal posteriors

We first assessed the probability that a connectivity parameter is above 0 ("$p>0$"), and the 95% confidence interval for that connectivity parameter. Table 1 gives $p>0$ for the bilinear parameter in the model as approximated by Laplace and MH. For the remaining parameters, the posterior probability was effectively one, for both methods.

Fig. 3 compares the Bayesian confidence intervals for neuronal parameters $A(1,1)$, $A(2,1)$, $B(2,1)$, $C(1,1)$, under MCMC (left hand of each pair) with those under the conventional Laplace approximation (right hand of each pair). Each pair allows for direct comparison of the mean and 95% confidence interval of these schemes for an identical data set (respective means in each pair are connected by a lateral line). Such comparisons are reported over realizations at two different noise levels, with SNR = 1 and SNR = 5. It is self evident that both the posterior means and the confidence intervals are very similar under the Laplace approximation and MH.

We examined the pairwise bivariate posterior of the parameters and found them to have subtle bivariate skew and nonlinear codependence. Such features are however very small in magnitude and, for practical purposes, linearity is an appropriate approximation (see Fig. 4 for the most extreme examples of skew we observed).

Finally, we asked whether these findings would extend to a different DCM model architecture. In particular we considered a model used recently in the experimental literature (Kumar et al., 2007) (see Fig. 5a). Briefly, we found that the posterior under this architecture has properties very similar to that under the architecture of Mechelli et al. (2003). In particular, we found good correspondence between the 95% confidence intervals (see Fig. 5b) and essentially identical probability-above-zero for MH vs. Laplace.

MH vs. Laplace posterior: uninformative priors

In the second set of analyses, we assessed the effect on posterior distributions and ensuing inference of increasing the prior variance[13]

[12] The conventional r-method (Gelman et al., 2004) might also be used to ensure that the chain was mixing well within a single convergence distribution According to this method, each chain is divided into n sub-chains of w iterations; a classical statistic is then computed that compares the within- vs. between-sub-chain variance. The approximate identity of these two variances is necessary (though never sufficient) to diagnose convergence.

[13] As we indicated above, having implemented a truncation of the parameter space to satisfy the eigenvalue constraint, we are free to choose any prior distributional form; Gaussian or otherwise. We have chosen a Gaussian with very high variance over the neuronal parameters because it is effectively indistinguishable from an uninformative prior.

Table 1
Approximate posterior probabilities that the bilinear parameter 'B{1}(2,1)' is above zero for the Laplacian approximate posterior vs. the MH sample

	Replicate 1		Replicate 2		Replicate 3		Replicate 4	
Parameter	Laplace	MH	Laplace	MH	Laplace	MH	Laplace	MH
'B{1}(2,1)'	0.8967	0.8939	0.9717	0.9738	0.9799	0.9840	0.9657	0.9611

The numbers are acquired by integration over the positive support of the normal in the Laplace case, and the proportion of samples above zero in the MH case. The inference $p>0$ is not compromised by the Laplace approximation. In all cases, it was within three decimal places of the number as inferred by MH. For all other neuronal parameters (other than 'B{1}(2,1)') at a SNR of one, and every parameter for a SNR of five, the discrepancy between this measure under MH vs. Laplace vanished and the two gave identical results (essentially all posterior mass being above zero).

on the neuronal parameters in A and B to 10,000. This makes the prior uninformative.

Each subplot in Fig. 6 gives four pairwise comparisons of the Bayesian 95% confidence intervals under informative (left hand of each pair) versus informative (right hand of each pair) priors, at a typical SNR of 2.5. Each pair allows for direct comparison of the 95% confidence intervals under these two priors, given an identical data set. Posterior inference is presented for each of the four neuronal parameters of interest ($A(2,1)$, $A(3,1)$, $B(2,1)$, $C(1,1)$).

In Fig. 7, we assess deviations from nonnormality as we change from strong to weak priors as expressed in the skew of the empirical distributions (skew being zero in the Gaussian case). We observed that in nearly all cases stronger Gaussian priors engender more Gaussian posteriors (i.e., posteriors with smaller skew). There was a trend for weak priors to be slightly more skewed, which is an example of the posterior dependence on the prior. In contrast, there was a good agreement between the posterior means under the two schemes (see Fig. 6). Furthermore, examination of the statistic $p>0$ yielded exact agreement between these two posteriors in all replications, at this intermediate SNR (2.5), despite very different priors.

Discussion

Exact inference with conventional priors

We report that under conventional priors (Friston et al., 2003) the Laplace approximation to the posterior yields sensible inferences under all conditions and for all replications examined in this work. These analyses illustrate the usefulness of MCMC for quality control on any approximation made in the service of expedient inference.

In particular, we observed nonlinearities and skew in the bivariate posterior samples that were small in magnitude, rendering linear correlation (or equivalently covariance) an adequate approximation. If we assess only gross attributes of the posterior, we find that ensuing inferences are robust to these slight discrepancies from normality. Critically, it is at this level that operational inferences are made in the neuroscience literature (e.g., Mechelli et al., 2003). In particular, it is conventional to infer a specific connection between two states, if the probability of connection parameter exceeding some threshold is greater than 95%. On implementing this test, via the MH inversion scheme, we find that our inferences concur with those of the Laplace

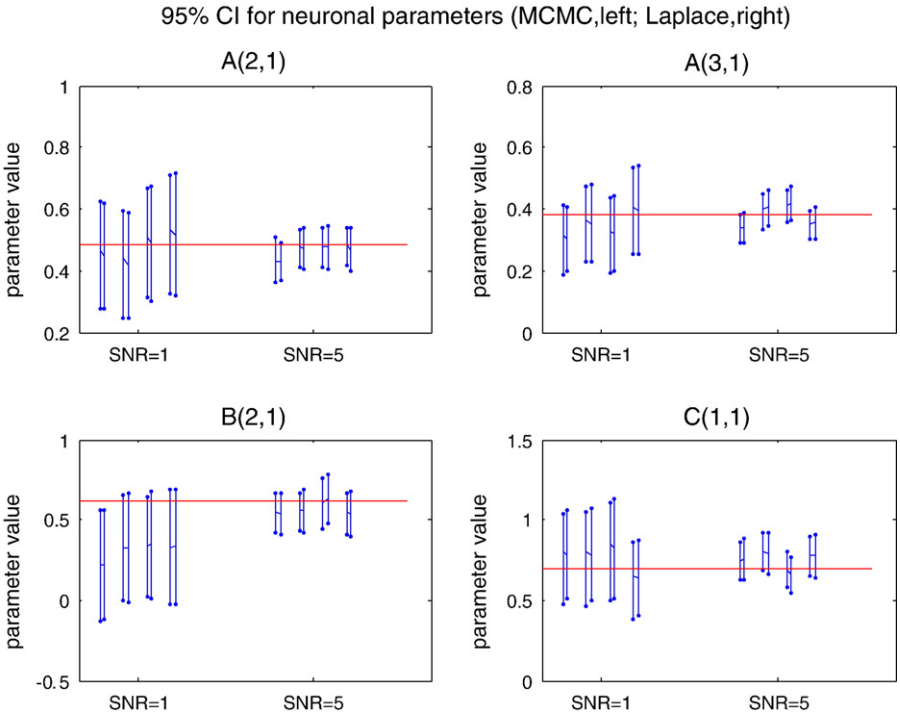

Fig. 3. In each of the 4 subplots, the posterior 95% CI has been constructed according to MH (left hand of each line pair) versus Laplace (right hand of each line pair), under four realizations at each of two signal/noise ratios (SNR=1 and SNR=5). The red line indicates the true parameter. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

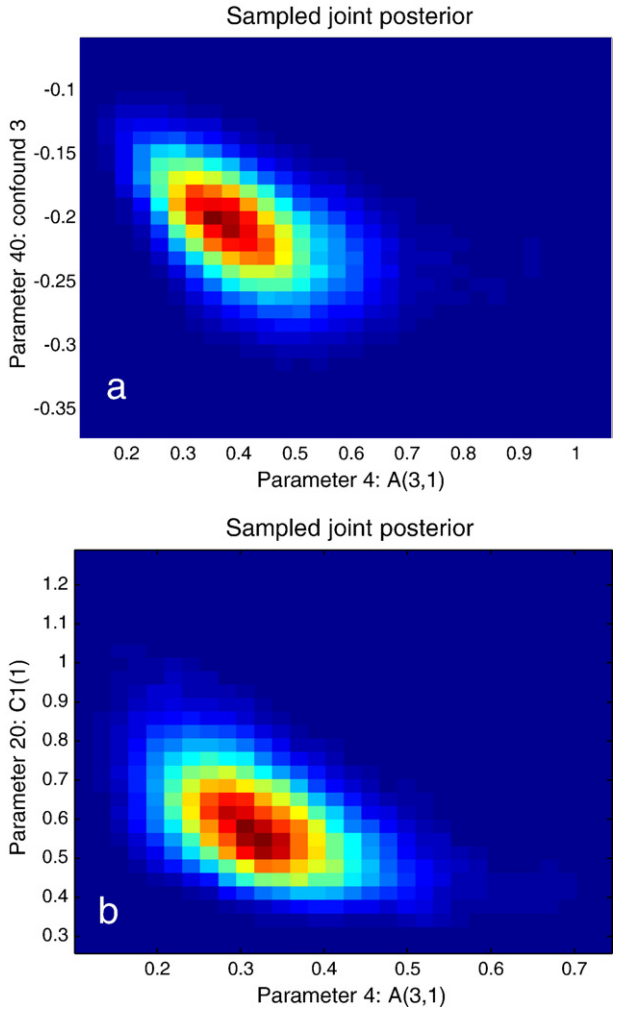

Fig. 4. Examples of posterior bivariate histogram, generated with rejection-MH where iso-contours are approximately, though not exactly, regular ellipses.

approximation on every count. Similarly, if we plot the 95% Bayesian confidence intervals as derived from MH, we find that departures of MH from the Laplace approximation are very small.

We have focused on the power of MCMC for assessing the quality of a posterior approximation. We suspect that a different approach is needed to examine the quality of model comparison approximation. We note that it is currently straightforward to compute an upper bound on the model evidence during MCMC simulations such as ours (by simply averaging all likelihoods accepted by the algorithm; Beal, 2003). While, in practice, this bound can be used for model comparison, it is not of itself a gold standard against which to assess the quality of extant model comparison approximations (e.g. Penny et al., 2004a,b).

Sensitivity to the prior

A characteristic of the conventional scheme is that it is not possible to fully gauge the sensitivity of the posterior to different prior distributions. To probabilistically satisfy the dissipation constraint by means of normal priors, the mean and variance of these priors must be fixed (i.e., shrinkage priors as in Friston et al., 2003). In our approach, we are not compelled to fix these values (nor even use a Gaussian to represent prior knowledge). We exploited this latitude, imposing hard constraints on the prior in combination with an uninformative density on the remaining parameter space. We found a trend towards increased posterior skew under such priors (Fig. 7). This indicates that the quality of the Laplace approximation is slightly diminished under informative priors. In contrast, we observed only small deviations in the mean under uninformative priors as compared with conventional priors, at an SNR of 2.5. Furthermore the important posterior measure $p>0$ was in complete agreement between weak and strong priors. This suggests that the informative parameters used in current DCM analyses are unnecessary and may need re-evaluation.

Our hard constraints forced zero prior probability on certain impossible parameter values a priori (values that imply exponentially explosive neuronal dynamics). In practice, these hard constraints are implemented in the rejection step in the algorithm. In general, we foresee complications with the use of hard constraints in a general MH scheme and more theoretical work is needed (see Appendix C). Pending resolution of these technical issues, and given our observations of rather Gaussian and prior-insensitive posteriors, it is clear that the conventional Laplace approximation and estimation algorithm (expectation maximization) is the most robust, speedy and pragmatic method available.

Acknowledgments

The Wellcome Trust funded this work. We would like to send a special thanks to David Lunn and Jean Daunizeau.

Appendix A. The MH Algorithm

For simplicity, we use the notation $p(\theta)$ to describe the value of the density at θ (but note that we always refer to continuous not discrete distributions). θ^* is a proposed vector of parameters, θ^{t-1} is the previously accepted parameter vector. The vectors, y, y^*, y^{t-1} are, respectively, the observed values for the time series, the fitted values under θ^* and the fitted values under θ^{t-1}. Λ is a covariance matrix chosen to give an acceptance ratio of around 20%.

For $t=1,K,n$

Step 1: if t equals 1; choose some initial values for the Markov chain: θ^{t-1},

Step 2: Generate a proposed parameter vector $\theta^* \sim N(\theta^{t-1}, \Lambda)$,

Step 3: Are all real eigenvalues of $A<0$?

If false, reject θ^* (do not count θ^* towards the sample)

If true, proceed to step 4

Step 4: With probability min(1, r), accept θ^* (count θ^* in the sample), where r is defined as

$$r = \frac{p(\theta^*|y)p(y)}{p(\theta^{t-1}|y)p(y)} = \frac{p(\theta^*,y)}{p(\theta^{t-1},y)}$$

$$= \frac{p(\theta^*,y|\Sigma_m,M,\Pi)}{p(\theta^{t-1},y|\Sigma_m,M,\Pi)}$$

$$= \frac{N(y|y^*,\Sigma_m)N(\theta^*|M,\Pi)}{N(y|y^{t-1},\Sigma_m)N(\theta^{t-1}|M,\Pi)}$$

Intuitively, the algorithm instructs us to accept a proposal with certainty if it is more probable in the true posterior distribution,

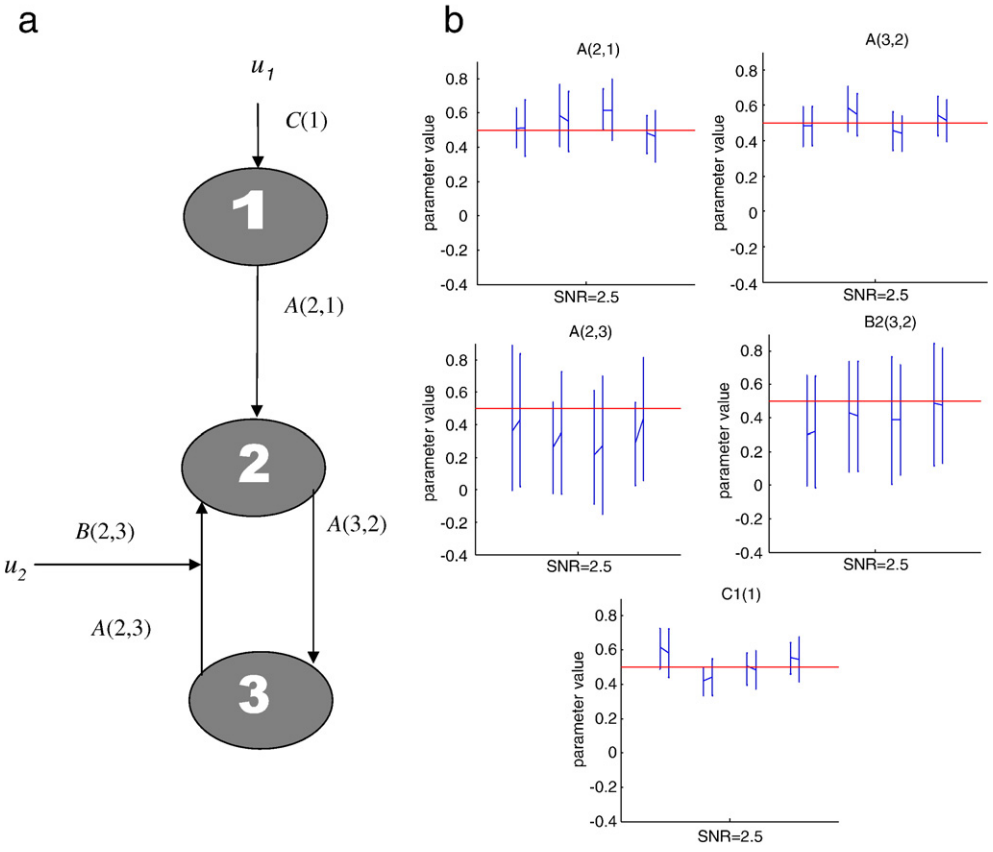

Fig. 5. (a) Model architecture (see Kumar et al., 2007). (b) In each of the 5 subplots, the neuronal 95% posterior CI has been constructed according to MH (left hand of each line pair) versus Laplace (right hand of each line pair), under four realizations at signal/noise ratio (SNR = 2. 5). The red line indicates the true parameter. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

otherwise accept it according to how much less probable it is (i.e., the ratio r above). The last two lines make explicit that inference is conditional on $\grave{I}$ and Π, the prior mean and covariance, which is fixed a priori. Inference is also conditional on Σ_m, the [restricted] maximum likelihood estimate of the variance of random effects in observables (as described in Friston et al., 2003).[14]

Appendix B. Priors on the coupling parameters

Consider any set of $l(l-1)$ inter-regional connections $a_{ij};i\neq j$ with sum of squared values $\xi=\sum a_{ij}^2$. For any given value of ξ, the biggest Lyapunov exponent λ_a obtains when the strengths are equal $a_{ij}=a$, in which case

$$\lambda_a = a - 1$$

$$\xi = l(l-1)a^2$$

This means that as the sum of squared connection strengths reaches $\xi=l/(l-1)$, the largest exponent attainable, approaches zero. Consequently, if ξ is constrained to be less than this threshold, we can set an upper bound on the probability that the principal exponent exceeds zero. ξ is constrained through the priors on a_{ij}. If each connection has a prior Gaussian density with zero expectation and variance v_a, then the sum of squares has a scaled Chi-squared distribution $\xi/v_a \sim \chi^2_{l(l-1)}$ with degrees of freedom $l(l-1)$. v_a is chosen to make $p(\xi>l(l-1))$ suitably small, i.e.

$$v_a = \frac{l/(l-1)}{\phi_\chi^{-1}(1-p)}$$

where ϕ_χ is the cumulative $\chi^2_{l(l-1)}$ distribution and p is the required probability. As the number of regions increases, the prior variance decreases.

Appendix C

It is of note that truncation of the parameter space from $\Re^p$ to some truncated space, say $\Im^p$, implies truncation of the proposal. Our proposal density must be defined exactly over this space. To define the proposal density adequately at θ^*, a renormalization of the joint proposal density over the new support is required. Such a proposal density would have the form

$$q(\theta^*|\Gamma,\Lambda) = \frac{p(\theta^*|\Gamma,\Lambda)}{I}$$

[14] More general (unconditional) results are possible. In particular, the above conditioning can be relieved by placing distributions over the parameter sets M, Π, Σ_m and integrating them out of a larger hierarchical model.

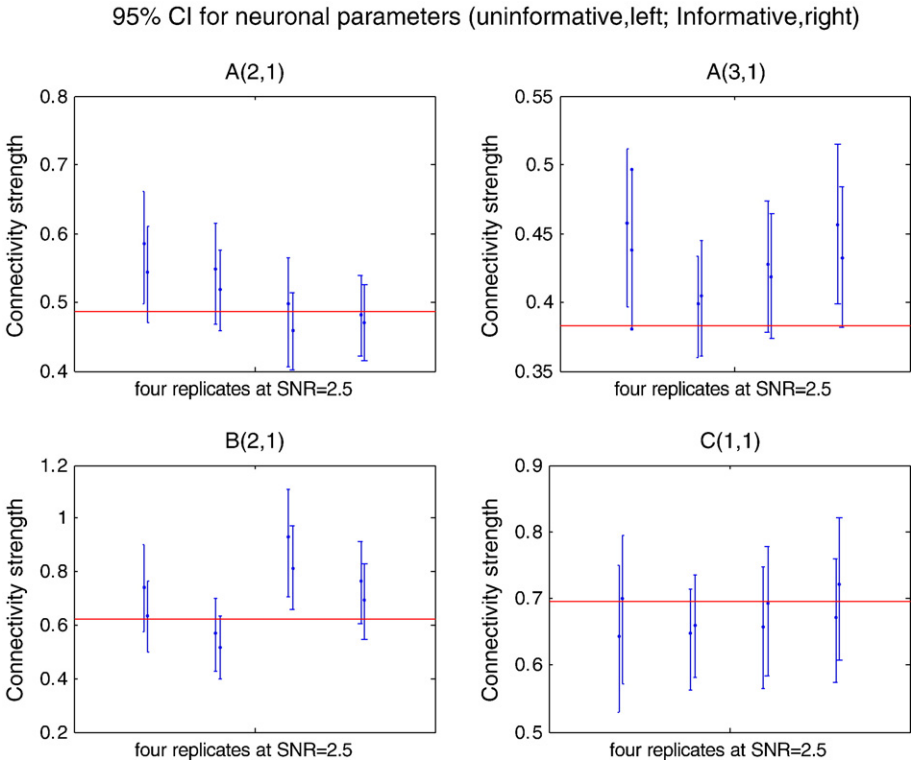

Fig. 6. In each of the 4 subplots, the posterior 95% CI has been reconstructed according to flat (left) versus conventional priors (right) under 4 noise realizations at an intermediate SNR of 2.5.

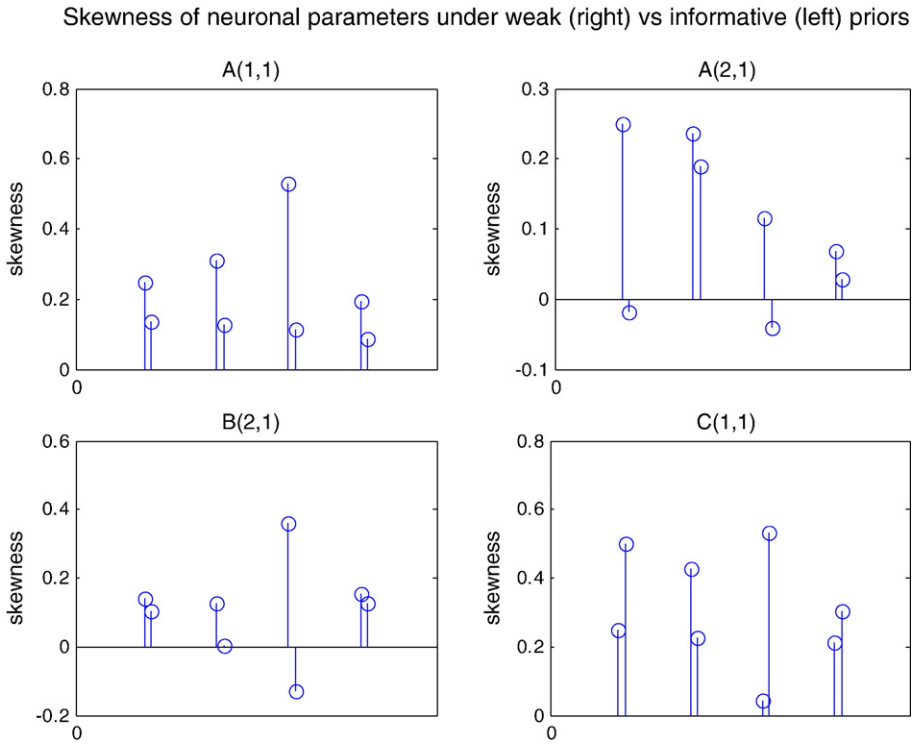

Fig. 7. Skewness of posteriors under uninformative (left of each pair) versus informative (right) at intermediate SNR (2.5). Within each couple, skewness has been calculated from identical data.

where p is normal with mean Γ and covariance matrix Λ, and I is the normalization constant obtained by the integration,

$$I = \int_{\Im^p} p(\theta^*|\Gamma, \Lambda) d\theta^*.$$

To assess the implications of this formulation of the proposal density, we will now examine the nature of I in the context of the MCMC sampling scheme. It is important to remember that iterative sampling relies on a sequence of proposal densities $\{S_i: i=1,\ldots,n\}$. A single (vector-valued) proposal θ_i^* is drawn from each S_i in the sequence. It is usual for sequences to be constructed such that they actively traverse the parameter–space in some manner (e.g., a random walk in which the proposal density S_i is considered normal and centered at the previous accepted parameter value: say, S_{i-1}). The proposal density is thus a function of its position in the sequence. For fixed truncation at the boundary of $\Im^p$, this implies that the normalizing constant I is also variable, say I_i; it is a function of the proposal's position in the sequence. Under these circumstances, in general $I_i \neq I_{i-1}$ and must both be calculated to correctly define

$$h = \frac{q(\theta^{t-1}|\Gamma, \Lambda)}{q(\theta^*|\Gamma, \Lambda)}$$

as required by the MH algorithm (this quantity appears as r, in Step 4 of the algorithm). Calculation of these integrals would clearly induce a costly step in the algorithm. Note, however, two possible remedies: (1) If the sequence of proposal densities $\{S_i: i=1,\ldots,n\}$ is identical, then $I_i = I_{i-1}$ for all i. As a consequence I in the numerator and denominator of h cancel for all h. Alternatively, (2) if the sequence of proposal densities $\{S_{i-1}, S_i\}$ is similar, we can ensure that if $I_i \approx I_{i-1}$. This latter will hold as long as the proposals are away from the boundaries of the truncated parameter–space and quite similar in location (relatively small step-size in the proposal).[15] There are no extant analytic guarantees regarding the general case where this does not hold. Of course, such complications would vanish when a simple MH scheme (with no rejection step) is used to reconstruct the true posterior, as is suitable when the dissipation constraint is enforced probabilistically (as in Friston et al., 2003). Regarding point (1), while "independence sampling" is documented in the MCMC literature, it is plagued by extremely low convergence rate.[16] This study has therefore implemented formulation (2).

References

Beal, M.J., 2003. Variational algorithms for approximate Bayesian inference. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003. (281 pages).

Bera, A.K., Jarque, C.M., Lee, L., 1984. Testing for the normality assumption in the limited dependent variable models. Int. Econ. 563–578.

Friston, 2002. Bayesian estimation of dynamical systems: an application to fMRI. NeuroImage 16, 513–530.

Friston, K.J., Harrison, L., Penny, W., 2003 (Aug). Dynamic causal modelling. NeuroImage 19 (4), 1273–1302.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W., 2007. Variational free energy and the Laplace approximation. NeuroImage 34 (1), 220–234.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2004. Bayesian Data Analysis. Chapman and Hall/CRC.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrica 57, 97–109.

Kumar, S., Stephan, K.E., Warren, J.D., Friston, K.J., Griffiths, T.D., 2007. Hierarchical processing of auditory objects in humans. PLoS Comput Biol. 3 (6) Jun.

Mechelli, A., Price, C.J., Noppeney, U., Friston, K.J., 2003. A dynamic causal modeling study on category effects: bottom-up or top-down mediation? J. Cogn. Neurosci. 15 (7), 925–934.

Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004a. Comparing dynamic causal models. NeuroImage 22 (3), 1157–1172.

Penny, W., Stefan, K.E., Michelli, A., Friston, K.J., 2004b. Comparing dynamic causal models. NeuroImage 22, 1157–1172.

Poirier, D.J., 1998 (August). Revising beliefs in non-identified models. Econ. Theory 14 (04), 483–509.

[15] The difference between the integral over two tail truncations is increasingly negligible if truncations at "similar" places (as defined by small step-size) and both are far into the tail of a normal.

[16] It has also been noted that to achieve a reasonable acceptance rate, the proposal density must have a form that approximately captures the correlation structure in this posterior ("target") distribution. While such an approximation already exists for (namely, joint normal approximations to the posterior estimated by expectation–maximization), we still observe terrible convergence.