NeuroImage 46 (2009) 1004–1017

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

# Bayesian model selection for group studies

Klaas Enno Stephan [a,b,*], Will D. Penny [a], Jean Daunizeau [a], Rosalyn J. Moran [a], Karl J. Friston [a]

[a] Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, 12 Queen Square, London, WC1N 3BG, UK
[b] Laboratory for Social and Neural Systems Research, Institute for Empirical Research in Economics, University of Zurich, Switzerland

ARTICLE INFO

*Article history:*
Received 1 September 2008
Revised 27 February 2009
Accepted 10 March 2009
Available online 20 March 2009

*Keywords:*
Random effects
Variational Bayes
Hierarchical models
Model evidence
Bayes factor
Model comparison
Dynamic causal modelling
DCM
fMRI
EEG
MEG
Source reconstruction

ABSTRACT

Bayesian model selection (BMS) is a powerful method for determining the most likely among a set of competing hypotheses about the mechanisms that generated observed data. BMS has recently found widespread application in neuroimaging, particularly in the context of dynamic causal modelling (DCM). However, so far, combining BMS results from several subjects has relied on simple (fixed effects) metrics, e.g. the group Bayes factor (*GBF*), that do not account for group heterogeneity or outliers. In this paper, we compare the *GBF* with two random effects methods for BMS at the between-subject or group level. These methods provide inference on model-space using a classical and Bayesian perspective respectively. First, a classical (frequentist) approach uses the log model evidence as a subject-specific summary statistic. This enables one to use analysis of variance to test for differences in log-evidences over models, relative to inter-subject differences. We then consider the same problem in Bayesian terms and describe a novel hierarchical model, which is optimised to furnish a probability density on the models themselves. This new variational Bayes method rests on treating the model as a random variable and estimating the parameters of a Dirichlet distribution which describes the probabilities for all models considered. These probabilities then define a multinomial distribution over model space, allowing one to compute how likely it is that a specific model generated the data of a randomly chosen subject as well as the exceedance probability of one model being more likely than any other model. Using empirical and synthetic data, we show that optimising a conditional density of the model probabilities, given the log-evidences for each model over subjects, is more informative and appropriate than both the *GBF* and frequentist tests of the log-evidences. In particular, we found that the hierarchical Bayesian approach is considerably more robust than either of the other approaches in the presence of outliers. We expect that this new random effects method will prove useful for a wide range of group studies, not only in the context of DCM, but also for other modelling endeavours, e.g. comparing different source reconstruction methods for EEG/MEG or selecting among competing computational models of learning and decision-making.



## Introduction

Model comparison and selection is central to the scientific process, in that it allows one to evaluate different hypotheses about the way data are caused (Pitt and Myung, 2002). Nearly all scientific reporting rests upon some form of model comparison, which represents a probabilistic statement about the beliefs in one hypothesis relative to some other(s), given observations or data. The fundamental Neyman–Pearson lemma states that the best statistic upon which to base model selection is simply the probability of observing the data under one model, divided by the probability under another model (Neyman and Pearson, 1933). This is known as a *log-likelihood ratio*. In a classical (frequentist) setting, the distribution of the log-likelihood ratio, under the null hypothesis that there is no difference between models, can be computed relatively easy for some models. Common examples include Wilk's Lambda for linear multivariate models and the *F*- and *t*-statistics for univariate models. In a Bayesian setting, the equivalent to the log-likelihood ratio is the log-evidence ratio, which is commonly known as a *Bayes factor* (Kass and Raftery, 1995). An important property of Bayes factors are that they can deal both with nested and non-nested models. In contrast, frequentist model comparison can be seen as a special case of Bayes factors where, under certain hierarchical restrictions on the models, their null distribution is readily available.

In this paper, we will consider the general case of how to use the model evidence for analyses at the group level, without putting any constraints on the models compared. These models can be non-linear, possibly dynamic and, critically, do not necessarily bear a hierarchical relationship to each other, i.e. they are not necessarily nested. The application domain we have in mind is the comparison of dynamic causal models (DCMs) for fMRI or electrophysiological data (Friston et al., 2003; Stephan et al., 2007a) that have been inverted for each subject. However, the theoretical framework described in this paper can be applied to any model, for example

* Corresponding author. Wellcome Trust Centre for Neuroimaging, Institute of Neurology, UCL, 12 Queen Square, London, WC1N 3BG, UK. Fax: +44 20 7813 1420.
*E-mail address:* k.stephan@fil.ion.ucl.ac.uk (K.E. Stephan).

1053-8119/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2009.03.025

when comparing different source reconstruction methods for EEG/MEG or selecting among competing computational models of learning and decision-making.

This paper is structured as follows. First, to ensure this paper is self-contained, particularly for readers without an in-depth knowledge of Bayesian statistics, we summarise the concept of log-evidence as a measure of model goodness and review commonly used approximations to it, i.e. the Akaike Information Criterion (*AIC*; Akaike, 1974), the Bayesian Information Criterion (*BIC*; Schwarz, 1978), and the negative free-energy (*F*). These approximations, which are described in Appendix A, differ in how they trade-off model fit against model complexity. Given any of these approximations to the log-evidence, we then consider model comparison at the group level. We address this issue both from a classical and Bayesian perspective. First, in a frequentist setting, we consider classical inference on the log-evidences themselves by treating them as summary statistics that reflect the evidence for each model for a given subject. Subsequently, using a hierarchical model and variational Bayes (VB), we describe a novel technique for inference on the conditional density of the models *per se*, given data (or log-evidences) from all subjects. This rests on treating the model as a random variable and estimating the parameters of a Dirichlet distribution, which describes the probabilities for all models considered. These probabilities then define a multinomial distribution over model space, allowing one to compute how likely it is that a specific model generated the data of a subject chosen at random.

We compare and contrast these random effects approaches to the conventional use of the group Bayes factor (*GBF*), an approach for model comparison at the between-subject level that has been used extensively in previous group studies in neuroimaging. For example, the *GBF* has been used frequently to decide between competing dynamic causal models fitted to fMRI (Acs and Greenlee, 2008; Allen et al., 2008; Grol et al., 2007; Heim et al., 2008; Kumar et al., 2007; Leff et al., 2008; Smith et al., 2006; Stephan et al., 2007b,c; Summerfield and Koechlin 2008) and EEG data (Garrido et al., 2007, 2008). While the *GBF* is a simple and straightforward index for model comparison at the group level, it assumes that all the subjects' data are generated by the same model (i.e. a fixed effects approach) and can be influenced adversely by violations of this assumption.

The novel Bayesian framework presented in this paper does not suffer from these shortcomings: it can quantify the probability that a particular model generated the data for any randomly selected subject, relative to other models, and it is robust to the presence of outliers. In the analyses below, we illustrate the advantages of this new approach using synthetic and empirical data. We show that computing a conditional density of the model probabilities, given the log-evidences for all subjects, can be superior to both the *GBF* and frequentist tests applied to the log-evidences. In particular, we found that our Bayesian approach is markedly more robust than either of the other approaches in the presence of outlying subjects.

## Methods

### *The model evidence and its approximations*

The model evidence $p(y|m)$ is the probability of obtaining observed data $y$ given a particular model $m$. It can be considered the holy grail of any model inversion and is necessary to compare different models or hypotheses. The evidence for some models can be computed relatively easily (e.g., for linear models); however, in general, computing the model evidence entails integrating out any dependency on the model parameters $\vartheta$:

$$p(y|m) = \int p(y|\vartheta,m)p(\vartheta|m)d\vartheta. \quad (1)$$

In many cases, this integration is analytically intractable and numerically difficult to compute. Usually, it is therefore necessary to use computationally tractable approximations to the model evidence (or the log-evidence[1]). A detailed description of some of the most common approximations is contained by Appendix A.

A systematic evaluation of the relative usefulness of different approximations to the log-evidence is not at the focus of this paper and will be presented in forthcoming work. This article deals with a different question, namely: given a particular approximation to the log-evidence and a number of inverted models, how can we infer which of several competing models is most likely to have generated the data from a group of subjects? In other words, how can we make inference on model space at the group level, taking into account potential heterogeneity across the group?

### *Inference on model space*

In this section, we consider inference at the group level, using subject-specific model-evidences obtained by inverting a generative model for each subject. We will first describe a classical approach, testing the null hypothesis that there are no differences among the relative log-evidences for various models over subjects. We then move on to more formal Bayesian inference on model space *per se*. In contrast to the *GBF*, which, as described above, represents a fixed effects analysis, both the classical and Bayesian approaches are random effects procedures and thus consider inter-subject heterogeneity explicitly.

#### *Classical (frequentist) inference*

A straightforward random effects procedure to evaluate the between-subject consistency of evidence for one model relative to others is to use the log-evidences across subjects as the basis for a classical log-likelihood ratio statistic, testing the null hypothesis that no single model is better (in terms of their log-evidences) than any other. This essentially involves performing an ANOVA, using the log-evidence as a summary statistic of model adequacy for each subject. This ANOVA then compares the differences among models to the differences among subjects with a classical *F*-statistic. If this statistic is significant one can then compare the best model with the second best using a *post hoc t*-test. Effectively, this tests for differences between models that are consistent and large in relation to differences within models over subjects. The most general implementation would be a repeated-measures ANOVA, where the log-evidences for the different models represent the repeated measure. At its simplest, the comparison of just two models over subjects reduces to a simple paired *t*-test on the log-evidences (or a one-sample *t*-test on the log-evidence differences). Log-evidences tend to be fairly well behaved, and the residuals of a simple ANOVA model, or tests of normality like Kolmogorov–Smirnoff, usually indicate that parametric assumptions are appropriate. In those cases when they are not, e.g. due to outlier subjects, one can use robust regression methods that are less sensitive to violations of normality (Diedrichsen and Shadmehr, 2005; Wager et al., 2005a,b) or non-parametric tests that do not make any distributional assumptions (e.g. a Wilcoxon signed rank test; see one of our examples below).

This classical random effects approach is simple to implement, straightforward and easily interpreted. In this sense, there seems little reason not to use it. However, as shown in the empirical examples below, this type of inference can be affected markedly by group heterogeneity, even when the distribution of log-evidence differences is normal. A more robust analysis is obtained by quantifying the

[1] Due to the monotonic nature of the logarithmic function, model comparisons yield equivalent results regardless whether one maximises the model evidence or the log-evidence. Since the latter is numerically easier, it is usually the preferred metric.

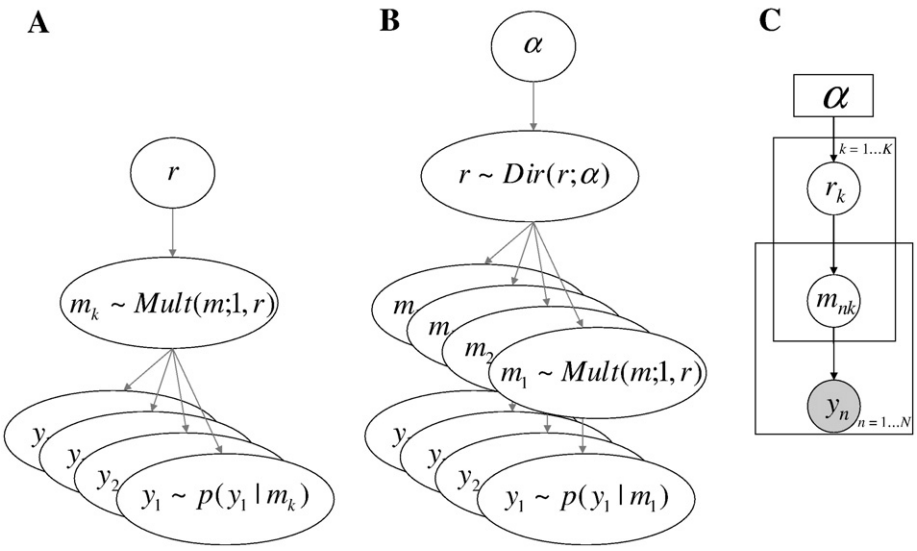


**Fig. 1.** Bayesian dependency graphs for fixed effects (A) and random effects generative models for multi-subject data (B, C). The graphical model in panels B and C are equivalent; we show both because 1B is more intuitive for readers unfamiliar with graphical models whereas 1C uses a more compact notation where rectangles denote deterministic parameters and shaded circles represent observed values. $\alpha$ = parameters of the Dirichlet distribution (number of model "occurrences"); $r$ = parameters of the multinomial distribution (probabilities of the models); $m$ = model labels; $y$ = observed data; $k$ = model index; $K$ = number of models; $n$ = subject index; $N$ = number of subjects.

density on model space itself, using a Bayesian approach as described in the next section.

*Bayesian inference on model space*

Previously, we have suggested the use of a group Bayes factor (*GBF*) that is simply the product of Bayes factors over $N$ subjects (Stephan et al., 2007b). This is equivalent to a fixed effects analysis that rests on multiplying the marginal likelihoods over subjects to furnish the probability of the multi-subject data, conditioned on each model:

$$GBF_{i,j} = \prod_{n=1}^{N} BF_{i,j}^{(n)}. \tag{2}$$

Here, the subscripts $i,j$ refer to the models being compared, and the bracketed superscript refers to the $n$-th subject. The reason one can simply multiply the probabilities (or add the log-evidences) is that the measured data can be regarded as conditionally independent samples over subjects. However, this does not represent a formal evaluation of the conditional density of a particular model given data from all subjects. Furthermore, it rests upon a very particular generative model for group data: first, select one of $K$ models from a multinomial distribution and then generate data, under this model, for each of the $N$ subjects. This is fundamentally different from a generative model which treats subjects as random effects: here we would select a model for each subject by sampling from a multinomial distribution, and then generate data under that subject-specific model. The distinction between these two generative models is illustrated graphically in Fig. 1.

In short, the *GBF* encodes the relative probability that the data were generated by one model relative to another, assuming the data were generated *by the same model* for all subjects. What we often want, however, is the density from which models are sampled to generate subject-specific data. In other words, we seek the conditional estimates of the multinomial parameters, i.e. the model probabilities $r = [r_1,\ldots,r_K]$, that generate switches or indicator variables, $m_n = [m_{n1},\ldots,m_{nK}]$, where $m_{nk} \in \{0,1\}$ for any given subject $n \in \{1,\ldots,N\}$, and only one of these switches is equal to one; i.e., $\sum_{k=1}^{K} m_{nk} = 1$. These indicator variables prescribe the model for the $n$-th subject; where $p(m_{nk} = 1) = r_k$. In the following, we describe a hierarchical Bayesian model that can be inverted to obtain an estimate of the posterior density over $r$.

*A variational Bayesian approach for inferring model probabilities*

We will deal with $K$ models with probabilities $r = [r_1,\ldots,r_K]$ that are described by a Dirichlet distribution:

$$p(r|\alpha) = Dir(r,\alpha) = \frac{1}{Z(\alpha)} \prod_k r_k^{\alpha_k - 1}$$

$$Z(\alpha) = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}. \tag{3}$$

Here, $\alpha = [\alpha_1,\ldots,\alpha_K]$ are related to the unobserved "occurrences" of models in the population; i.e. $\alpha_k - 1$ can be thought of as the effective number of subjects in which model $k$ generated the observed data. Given the probabilities $r$, the distribution of the multinomial variable $m_n$ describes the probability that model $k$ generated the data of subject $n$:

$$p(m_n|r) = \prod_k r_k^{m_{nk}}. \tag{4}$$

For any given subject $n$, we can sample from this multinomial distribution to obtain a particular model $k$. The marginal likelihood of the data in the $n$-th subject, given this model $k$, is then obtained by integrating over the parameters of the model selected:

$$p(y_n|m_{nk}) = \int p(y|\vartheta) p(\vartheta|m_{nk}) d\vartheta. \tag{5}$$

The graphical model summarising the dependencies among $r$, $m$ and $y$ as described by Eqs. (3)–(5) is shown in Figs. 1B and C. Our goal is to invert this hierarchical model and estimate the posterior distribution over $r$.

Given the structure of the hierarchical model in Fig. 1, the joint probability of the parameters and the data $y$ can be written as:

$$\begin{aligned} p(y,r,m) &= p(y|m)p(m|r)p(r|\alpha_0) \\ &= p(r|\alpha_0)\left[\prod_n p(y_n|m_n)p(m_n|r)\right] \\ &= \frac{1}{Z(\alpha_0)}\left[\prod_k r_k^{\alpha_{0k}-1}\right]\left[\prod_n p(y_n|m_n)\prod_k r_k^{m_{nk}}\right] \\ &= \frac{1}{Z(\alpha_0)}\prod_n\left[\prod_k [p(y_n|m_{nk})r_k]^{m_{nk}} r_k^{\alpha_{0k}-1}\right]. \end{aligned} \tag{6}$$

The log joint probability is therefore given by:

$$\ln p(y,r,m) = -\ln Z(\alpha_0) + \sum_n \sum_k ((\alpha_{0k} - 1)\ln r_k + m_{nk}(\ln\ p(y_n|m_{nk}) + \ln r_k)). \quad (7)$$

The inversion of our hierarchical model relies on the following variational Bayesian (VB) approach in which we assume that an approximate posterior density $q$ can be described by the following mean-field factorisation:

$$\begin{aligned} q(r,m) &= q(r)q(m) \\ q(r) &\propto \exp(I(r)) \\ q(m) &\propto \exp(I(m)) \\ I(r) &= \langle \ln\ p(y,r,m)\rangle_{q(m)} \\ I(m) &= \langle \ln\ p(y,r,m)\rangle_{q(r)}. \end{aligned} \quad (8)$$

Here, $I(r)$ and $I(r)$ are variational energies for the mean-field partition. Note that throughout the paper we use "log" and "ln" interchangeably to refer to the natural logarithm. The mean-field assumption in Eq. (8) means that the VB posterior will only be approximate but, as we shall see, it provides a particularly simple and intuitive algorithm (c.f. Eq. (14)). This algorithm provides precise estimates of the parameters $\alpha$ defining the approximate Dirichlet posterior $q(r)\approx p(r|y)$; this was verified by comparisons with a sampling method which is described in Appendix B.

To obtain the approximate posterior $q(m)\approx p(m|y)$, we have to do two things: first, compute $I(m)$ and second, determine the normalizing constant or partition function for $\exp(I(m))$, which renders $q(m)$ a probability density. Making use of the log joint probability in Eq. (7) and omitting terms that do not depend on $m$, the variational energy is:

$$\begin{aligned} I(m) &= \int q(r)\ln p(y,r,m)dr \\ &= \sum_n \sum_k m_{nk}(\ln p(y_n|m_{nk}) + \int q(r_k)\ln r_k dr_k) \\ &= \sum_n \sum_k m_{nk}(\ln p(y_n|m_{nk}) + \Psi(\alpha_k) - \Psi(\alpha_S)). \end{aligned} \quad (9)$$

Here, $\alpha_S = \sum_k \alpha_k$ and $\Psi$ is the digamma function.[2]

$$\Psi(\alpha_k) = \frac{\partial \ln\ \Gamma(\alpha_k)}{\partial \alpha_k}. \quad (10)$$

The next step is to obtain the approximate posterior, $q(m)$: If $g_{nk}$ is our (normalized) posterior belief that model $k$ generated the data from subject $n$, i.e. $g_{nk} = q(m_{nk} = 1)$, then Eq. (9) tells us that:

$$\begin{aligned} g_{nk} &= \frac{u_{nk}}{u_n} \\ u_{nk} &= \exp(\ln p(y_n|m_{nk}) + \Psi(\alpha_k) - \Psi(\alpha_S)) \\ u_n &= \sum_k u_{nk} \end{aligned} \quad (11)$$

where $u_{nk}$ is the equivalent (non-normalized) belief and $u_n$ is the partition function for $\exp(I(m))$ that ensures that the posterior probabilities sum to one.

We now repeat the above procedure but this time for the approximate posterior over $r$. By substituting in the log joint probability from Eq. (7) and omitting terms that do not depend on $r$, we have:

$$\begin{aligned} I(r) &= \int q(m)\ln\ p(y,r,m)dm \\ &= \sum_k \left[(\alpha_{0k} - 1)\ln r_k + \sum_n g_{nk}\ln\ r_k\right] \\ &= \sum_k (\alpha_{0k} + \beta_k - 1)\ln r_k. \end{aligned} \quad (12)$$

Here, $\beta_k = \Sigma g_{nk}$ is the expected number of subjects whose data we believe were generated by model $k$. Now, from Eq. (8) we have $\ln q(r) = I(r) + \ldots$ and from Eq. (3) we see that the log of a Dirichlet density is given by $\ln Dir(r;a) = \sum_k (\alpha_k - 1)\ln r_k + \ldots$. Hence, by comparing terms we see that the approximate posterior $q(r) = \mathrm{Dir}(r;\alpha)$ where:

$$\alpha = \alpha_0 + \beta. \quad (13)$$

In short, Eq. (13) simply adds the 'data counts', $\beta$, to the 'prior counts', $\alpha_0$. This is an example of a free-form VB approximation, where the optimal form of the approximate posterior (in this case a Dirichlet), has been derived rather than assumed before-hand (c.f. fixed-form VB approximations; Friston et al., 2007). It should be stressed, however, that due to the mean-field assumption used by our VB approach (see Eq. (8)), $q(r)$ is only an approximate posterior and the true posterior distribution $p(r|y)$ does not necessarily have the exact form of a Dirichlet distribution.

The above equations can be implemented as an optimisation algorithm which updates estimates of $\alpha$ iteratively until convergence. By combining Eqs. (11), (12) and (13) we get the following pseudo-code of a simple algorithm that gives us the parameters of the conditional density we seek, i.e. $q(r) = \mathrm{Dir}(r;\alpha)$:

$\alpha = \alpha_0$.

Until convergence:

$$\begin{aligned} u_{nk} &= \exp\left(\ln p(y_n|m_{nk}) + \Psi(\alpha_k) - \Psi\left(\sum_k \alpha_k\right)\right) \\ \beta_k &= \sum_n \frac{u_{nk}}{\sum_k u_{nk}} \\ \alpha &= \alpha_0 + \beta \end{aligned} \quad (14)$$

end.

We make the usual assumption that, *a priori*; no models have been "seen" (i.e. the Dirichlet prior is $\alpha_0 = [1,\ldots,1]$).[3] Critically, this scheme requires only the log-evidences over models and subjects (c.f. Eq. (11)).

*Using the Dirichlet density p(r|y; α) for model comparison*

After the above optimisation of the Dirichlet parameters, $\alpha$, the Dirichlet density $p(r|y;\alpha)$ can be used for model comparisons at the group level. There are several ways to report this comparison that

[2] See Appendix B in Bishop (2006) concerning the use of the digamma function in Eq. 10.

[3] Note that this choice of Dirichlet prior is a "flat" prior, assigning uniform probabilities to all models. In contrast, a Dirichlet prior with elements below unity results in a highly concave probability density that concentrates the probability mass around zero and one, respectively.

result in equivalent model rankings. The simplest option is to report the estimates of the Dirichlet parameter estimates $\alpha$. Another possibility is to use those estimates to compute the expected multinomial parameters $\langle r_k \rangle$ and thus the expected likelihood of obtaining the $k$-th model, i.e. $p(m_{nk}=1|r)=\mathrm{Mult}(m;\ 1,r)$, for any randomly selected subject[4]:

$$\langle r_k \rangle_q = \alpha_k / (\alpha_1 + \ldots + \alpha_K). \tag{15}$$

A third option is to use the conditional model probability $p(r|y;\alpha)$ to quantify an *exceedance probability*, i.e. our belief that a particular model $k$ is more likely than any other model (of the $K$ models tested), given the group data:

$$\forall j \in \{1 \ldots K | j \neq k\}:$$
$$\varphi_k = p\left(r_k > r_j | y;\alpha\right). \tag{16}$$

The exceedance probabilities $\varphi_k$ sum to one over all models tested. They are particularly intuitive when comparing two models (or model subsets, see below). In this case, because the conditional probabilities of the models $\langle r_k \rangle$ also sum to one, the exceedance probability of one model, compared to another, can be written as:

$$\begin{aligned}\varphi_1 &= p(r_1 > r_2 | y;\alpha) \\ &= p(r_1 > 0.5 | y;\alpha).\end{aligned} \tag{17}$$

The analyses of empirical data below include several examples where two models are compared; the associated exceedance probabilities are shown in Figs. 3, 6, 9 and 13.

Either the Dirichlet parameter estimates $\alpha$, the conditional expectations of model probabilities $\langle r_k \rangle$ or the exceedance probabilities $\varphi_k$ can be used to rank models at the group level. In the next section, we present several practical examples of our method, applying it to both synthetic and empirical data. In this paper, we focus on comparing two models (or two model subsets) and largely rely on exceedance probabilities when discussing the results of our analyses. However, for each analysis we also report the estimates of $\alpha$ and the conditional expectations of model probabilities, $\langle r_k \rangle$; these are shown in the figures.

*Model space partitioning*

A particular strength of the approach presented in this paper is that it cannot only be used to compare specific models, but also to compare particular classes or subsets of models, resulting from a partition of model space. For example, one may want to compute the probability that a specific model attribute, say the presence vs. absence of a particular connection in a DCM, improves or reduces model performance, regardless of any other differences among the models considered. This type of inference rests on comparing two (or more) subsets of model space, pooling information over all models in these subsets. This effectively removes uncertainty about any aspect of model structure, other than the attribute of interest (which defines the partition). Heuristically, this sort of analysis can be considered a Bayesian analogue of tests for "main effects" in classical ANOVA.

[4] For the special case of "drawing" a single "sample" (model), the multinomial distribution of models reduces to $p(m_{nk}=1|r)=r_k$. Therefore, for any given subject, $\langle r_k \rangle$ represents the conditional expectation that the $k$-th model generated the subject's data.

Within our framework this type of analysis can be performed by exploiting the agglomerative property of the Dirichlet distribution. Generally, for any partition of model space into $J$ disjoint subsets, $N_1$, $N_2$,…,$N_J$, this property ensures that:

$$\begin{aligned}&(r_1,r_2,\ldots,r_K) \sim Dir(\alpha_1,\alpha_2,\ldots,\alpha_K) \\ &\Rightarrow r_1^* = \sum_{k\in N_1} r_k, r_2^* = \sum_{k\in N_2} r_k, \ldots, r_J^* = \sum_{k\in N_J} r_k \\ &\sim Dir\left(\alpha_1^* = \sum_{k\in N_1} \alpha_k, \alpha_2^* = \sum_{k\in N_2} \alpha_k, \ldots, \alpha_J^* = \sum_{k\in N_J} \alpha_k\right).\end{aligned} \tag{18}$$

In other words, once we have estimates of the Dirichlet parameters $\alpha_k$ for all $K$ models, it is easy to evaluate the relative importance of different model subspaces: for any given partition of model space, a new Dirichlet density reflecting this partition can be defined by simply adding $\alpha_k$ for all models $k$ belonging to the same subset. The resulting Dirichlet can then be used to compare different subsets of model space in exactly the same way as one compares individual models, e.g. using exceedance probabilities. An example of this application is shown in Figs. 12 and 13.

## Results

In what follows, we compare classical inference, the *GBF* (fixed effects) and inference on model space (random effects) using both synthetic and real data. These data have been previously published and have been analysed in various ways, including group level model inference using *GBF*s (Stephan et al., 2007b,c, 2008).

*Synthetic data: nonlinear vs. bilinear modulation*

To demonstrate the face validity of our method, we used simulated data, where the true model was known. Specifically, we used one of

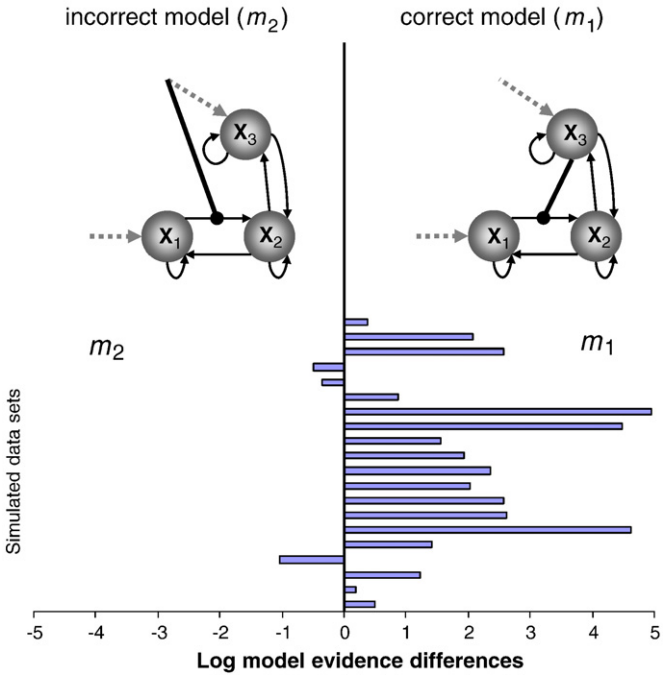


**Fig. 2.** Synthetic data consisting of twenty time-series that were generated using a three-area nonlinear DCM and adding random observation noise (see Stephan et al., 2008 for details). To each of these time-series, two models were fitted and compared: (i) a nonlinear DCM with the same structure as the model that generated the data ("correct model" $m_1$), and a bilinear model ("incorrect model" $m_2$). The difference in log-evidences for all twenty data sets is plotted as a bar chart.

the synthetic data sets described by Stephan et al. (2008), consisting of twenty synthetic BOLD time-series that were generated using a three-area nonlinear DCM with fixed parameters and adding Gaussian observation noise to achieve a signal-to-noise ratio (SNR) of two. Each time-series consisted of 100 data points that were obtained by sampling the model output at a frequency of 1 Hz over a period of 100 s. For each time-series, we fitted (i) a nonlinear DCM with the same model structure as the model that generated the data ("correct model" in Fig. 2, model $m_1$), and (ii) a second DCM that was similar in structure but included a bilinear (instead of a nonlinear) modulatory influence ("incorrect model" in Fig. 2, model $m_2$). Using the negative free-energy approximation to the log-evidence, the differences in log-evidences for all twenty time-series are plotted in the lower part of Fig. 2. It can be seen that in 17 out of 20 cases the nonlinear model was correctly identified as the more likely model. The overall *GBF* ($9\times10^{14}$) was also clearly in favour of the correct model.

Here, we revisit this synthetic data set using random effects BMS procedures. We first used classical inference, applying a paired *t*-test to the log-evidences of the two models. This test rejected the null hypothesis of no difference in model goodness ($t=4.615$, $df=19$, $p<10^{-4}$). Applying the novel hierarchical BMS approach gave an even clearer (and arguably also more useful) answer: the exceedance probability $\varphi_1$, i.e. the probability of $m_1$ being a more likely model than $m_2$, was 100% (Fig. 3). In other words, using the exceedance probability $\varphi$ as a criterion, the correct model was identified perfectly, given all twenty data sets and the chosen level of noise. To further corroborate this result, we compared the result from our VB algorithm to an independent method which estimates the parameters $\alpha$ by sampling from the approximate Dirichlet posterior $q(r)\approx p(r|y)$. This comparison showed that the VB estimate of $\alpha$ resulted in an estimate of the negative free-energy $F(y,\alpha)\leq\ln p(y|\alpha)$ that was consistent with the results from the sampling approach (Fig. 4). This provides an additional validation of our VB technique. We used this sampling approach to verify the correctness of our VB estimates in all subsequent analyses.

It should be noted that this simulation study concerned the extreme case that only one model had generated all data, i.e. $r_1=100\%$ and $r_2=0\%$, making it easy to intuitively understand the performance of the proposed model selection procedure. However, this simulation did not probe the robustness of our method when randomly sampling from a heterogeneous population of subjects whose data had been generated by different models. We will revisit this scenario in a later section of this paper once we have introduced and compared two alternative DCMs of inter-hemispheric interactions using empirical data.

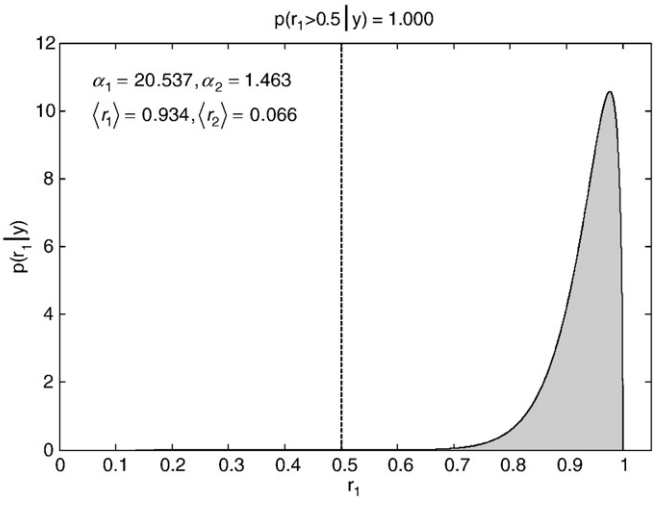


**Fig. 3.** The Dirichlet density describing the probability of the nonlinear model $m_1$ in Fig. 2 given the synthetic data across the 20 realisations. The shaded area represents the exceedance probability $\varphi_1$ of $m_1$ being a more likely model than the (incorrect) bilinear model $m_2$ (compare Fig. 2). $\alpha$ = VB estimates of the Dirichlet parameters; $\langle r_1\rangle$, $\langle r_2\rangle$ = conditional expectations of the probabilities of the two models.

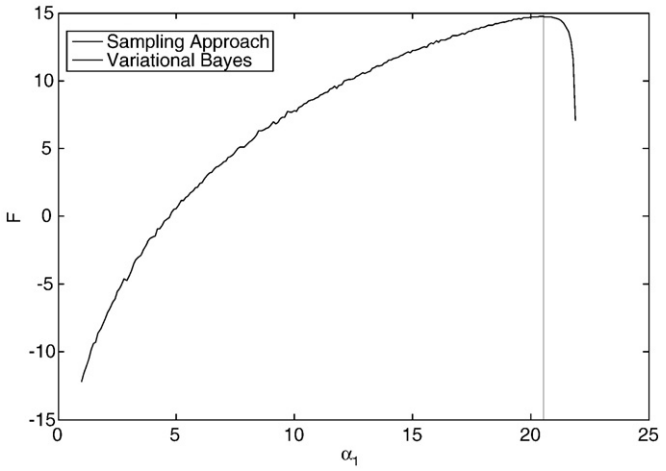


**Fig. 4.** Confirmation of our VB estimate for $\alpha_1$ (vertical dotted line) in Fig. 3 by comparing it against the result obtained by a sampling approach (solid line); see main text for details.

*Comparing different six-area DCMs of the ventral visual stream*

As a first empirical application, we investigated a case we had encountered in our previous research (Stephan et al., 2007b) and which had actually triggered our interest in developing more powerful group level inference about models. This model comparison concerned DCMs describing alternative mechanisms of inter-hemispheric integration in terms of context-dependent modulation of connections. In one of the analyses of the original report (Stephan et al., 2007b), competing DCMs had been constructed for the ventral stream of the visual system by systematically changing which of the experimentally controlled conditions modulated the intra- and/or the inter-hemispheric connections.

First, we focused on the six-area model of the ventral stream, comprising the lingual gyrus (LG), middle occipital gyrus (MOG) and fusiform gyrus (FG) in both hemispheres, and revisited the comparison of the best two models as indexed by the *GBF*. In the first model, $m_1$, inter-hemispheric connections were modulated by a letter decision task, but conditional on the visual field of stimulus presentation (LD|VF); intra-hemispheric connections were modulated by LD alone (see right side of Fig. 5). In the second model, $m_2$, these modulations were reversed: inter-hemispheric connections were modulated by LD and intra-hemispheric connections were modulated by LD|VF (see left side of Fig. 5). The distribution of log-evidence differences (approximated by *AIC*/*BIC*, following the procedure suggested by Penny et al., 2004) is shown in the centre of Fig. 5: Although $m_1$ was robustly superior in 11 of the 12 subjects, a single outlier was so extreme that the *GBF* indicated an overall superiority of $m_2$ (*GBF* = 15 in favour of $m_2$). In contrast, model comparison using our novel Bayesian method was not affected by this outlier: the exceedance probability in favour of $m_1$ was very high ($\varphi_1=99.7\%$), and the conditional expectation $\langle r_1\rangle$ that $m_1$ generated the data of any randomly selected subject was 84.3% (Fig. 6). The estimates of our VB method were confirmed by the sampling approach (Fig. 7).

For comparison, we also applied frequentist statistics to the log-evidences as described above. The single outlier subject made the distribution of the log-evidence differences non-normal (Kolmogorov–Smirnov test: $p<10^{-7}$, $D_N=0.822$), and thus prevented

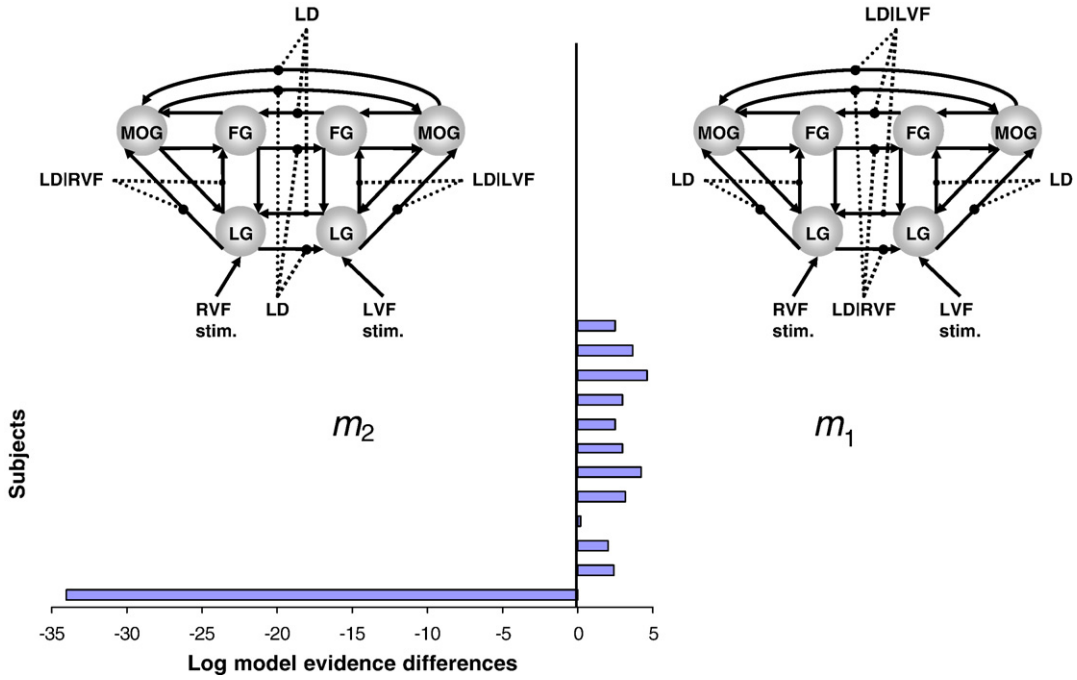


**Fig. 5.** Comparison of DCMs describing alternative mechanisms of inter-hemispheric integration in terms of context-dependent modulation of connections (Stephan et al., 2007b). Two variants of a six-area model of the ventral stream, comprising the lingual gyrus (LG), middle occipital gyrus (MOG) and fusiform gyrus (FG) in both hemispheres were compared. In the first model, $m_1$, inter-hemispheric connections were modulated by a letter decision (LD) task, but conditional on the visual field of stimulus presentation (LD|VF); intra-hemispheric connections were modulated by LD alone. In the second model, $m_2$, these modulations were reversed: inter-hemispheric connections were modulated by LD and intra-hemispheric connections were modulated by LD|VF alone. The distribution of log-evidence differences across the 12 subjects is shown at the bottom: although $m_1$ was superior in 11 of the 12 subjects, a single outlier was so extreme that model comparison based on the *GBF* favoured $m_2$ (*GBF* = 15 in favour of $m_2$).

detection of a significant difference between the two models by a one-tailed paired *t*-test ($t = 0.073$, $df = 11$, $p = 0.471$). Given this deviation from normality, we applied a nonparametric Wilcoxon signed rank test which makes no distributional assumptions; this test was indeed able to find a significant difference between the models ($p = 0.034$).

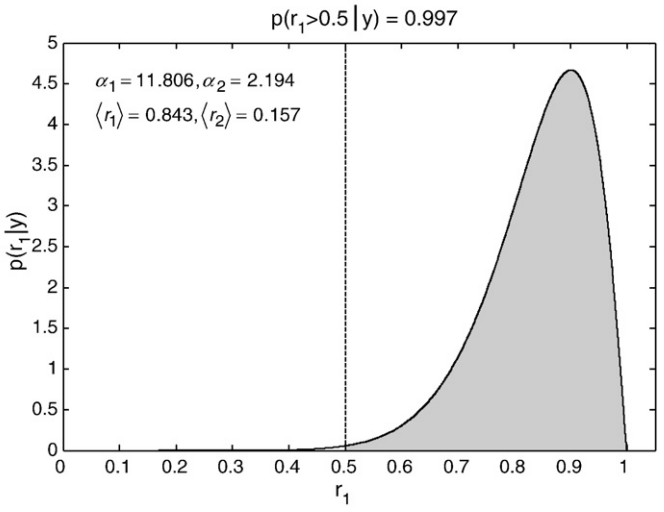


**Fig. 6.** The Dirichlet density describing the probability of model $m_1$ in Fig. 5 given the measured data across the group. The shaded area represents the exceedance probability $\varphi_1 = p(r_1 > 0.5|y; \alpha)$ of $m_1$ being a more likely model than the alternative model $m_2$ (compare Fig. 5). In contrast to the conventional *GBF* or inference based on frequentist statistics, our variational Bayesian method was not affected by the strong outlier subject shown by Fig. 5: the exceedance probability in favour of $m_1$ was $\varphi_1 = 99.7\%$.

*Comparing different four-area DCMs of the ventral visual stream*

Next, we investigated a variant of the previous case where the distribution of log-evidences across subjects was more heterogeneous. This model comparison was essentially identical to the previous one, except that the models in question only contained four areas (LG and FG in both hemispheres), instead of six. Visual inspection of the distribution of log-evidence differences (Fig. 8) shows that the same subject as in the previous example favoured $m_2$, albeit far less strongly; in addition, three more subjects showed evidence in favour of $m_2$, albeit only weakly. Given this constellation, the original analysis by Stephan et al. (2007b) only found a relatively

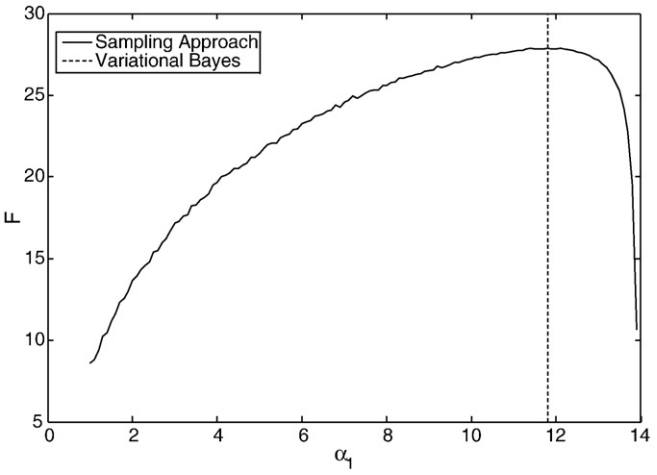


**Fig. 7.** Confirmation of our VB estimate for $\alpha_1$ (vertical dotted line) in Fig. 6 by comparing it against the result obtained by a sampling approach (solid line); see main text for details.

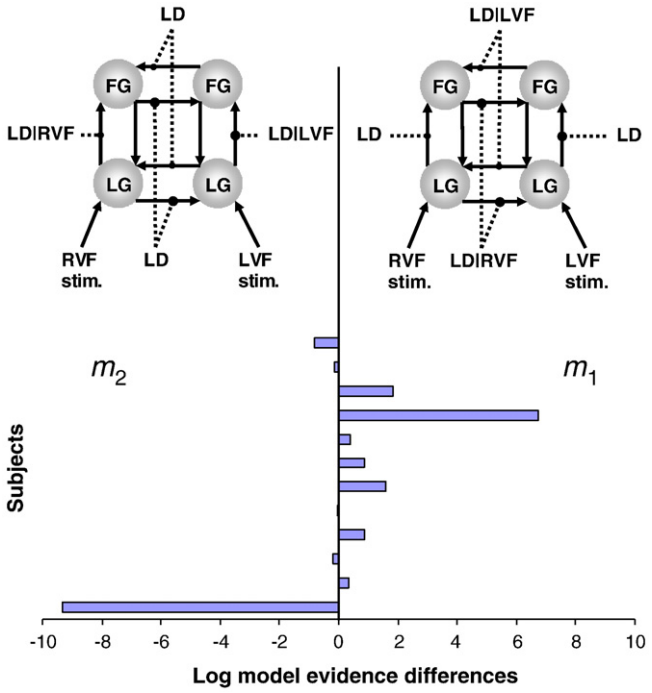


**Fig. 8.** A variant of the model comparison shown by Fig. 5; here the models in question contained four areas (LG and FG in both hemispheres). The distribution of log-evidence differences shows that the same subject as in Fig. 5 constituted an outlier; in addition three more subjects showed weak evidence in favour of $m_2$.

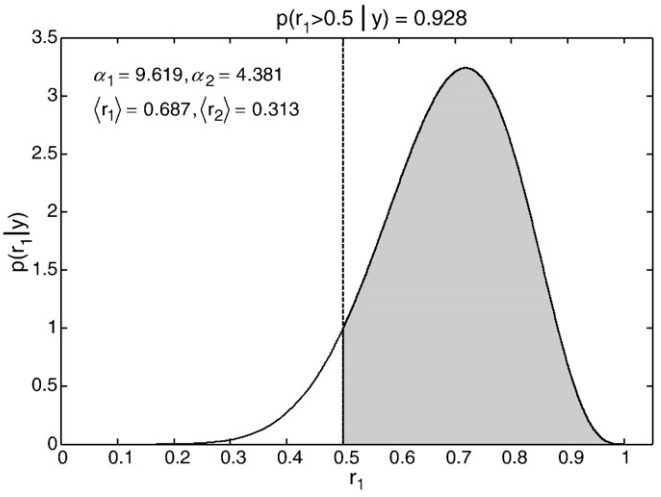


**Fig. 9.** The Dirichlet density describing the probability of model $m_1$ in Fig. 8 given the measured data across the group. The shaded area represents the exceedance probability $\varphi_1 = p(r_1 > 0.5|y; \alpha)$ of $m_1$ being a more likely model than the alternative model $m_2$ (compare Fig. 8). Despite the strong outlier subject shown by Fig. 8, the exceedance probability of $\varphi_1 = 92.8\%$ was favouring $m_1$ as a more likely model than $m_2$.

weak superiority of $m_1$ ($GBF = 8$). In contrast, the VB method gave an exceedance probability of $\varphi_1 = 92.8\%$ in favour of $m_1$, indicating more clearly that $m_1$ is a superior model (Fig. 9). As above, the estimates of our VB method were confirmed by sampling (Fig. 10).

When comparing this result to the frequentist random effects approach, a one-tailed paired $t$-test was unable to detect a significant difference between the two models ($t = 0.165$, $df = 11$, $p = 0.436$). In contrast to the previous example, this failure was not due to outlier-induced deviations from normality: a Kolmogorov–Smirnov test applied to the log-evidences was unable to reject the null hypothesis that they were normally distributed ($p = 0.743$). Here, the between-subject variability, while in accordance with normality assumptions, was simply too large to reject the null hypothesis with the classical $t$-test. A nonparametric Wilcoxon signed rank test did not fare any better ($p = 0.266$).

*Synthetic data: randomly sampling from a heterogeneous population*

In a second simulation study, we examined the robustness of our method when randomly sampling from a heterogeneous population of subjects. Specifically, we dealt with a population in which 70% of subjects showed brain responses as generated by model $m_1$ shown in Fig. 8, whereas brain activity in the remaining 30% of the population was generated by model $m_2$. We randomly sampled 20 subjects from this population and generated synthetic fMRI data by integrating the state equations of the associated models with fixed parameters and inputs[5] and adding Gaussian observation noise to achieve an SNR of two. Each synthetic data set had exactly the same structure as the empirical data described in the previous section (700 data points, TR = 3 s). Both $m_1$ and $m_2$ were then fitted to all 20 synthetic data sets, and the resulting log-evidences were used to perform both fixed effects BMS and random effects BMS, using the VB method described in this paper. This sampling and data generation procedure was repeated 20 times, resulting in a total of 400 generated data sets and 800 fitted models. For each of the 20 sets of 20 subjects, we computed the different indices provided by random effects BMS (i.e., $\alpha$, $\langle r \rangle$, $\varphi$) and fixed effects BMS (log GBF). The means of these indices are plotted in Fig. 11, together with 95% confidence intervals (CI). If our random effects BMS method were perfect in uncovering the underlying structure of the population we sampled from, one would expect to find the following average estimates: (i) $\alpha_1 = 22 \times 0.7 = 15.4$, $\alpha_2 = 22 \times 0.3 = 6.6$ for the Dirichlet parameters, (ii) $\langle r_1 \rangle = 0.7$, $\langle r_2 \rangle = 0.3$ for the posterior expectations of model probabilities, and (iii) $\varphi_1 = 1$, $\varphi_2 = 0$ as exceedance probabilities (note that the exceedance probability is not the posterior model probability itself, but a statement of belief about the posterior probability of one model being higher than the posterior probability of any other model). The actual estimates of the BMS indices for the simulated data were (i) $\alpha_1 = 15.4$ (CI: 14.1–16.7) and $\alpha_2 = 6.6$ (CI: 5.3–7.9), (ii) $\langle r_1 \rangle = 0.7$ (CI: 0.64–0.76) and $\langle r_2 \rangle = 0.3$ (CI: 0.24–0.36), and (iii) $\varphi_1 = 0.89$ (CI: 0.83–0.96) and $\varphi_1 = 0.11$ (CI: 0.04–0.17). For

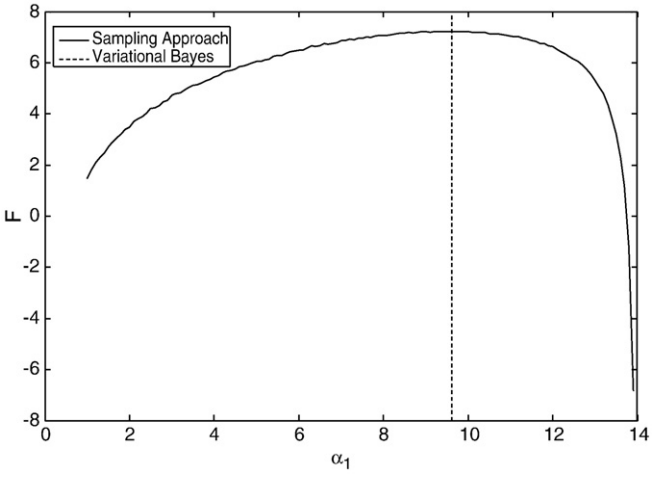


**Fig. 10.** Confirmation of our VB estimate for $\alpha_1$ (vertical dotted line) in Fig. 9 by comparing it against the result obtained by a sampling approach (solid line); see main text for details.

[5] The coupling parameters of all endogenous connections were set to 0.1 $s^{-1}$, except for the inhibitory self-connections whose strengths were set to $-1$ $s^{-1}$. Furthermore, the strengths of all modulatory and driving inputs were set to 0.3 $s^{-1}$. The input functions were the same as in the empirical dataset described above.

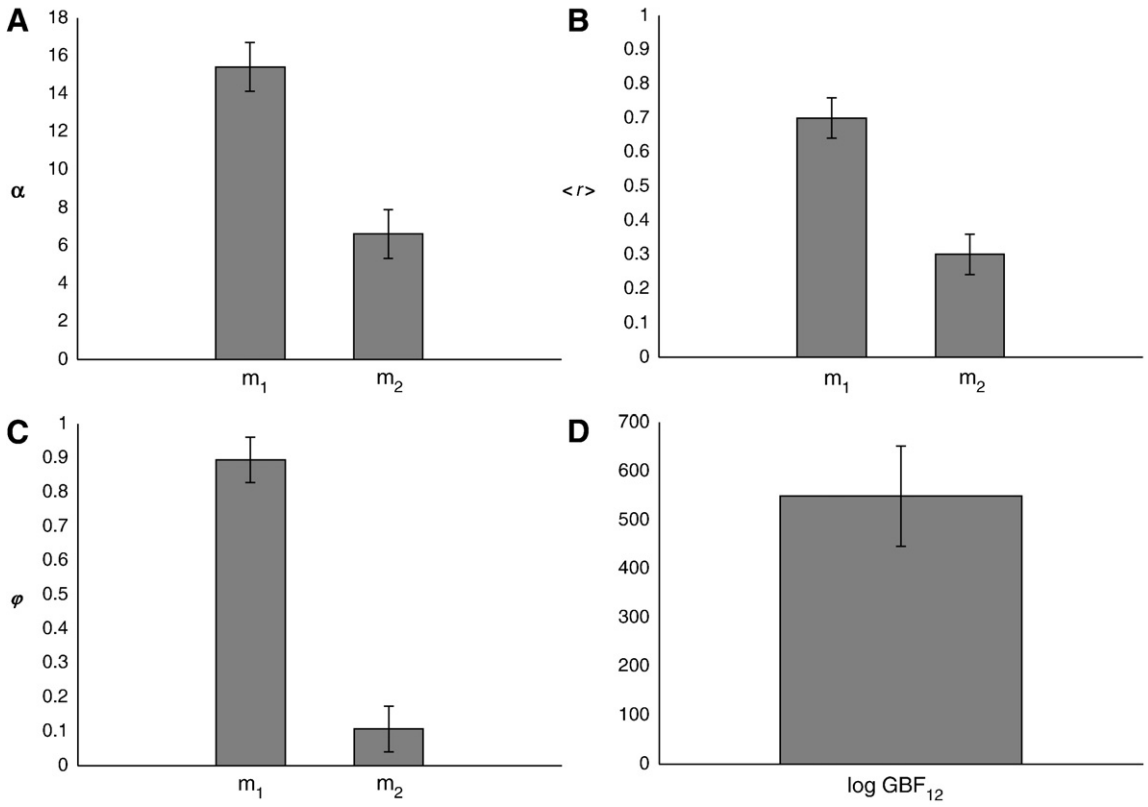


**Fig. 11.** Summary of the results from a simulation study in which we examined the robustness of our method when randomly sampling from a heterogeneous population of subjects. Specifically, we dealt with a population in which 70% of the subjects showed brain responses as generated by model $m_1$ shown in Fig. 8, whereas brain activity in the remaining 30% of the population was generated by model $m_2$. We randomly sampled 20 subjects from this population and generated synthetic fMRI data by integrating the state equations of the associated models with fixed parameters and inputs and adding Gaussian observation noise to achieve an SNR of two. Both $m_1$ and $m_2$ were then fitted to all 20 synthetic data sets. This sampling and data generation procedure was repeated 20 times, resulting in a total of 400 generated data sets and 800 fitted models. For each of the 20 sets of 20 subjects, we computed the different indices provided by random effects BMS (i.e., $\alpha$, $\langle r \rangle$, $\varphi$) and fixed effects BMS (log GBF). This figure shows the mean of these indices together with their 95% confidence intervals (CI).

comparison, the average log GBF in favour of model $m_1$ was 548.9 (CI: 446.2–651.6).

In conclusion, while our random effects BMS method provides a slightly overconservative estimate of exceedance probabilities for the chosen sample size, it shows very good performance overall, providing BMS indices that accurately reflect the structure of the population we sampled from. In particular, the Dirichlet parameters and posterior expectations of model probabilities (which represent the expected probability of obtaining the $k$-th model when randomly selecting a subject) were estimated very precisely. This result not only validates the results obtained for the empirical data set described above, but demonstrates more generally that our BMS procedure is robust when randomly sampling from a heterogeneous population of subjects.

*Comparing different hemodynamic models by model space partitioning*

Finally, we revisited a comparison of DCMs, which were identical in network architecture (the same as $m_1$ in Fig. 8) but differed in the hemodynamic forward model employed (Stephan et al., 2007c). A three-factor design was used to construct 8 different models: (i) nonlinear vs. linear BOLD equations, (ii) classical vs. revised coefficients of the BOLD equation, and (iii) free vs. fixed parameter ($\varepsilon$) for the ratio of intra- and extravascular signal changes. In the original analysis by Stephan et al. (2007c), the *GBF* (based on the negative free-energy approximation) was used to establish the best among the eight models. The best model, abbreviated as $\mathrm{RBM_N}(\varepsilon)$ in Fig. 12, was characterised by (i) a nonlinear BOLD equation, (ii) revised coefficients of the BOLD equation, and (iii) free $\varepsilon$. The difference of its summed log-evidence compared to the second-best model, its linear counterpart $\mathrm{RBM_L}(\varepsilon)$, was 5.26, corresponding to a *GBF* of 192 in favour of the nonlinear model. The summed log-evidences for all 8 models are shown in Fig. 12A.

Here, we demonstrate how one can use the agglomerative property of the Dirichlet distribution (Eq. (18)) to go beyond selective comparisons of specific models and instead examine the relative importance of particular model attributes or model subspaces. Given the three factors above, we focused on the importance of nonlinearities: what is the posterior probability that nonlinear BOLD equations improve the model compared to linear BOLD equations, regardless of any other dimensions of model space (i.e., classical vs. revised coefficients and free vs. fixed $\varepsilon$)?

Following Eq. (18), this question is addressed easily. In a first step, the VB procedure was applied to the entire set of eight models, yielding posterior estimates of the Dirichlet parameters $\alpha_1,\ldots,\alpha_8$ (see Fig. 12B). Subsequently, a new Dirichlet density reflecting the partition of model space into nonlinear and linear subspaces was computed by summing $\alpha_k$ separately for the nonlinear and linear models (Fig. 12C; for simplicity the ordering of the models in Fig. 12 has been chosen such that the first four models are nonlinear [left of the dashed line], whereas the last four models are linear [right of the dashed line]) The resulting Dirichlet can then be used to compare nonlinear and linear models in exactly the same way as one compares two models; e.g. using exceedance probabilities. Fig. 13 shows the result of this comparison: the probability that nonlinear hemodynamic models are better than linear models, regardless of other model attributes, was $\varphi_1 = 98.6\%$.

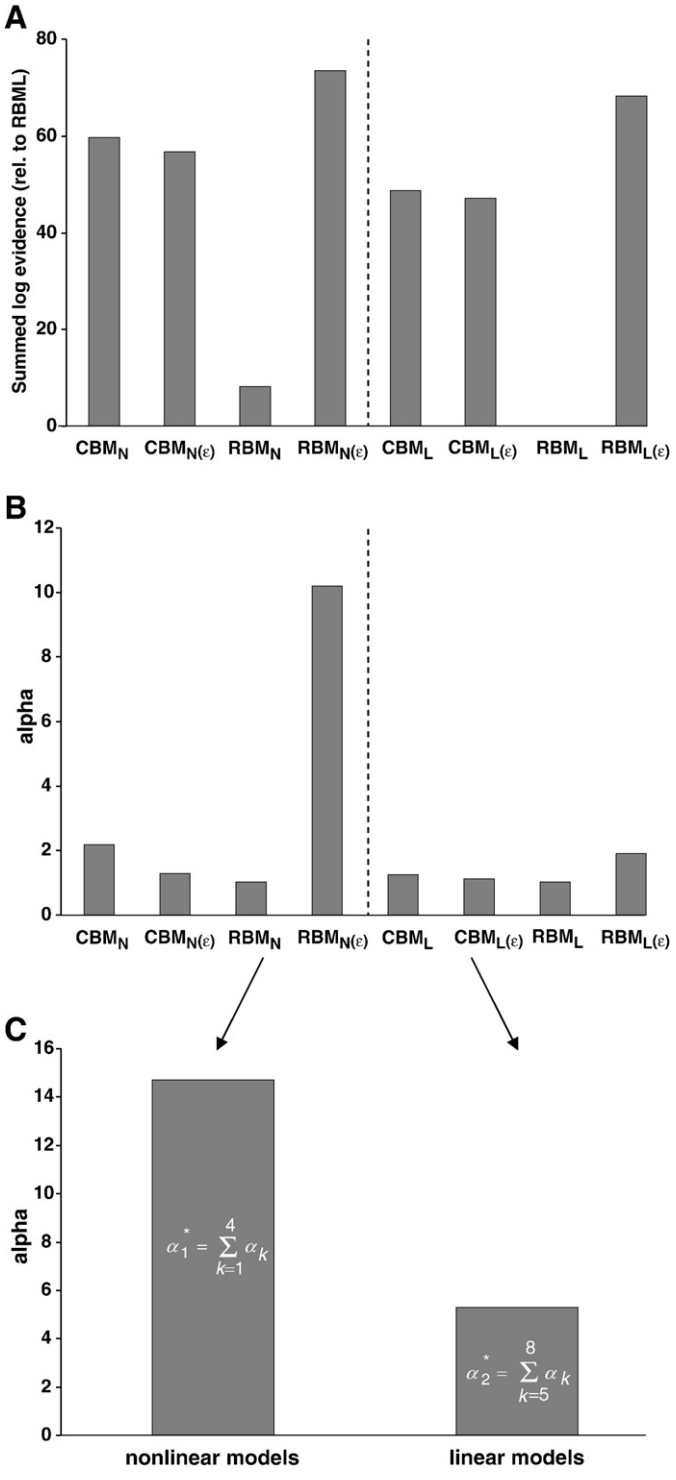


**Fig. 12.** An example of model space partitioning applied to the case of DCMs which were identical in network architecture (the same as $m_1$ in Fig. 8) but differed in the hemodynamic forward model employed (for details, see Stephan et al., 2007c). (A) Eight different models were constructed by means of a three-factorial process: (i) nonlinear vs. linear BOLD equations (subscript N), (ii) classical (CBM) vs. revised (RBM) coefficients of the BOLD equation, and (iii) free vs. fixed parameter ($\varepsilon$) for the ratio of intra- and extravascular signal changes. The bar plot shows the summed log-evidences for all eight models, relative to the worst model ($RBM_L$). The dashed line separates the nonlinear models (on the left) from the linear models (on the right). (B) VB estimates of the Dirichlet parameters for all eight models. (C) VB estimates of the Dirichlet parameters for nonlinear and linear partitions of model space.

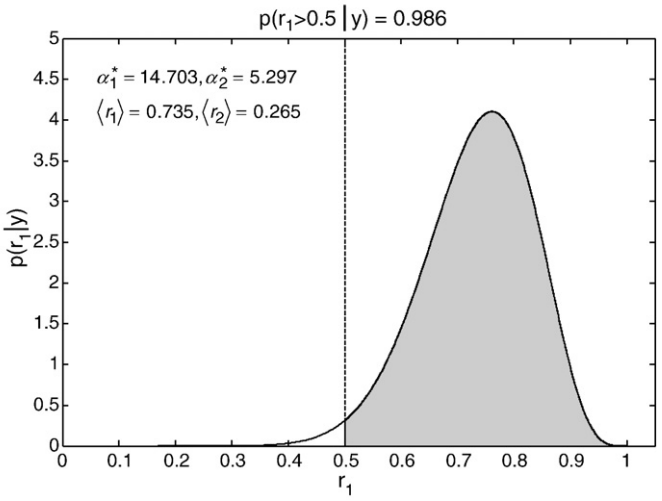


**Fig. 13.** The Dirichlet density for the nonlinear subpart of model space, defined by the parameter estimates shown by Fig. 12C. The exceedance probability of $\varphi_1 = 98.6\%$ (shaded area) indicates the probability that nonlinear hemodynamic models were better than linear models, regardless of any other aspect of model structure.

For comparison, we also used classical inference, applying a repeated-measure ANOVA (with Greenhouse–Geisser correction for non-sphericity) to the log-evidences of the eight models. The result of this test was compatible with the above analysis, rejecting the null hypothesis that linear and nonlinear models were equal in log-evidence ($F = 24.330$, $df = 1,11$, $p < 0.0004$).

## Discussion

In this paper, we have introduced a novel approach for model selection at the group level. Provisional experience suggests that this approach represents a more powerful way of quantifying one's belief that a particular model is more likely than any other at the group level, relative to the conventional *GBF*. Critically, this variational Bayesian approach rests on treating the model switches $m_i$ as a random variable, within a full hierarchical model for multi-subject data (see Fig. 1), and thus accommodates random effects at the between-subject level. Notably, this inference procedure needs only the log-evidences for each model and subject.

In the empirical examples above, we showed two cases where frequentist tests failed to indicate clear differences between models, while the novel Bayesian approach succeeded. In one case (the six-area ventral stream model), a strong outlier subject made the distribution of log-evidences non-normal and thus rendered the *t*-test (but not a non-parametric test) unable to find a significant difference between models. In another case (the four-area ventral stream model), the distribution of log-evidences was normal, but with a between-subject variance that was big enough to prevent significant results by frequentist tests (parametric or non-parametric). It should be noted, however, that the frequentist and Bayesian approaches do not test the same thing. The frequentist approach tries to reject the null hypothesis that there are no differences in log-evidence across models. In contrast, the Bayesian approach estimates the models' probabilities, given the data, and enables inference in terms of exceedance probabilities: the exceedance probability $\varphi_k$ is the probability that a given model $k$ is more likely than any other model (of the $K$ models tested). Furthermore, we can compute the posterior probabilities of the models themselves: $\langle r_k \rangle$ is the expected probability that the $k$-th model generated the data for a randomly selected subject.

The exceedance probability of a model differs in a subtle but important way from the conventional posterior probability of a model

in Bayesian model comparison: because we have a hierarchical model, the posterior probability that any particular model caused the data from a subject chosen at random, is itself a random variable ($r$ in the derivations above). This means that the exceedance probability is a statement of belief about the posterior probability, not the posterior probability itself. So, for example, when we say that the exceedance probability is 98%, we mean that we can be 98% confident that the favoured model has a greater posterior probability than any other model tested. This is not the same as saying that the posterior probability of the favoured model is 98%. The advantage of using exceedance probabilities is that they are sensitive to the confidence in the posterior probability and easily interpretable (since they sum to unity over all models tested).

As can be seen from Eqs. (9) and (11), our method is sensitive to both the distribution and the magnitude of log-evidence differences. The same is true for frequentist tests applied to log-evidence differences, e.g. $t$-tests. However, a critical difference between these frequentist approaches and the VB method is that for the latter the influence of outliers has a natural bound. There is a simple and intuitive reason for this nice property of the VB method: if we keep increasing the log-evidence of model $k$ for a particular subject $n$, our posterior belief that $k$ generated the data of subject $n$ (that is, $g_{nk}=q$ ($m_{nk}=1$); see Eq. (11)) will asymptote to one. Once it has reached unity (which corresponds to complete certainty), any further increase in the log-evidence of model $k$ for subject $n$ has no further influence. This is because the model probabilities are distributed according to the approximate posterior Dirichlet $\mathrm{Dir}(r;\ \alpha_0+\beta)=q(r)$, where $\beta_k$ represents the conditional expectation of the number of subjects whose data we believe were generated by model $k$ and is simply the sum of the subject-specific posterior probabilities that model $k$ generated their individual data. In contrast, frequentist tests like $t$-tests do not show this bounded behaviour with regard to outliers. This is because the sample variance increases monotonically with the magnitude of the outlier, leading to a monotonic decrease of the $t$-statistic. We demonstrated this difference between frequentist approaches and our VB method by two empirical examples with outliers.

Another important advantage of the method proposed here is that it can go beyond the selective comparison of specific models and enables one to assess the importance of changes along any specific dimension of model space. This type of inference, which could be seen as a Bayesian analogue of testing for "main effects" in classical ANOVA, rests on comparing two (or more) subsets of models (i.e., model subspaces). These partitions would typically reflect those components of model structure that one seeks inference about; e.g. whether a specific connection should be included in the model or not, whether a particular connection is modulated by one experimental condition or another, or whether certain effects are linear or nonlinear. We used this approach to demonstrate that hemodynamic models with nonlinear BOLD equations are superior to those with linear ones. This result is in accordance with previous studies that highlight the importance of nonlinearities in the BOLD signal (Deneux and Faugeras, 2006; Friston et al., 2000; Miller et al., 2001; Stephan et al., 2007c; Vazquez and Noll, 1998; Wager et al., 2005a,b). However, in these earlier studies, this conclusion was based on comparisons of specific and single instances of linear and nonlinear hemodynamic models. The inferential advance achieved by the present method is that an arbitrarily large set of models can be considered together, allowing one to integrate out uncertainty over any aspect of model structure, other than the one of interest.

At first glance, it may appear surprising that the hierarchical model described above has been introduced as a generative model for the data $y$, given its inversion does not need the data but the model evidence, $p(y|m)$. This apparent contradiction could be resolved by noting that the log-evidence is a function of the data and represents a sufficient 'summary statistic'. To generate data, one would need to introduce the model parameters $\vartheta_k$ to the graphical model shown in Figs. 1B,C. In the context of DCM, for example, once one has drawn a model $k$ from the multinomial distribution for a specific subject $n$ (i.e., generated a label $m_{nk}=1$), one could generate fMRI time-series by drawing model parameters $\vartheta_k$ from their prior distributions and adding some observation error. However, because the model evidence $p(y|m)$ results from integrating out the influence of the parameters $\vartheta_k$ on the data $y$ (see Eq. (1)), this component is unnecessary during inversion of the generative model.

One property of the method proposed in this paper is that for each subject $n$ our posterior beliefs about model $k$ having generated their data sum to one over all models that are considered, that is $\forall n: \sum_{k=1}^{K} g_{nk}=1$ (c.f. Eq. (11)). In other words, our posterior belief about which model $k$ is most likely to have generated the data for a given subject $n$ is a function of the entire set of models considered. This means that reducing or extending model space can change our inference about which model is most likely at the group level. Although this is a fairly trivial corollary, it should not be forgotten when using this method in practice. In short, one should infer the most likely model by comparing the entire set of plausible models at once, instead of selectively analysing subparts of model space.

To our knowledge, there has been relatively little work on group level methods for Bayesian model comparison so far. In addition to the *GBF* (Stephan et al., 2007b), we had previously suggested a metric called the "positive evidence ratio" (*PER*; Stephan et al., 2007b,c). Based on the conventional definition of "positive evidence" as a Bayes factor larger than three (Kass and Raftery, 1995), the *PER* is simply the number of subjects where there is positive (or stronger) evidence for model 1 divided by the number of subjects with positive (or stronger) evidence for model 2. While the *PER* is insensitive to outliers, it is also insensitive to the magnitude of the differences across subjects. More importantly, however, it is only a descriptive index that does not allow for probabilistic inference in a straightforward manner. In the approach described in this paper, the sufficient statistics for the model frequencies are the posterior estimates of the Dirichlet parameters ($\alpha$). When the differences in model evidences are very strong, these simply boil down to the number of subjects with positive (and more) evidence in favour of a particular model. In that case where for each subject there is one highly superior model, the expected model frequencies become identical to the PER. From this perspective, the present approach can be considered a (probabilistic) generalisation of the PER.

The only other work on group level methods for Bayesian model comparison that we are aware of is a recent paper by Li et al. (2008) who suggested a "group-level *BIC* score". This score is derived by summing the *BIC* for each model across subjects. As explained in Appendix A, the *BIC* is a well-known approximation to the log-evidence (Schwarz, 1978). The group-level *BIC* score by Li et al. (2008) thus approximates the sum of log-evidences and simply corresponds to the log *GBF*. Effectively, the analysis by Li et al. (2008) thus used a fixed effects analysis across models that is formally identical to that used in reports of DCM studies (e.g. Acs and Greenlee, 2008; Allen et al., 2008; Grol et al., 2007; Heim et al., 2008; Kumar et al., 2007; Smith et al., 2006; Stephan et al., 2007a,b; Summerfield and Koechlin, 2008).

Finally, it should be noted that a random effects model selection approach is not necessarily preferable to a fixed effects approach. The choice between fixed and random effects BMS depends on the specific scientific question addressed. In the context of basic mechanisms that are unlikely to differ across subjects, the conventional GBF is both sufficient and appropriate. For example, it is unlikely that subjects differ with regard to basic physiological mechanisms such as the involvement of sodium ion channels in action potential generation or the presence of certain types of connections in the brain. In this context, it is perfectly tenable to assume that all subjects generate data under the same model; and the data from all subjects can be pooled to select this model in the usual way. In contrast, whenever subjects can

exhibit different models or functional architectures, the random effects BMS technique presented in this paper is a more appropriate method. For example, there is evidence that many higher cognitive functions can rely on more than one neurobiological system (Price and Friston, 2002). Also, it is likely that in some mental diseases, e.g. schizophrenia, patients with identical symptoms show heterogeneity with regard to the pathophysiological processes involved (Stephan et al., 2006).

In summary, in contrast to the *GBF* and other established approaches for group-level model comparison, the approach suggested in this paper rests on a hierarchical model for multi-subject data that accommodates random effects at the between-subject level (Fig. 1) and thus provides a generic framework for hypothesis testing. We expect this method to be a useful tool for group studies, not only in the context of dynamic causal modelling, but also for a range of other modelling endeavours; for example, comparing different source reconstruction methods for EEG/MEG at the group level (Henson et al., 2007; Litvak and Friston, 2008; Mattout et al., 2007), or selecting among competing computational models of learning and decision-making, given data from a group of subjects (Brodersen et al., 2008; Hampton et al., 2006).

*Software note*

The method described in this paper is freely available to the community as part of the open-source software package Statistical Parametric Mapping (SPM8; http://www.fil.ion.ucl.ac.uk/spm).

## Acknowledgments

This work was funded by the Wellcome Trust (KES, WDP, RJM, KJF) and the University Research Priority Program "Foundations of Human Social Behaviour" at the University of Zurich (KES). JD is funded by Marie Curie Fellowship. We are very grateful to Marcia Bennett for helping prepare this manuscript, to the FIL Methods Group, particularly Justin Chumbley, for useful discussions and to Jon Roiser and Dominik Bach for helpful comments on practical applications. Finally, we would like to thank the two anonymous reviewers for their constructive comments which have greatly helped to improve this paper.

## Appendix A. Approximations to the log model evidence

With the exception of some special cases (e.g., linear models), the integral expression for the model evidence (Eq. (1)) is analytically intractable and numerically difficult to compute. Under these circumstances, people generally adopt a bound approach where, instead of evaluating the integral above, one optimises a bound on the integral using iterative sampling or analytic techniques. The most common approach of the latter kind is variational Bayes. In this framework, one posits an approximating conditional or posterior density on the unknown parameters, $q(\vartheta)$, and optimises this density with respect to a free-energy bound, $F$, on the log-evidence[6]:

$$F = \log p(y|m) - KL[q(\vartheta), p(\vartheta|y,m)]. \tag{A.1}$$

Because of its relation to variational calculus and Gibb's free-energy in statistical physics, this free-energy bound $F$ is often referred to as the "negative free-energy" or "variational free-energy" (Friston et al., 2007; MacKay, 2003; Neal and Hinton, 1998). Its second term is the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951) between the approximating posterior density $q(\vartheta)$ and the true posterior $p(\vartheta|y,m)$, which is always positive (or zero when $q(\vartheta)$ becomes identical to $p(\vartheta|y,m)$). By iterative optimisation, the negative free-energy $F$ is made an increasingly tighter lower bound on the desired log-evidence, $\ln p(y|m)$; as a consequence, the KL divergence between the approximating and true posterior is minimised. There are a number of approximations that are used when specifying the form of $q(\vartheta)$. These include the ubiquitous mean-field approximation, where various sets of unknown parameters are assumed to be independent, so that the conditional density can be factorised. A common example here would be a bipartition into the regression coefficients of a general linear model and the parameters controlling random effects or error variance. Another common approximation within the mean-field framework is to assume that the conditional density is multivariate Gaussian. This is also known as the Laplace approximation, a full treatment of which can be found in Friston et al. (2007).

For any approximation to the conditional density, the free-energy bound on the log-evidence can be re-written as a mixture of accuracy and complexity:

$$F = \langle \log p(y|\vartheta,m) \rangle_q - KL[q(\vartheta), p(\vartheta|m)]. \tag{A.2}$$

The *accuracy* (first term) is simply the log-likelihood of the data expected under the conditional density. The *complexity* (second term) is the Kullback–Leibler divergence between the approximating posterior and prior density. In other words, it reflects the amount of information obtained about the model parameters, from the data. Clearly, model complexity will increase with the number of parameters (provided that they can be estimated precisely and that they diverge from their prior values). However, model complexity depends on factors other than the mere number of parameters, e.g. how much these parameters are dependent on each other, both *a priori* and *a posteriori*. This is seen easily under the Laplace approximation, i.e. assuming that the conditional density is multivariate Gaussian. In this case, the complexity can be written as follows (see the Appendix of Penny et al., 2004):

$$KL[q(\vartheta), p(\vartheta|m)] = \tfrac{1}{2}|C_\vartheta| - \tfrac{1}{2}|C_{\vartheta|y}| + \tfrac{1}{2}\left(\mu_{\vartheta|y} - \mu_\vartheta\right)^T C_\vartheta^{-1}\left(\mu_{\vartheta|y} - \mu_\vartheta\right). \tag{A.3}$$

Here, $|C_\vartheta|$ and $|C_{\vartheta|y}|$ are the determinants of the prior and posterior covariance matrices and $\mu_{\vartheta|y}$ and $\mu_\vartheta$ are the posterior and prior means, respectively. The first term shows that the penalty conveyed by model complexity increases the more independent the parameters are *a priori*;[7] this is equivalent to saying that the penalty increases with the effective degrees of freedom of the model. Conversely, additional parameters whose effects are redundant in relation to existing parameters *do not* increase model complexity. The second term says that complexity decreases with the degree of independence that the parameters have *a posteriori*. This accords with the general notion that the parameter estimates of a good model should be as precise and uncorrelated as possible. The final term shows that the complexity increases with the distance between the prior and posterior means. In other words, model goodness decreases if one makes bad assumptions about the parameter values *a priori* (i.e., using suboptimal priors), thus forcing the posterior estimates to diverge markedly from the prior means.

In addition to the free-energy bound approximation, there are two other commonly used approximations to the log-evidence, which appeal to the behaviour of the complexity term as the number of observations becomes infinite. We will call these limit-approximations. These include the *AIC* and *BIC* (see Penny et al., 2004). The key difference between the free-energy bound and these limit approximations is that

[6] Because of the monotonic nature of the logarithm, one can maximise the model evidence or the log-evidence; the latter, however, is numerically more convenient to deal with. Please note that for simplicity and clarity we have removed constant terms from the definition of all approximations to the log-evidence discussed in this paper.

[7] It is helpful to note that the determinant of a covariance matrix can be treated as a measure of the volume spanned by a set of vectors (Woodruff 2005). This volume increases with the degree of independence amongst the vectors.

the latter assume a much simpler approximation to the complexity. Under Gaussian assumptions about the error:

$$BIC = \langle \log p(y|\vartheta,m)\rangle_q - \frac{p}{2}\log n$$

$$AIC = \langle \log p(y|\vartheta,m)\rangle_q - p. \tag{A.4}$$

It can be seen that the *AIC* and *BIC* approximate the complexity with the number of parameters or the number of parameters $p$, scaled by the log of the number of observations, $n$. These can be useful approximations when it is difficult to invert the model or optimise the free-energy bound, because one only needs to compute the accuracy or fit of the model to provide an estimate of the log-evidence. However, comparing the complexity terms in these expressions to Eq. (A.3), shows that both the *AIC* and *BIC* will fail in various situations. An obvious example is redundant parameterisation; the true complexity will not change when we add a parameter whose effect is identical to another parameter in measurement space. While the free-energy bound would take this redundancy into account, retaining the same complexity, the *AIC* and *BIC* approximations would indicate that complexity has increased. In practice, many models show partial dependencies amongst parameters, meaning that *AIC* and *BIC* routinely over-estimate the effect that adding or removing parameters has on model complexity.

## Appendix B. Sampling approach to estimating the Dirichlet parameters

In this appendix, we introduce a sampling procedure that provides an approximation to the negative free energy $F(y,\alpha)\leq \ln p(y|\alpha)$ which is independent from the VB approach described in the main text. This sampling procedure can be used to demonstrate the correctness of the proposed VB procedure by verifying that the algorithm described by Eq. (14) provides an accurate solution for the variational energies in the mean-field approximation of Eq. (8). In this context, it should be noted that we are assuming that the exact posterior $p(r|y)$ can be adequately approximated by a Dirichlet density $q(r)$; therefore, the procedure proposed in this appendix samples from the approximate posterior $q(r)$, not from the exact posterior $p(r|y)$.

We seek the posterior density on the multinomial parameters $r=[r_1,\ldots,r_K]$ that generate switches or indicator variables, $m_{nk}\in\{0,1\}$, prescribing the $n$-th subject's model; i.e., $p(m_{nk}=1)=r_k$. To simplify things, we will assume an approximating form, $q(r;\alpha)$ for this density, with sufficient statistics $\alpha$. Specifically, we assume a Dirichlet density:

$$q(r;\alpha) = D(\alpha) = \frac{1}{Z(\alpha)}\prod_{k=1}^{K} r_k^{\alpha_k - 1}$$

$$\Rightarrow \ln q(r;\alpha) = -\ln Z(\alpha) + \sum_k (\alpha_k - 1)\ln r_k \tag{B.1}$$

where the expected multinomial parameters (i.e., conditional expectation that the $k$-th model will be selected at random) are:

$$\langle r_k\rangle_q = \frac{\alpha_k}{\alpha_S}$$

$$\alpha_S = \sum_{k=1}^{K}\alpha_k. \tag{B.2}$$

Note that a Dirichlet form ensures that $\sum_{k=1}^{K} r_k = 1$. The normalising or partition coefficient in Eq. (B.1) is:

$$Z(\alpha) = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\alpha_S)} \Rightarrow \ln Z(\alpha) = -\ln\Gamma(\alpha_S) + \sum_k \ln\Gamma(\alpha_k). \tag{B.3}$$

We can now construct a free-energy bound in the usual way, assuming Dirichlet priors $\alpha_0$ (which would usually be $\alpha_0=[1,\ldots,1]$ unless one had prior beliefs about which model is more likely to be selected):

$$F(y,\alpha) = \langle \ln p(y_1|r) + \ldots + \ln p(y_N|r) + \ln p(r|\alpha_0) - \ln q(r|\alpha)\rangle_q. \tag{B.4}$$

This can be decomposed into three terms:

$$\langle \ln p(y_n|r)\rangle_q = \langle \ln \sum_k p(y_n|m_{nk}=1) r_k\rangle_q$$

$$\langle \ln q(r|\alpha)\rangle_q = -\ln Z(\alpha) + \sum_k (\alpha_k - 1)[\Psi(\alpha_k) - \Psi(\alpha_S)]$$

$$\langle \ln p(r|\alpha_0)\rangle_q = -\ln Z(\alpha_0) + \sum_k (\alpha_{0k} - 1)[\Psi(\alpha_k) - \Psi(\alpha_S)]. \tag{B.5}$$

The last two terms only depend on the priors $\alpha_{0k}$ and the parameters $\alpha$ of the Dirichlet and can thus be computed directly. The first term can be computed numerically by drawing a large number of samples from $q(r;\alpha)$. In this paper, we gridded the possible range for values of $\alpha_k$, i.e. $[1\ldots K+1]$, using a bin size of 0.1, and then drew 1000 samples per bin, exploiting a relationship between Gamma and Dirichlet distributions described by Ferguson (1973). Given those samples, the Dirichlet parameters are those that maximise $F$:

$$\alpha = \arg\max_{\alpha} F(y,\alpha). \tag{B.6}$$

As a final note, we would like to point out that one could also use Jensen's inequality to simplify the first term in Eq. (B.5):

$$\begin{aligned}\langle \ln p(y_n|r)\rangle_q &= \langle \ln \sum_k p(y_n|m_{nk}=1) r_k\rangle_q \\ &\geq \langle \sum_k r_k \ln\ p(y_n|m_{nk}=1)\rangle_q \\ &= \sum_k \frac{\alpha_k}{\alpha_S} \ln\ p(y_n|m_{nk}=1).\end{aligned} \tag{B.7}$$

This effectively provides a lower-bound on a lower-bound, which can be simplified to give

$$\tilde{F}(y,\alpha) = \sum_k \left(\frac{\alpha_k}{\alpha_S}\sum_n \ln p(y_n|m_{nk}=1) - \left(\alpha_k - \alpha_k^0\right)[\Psi(\alpha_k) - \Psi(\alpha_S)] + \ln\Gamma(\alpha_k)\right) - \ln\Gamma(\alpha_S). \tag{B.8}$$

Given the priors, $\alpha_0$, and the log-evidences $\ln p(y_n|m_{nk}=1)$ for each subject and model, $\tilde{F}$ could be used as an alternative method to estimate the Dirichlet parameters $\alpha$ using conventional nonlinear optimisation. In practice, however, we have found the VB method described in the main text to be superior.

## References


Acs, F., Greenlee, M.W., 2008. Connectivity modulation of early visual processing areas during covert and overt tracking tasks. NeuroImage 41, 380–388.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automatic Control 19, 716–723.

Allen, P., Mechelli, A., Stephan, K.E., Day, F., Dalton, J., Williams, S., McGuire, P.K., 2008. Fronto-temporal interactions during overt verbal initiation and suppression. J. Cogn. Neurosci. 20, 1656–1669.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, Berlin.

Brodersen, K.H., Penny, W.D., Harrison, L.M., Daunizeau, J., Ruff, C.C., Duzel, E., Friston, K.J., Stephan, K.E., 2008. Integrated Bayesian models of learning and decision making for saccadic eye movements. Neural Netw. 21, 1247–1260.

Deneux, T., Faugeras, O., 2006. Using nonlinear models in fMRI data analysis: model selection and activation detection. NeuroImage 32, 1669–1689.

Diedrichsen, J., Shadmehr, R., 2005. Detecting and adjusting for artifacts in fMRI time series data. NeuroImage 27, 624–634.

Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. Ann. Stat. 1, 209–230.
Friston, K.J., Mechelli, A., Turner, R., Price, C.J., 2000. Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. NeuroImage 12, 466–477.
Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. NeuroImage 19, 1273–1302.
Friston, K.J., Mattout, J., Trujillo-Barreto, N., Ashburner, A., Penny, W.D., 2007. Variational free-energy and the Laplace approximation. NeuroImage 34, 220–234.
Garrido, M.I., Kilner, J.M., Kiebel, S.J., Stephan, K.E., Friston, K.J., 2007. Dynamic causal modelling of evoked potentials: a reproducibility study. NeuroImage 36, 571–580.
Garrido, M.I., Friston, K.J., Kiebel, S.J., Stephan, K.E., Baldeweg, T., Kilner, J.M., 2008. The functional anatomy of the MMN: a DCM study of the roving paradigm. NeuroImage 42, 936–944.
Grol, M.J., Majdandzić, J., Stephan, K.E., Verhagen, L., Dijkerman, H.C., Bekkering, H., Verstraten, F.A., Toni, I., 2007. Parieto-frontal connectivity during visually guided grasping. J. Neurosci. 27, 11877–11887.
Hampton, A.N., Bossaerts, P., O Doherty, J.P., 2006. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J. Neurosci. 26, 8360–8367.
Heim, S., Eickhoff, S.B., Ischebeck, A.K., Friederici, A.D., Stephan, K.E., Amunts, K., 2008. Effective connectivity of the left BA 44, BA 45, and inferior temporal gyrus during lexical and phonological decisions identified with DCM. Hum. Brain Mapp. 30, 392–402.
Henson, R.N., Mattout, J., Singh, K.D., Barnes, G.R., Hillebrand, A., Friston, K.J., 2007. Population-level inferences for distributed MEG source localization under multiple constraints: application to face-evoked fields. NeuroImage 38, 422–438.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795.
Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22, 79–86.
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, e100.
Leff, A.P., Schofield, A.M., Stephan, K.E., Crinion, J.T., Friston, K.J., Price, C.J., 2008. The cortical dynamics of intelligible speech. J. Neurosci. 28, 13209–13215.
Li, J., Wang, J., Palmer, S.J., McKeown, M.J., 2008. Dynamic Bayesian network modelling of fMRI:A comparison of group-analysis methods. NeuroImage 41, 398–407.
Litvak, V., Friston, K., 2008. Electromagnetic source reconstruction for group studies. NeuroImage 42, 1490–1498.
MacKay, D.J.C., 2003. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge.
Mattout, J., Henson, R.N., Friston, K.J., 2007. Canonical source reconstruction for MEG. Comput. Intell. Neurosci. 2007, 67613.
Miller, K.L., Luh, W.M., Liu, T.T., Martinez, A., Obata, T., Wong, E.C., Frank, L.R., Buxton, R.B., 2001. Nonlinear temporal dynamics of the cerebral blood flow response. Hum. Brain Mapp. 13, 1–12.
Neal, R.M., Hinton, G.E., 1998. A view of the EM algorithm that justifies incremental sparse and other variants. In: Jordan, M.I. (Ed.), Learning in Graphical Models. Kluwer Academic Publishers, Dordrecht.
Neyman, J., Pearson, E., 1933. On the problem of the most efficient tests of statistical hypotheses. Philos. Trans. R. Soc. Lond. A 231, 289–337.
Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004. Comparing dynamic causal models. NeuroImage 22, 1157–1172.
Pitt, M.A., Myung, I.J., 2002. When a good fit can be bad. Trends Cogn. Sci. 6, 421–425.
Price, C.J., Friston, K.J., 2002. Degeneracy and cognitive anatomy. Trends Cogn. Sci. 6, 416–421.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.
Smith, A.P., Stephan, K.E., Rugg, M.D., Dolan, R.J., 2006. Task and content modulate amygdala-hippocampal connectivity in emotional retrieval. Neuron 49, 631–638.
Stephan, K.E., Baldeweg, T., Friston, K.J., 2006. Synaptic plasticity and dysconnection in schizophrenia. Biol. Psychiatry 59, 929–939.
Stephan, K.E., Harrison, L.M., Kiebel, S.J., David, O., Penny, W.D., Friston, K.J., 2007a. Dynamic causal models of neural system dynamics: current state and future extensions. J. Biosci. 32, 129–144.
Stephan, K.E., Marshall, J.C., Penny, W.D., Friston, K.J., Fink, G.R., 2007b. Interhemispheric integration of visual processing during task-driven lateralization. J. Neurosci. 27, 3512–3522.
Stephan, K.E., Weiskopf, N., Drysdale, P.M., Robinson, P.A., Friston, K.J., 2007c. Comparing hemodynamic models with DCM. NeuroImage 38, 387–401.
Stephan, K.E., Kasper, L., Harrison, L.M., Daunizeau, J., den Ouden, H.E.M., Breakspear, M., Friston, K.J., 2008. Nonlinear dynamic causal models for fMRI. NeuroImage 42, 649–662.
Summerfield, C., Koechlin, E., 2008. A neural representation of prior information during perceptual inference. Neuron 59, 336–347.
Vazquez, A.L., Noll, D.C., 1998. Nonlinear aspects of the BOLD response in functional MRI. NeuroImage 7, 108–118.
Wager, T.D., Vazquez, A., Hernandez, L., Noll, D.C., 2005a. Accounting for nonlinear bold effects in fMRI: parameter estimates and a model for prediction in rapid event-related studies. NeuroImage 25, 206–218.
Wager, T.D., Keller, M.C., Lacey, S.C., Jonides, J., 2005b. Increased sensitivity in neuroimaging analyses using robust regression. NeuroImage 26, 99–113.
Woodruff, D.L., 2005. General purpose metrics for solution variety. In: Rego, C., Alidaee, B. (Eds.), Metaheuristic Optimization Via Memory and Evolution: Tabu Search and Scatter Search. Springer, Berlin.