NeuroImage 84 (2014) 971–985

Contents lists available at ScienceDirect

NeuroImage

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

Bayesian model selection for group studies — Revisited

L. Rigoux [a], K.E. Stephan [b,c], K.J. Friston [b], J. Daunizeau [a,b,*]

[a] Brain and Spine Institute, Paris, France
[b] Wellcome Trust Centre for Neuroimaging, University College London, UK
[c] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich & ETH Zurich, Switzerland

ARTICLE INFO

Article history:
Accepted 29 August 2013
Available online 7 September 2013

Keywords:
Statistical risk
Exceedance probability
Between-condition comparison
Between-group comparison
Mixed effects
Random effects
DCM

ABSTRACT

In this paper, we revisit the problem of Bayesian model selection (BMS) at the group level. We originally addressed this issue in Stephan et al. (2009), where models are treated as random effects that could differ between subjects, with an unknown population distribution. Here, we extend this work, by (i) introducing the Bayesian omnibus risk (BOR) as a measure of the statistical risk incurred when performing group BMS, (ii) highlighting the difference between random effects BMS and classical random effects analyses of parameter estimates, and (iii) addressing the problem of between group or condition model comparisons. We address the first issue by quantifying the chance likelihood of apparent differences in model frequencies. This leads to the notion of *protected* exceedance probabilities. The second issue arises when people want to ask "whether a model parameter is zero or not" at the group level. Here, we provide guidance as to whether to use a classical second-level analysis of parameter estimates, or random effects BMS. The third issue rests on the evidence for a difference in model labels or frequencies across groups or conditions. Overall, we hope that the material presented in this paper finesses the problems of group-level BMS in the analysis of neuroimaging and behavioural data.

Introduction

Any statistical measure of empirical evidence rests on some form of model comparison. In a classical setting, one typically compares the null with an alternative hypothesis, where the former is a model of how chance could have generated the data. Theoretical results specify the sense in which model comparison can be considered optimal. For example, the Neyman–Pearson lemma essentially states that statistical tests based on the likelihood ratio (such as a simple *t*-test) are the most powerful, i.e., they have the best chance of detecting an effect (see e.g., Casella and Berger, 2001). From this perspective, Bayesian model comparison can be seen as a simple extension to likelihood tests, in that it allows for the comparison of more than two models. In fact, likelihood ratios are used in a Bayesian setting, under the name of *Bayes factors* (Kass and Raftery, 1995). These are just the ratio of experimental evidence in favour of one model relative to another. Having said this, established classical and Bayesian techniques may give different answers to the same question — a difference that has entertained generations of statisticians (see e.g., Fienberg, 2006).

In this paper, we consider the problem of performing random effects Bayesian model selection (BMS) at the group level. This was originally addressed in Stephan et al. (2009), where models were treated as random effects that could differ between subjects and have a fixed (unknown) distribution in the population. The implicit hierarchical model is then inverted using variational or sampling techniques (see Penny et al., 2010), to provide conditional estimates of the frequency with which any model prevails in the population. This random effects BMS procedure complements fixed effects procedures that assume that subjects are sampled from a homogenous population with one (unknown) model (cf. the log group Bayes factor that sums log-evidences over subjects; Stephan et al., 2007). Stephan et al. (2009) also introduced the notion of *exceedance probability*, which measures how likely it is that any given model is more frequent than all other models in the comparison set. These two summary statistics typically constitute the results of random effects BMS (see, for example, den Ouden et al., 2010).

While the random effects BMS procedure suggested in Stephan et al. (2009) and Penny et al. (2010) has proven useful in practice — and has been employed by more than hundred published studies to date, some conceptual issues are still outstanding. In this paper, we extend the approach described in Stephan et al. (2009) in three ways: (i) we provide a complete picture of the statistical risk incurred when performing group BMS, (ii) we examine the formal difference between random effects BMS and classical random effects analyses of parameter estimates, when asking whether a particular parameter is zero or not, and (iii) we address the problem of between-group and between-condition comparisons.

* Corresponding author at: Motivation, Brain and Behaviour Group, Brain and Spine Institute, 47, bvd de l'Hopital, 75013, Paris, France. Fax: +33 1 57 27 47 94.
E-mail address: jean.daunizeau@gmail.com (J. Daunizeau).
URL: http://sites.google.com/site/jeandaunizeauswebsite (J. Daunizeau).

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.08.065

Section 2 revisits random effects BMS, providing a definition of the null at the group level. This allows us to quantify the statistical risk incurred by performing random effects BMS, i.e. how likely it is that differences in model evidences are due to chance. *En passant*, we clarify the interpretation of exceedance probabilities and provide guidance with regard to summary statistics that should be reported when using random effects BMS.

Section 3 addresses the difference between random effects BMS and classical random effects analyses of parameter estimates. In principle, group effects can be assessed using a classical random effects analysis of the parameter estimates across subjects (e.g., using t-tests), or using random effects BMS (reduced versus full model). However, these approaches do not answer the same question (and therefore may not give the same answer). Here, we explain the nature of this difference and identify the situations that would yield identical or different conclusions.

Section 4 introduces a simple extension to the original framework proposed in Stephan et al. (2009). In brief, we propose a test of whether two (or more) groups of subjects come from the same population. We also address the related issue of between condition comparisons. The key idea behind these procedures is a generalization of the intuition that underlies classical paired t-tests; i.e. one has to quantify the evidence for a difference — as opposed to the difference of evidences.

For all three issues, we use Monte-Carlo simulations to assess the performance of random effects BMS in the context of key applications, e.g. Dynamic Causal Modeling (see Daunizeau et al., 2011a for a recent review).

On the statistical risk of group BMS

In this section, we first revisit the approach to random effects BMS proposed in Stephan et al. (2009), recasting it as an extension of Polya's urn model. This serves to identify the nature of the risk associated with model selection. In brief, we focus on the risk of stating that a given model is a better explanation for the data than other models, given that chance could have favoured this particular model. In turn, we propose a simple Bayesian "omnibus test", to exclude chance as a likely explanation for an apparent difference in model frequencies.

Polya's urn model

The random effects BMS can be viewed as a simple extension of the so-called Polya's urn model (see, e.g., Johnson and Kotz, 1977), which we will revisit here. Consider an infinite urn, containing K different sorts of marbles. Let r_k be the frequency of marbles of type $k \in [1,K]$ in the urn. The marble frequencies satisfy: $0 \leq r_k \leq 1$ and $1 = \sum_{k=1}^{K} r_k$. Let us randomly draw n marbles from the urn. Let m_i be the outcome of the ith sample, where $i \in [1,n]$. The probability of observing any given outcome m_i is determined by the respective frequency r_k of each type of marble and has the following multinomial distribution:

$$p(m_i|r_k) = \prod_{k=1}^{K} r_k^{m_{ik}}$$
$$m_{ik} = \begin{cases} 1 & \text{if } k = l \\ 0 & \text{otherwise} \end{cases} \forall k \in [1,K] \quad (1)$$

where $m_i \in [0,1]$ is a one-in-K vector, i.e. the index $l \in [1,K]$ of the non-zero entry encodes the marble's type. Given a set of n observed marbles, one can ask questions about the unknown marble frequencies in the urn. Within a Bayesian approach, Eq. (1) expresses the likelihood function, which is completed with priors $p(r|H)$ on marble frequencies to form a posterior density over marble frequencies $p(r|m,H)$, as follows:

$$p(r|m,H) = \frac{p(r|H)}{p(m|H)} \prod_{i=1}^{n} p(m_i|r_k) = \frac{p(r|H)}{p(m|H)} \prod_{k=1}^{K} r_k^{\sum_{i=1}^{n} m_{ik}}$$
$$p(m|H) = \int p(r|H) \prod_{k=1}^{K} r_k^{\sum_{i=1}^{n} m_{ik}} dr \quad (2)$$

where $p(m|H)$ is the (Polya's urn) model evidence, under the prior assumption H. A "reasonable" prior assumption H_1 is that, a priori, the urn is expected to be unbiased, i.e.: $E[r_k|H_1] = 1/K$. This prior assumption can be captured using the following Dirichlet probability density function:

$$p(r|H_1) = \frac{\Gamma(K\alpha_0)}{\Gamma(\alpha_0)^K} \prod_{k=1}^{K} r_k^{\alpha_0 - 1} \quad (3)$$

where Γ is the gamma function and α_0 is the so-called *concentration parameter* (it controls the prior variance of marble frequencies). Usually, one invokes uninformative (flat) priors on marble frequencies, by setting $\alpha_0 = 1$. Under H_1, one can explain differences in the observed frequencies of marbles with a difference in the "true" (but unknown) frequencies of marbles. This will be expressed in the posterior distribution $p(r|m, H_1)$, which will deviate from the prior, i.e.: $E[r_k|m, H_1] \neq 1/K$. One can also derive the so-called *exceedance probability* (EP) φ_k — the probability that the kth marble type is more frequent in the urn than any other type (given observed marbles): $\varphi_k = P(r_k \geq r_{k' \neq k}|m, H_1)$. As with marble frequencies, the EPs satisfy: $0 \leq \varphi_k \leq 1$ and $1 = \sum_{k=1}^{K} \varphi_k$. They express a degree of (posterior) confidence on the difference between marble frequencies; we will discuss EPs in detail below. At this point, it suffices to say that all conclusions drawn from these sufficient statistics are valid, *under* H_1.

However, one may want to consider another prior assumption, which arises at the infinite concentration limit, i.e.: $H_1 \rightarrow^{\alpha_0 \rightarrow \infty} H_0$. Under *the null* H_0, the marble frequencies are all equal to each other, i.e.: $r_k = 1/K$. This is typically encoded through a delta-Dirac distribution, as follows:

$$p(r|H_0) = \begin{cases} 1 & \text{if } r_k = 1/K \ \forall k \in [1,K] \\ 0 & \text{otherwise} \end{cases}. \quad (4)$$

Eq. (4) means that H_0 differs from H_1 in that the actual marble frequencies r are fixed (their prior variance is zero). Under the null, any apparent difference in the frequencies is simply due to chance. This makes the null a candidate explanation for the observed marbles. This is important, because it means that any inference based upon sufficient statistics derived under H_1 implicitly assumes that the null is a (comparatively) less plausible assumption. Crucially, should the null turn out to be a viable assumption, this would invalidate the conclusions drawn under H_1. In other terms, the risk we take in relying upon the posterior density $p(r|m, H_1)$ can be defined in terms of the probability P_o of having erroneously chosen H_1 against H_0, given the observed marbles m. This is simply the posterior probability of H_0 versus H_1 (see Daunizeau et al., 2011b for a formal decision theoretic derivation of model selection error risk). Under flat priors on H, P_o is given by:

$$P_o = \frac{1}{1 + \frac{p(m|H_1)}{p(m|H_0).}} \quad (5)$$

Eq. (5) evaluates the probability that the observed sample may have occurred by chance. The above *Bayesian omnibus risk* (*BOR*) P_o can be compared to any desired error rate, e.g. 5%, in analogy to classical p-values of "omnibus tests". The statistically-literate reader might notice that the BOR is but a normalization of the Bayes factor $p(m|H_1)/p(m|H_0)$. Although introducing the BOR might thus appear superfluous, we believe that it is useful to think in terms of the statistical risk incurred when relying on EPs (see next section). In addition, we will eventually use the BOR to define *protected* exceedance probabilities (see the Implications for group BMS section).

Testing whether any (marble) model is "significantly more frequent than any other"

Through the Bayesian comparison of H_1 and H_0, the BOR directly quantifies the probability that model frequencies are all equal to each other. However, it may also be tempting to interpret (one minus) the maximum exceedance probability (EP) as some form of "Bayesian *p*-value" — in the sense that a departure of the maximum EP from $1/K$ expresses evidence in favour of H_1 (against H_0). This intuition deserves careful scrutiny: one can show (see Appendix 1) that, although the qualitative behaviour of the maximum EP is similar to the Bayesian omnibus risk, the impact of a difference in the marble counts differs. In brief, for the maximum EP, it is proportional to the square root $\sqrt{n}$ of the number of marbles, whereas it is simply proportional to n for the (log-transformed) BOR. Although this partly justifies the intuition behind the interpretation of EPs, this also begs the question: which of these two statistics should be used to detect a difference in marble frequencies?

To address this question, we conducted a Monte-Carlo simulation study where $K = 4$ models were compared, given a group of $n \in \{16,64\}$ subjects. In brief, Polya's urn counts m were simulated under H_0 and H_1, respectively. Then both EPs[1] and BOR omnibus statistics were computed for both types of datasets. This procedure was repeated 1024 times, in order to perform a Receiver Operating Characteristic (ROC) analysis. Fig. 1 summarizes the comparison of EPs and BOR, with respect to their relative ability to disambiguate chance from real differences in marble frequencies.

First, one can see that the maximum EP and 1-BOR are correlated across Monte-Carlo simulations (upper panels in Fig. 1). This means that they both reflect the strength of evidence for or against H_0. Second, it is clear that the net effect of increasing the number of subjects n is to improve the discriminability of H_0 and H_1 for both statistics. Their empirical histograms suggest that – on average – H_0 and H_1 are better discriminated using BOR (lower panels). This was confirmed by deriving the area under the ROC curve A_{ROC}. For $n = 16$ subjects, $A_{ROC} = 0.90$ for BOR and $A_{ROC} = 0.81$ for EPs. For $n = 64$ subjects, these scores increase to $A_{ROC} = 0.98$ and $A_{ROC} = 0.90$, respectively.

As a further quantitative comparison, we examined the statistical power (true positive rate) for thresholds that yield a false positive rate of 5%. For $n = 16$ subjects, the power was 65.3% for BOR and 47.5% for EPs. For $n = 64$ subjects, power increases to 94.6% and 68.2%, respectively. Finally, we examined the total error rate (the sum of type I and type II error rates) that indicates the probability of confusing H_0 and H_1. We determined the disambiguation threshold (on either φ and P_0), at which the probability of confusing H_0 and H_1 is minimal and thus best discriminates between H_1 and H_0 (under equal costs for type I and type II errors). For $n = 16$ subjects, the disambiguation thresholds were 0.55 for BOR and 0.74 for EPs, whereas for $n = 64$, these were 0.49 and 0.79, respectively. The associated total error rates were: 0.38 ($n = 16$) and 0.09 ($n = 64$) for BOR, 0.52 ($n = 16$) and 0.34 ($n = 64$) for EPs. In other words, one can interpret $1 - P_0 \approx 0.75$ as strong evidence in favour of H_1, whereas $\varphi \approx 0.75$ provides little evidence in favour or against H_1. Taken together, these results demonstrate a slight overconfidence bias for EPs. This implies that the risk of wrongly declaring a model "more frequent than any other one" (above and beyond chance) is better assessed in terms of BOR than in terms of EPs.

Implications for group BMS

Random effects BMS, as described in Stephan et al. (2009), is a simple extension of Polya's urn model, where label variables m are observed indirectly, through subject-wise log model evidences $L_{ik} = \log p(y_i|m_{ik} = 1)$, which encode how likely the ith subject's dataset y_i is under the kth model. This induces a hierarchical probabilistic model that can be inverted using either sampling (e.g. Gibbs) or variational approaches, to yield a posterior density $p(r|y,H_1)$ over model frequencies. As with Polya's urn model above, a Bayesian omnibus risk P_o can be derived, which evaluates the chance likelihood of observed subject-specific data $y = (y_1, \ldots, y_n)$:

$$
\begin{aligned}
P_o &= \frac{p(y|H_0)}{p(y|H_0) + p(y|H_1)} \\
p(y|H) &= \sum_m \int p(y|m,H)p(m|r,H)p(r|H)dr \\
p(y|m,H) &= \exp\left(\sum_{i=1}^{n}\sum_{k=1}^{K} m_{ik}L_{ik}\right)
\end{aligned}
\tag{6}
$$

where $p(m|r,H)$ and $p(r|H)$ depend on $H \in \{H_0,H_1\}$ and are given by Eqs. (1), (3) and (4). The derivation of $P(y|H)$ given within-subject model evidences L_{ik} is described in Appendix 2.

In most recent studies relying on random effects BMS, researchers typically report EPs φ_k in a similar way to classical p-values; i.e., as a quantitative measure of the amount of evidence for the "best" model (e.g., Boly et al., 2011; Daw et al., 2011; den Ouden et al., 2010; Fleming et al., 2010; Tricomi et al., 2010). However, the argument above suggests that EPs may not be ideally suited for such a purpose. This is because their derivation is conditional upon H_1 and does not consider that apparent differences in model frequencies may be due to chance. Having said this, using the BOR alone does not tell us which model (if any) is the "best" — because it is an omnibus statistic.

To facilitate inferences about specific models (as opposed to omnibus testing), we now introduce a *protected* exceedance probability $\tilde{\varphi}_k$ that uses the BOR to compute a Bayesian model average of the exceedance probability. This average accounts for the fact that the observed variability in (log-) model evidences could be due to chance by marginalizing the exceedance probabilities over H_1 and H_0, as follows:

$$
\begin{aligned}
\tilde{\varphi}_k &= P(r_k \geq r_{k'\neq k}|y) \\
&= P(r_k \geq r_{k'\neq k}|y,H_1)P(H_1|y) + P(r_k \geq r_{k'\neq k}|y,H_0)P(H_0|y) \\
&= \varphi_k(1-P_0) + \frac{1}{K}P_0
\end{aligned}
\tag{7}
$$

Here, we have used the limit definition of H_0 (i.e.: $H_1 \xrightarrow{\alpha_0 \to \infty} H_0$), to derive the exceedance probability under H_0.[2] Eq. (7) is a direct application of Bayesian model averaging (Madigan et al., 1996), where we have averaged over H_0 and H_1. Note that these protected

[1] Note that here, we inspect the behaviour of the maximum EP ($\max_k \varphi_k$), as this is a typical summary statistics of applications of random effects BMS.

[2] At the limit $\alpha_0 \to \infty$, the posterior counts of the frequency Dirichlet density are dominated by the prior counts ($\alpha_k \to \alpha_0$), irrespective of the likelihood term. As a consequence, the posterior tends to an equi-frequency belief.

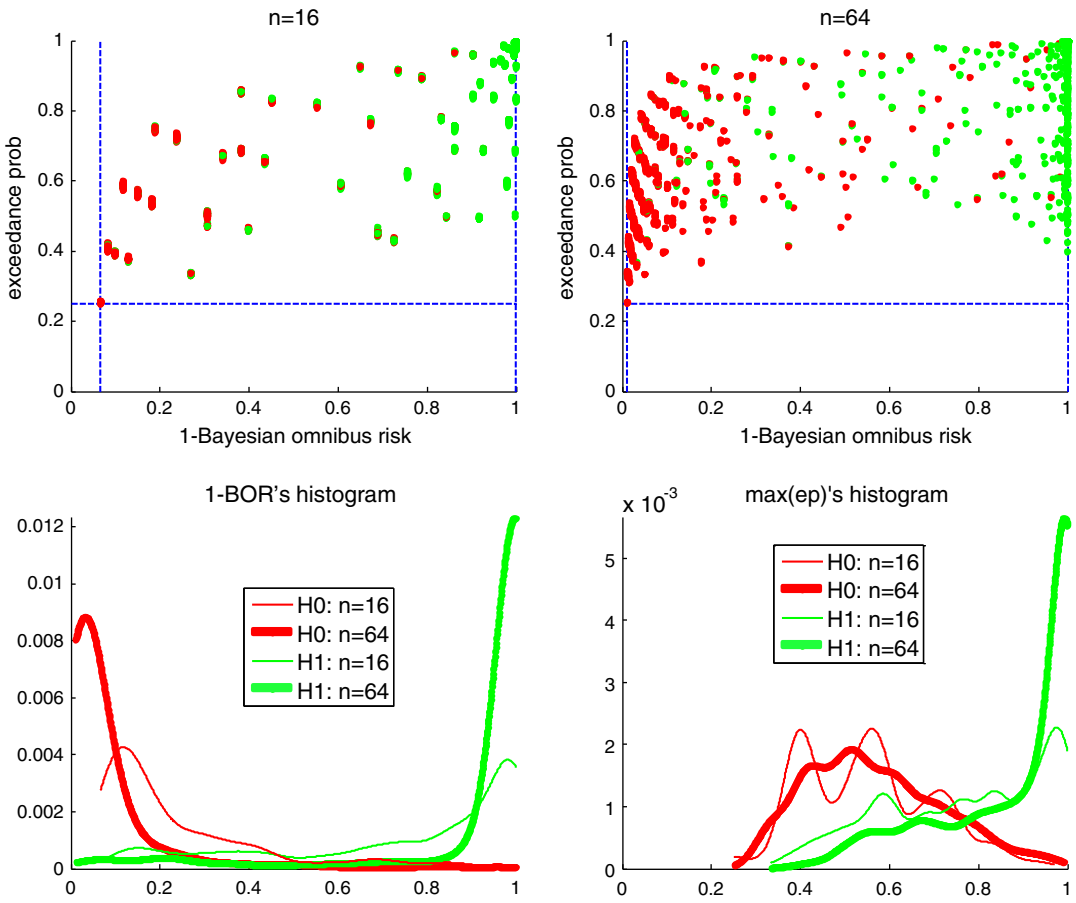

Fig. 1. Exceedance probabilities and the Bayesian omnibus risk. This figure depicts the distribution of the maximum exceedance probability (EP) and the Bayesian omnibus risk (BOR) under H_0 (in red) and under H_1 (in green). Upper-left: EPs (y-axis) is plotted against 1-BOR (x-axis) for $n = 16$ subjects. Each dot is one (out of 1024) Monte-Carlo simulation. Dotted blue lines indicate the lower and upper bounds for the BOR and the maximum EP. Upper-right: same format, with $n = 64$ subjects. Lower-left: Monte-Carlo histogram of 1-BOR (thick line: $n = 64$, thin line: $n = 16$). Upper-right: same format, for the maximum EP.

exceedance probabilities still sum to one. In short, protected EPs quantify the probability that any one model is more frequent than the others, above and beyond chance. To demonstrate the gain in terms of statistical risk, we repeated the ROC analysis of the previous section (Testing whether any (marble) model is "significantly more frequent than any other" section). For protected EPs, $A_{ROC} = 0.88$ for $n = 16$ subjects and $A_{ROC} = 0.95$ for $n = 64$ subjects. The power was 59.3% for $n = 16$ subjects and 94.8% for $n = 64$ subjects. The disambiguation threshold was 0.48 for $n = 16$ subjects and 0.51 for $n = 64$, with associated total error rates of 0.35 and 0.09. This represents a considerable improvement over the unprotected EP. For completeness, Table 1 summarizes the results for BOR, unprotected and protected EPs (cf. Testing whether any (marble) model is "significantly more frequent than any other" section above).

Table 1
Detecting differences in marble frequencies: ROC analysis of BOR (top), unprotected (middle) and protected EPs (bottom). The area under the ROC curve (A_{ROC}), power (at 5% false positive rate), disambiguation threshold and its associated total error rate (TER) are given for both $n = 16$ and $n = 64$ sample sizes.

	$n = 16$				$n = 64$			
	A_{ROC}	Power	Threshold	TER	A_{ROC}	Power	Threshold	TER
P_0	0.90	65.3%	0.55	38%	0.98	94.6%	0.49	9%
$\max_k \varphi_k$	0.81	47.5%	0.74	52%	0.90	68.2%	0.79	34%
$\max_k \tilde{\varphi}_k$	0.88	59.3%	0.48	35%	0.95	94.8%	0.51	9%

Application to Dynamic Causal Modeling

A key application of random effects BMS is Dynamic Causal Modeling (DCM), which was introduced to study the effective connectivity among brain regions using neuroimaging data. At the core of DCM are biophysical models that describe how the brain is wired and how it responds to different stimuli. DCM then embeds these models into a formal (Bayesian) statistical framework that allows for parameter estimation and model comparison when analyzing neuroimaging time series. We have summarized the relevant mathematical details in Appendix 3 of this manuscript (see also Daunizeau et al., 2011a for a recent review).

In this section, we focus on a simple DCM three-region network comparison, namely: parallel (m_1) versus serial (m_2) connectivity structures (see Fig. 2). The main difference between this example and the above Polya's urn treatment comes from the uncertain nature of within-subject (relative) empirical evidence in favour of candidate models. In brief, we expect natural variations in within-subject log-Bayes factors to induce a higher model selection error risk at the group level. We simulated synthetic fMRI time series y under m_1 and m_2 (64 Monte-Carlo repetitions for each model, session duration: 10 min; TR = 2 s; SNR = −20 dB). Each dataset was then inverted under both models, yielding $64 \times 2 = 128$ model evidences.

First, we checked that models could be disambiguated on the basis of 'within-subject' Bayesian model comparison. Let $LBF = \log p(y|m_1) - \log p(y|m_2)$ be the log Bayes factor that measures the relative evidence in favour of m_1 against m_2. Fig. 2 shows the Monte-Carlo empirical distributions of LBF for data simulated

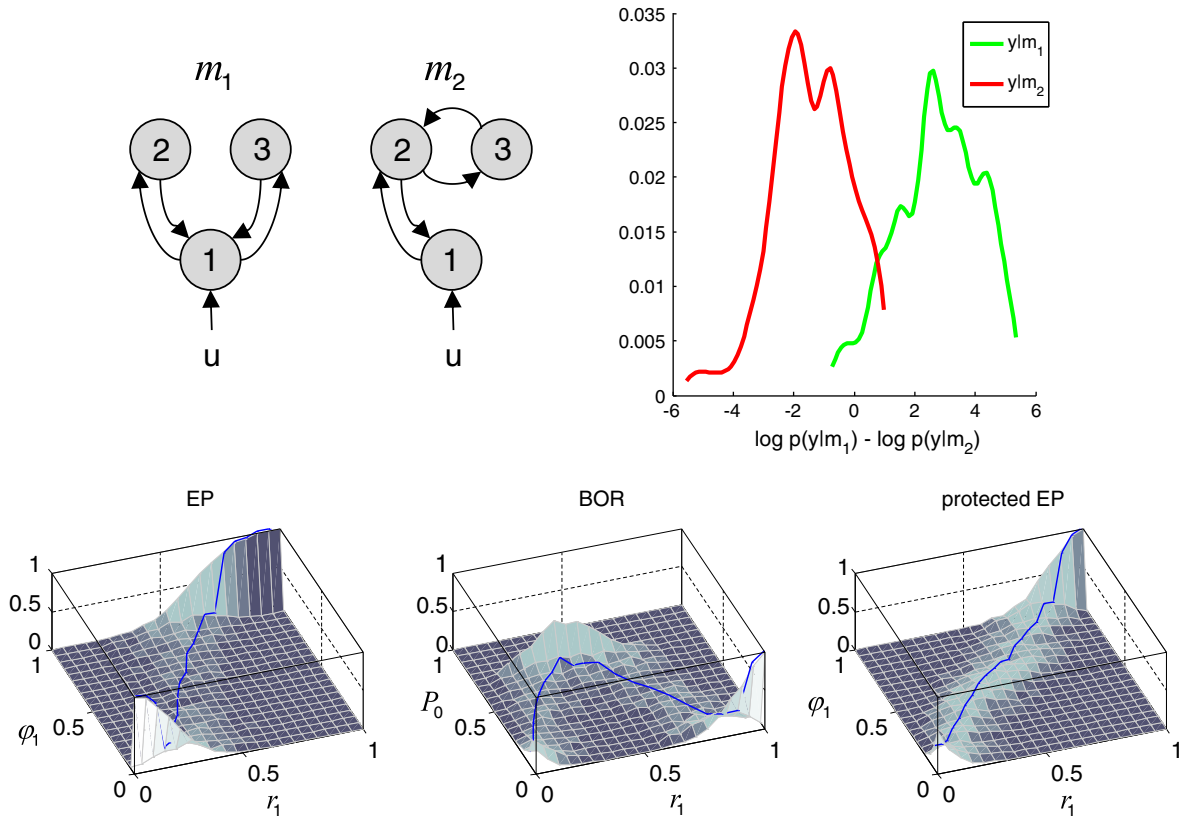

Fig. 2. Group-BMS: application to DCM. Upper-left: The two candidate three-region network models that will be compared at the group level (m_1: parallel architecture, m_2: serial architecture). Upper-right: Mote-Carlo distribution of – subject-level – log Bayes factors *LBF*, given data generated either under model m_1 (green) or under model m_2 (red). Lower-left: Monte-Carlo histogram (z-axis) of unprotected EP φ_1 (y-axis) as a function of frequency r_1 of model m_1 in the population (x-axis). The blue line indicates the Monte-Carlo average. The group-level null H_0 is situated at $r_1 = 1/2$. Lower-middle: same format, for BOR P_0. Lower-right: same format, for protected EP $\tilde{\varphi}_1$.

either under m_1 or under m_2. Statistical analysis confirms that the mean *LBF* is significantly positive (resp. negative) for data simulated under m_1 (resp. m_2), at $p = 10^{-4}$. However, one can see that the variability around these mean effects is likely to induce some confusion, and thus increase the posterior uncertainty at the group level.

We then wanted to assess the impact of the population profile, in terms of model frequencies. We thus spanned the frequency r_1 of m_1 from 0 to 1. For each frequency, we then randomly draw 256 groups of subjects (sample size: $n = 16$), from the multinomial distribution given in Eq. (1). For each group of subjects, we derived the BOR, as well as the protected and unprotected EP of model m_1 using the group-BMS approach. Their empirical Monte-Carlo distributions can be eyeballed in Fig. 2. One can see that unprotected EP's distribution tends to extreme values for relatively small departures from H_0 (equal model frequencies, i.e.: $r_1 = r_2 = 1/2$). In addition, its distribution under the null is almost flat. As expected, the BOR is maximal around the null, and minimal for extreme values of the true model frequency (i.e. $r_1 \to 0$ or $r_1 \to 1$). This eventually refocuses the distribution of the protected EP under the null, which is centered on $\tilde{\varphi} = 1/2$ and shows no extreme value. This means that observing an extreme protected EP is strong evidence for H_1, which is the main difference between protected and unprotected EPs.

Finally, we reproduced the ROC analysis of Testing whether any (marble) model is "significantly more frequent than any other" section, by splitting samples into H_0 (equal model frequencies, i.e.: $r_1 = r_2$) and H_1 (pooled samples for all $r_1 \neq r_2$). Table 2 summarizes the comparison of (protected and unprotected) maximum EPs and BOR, with respect to their relative ability to disambiguate between H_0 and H_1.

Although power and area under the ROC curve are similar, one can see how strikingly different are the disambiguation thresholds. In brief, unprotected exceedance probabilities have to be of the order of 0.99 to indicate strong evidence in favour of H_1. This is because extreme EP values are likely under H_0 (cf. Fig. 2). For example, an unprotected exceedance probability of 0.75 yields a total error rate of about 80%. Protected exceedance probabilities do not suffer from such overconfidence bias (e.g., for the same numerical value, the total error rate is less than 60%).

In the next section, we turn to the more pragmatic issue of how to pool evidence from multiple subjects to determine the form of the model that best explains their responses and how this relates to testing for the role of specific model parameters.

Random effects BMS and classical random effects analysis of parameter estimates

In this section, we focus on a specific question, namely "whether a model parameter is zero or not" at the group level. In a classical setting, this is typically addressed using a two-sided *t*-test on the parameter of interest. Effectively, this relies on the parameter estimate — from each subject — as a summary statistic to perform a random effects analysis;

Table 2
Detecting differences in model frequencies: ROC analysis of BOR (top), EPs (middle) and protected EPs (bottom). The area under the ROC curve (A_{ROC}), power (at 5% false positive rate), disambiguation threshold and its associated minimal total error rate (TER) are given for groups of $n = 16$ subjects.

	A_{ROC}	Power	Threshold	TER
P_0	0.76	48.4%	0.67	56%
$\max_k \varphi_k$	0.75	47.5%	0.99	56%
$\max_k \tilde{\varphi}_k$	0.76	48.4%	0.84	56%

testing whether the group mean is significantly different from zero. However, one could also perform a group BMS with two models (with and without the parameter of interest) and report the protected EP of the full model; i.e., the probability that the parameter is more frequently present than absent in the population. The difference between the two approaches is fundamental. In brief, classical random-effect analysis detects whether parameter estimates are consistent across subjects. In contrast, the group-BMS approach identifies the proportion of subjects, who are best described in terms of the full model. Critically, this is not a statement about the consistency of parameter estimates over subjects. In this section, we compare both approaches and identify the conditions in which they will yield similar (resp. different) conclusions, using simple numerical simulations. We then demonstrate how this translates to the context of Dynamic Causal Modeling (DCM) — a key application domain for random effects BMS. In particular, we investigate the impact of SNR, sample size, magnitude of the group mean and group variability.

Linear mixed-effect analysis

Recall the form of the linear mixed-effects model (see, for example, Friston et al., 2007a for an application to fMRI data analysis):

$$\begin{aligned} y_i &= X\beta_i + e_i^{(1)} \\ \beta_i &= \beta + e_i^{(2)} \end{aligned} \tag{8}$$

where y_i is subject i's data, X encodes some experimental factors that are weighted by (unknown) parameters β_i, $e_i^{(1)}$ are the model's residuals, β is the group mean and $e_i^{(2)}$ model random effects over subjects. In a classical setting, this model can be used to assess whether there is an effect of X at the group level, by testing $\beta = 0$. Under the assumption that the variance of $e_i^{(1)}$ is roughly the same over subjects, this is simply done using a t-test on the within-subject estimates $\hat{\beta}_i$ (e.g., least-square solution to the first line in Eq. (8)). This is the summary statistic approach to mixed effects, where the parameter estimates are used to summarize subject specific effects (Friston et al., 2005; Holmes and Friston, 1998).

In a Bayesian setting, one could invert the whole random-effect model given in Eq. (8), and quantify the evidence in favour of a non-zero group-mean β (using model comparison). This is known to be qualitatively similar to the above summary statistic approach (Penny et al., 2007). Alternatively, one could use group-BMS to compare two models: the full model m_1 (first line of Eq. (8)) and a reduced (null) model m_0, which assumes $\beta_i = 0$ (i.e. a prior mean of zero with infinite prior precision). The latter approach culminates in the derivation of the protected EP $\tilde{\varphi}$ for the full model m_1 compared to the reduced model m_0.

In practice, these two procedures may not give the same answer. Fig. 3 summarizes a simple set of simulations that reveal the crucial difference between classical random effects analysis and group-BMS. In brief, we simulated data under the model in Eq. (8), for $n = 16$ subjects, where we controlled the distribution of the within-subject effects β_i, in terms of how consistent (across subjects) and strong (compared to residuals $e_i^{(1)}$) they are. We now consider four scenarios of agreement and disagreement between classical mixed effects analysis of parameter estimates and random effects BMS:

- Scenario A: within-subject effects are both consistent and strong. More precisely, we simulated within-subject data (cf. first line of Equation 8) with $\beta_i = 1$ for all subjects (SNR = 1 dB). Parameter estimates for each subjects are shown on Fig. 3. In this case, both the classical ($p < 10^{-5}$) and the Bayesian ($\tilde{\varphi} \approx 1$) inference agree on the presence of the effect at the group level.
- Scenario B: within-subject effects are strong but inconsistent (i.e.: half of the subjects have a positive effect – $\beta_i = 1$ – and the others have a negative effect — $\beta_i = -1$; SNR = 1 dB). In this case, the classical approach finds no effect ($p = 0.98$) but random effects BMS tells us that there is an effect ($\tilde{\varphi} \approx 1$).
- Scenario C: within-subject effects are consistent but (half are) weak (i.e. $\beta_i = 0$). In this case, the classical approach tells us: "there is an effect" ($p < 10^{-3}$) but random effects BMS disagrees ($\tilde{\varphi} \approx 0.5$).
- Scenario D: within-subject effects are both inconsistent and weak ($\beta_i = 0$: 8 subjects, $\beta_i = -1$: 4 subjects, $\beta_i = 1$: 4 subjects). In this case, both approaches agree with each other in finding no effect ($p = 0.99$ and $\tilde{\varphi} \approx 0.5$).

These examples highlight the difference between the two approaches. In brief, classical tests are sensitive to the consistency of the signed effect (parameter estimate) across subjects. In contrast, random effects BMS cares about the proportion of subjects who show strong evidence for this parameter, irrespective of its sign.

Application to Dynamic Causal Modeling

Since VB model inversion provides both parameter estimates and model evidence, both classical and Bayesian forms of group-level inference can be found in the DCM literature. In this section, we ask

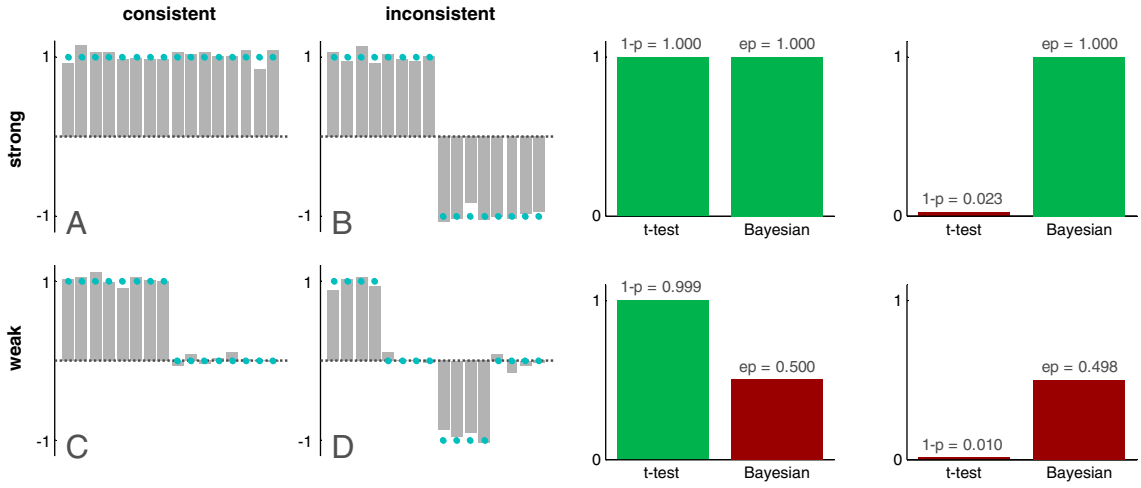

Fig. 3. The difference between group-BMS and classical random effect analysis. This figure summarizes the impact of two distinct features of group data, namely: (i) whether there is clear within-subject evidence for the effect of interest (upper versus lower panels), and (ii) whether the effect of interest is consistent across subjects (left versus right panels). Left: parameter estimates for each of the $n = 16$ subjects, for scenarios A, B, C and D (see main text). Simulated parameters are depicted using green circles. Right: Corresponding p-values (classical t-test for the mean) and EP. These test statistics are highlighted in green when the test is positive and in red otherwise.

how these techniques compare in terms of standard statistical risk (specificity and sensitivity measures) using Monte Carlo simulations of synthetic fMRI data. Our hope here was to provide an extensive comparison, across a wide range of experimental conditions. This serves to identify the factors (e.g., within- and between-subject variability, number of subjects) that influence the performance of these techniques.

We focus on addressing the question of whether or not some input has a modulatory effect on a network connection. In brief, we used a simplified version of the exemplar analysis in Friston et al. (2003) that tries to identify the impact of experimental factors (photic stimulation and motion) in a three-region network (V1, V5 and SPC[3]). In particular, we address whether motion modulates the connection from V1 (a visual input region) to V5 (a motion sensitive area): see Fig. 4.

Let b be the modulatory effect, and m_0 and m_1 the models under which $b = 0$ and $b \neq 0$, respectively. We simulated synthetic fMRI time series y under m_0 and m_1 (session duration: 10 min; TR = 2 s). When simulating data under m_1, we controlled the mean and variance ($\eta_b = 0$, 0.3 or 0.6; and $\sigma_b = 0.2$ or 0.5) of the population distribution of modulatory effects. Note that the population mean η_b can be 0 under m_1 (but, in contradistinction to m_0, the population variance σ_b is non-zero). The summary statistics of the population distribution constitute the first factor of our design ($1 + 3 \times 2$ levels: m_0 or m_1, with three population means and two population variances). We also varied the number of subjects in the group ($n = 2, 4$ or 8 subjects), and the signal-to-noise ratio (SNR $= -40$ dB to 0 dB). These constitute the two other factors of our $(1 + 3 \times 2) \times 3 \times 2$ factorial design, for which we performed 100 Monte-Carlo simulations.

Let $LBF_i = \log p(y_i|m_1) - \log p(y_i|m_0)$ be the log Bayes factor that measures the (within-subject) evidence in favour of the presence of the modulatory effect (m_1 against m_0). The derivation of $P(y|H_1)$ and $P(y|H_0)$ given subject-wise model evidences L_{ik} are described in Appendix 2.

We now list all the group-level inference approaches (and exemplar papers using them) that we have compared:

- Random effects group-BMS (simply abbreviated as *BMS* in the following). Here, using the protected exceedance probability.
- Classical random-effects analysis of parameter estimates This is simply the second-level analysis of the mixed-effects model above, which consists in rejecting the null if $b \neq 0$ according to a t-test on the subject-wise parameter estimates (see, for example, Leff et al., 2008).
- Positive evidence ratio (*PER*). This idea was introduced in Stephan et al. (2007), based on the definition of "positive evidence" in favour of a given model (Kass and Raftery, 1995). The PER is simply the number of subjects with positive evidence in favour of m_1 (i.e., $LBF_i > \log(3)$) divided by the number of subject with positive evidence in favour of m_0 (i.e., $LBF_i < -\log(3)$). One would then reject m_0 if $PER > 1$. Other variants of the same idea have used a binomial test on the number of subjects showing positive evidence (Ethofer et al., 2006).
- Classical random effects analysis of log-evidences. Here, the idea is to test how consistent the evidence is in favour of m_1, across subjects. One way to do this is to test the null hypothesis that $LBF = 0$ according to a one-sided t-test on the subject-wise (log) Bayes factors; this was suggested by Stephan et al. (2009) as the classical complement for random effects BMS (for a practical application see Chen et al., 2009). For nested models, some authors have also used a classical chi-squared test on the group's log Bayes factor $\sum_i LBF_i$ (Vuong, 1989). This needs to be distinguished from a *fixed-effect BMS* analysis that sums the subject-wise (log) Bayes factors to yield the pooled evidence for m_1, under the assumption that all subjects present the same model. One could then use the usual definition of "positive evidence" at the group level, i.e. reject m_0 if $\sum_i LBF_i > \log(3)$ (see e.g., Garrido et al., 2007).

[3] SPC = superior parietal cortex.

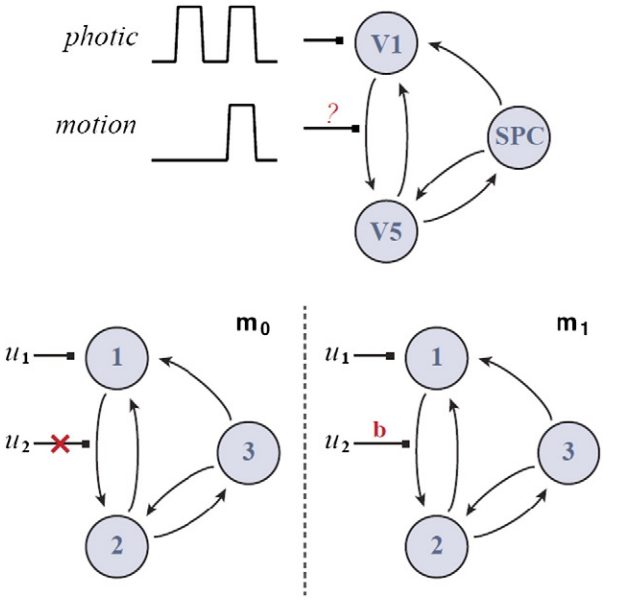

Fig. 4. Simulation set-up and model space. Top: summary of the exemplar analysis in Friston et al. (2003). The question we address is whether u_2 modulates the connection from V1 (region 1) to V5 (region 2). Bottom: corresponding dynamic models for the two competing hypotheses ($m_0 : b = 0$ and $m_1 : b \neq 0$).

First, we ensured that the modulatory effect was identifiable under m_1. Fig. 5 summarizes an exemplar inversion, in terms of the first two moments (μ and Σ) of the approximate posterior density q on DCM parameters. One can see that the modulatory effect is not confounded by hemodynamic changes (although it is only partially identifiable from the connectivity at rest).

Second, we checked that subject-wise Bayes factors behaved as expected. Fig. 6 summarizes one Monte-Carlo simulation, in terms of the ($n = 8$) within-subject Bayes factors, as a function of both the mean (η_b) and the variance (σ_b) of the population distribution of the modulatory effect. Overall, one can see that increasing the population mean η_b increases the average log Bayes factors, whereas increasing the population variance σ_b increases the variability of log Bayes factors across subjects.

We then conducted a Receiver Operating Characteristic (ROC) analysis to compare the different group-level approaches in terms of their ability to discriminate between m_0 and m_1. Fig. 7 depicts the resulting ROC curves, revealing how the sensitivity/specificity trade-off of each approach varies as the corresponding threshold on the test statistic changes.

These results show that the main effect of decreasing SNR is to decrease the area under the ROC curve, which measures the overall ability to discriminate between m_0 and m_1. In addition, increasing the group size (n) improves performance. Finally, there is an interaction between the effects of the population mean (η_b) and variance (σ_b) on the area under the ROC curve. In brief, the performance improvement due to an increase in η_b is dampened by an increase in σ_b. We now turn to a quantitative comparison between the group-level inferences.

Fig. 8 summarizes the effect of group size and SNR on the area under the ROC curve, after having pooled all the data simulated under m_1. There is a significant interaction between group size, SNR and method on the area under the ROC curve. More precisely, at high SNR (SNR = 0 dB), all methods (except the test on parameter estimates) perform equally well (and almost perfectly). At low SNR

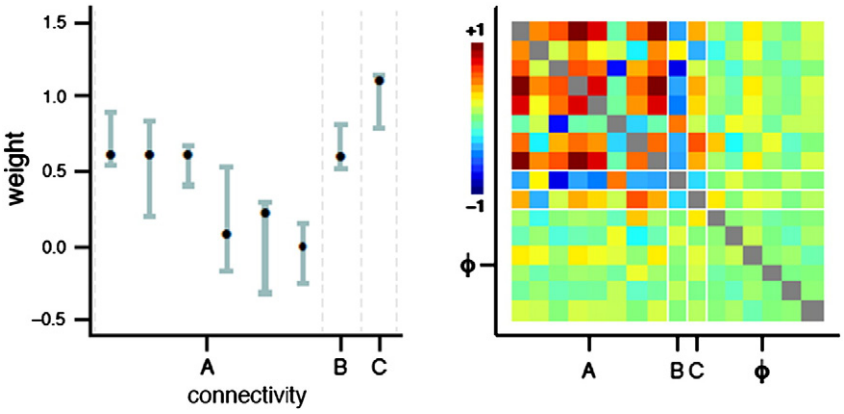

Fig. 5. Example of model simulation and inversion. Left: simulated parameters (dots) and their estimated values (95% credible interval in gray lines) for a typical Monte-Carlo simulation. Block A corresponds to network connectivity weights, B is the modulatory effect and C is the driving effect of u_1 onto region 1. Right: posterior correlation matrix of the DCM parameters (green suggests there is no identifiability issue).

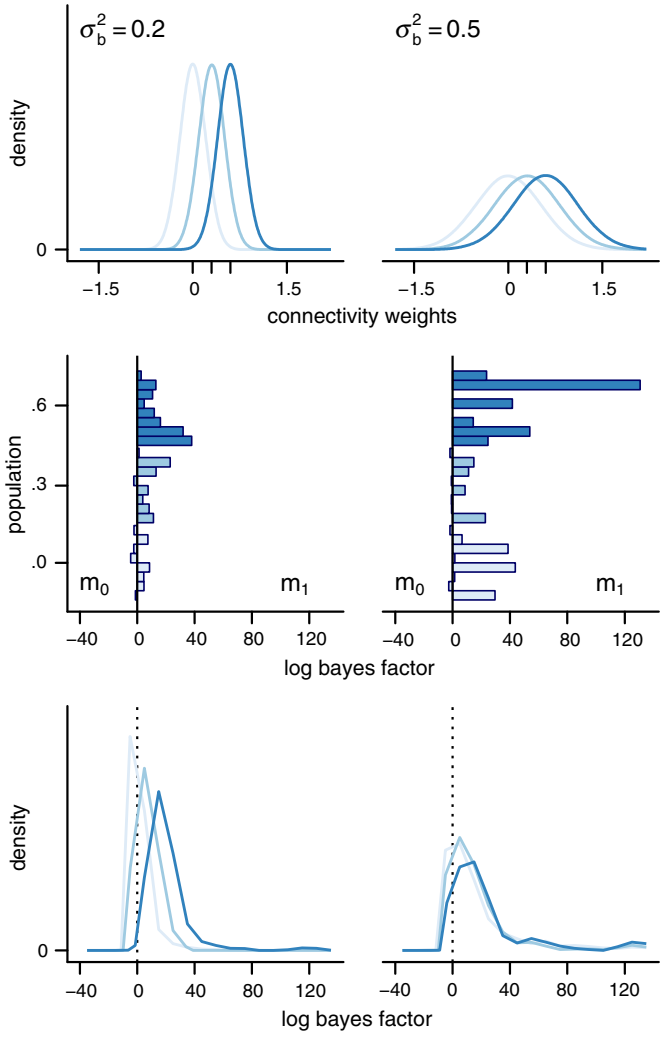

Fig. 6. Effect of the population distribution moments onto within-subjects Bayes factors. Left: small population variance ($\sigma_b{}^2 = .2$); Right: high population variance ($S_b{}^2 = .5$). In all panels, the color indicates the population mean (gray: $\eta_b = 0$, cyan: $\eta_b = .3$, blue: $\eta_b = .6$). Upper panels: population distributions over the modulatory effect b. Middle panels: Within-subjects (log) Bayes factors (x-axis) for a typical Monte-Carlo simulation with $n = 8$ subjects (y-axis) and SNR = 0 dB. Lower panels: Monte-Carlo distribution of within-subjects (log) Bayes factors (SNR = 0 dB).

(SNR = −40 dB), all methods are at chance level. At intermediate SNR levels (SNR = −20 dB), there is a clear ranking of all methods (and random effects BMS performs best). Note that for all SNR levels, the test on parameter estimates always exhibits the poorest performance level.

In addition, these results suggest that all methods perform rather poorly, when compared to *BMS* (at least for intermediate SNR levels). These simulation results replicate previous empirical results in Stephan et al. (2009), which highlighted the superior performance of random effects BMS compared to a classical random effects analysis of log model evidences. This global tendency is confirmed when inspecting the sensitivity (statistical power) of each method when fixing its type I error rate (false positive rate) to 5%.

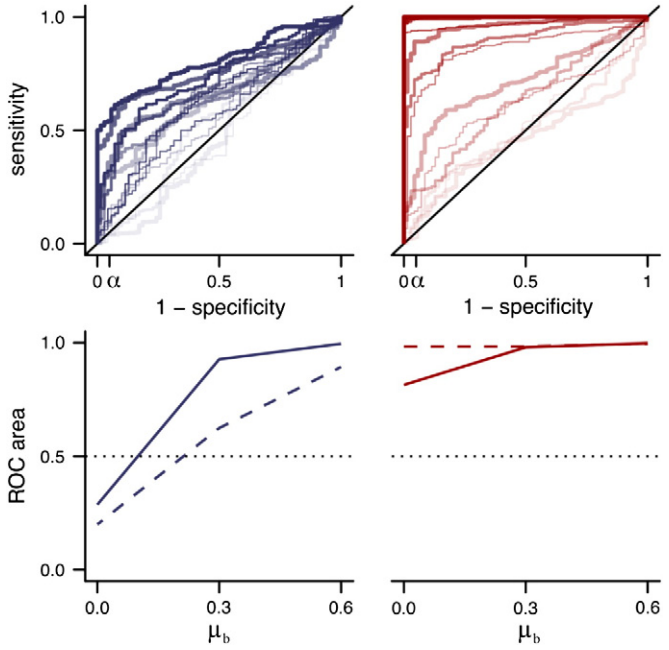

Fig. 7. ROC comparison of group-level approaches. This figure focuses on the comparison of classical RFX (Left) and group-BMS (Right) approaches. Upper panels: ROC curves (x-axis: 1-specificity, y-axis: sensitivity), averaged across population profiles. A darker color (resp. a thicker line) indicates an increase in SNR (resp. in group size n). Lower panels: Areas under the ROC curve (y-axis) as a function of the population mean (x-axis) and the population variance (solid lines: $\sigma_b{}^2 = .2$; dashed lines: $\sigma_b{}^2 = .5$), for SNR = −20 dB and $n = 8$ subjects.

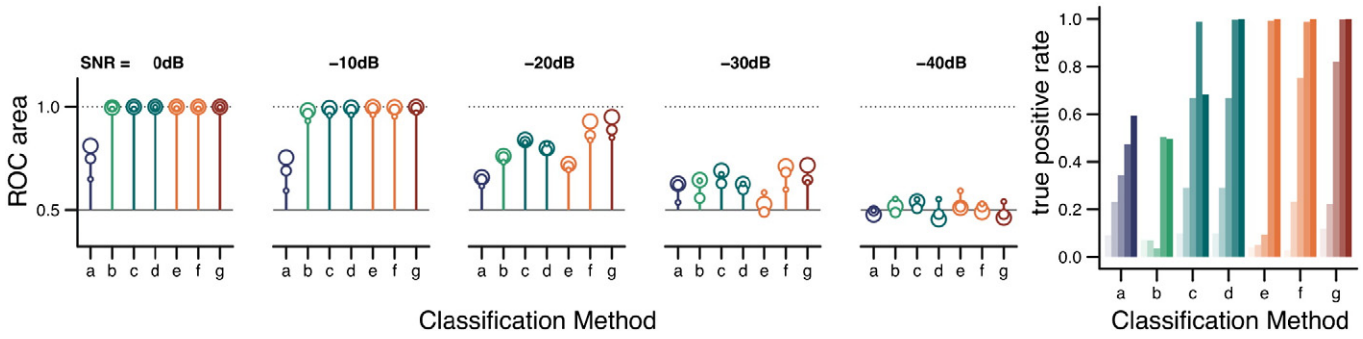

Fig. 8. Global ROC comparison of group-level approaches. Left panels: areas under the ROC curve (y-axis), averaged across SNR and group size n for all tested methods: (a) t-test on parameter estimates, (b) t-test on log-evidence, (c) Chi-squared test (d) positive evidence for the group Bayes factor, (e) ratio and (f) binomial probability of positive evidence counts, (g) group BMS as summarized by the protected exceedance probability. The color indicates the method (violet: classical RFX, green: RFX-LE, blue: PER, orange: SBF2-3, red: BMS). Simulated SNR decreases from left to right (10 dB per panel). Right panel: Statistical power (y-axis) at a Type I error rate of 5% (cf. marker α on upper panels of Fig. 6).

Between-group and between-condition BMS

In this section, we address the relationship between different treatment conditions and groups; for example, dealing with one group of subjects measured under two conditions,[4] or two groups of subjects. Until now, condition and group effects have been addressed by performing random effects BMS independently for the different conditions or groups, and then checking anecdotally to see whether the results of random effects BMS were consistent (see, e.g., van Leeuwen et al., 2011). This approach is limited, because it does not test the hypothesis that the same model describes the two conditions or groups. In this section, we address the issue of evaluating the evidence for a difference – in terms of models – between conditions (or groups).

Between-conditions comparison

In the following, we assume that the experimental design includes p conditions, to which a group of n subjects were exposed. Let y_{ij} be the ith subject's response to the jth condition, which are to be interpreted with K alternative models. We assume that a Bayesian subject-level analysis has provided us with the log-evidence $L_{ijk} = \log p(y_{ij}|m_{ijk})$ of the kth model, for the ith subject under the jth condition, with $i = 1,..., n$, $j = 1,..., p$ and $k = 1,..., K$. One can think of the p conditions as inducing an augmented model space composed of K^p p-tuples t_i that encode all combinations of candidate models and conditions. Here, any p-tuple t_i identifies the models associated with each condition (which may or may not be the same). The log-evidence $\widetilde{L}_{ih} = \log p(y_i|t_{ih})$ of the hth tuple, for the ith subject can be derived by summing up the log evidences over the appropriate conditions.

Random effects BMS can then be used to identify the best of these p-tuples at the group level, by passing the log-evidences $\widetilde{L}_{ih}$ to the random-effect BMS. To assess the probability that the same model underlies all conditions, one could use "family" inference (Penny et al., 2010) on a partition of the K^p tuples that divides them into a first subset, in which the same model underlies all conditions, and a second subset containing the remaining tuples (with distinct condition-specific models). The ensuing protected EPs can then be used to test whether different conditions correspond to different models.

Fig. 9 shows an example with $K = 2$ models and $p = 2$ conditions. This induces four different 2-tuples (t_1 to t_4) that differ in which model (m_1 or m_2) is assumed to generate data under each condition ($y_1$and y_2). The log-evidence of these 2-tuples is derived as follows (dropping the subject's index):

$$\begin{cases} \widetilde{L}_1 = \log p(y|t_1) = \log p(y_1|m_1) + \log p(y_2|m_1) = L_{11} + L_{21} \\ \widetilde{L}_2 = \log p(y|t_2) = \log p(y_1|m_1) + \log p(y_2|m_2) = L_{11} + L_{22} \\ \widetilde{L}_3 = \log p(y|t_3) = \log p(y_1|m_2) + \log p(y_2|m_1) = L_{12} + L_{21} \\ \widetilde{L}_4 = \log p(y|t_4) = \log p(y_1|m_2) + \log p(y_2|m_2) = L_{12} + L_{22} \end{cases} \quad (9)$$

where L_{jk} refers to the log-evidence $\log p(y_j|m_k)$ of the kth model under the jth condition. The set of candidate 2-tuples can then be partitioned into two families, which differ in terms of whether the same model underlies both conditions ($f_= = \{t_1,t_4\}$) or not ($f_{\neq} = \{t_2,t_3\}$). The protected EP of family $f_=$ (resp. $f_{\neq}$) quantifies the probability that the two conditions rely more (resp. less) frequently on the same model than on different models. In what follows, we reproduce the analysis of the Application to Dynamic Causal Modeling section, in the aim of demonstrating the between-condition (group-level) BMS.

As in the Application to Dynamic Causal Modeling section, we focus on the comparison of parallel versus serial DCMs (model m_1 and m_2, respectively; cf. Fig. 2). Now, being exposed to two conditions, each subject is described by a particular 2-tuple (as opposed to models). We want to assess the impact of the frequency $r_=$ of the 2-tuples family $f_=$, which we systematically vary from 0 to 1. For each frequency, we randomly draw 256 groups of subjects (sample size: $n = 16$). Subjects belonging to family $f_=$ were equally likely to be associated with 2-tuples t_1 or t_4 (i.e., models m_1 and m_2, respectively). For each group of subjects, we derived the BOR, as well as the protected and unprotected EP ($\varphi_=$ and $\varphi_{\neq}$, respectively) of family $f_=$ using the group-BMS approach. Their empirical Monte-Carlo distributions can be eyeballed in Fig. 9.

One can see that these are qualitatively similar to those of the Application to Dynamic Causal Modeling section. For example, observing an extreme value (0 or 1) around the null is much less likely for protected than for unprotected EPs. However, when compared to "simple" model comparison (cf. Fig. 2), the BOR is less sensitive to deviations from the null. This is a non-trivial consequence of a slight subject-level model identifiability issue. In brief, group-level evidence in favour or against $f_=$ is obscured by subject-level model selection errors. Finally, we performed a ROC analysis, in the aim of assessing the ability of the scheme to detect whether family $f_=$ was more frequent than $f_{\neq}$. This was done by splitting the samples according to $r_= \leq r_{\neq}$ and $r_= \geq r_{\neq}$. Table 3 summarizes the comparison of protected and unprotected EPs, with respect to their relative ability to disambiguate between heterogeneous ($r_= \leq r_{\neq}$) and homogeneous ($r_= \geq r_{\neq}$) conditions.

[4] Here, we do not refer to experimental conditions that co-exist within the same measurement session, such as changes in task demands. Instead, we refer to conditions that differ across two measurements within the same subject, i.e., a session-wise difference such as drug application.

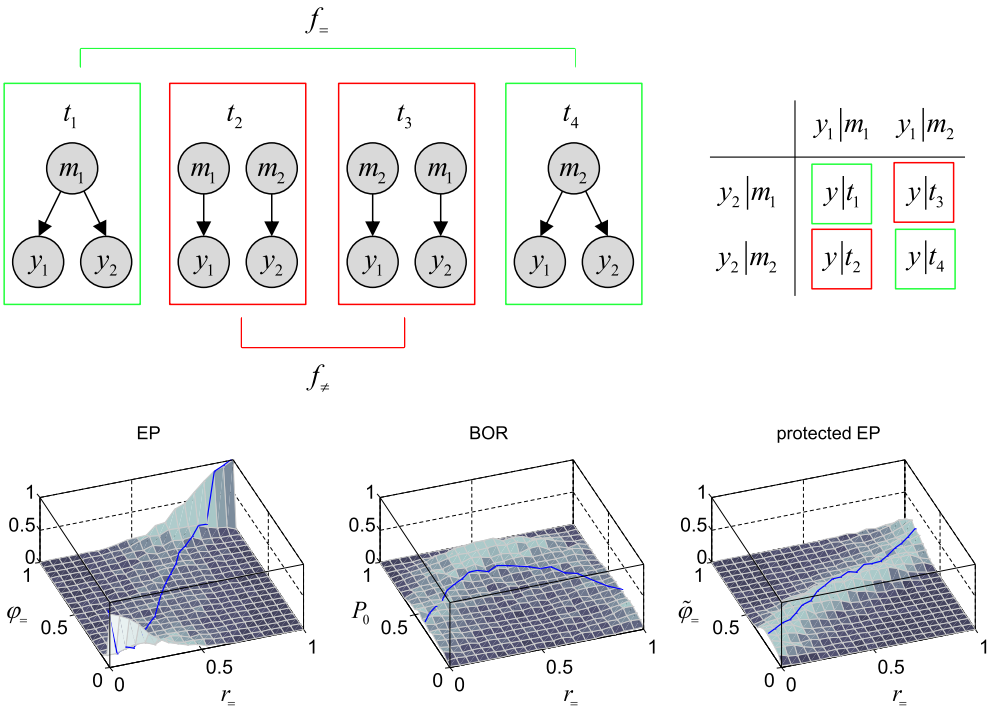

Fig. 9. Between-conditions BMS: application to DCM. Upper-left: The four 2-tuples t_1 to t_4 are shown, in terms of the possible combinations of associations of models (m_1 or m_2) and conditions (y_1 or y_2). These are partitioned into families $f_=$ (green) and $f_\neq$ (red), which correspond to homogeneous and heterogeneous conditions, respectively (see main text). Upper-right: This table summarizes the definition of the four 2-tuples in terms of the possible combinations of models (m_1 or m_2) for each condition (y_1 or y_2). Lower-left: Monte-Carlo histogram (z-axis) of unprotected EP $\varphi_=$ (y-axis) as a function of frequency $r_=$ of family $f_=$ in the population (x-axis). The blue line indicates the Monte-Carlo average. Lower-middle: same format, for BOR P_0. Lower-right: same format, for protected EP $\tilde{\varphi}_=$.

First, note that we did not include the BOR in this analysis, as it cannot discriminate between heterogeneous and homogeneous conditions (in fact, its area under the ROC curve is about 0.50). Second, we do not report the statistical power at 5%, because it does not make sense to break the symmetry between the two hypotheses in this case. Now, one can see that, overall, the discrimination ability of protected and unprotected EPs is identical and very high. However, the overconfidence of unprotected EPs expresses itself as a surprisingly high disambiguation threshold (which nonetheless yields identical total error rate). Taken together, our Monte-Carlo simulations provide face validity to the above between-condition group-BMS approach.

Between-group comparison

Assessing between-group model comparison in terms of random effects amounts to asking whether model frequencies are the same or different between groups. Let us partition all subjects into S subgroups, indexed by the variable s — for example, subgroups that have been exposed to different treatments. Let I_s be the sets of indices of subjects belonging to the sth subgroup, and $y_s = \{y_i\}_{i\in I_s}$ be the corresponding subset of data. As above, K alternative models are considered in relation to these subject-specific responses. This section addresses the question of disambiguating the two following hypotheses (at the group level):

- $H_=$: $\{y_1,..., y_S\}$ come from the same population, i.e. model frequencies r are the same for all subgroups.
- $H_\neq$: $\{y_1,..., y_S\}$ come from different populations, i.e. they have distinct model frequencies $r^{(s)}$.

Again, we use subject-specific (log-) model evidences $L_{ik} = \log p(y_i|m_{ik})$. These can be used to derive the evidence $p(H_{=/\neq}|y)$ of assumptions $H_=$ and $H_\neq$ at the group level. Under $H_=$, the datasets y_s can be pooled in the usual way to perform a standard random effects BMS, yielding a single evidence $p(y|H_=) = p(y|H_1)$, where $p(y|H_1)$ is given by Eq. (6). Under $H_\neq$, datasets y_s are marginally independent. In this case, the evidence $p(y|H_\neq)$ is the product of group-specific evidences. This implies that the posterior probability of $H_\neq$ is:

$$p(H_\neq|y) = \frac{p(y|H_\neq)}{p(y|H_\neq) + p(y|H_=)}$$
$$p(y|H_\neq) = \prod_{s=1}^{S} p(y_s|H_1) \quad (10)$$
$$p(y|H_=) = p(\bigcup_{s=1}^{S} y_s|H_1).$$

This statistic quantifies the likelihood that the subgroups have distinct model frequencies.

We now adapt the analysis of the Between-conditions comparison section above to the situation of between-group comparison (parallel versus serial DCMs, two groups of sample size $n = 16$ each). We want to assess the impact of the difference in the model frequency profile of the two groups. We thus varied the frequency r_1 of model m_1 from 0 to 1 for both groups, in a factorial way. For each pair of frequencies

Table 3
Detecting differences in model frequencies: ROC analysis of unprotected EPs (top) and protected EPs (bottom). The area under the ROC curve (A_{ROC}), disambiguation threshold and its associated minimal total error rate (TER) are given for groups of $n = 16$ subjects.

	A_{ROC}	Threshold	TER
$\max_k \varphi_k$	0.95	0.65	24%
$\max_k \tilde{\varphi}_k$	0.93	0.50	25%

$(r_1^{(1)},r_1^{(2)})$, we randomly draw two groups of subjects, 256 times. Each subject in each group was given either model m_1 or model m_2, according to the appropriate model frequency in the group (cf. Eq. (1)). For each pair of groups, we then derived the posterior probability $p(H_{\neq}|y)$ (cf. Eq. (10)). The first two moments of its empirical Monte-Carlo distribution, as a function of both group frequency profiles, can be eyeballed in Fig. 10.

As we expected, $p(H_{\neq}|y)$ reaches unity for very different frequency profiles ($r_1^{(1)} \approx 1 - r_1^{(2)}$), and is minimal for similar frequency profiles ($r_1^{(1)} \approx r_1^{(2)}$). However, one finds weaker statistical evidence for the latter situation, which yields $p(H_{\neq}|y) \approx 0.1$ on average. This is because even when the underlying models are all identical, natural variations in within-subject log-Bayes factors (cf. Fig. 2 and Application to Dynamic Causal Modeling section) induce some model selection errors. This eventually compromises the evidence in favour of $H_{=}$. In brief, partial non-identifiability issues make it more difficult to conclude in favour of $H_{=}$. Let us now focus on the Monte-Carlo standard deviation of $p(H_{\neq}|y)$. First, $p(H_{\neq}|y)$ shows minimal variability for either very different or very similar frequency profiles. Second, one can see that, for similar frequency profiles ($r_1^{(1)} \approx r_1^{(2)}$), its standard deviation increases around the null ($r_1^{(1)} \approx r_1^{(2)} \approx 1/2$). This is due to the bigger within-group variability in terms of models, which increases the chance probability of observing (seemingly) different frequency profiles. This eventually also impacts on the mean $p(H_{\neq}|y)$ for similar frequency profiles ($r_1^{(1)} \approx r_1^{(2)}$), which increases as the frequency profile tends towards the null.

Clearly, the question of whether it is useful to consider differences in model frequencies as diagnostic of a treatment effect deserves careful consideration. However, in situations where group or condition differences are expressed in terms of categorical differences between models, the approaches described above provide a principled way of making suitable inferences.

Discussion

In this work, we introduced three extensions of our original approach to random effects BMS (Stephan et al., 2009). First, we have described a *protected* exceedance probability that any model is more frequent than the others (above and beyond chance). Second, we have presented systematic simulations of various approaches to address questions about specific treatment effects on model parameters using group studies. Third, we considered approaches to between-condition and between-group BMS inference on models.

A major contribution of this paper is the re-evaluation of exceedance probabilities (EP), in terms of the statistical risk incurred when performing random effects BMS. We conclude that EP cannot be used to assess this statistical risk. More precisely, EPs are slightly overconfident — for example, if the best model has an EP of $\varphi = 0.95$, the probability that there is no difference in model frequencies is greater than 0.05. This is because the definition of EP does not consider a null model at the group level. In other words, chance is discounted as a potential explanation for the data. Although this does not invalidate EP-based ranking of candidate models, it means that one should not equate it with classical 5% significance thresholds.

Our reading of the literature suggests that there may have been a potential misunderstanding about the nature of exceedance probabilities. Recall that, as is evident from the Polya's urn treatment, one can think of r as the *frequency* (or proportion) of models within the population. Since r is not known with infinite precision, we quantify our (Bayesian) belief about it using a probability density function. After having observed the data, we can interrogate the posterior density over model frequencies in many ways. For example, its first-order moment is useful to define a posterior estimate $\langle r \rangle = E[r|y,H_1]$ of model frequencies, under H_1 (which refers to the prior assumption of "no bias"; cf. the Polya's urn model section). Error bars on this estimate can be derived from the posterior variance of r. In this context, EPs are simply the posterior probability $P(r_k > r_{k' \neq k}|y,H_1)$ that each model is more frequent than others, under H_1 — that frequencies can differ. From this perspective, it becomes natural to consider other assumptions regarding model frequencies, one of which being the null H_0 — that frequencies do not differ. This motivates *protected* EPs, which rely upon Bayesian model averaging to account for the possibility that there may be no difference in model frequencies.

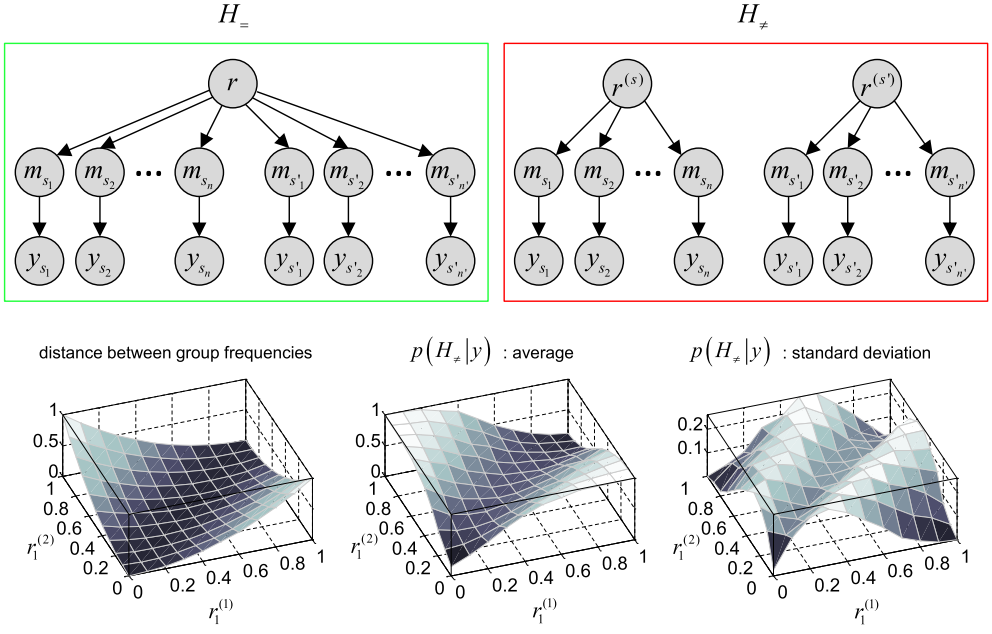

Fig. 10. Between-group BMS: application to DCM. Upper-left: The two group-level hypotheses are shown. $H_{=}$ assumes that models underlying both group datasets (y_s and $y_{s'}$) are samples from the same population frequency profile r. $H_{\neq}$ assumes that models underlying both group datasets (y_s and $y_{s'}$) are samples from different population frequency profiles $r^{(s)} \neq r^{(s')}$. Lower-left: quadratic distance (z-axis) between group-specific frequency profiles, as a function of $r^{(1)}$ (x-axis) and $r^{(2)}$ (y-axis). $H_{=}$ is situated on the main diagonal $r_1^{(1)} = r_1^{(2)}$. The group-specific null assumptions (H_0) correspond to $r_1^{(1)} = 1/2$ and $r_1^{(2)} = 1/2$, respectively. Lower-middle: same format, for the Monte-Carlo average of the evidence for a difference $p(H_{\neq}|y)$. Lower-right: same format, for the Monte-Carlo standard deviation of the evidence for a difference $p(H_{\neq}|y)$.

Effectively, this eliminates an overconfidence bias of exceedance probability, when interpreted in a frequentist sense; i.e., in terms of sensitivity and specificity.

The motivation for the protected EP was to correct for its slight overconfidence, which arises when one uses the maximum EP as an index of the evidence against the null. This bias is a consequence of discounting the null as a potential explanation for apparent differences in model frequencies. From this perspective, one has to consider H_1 as one candidate scenario for explaining the observed variability in subject-wise evidences. In fact, there is another scenario that we did not consider in detail, namely: the same model could have generated all the data. This is the assumption that underlies *fixed-effects* BMS (this is discussed in Stephan et al., 2009). The evidence for this model could also be quantified (results not shown). Note that considering fixed-effects in addition to H_1 and H_0 might increase the Bayesian omnibus risk. In other words, as they stand, protected EPs are not corrected for the possibility of fixed-effects. Having said this, we expect this correction to be less severe, because fixed-effects correspond to the limiting situation where model frequencies are all zero with the exception of one model. This means that neglecting fixed-effects in the derivation of protected EPs is conservative. The question of which scenarios or prior assumptions about model frequencies should be entertained is – of course – a question about prior beliefs about the random behaviour of models. In effect, the protected EP described in this paper is a special case in which, a priori, model frequencies cannot be zero or one — all these scenarios are considered, a priori, implausible.

In terms of the insights provided by the simulation results of the last section — we demonstrated that the classical random effects tests of parameter estimates are not suited to test for consistent *evidence* for the presence of model parameters across subjects. However, it is a perfectly valid method for asking whether an effect is consistent across subjects. In fact, our Monte-Carlo simulations demonstrate that its usefulness as the test for consistency generalizes to all DCM parameter estimates (results not shown). The point here is that these inferences do not compete with each other: they are essentially complementary and address different questions. In other words, it is perfectly reasonable to ask both types of questions, and thus use both approaches.

In the context of DCM, Bayesian model selection rests on the free energy approximation to the model evidence (Stephan et al., 2009). However, other approximations (that also penalize model complexity) have been used (Penny, 2012). The so-called *Bayesian Information Criterion* (Schwarz, 1978) and *Akaike's information criterion* (Akaike, 1973) are two common examples. We assessed the impact of model evidence approximations on the efficiency of random effects BMS (results not shown). In brief, group-BMS performs significantly worse with AIC and/or BIC indices than with the Free Energy. More precisely, we found statistical evidence that, in comparison to the Free Energy, BIC over estimates model complexity, whereas AIC underestimates it. This replicates the results in Penny (2012).

Finally, we would like to emphasize our take-home message. One may ask which summary statistics to report from a (potentially complex) group-level analysis. In brief, one should report the statistics that directly quantifies the statistical risk incurred when stating the assertion of interest (e.g. “there is no difference between group A and group B”). This is because readers can then both evaluate the strength of evidence, and compare the risk to any acceptable error rate. With this work, we hope to have shown that — in most cases — this risk can be evaluated in a relatively straightforward way.

Acknowledgments

This work was supported by the European Research Council (JD), by the Ville de Paris (LR), and by the IHU-A-ICM (JD, LR). KES acknowledges support by the René and Susanne Braginsky Foundation and KJF acknowledges support from the Wellcome Trust.

Conflict of interest

The authors declare that there are no conflicts of interest.

Appendix 1. Limit results for the Polya's urn model

In this section, we quantify the impact of a small difference in marble counts on the Bayesian omnibus risk (BOR) and on the exceedance probability (φ), in the context of Polya's urn model.

The Bayesian omnibus risk (BOR)

Recall that the Bayesian omnibus risk (BOR) can be defined in terms of the log-Bayes factor (LBF) that measures the evidence for H_1 (against H_0):

$$\begin{aligned} BOR &= \frac{1}{1+\exp(LBF)} \\ LBF &= \log\frac{p(m|H_1)}{p(m|H_0)} \\ &= \log\frac{K^n\ \Gamma(K\alpha_0)\prod_{k=1}^{K}\Gamma\left(\alpha_0+\sum_{i=1}^{n}m_{ik}\right)}{\Gamma(\alpha_0)^K\Gamma(K\alpha_0+n)}. \end{aligned} \tag{A1}$$

Consider the behaviour of LBF around the limiting case, where the observed number of marbles of each type is the same. Let us assume that we observe n/K marbles of each type, except for two (arbitrary) types, whose number of marbles are $n/K+\varepsilon$ and $n/K-\varepsilon$, respectively. Here, ε is a small perturbation around equal counts. Under flat priors on the marble frequencies ($\alpha_0=1$), one can express the LBF as a function of the perturbation ε:

$$\begin{aligned} LBF(\varepsilon) &= \log\frac{K^n\ (K-1)!\Gamma\left(1+\frac{n}{K}\right)^{K-2}\Gamma\left(1+\frac{n}{K}+\varepsilon\right)\Gamma\left(1+\frac{n}{K}-\varepsilon\right)}{\Gamma(K+n)} \\ &= \log\frac{K^n(K-1)!\left(\left(\frac{n}{K}\right)!\right)^{K-2}\left(\frac{n}{K}+\varepsilon\right)!\left(\frac{n}{K}-\varepsilon\right)!}{(K+n-1)!}. \end{aligned} \tag{A2}$$

We can make use of Stirling's approximation (Dan Romik, 2000) to yield the asymptotic behaviour of the log-Bayes factor:

$$\begin{aligned} LBF(\varepsilon)\approx & n\log K+(K-1)\log(K-1)-(K-1) \\ & +(K-2)\left(\frac{n}{K}\log\frac{n}{K}-\frac{n}{K}\right) \\ & +\left(\frac{n}{K}+\varepsilon\right)\log\left(\frac{n}{K}+\varepsilon\right)-\left(\frac{n}{K}+\varepsilon\right) \\ & +\left(\frac{n}{K}-\varepsilon\right)\log\left(\frac{n}{K}-\varepsilon\right)-\left(\frac{n}{K}-\varepsilon\right) \\ & -(K+n)\log(K+n)+(K+n). \end{aligned} \tag{A3}$$

One can now use a Taylor expansion around $\varepsilon=0$ to arrive at an approximation of the omnibus Bayes factor:

$$\begin{aligned} LBF(\varepsilon)\approx & LBF(0)+\varepsilon\frac{\partial}{\partial\varepsilon}LBF|_{\varepsilon=0}+\frac{1}{2}\varepsilon^2\frac{\partial^2}{\partial\varepsilon^2}LBF|_{\varepsilon=0} \\ = & K\log\frac{K-1}{K+n}+n\log\frac{n}{K+n}-\log(K-1)+1+\frac{K}{n}\varepsilon^2 \end{aligned} \tag{A4}$$

where the linear term in the Taylor expansion vanishes. The limiting

behaviour of Eq. (A4) can now be used to derive the behaviour of the Bayesian omnibus risk as a function of ε ($BOR(\varepsilon)$, cf. Eq. (A1)).

First, note that: $BOR(0) \xrightarrow{n\to\infty} 1$, i.e., equal counts asymptotically yield unambiguous evidence for the null. Second, note that $BOR(\varepsilon) \leq BOR(0)$, i.e., the net effect of the perturbation ε is to decrease the evidence for the null. However, the contribution of the perturbation term is inversely proportional to the number of marbles n.

The exceedance probability

Recall that the exceedance probability is defined as follows: $\varphi_k = P(r_k > r_{k'\neq k} | m, H_1)$. As above, we are interested in the asymptotic behavior of $\max_k \varphi_k$, given small perturbations ε around equal counts. For the sake of simplicity, we will focus on the $K = 2$ case. In this case, the exceedance probability can be written as:

$$\begin{aligned}\varphi_1 &= \int_{1/2}^{1} p(r_1|m,H_1)dr_1 \\ &= \frac{\Gamma(2\alpha_0+n)}{\Gamma\left(\alpha_0+\sum_{i=1}^{n} m_{i2}\right)\Gamma\left(\alpha_0+\sum_{i=1}^{n} m_{i1}\right)} \underbrace{\int_{1/2}^{1} r^{\alpha_0+\sum_{i=1}^{n} m_{i1}-1}(1-r)^{\alpha_0+\sum_{i=1}^{n} m_{i2}-1} dr}_{I(m)}\end{aligned} \tag{A5}$$

where $I(m)$ is an integral that depends on the marbles' counts m. Again, let us assume that we observe $n/2 + \varepsilon$ marbles of type 1 and $n/2 - \varepsilon$ marbles of type 2, respectively. Under flat priors on the marble frequencies ($\alpha_0 = 1$), one can express the exceedance probability as a function of the perturbation ε:

$$\begin{aligned}\varphi_1(\varepsilon) &= \frac{\Gamma(2+n)}{\Gamma\left(1+\frac{n}{2}-\varepsilon\right)\Gamma\left(1+\frac{n}{2}+\varepsilon\right)} I(\varepsilon) \\ &= \frac{(n+1)!}{\left(\frac{n}{2}+\varepsilon\right)!\left(\frac{n}{2}-\varepsilon\right)!} I(\varepsilon).\end{aligned} \tag{A6}$$

Using integration by parts, one can derive a recurrence relation for the integral term $I(\varepsilon)$, as follows:

$$\begin{aligned}I(\varepsilon) &= \int_{1/2}^{1} r^{\frac{n}{2}+\varepsilon}(1-r)^{\frac{n}{2}-\varepsilon} dr \\ &= \frac{1}{\frac{n}{2}+1+\varepsilon}\left(\left[r^{\frac{n}{2}+1+\varepsilon}(1-r)^{\frac{n}{2}-\varepsilon}\right]_{1/2}^{1} + \left(\frac{n}{2}-\varepsilon\right)\int_{1/2}^{1} r^{\frac{n}{2}+\varepsilon+1}(1-r)^{\frac{n}{2}-\varepsilon-1} dr\right) \\ &= \frac{1}{\frac{n}{2}+1+\varepsilon}\left(-2^{-n-1} + \left(\frac{n}{2}-\varepsilon\right)\int_{1/2}^{1} r^{\frac{n}{2}+\varepsilon+1}(1-r)^{\frac{n}{2}-\varepsilon-1} dr\right) \\ &= \frac{1}{\frac{n}{2}+1+\varepsilon}\left(-2^{-n-1} + \left(\frac{n}{2}-\varepsilon\right) I(\varepsilon+1)\right).\end{aligned} \tag{A7}$$

Note that $I(n/2)$ can be solved analytically, i.e.:

$$I\left(\frac{n}{2}\right) = \int_{1/2}^{1} r^n dr = \frac{1}{n+1}\left[r^{n+1}\right]_{1/2}^{1} = \frac{1}{n+1}\left(1-2^{-n-1}\right). \tag{A8}$$

From Eq. (A7), backwards induction yields:

$$I\left(\frac{n}{2}-k\right) = k!(n-k)!\left(\frac{1}{(n+1)!} - 2^{-n-1}\sum_{i=0}^{k}\frac{1}{i!(n-i+1)!}\right) \tag{A9}$$

where k is an arbitrary integer. Now setting $k = n/2 - \varepsilon$ solves the integral calculus:

$$I(\varepsilon) = \left(\frac{n}{2}-\varepsilon\right)!\left(\frac{n}{2}+\varepsilon\right)!\left(\frac{1}{(n+1)!} - 2^{-n-1}\sum_{i=0}^{n/2-\varepsilon}\frac{1}{i!(n-i+1)!}\right). \tag{A10}$$

Eq. (A10) can now be inserted into Eq. (A6) to derive the following analytic expression for the exceedance probability:

$$\begin{aligned}\varphi_1(\varepsilon) &= (n+1)!\left(\frac{1}{(n+1)!} - 2^{-n-1}\sum_{i=0}^{n/2-\varepsilon}\frac{1}{i!(n-i+1)!}\right) \\ &= 1 - 2^{-n-1}\sum_{i=0}^{n/2-\varepsilon}\frac{(n+1)!}{i!(n-i+1)!} \\ &= 1 - 2^{-n-1}\sum_{i=0}^{n/2-\varepsilon}\binom{n+1}{i}\end{aligned} \tag{A11}$$

where $\binom{n+1}{i}$ is the binomial coefficient that counts the number of ways to sample i marbles from an urn containing $n + 1$ marbles. From Eq. (A11), one can derive the natural bounds of the exceedance probability, i.e.: (i) if there are equal counts of each marble type ($\varepsilon = 0$) then the exceedance probability is indecisive ($\varphi_1(0) = 1/2$), and (ii) if all sampled marbles are of the first type ($\varepsilon = n/2$), then the exceedance probability is unambiguous ($\varphi_1(n/2) = 1$). The lower bound follows from the theorem on the sum of binomial coefficients (Abramowitz and Stegun, 1968), i.e.:

$$\begin{aligned}\varphi_1(0) &= 1 - \frac{1}{2^{n+1}}\sum_{i=0}^{n/2}\binom{n+1}{i} \\ &= 1 - \frac{2^n}{2^{n+1}} \\ &= 1/2.\end{aligned} \tag{A12}$$

Now, let us approximate the derivative of the exceedance probability through the limiting case of finite differences (for the smallest difference in marble counts):

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}\varphi_1(\varepsilon) &\approx \lim_{\Delta\varepsilon\to 1}\frac{\varphi_1(\varepsilon+\Delta\varepsilon)-\varphi_1(\varepsilon)}{\Delta\varepsilon} \\ &= 2^{-n-1}\binom{n+1}{\frac{n}{2}-\varepsilon}.\end{aligned} \tag{A13}$$

This can now enter a Taylor expansion, to yield a linear approximation to the exceedance probability around small imbalance in the marble counts ($\varepsilon/n \to 0$):

$$\begin{aligned}\varphi_1(\varepsilon) &\approx \varphi_1(0) + \frac{\partial\varphi_1}{\partial\varepsilon}\Big|_0 \varepsilon + O\left(\varepsilon^2\right) \\ &= \frac{1}{2} + 2^{-n-1}\underbrace{\binom{n+1}{\frac{n}{2}}}_{\approx 2^{n+1}/\sqrt{\pi n}}\varepsilon \\ &= \frac{1}{2} + \frac{1}{\sqrt{\pi n}}\varepsilon\end{aligned} \tag{A14}$$

where the second line derives from Stirling's approximation. Note that Eq. (A14) can easily be generalized to the maximum exceedance probability, which, by definition, benefits from the difference ε in marble counts:

$$\left.\begin{aligned}\max_k \varphi_k &= \varphi_1 \quad \text{if } \varepsilon > 0 \\ \max_k \varphi_k &= 1-\varphi_1 \quad \text{if } \varepsilon < 0\end{aligned}\right\} \Rightarrow \max_k \varphi_k = \frac{1}{2} + \frac{1}{\sqrt{\pi n}}|\varepsilon| \tag{A15}$$

Overall, this means that the qualitative behaviour of the maximum exceedance probability is similar to the Bayesian omnibus risk. However,

in contradistinction to BOR, the contribution of the perturbation term to the maximum exceedance probability is inversely proportional to the square root $\sqrt{n}$ of the number of marbles. This means that the two statistics may not show the same statistical power. In fact, we demonstrate this difference using Monte-Carlo simulations (see main text).

Appendix 2. Model evidence for the random-effect model

In this section, we derive an approximation to the Bayesian omnibus risk, given the VB treatment of random effects BMS, as described in Stephan et al. (2009).

First, we derive the group evidence for the null $p(y|H_0)$: under the null, model frequencies are a priori assumed to be fixed to $1/K$, where K is the number of models to be compared. This means that the prior on model labels (Eq. (5) in the main text) simplifies to $p(m_i|H_0) = \prod_{k=1}^{K} 1/K^{m_{ik}}$. This induces the following Free energy:

$$\begin{aligned} F_0 &= \left\langle \sum_{i=1}^{n}\sum_{k=1}^{K} m_{ik}\log p(y_i|m_{ik}=1)\right\rangle + \left\langle \sum_{i=1}^{n}\sum_{k=1}^{K} m_{ik}\log p(m_{ik}=1|H_0)\right\rangle + S(q) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} L_{ik}w_{ik} - \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log K - \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log w_{ik} \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(L_{ik} - \log K - \log w_{ik}) \end{aligned} \tag{A17}$$

where q is the (multinomial; see above) variational posterior on model labels m_i, whose sufficient statistics is w_{ik}, i.e. the probability that the ith subject is best described by the kth model (under H_0). Maximizing F_0 with respect to q yields the posterior density on model labels:

$$\begin{aligned} \log q(m) &= \sum_{i=1}^{n}\sum_{k=1}^{K}\left[\log p(y_i|m_{ik}=1)\right]^{m_{ik}} + \sum_{i=1}^{n}\log p(m_i|H_0) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} m_{ik}L_{ik} - \sum_{i=1}^{n}\sum_{k=1}^{K} m_{ik}\log K = \sum_{i=1}^{n}\sum_{k=1}^{K} m_{ik}(L_{ik} - \log K) \end{aligned} \tag{A18}$$

where we have omitted constant terms for clarity. One can see that the posterior density on model labels has a multinomial form, whose sufficient statistics w_{ik} are given by:

$$\begin{aligned} q(m) &= \prod_{i=1}^{n} q(m_i) \\ q(m_i) &= \prod_{k=1}^{K} {w_{ik}}^{m_{ik}} \\ w_{ik} &= \frac{\exp(L_{ik})}{\sum_{k'=1}^{K}\exp(L_{ik'})}. \end{aligned} \tag{A19}$$

Substituting Eq. (A19) into Eq. (A17) yields the evidence for the null $F_0 = \log p(y|H_0)$.

Let us now focus on H_1, i.e. the random-effects group-BMS model we introduced in Stephan et al. (2009). Note that the Dirichlet prior on model frequencies (Eq. (3)) induces the following Free Energy bound:

$$\begin{aligned} F_1 &= \left\langle \sum_{i=1}^{n}\sum_{k=1}^{K} m_{ik}\log p(y_i|m_{ik}=1)\right\rangle + \left\langle \sum_{i=1}^{n}\sum_{k=1}^{K} m_{ik}\log p(m_{ik}=1|r_k)\right\rangle \\ &\quad + \langle \log p(r|H_1)\rangle + S(q) = \sum_{i=1}^{n}\sum_{k=1}^{K} L_{ik}z_{ik} + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\langle \log r_k\rangle \\ &\quad + \sum_{k=1}^{K}(\alpha_0 - 1)\langle \log r_k\rangle + \log\Gamma(K\alpha_0) - K\log\Gamma(\alpha_0) - \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log z_{ik} \\ &\quad + \sum_{k=1}^{K}\log\Gamma(\alpha_k) - \log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) - \sum_{k=1}^{K}(\alpha_k - 1)\langle \log r_k\rangle\langle \log r_k\rangle \\ &\quad = \psi(\alpha_k) - \psi\left(\sum_{k'=1}^{K}\alpha_{k'}\right) \end{aligned} \tag{A20}$$

where ψ is the digamma function and the remaining expectation is taken under the Dirichlet posterior over model frequencies $q(r)$, whose sufficient statistics are $\{\alpha_k\}_{k=1,\ldots,K}$. Here, z_{ik} is the probability that the ith subject is best described by the kth model under H_1. The VB treatment of this model under a mean-field assumption is described in Stephan et al. (2009), yielding the following VB update rules:

$$\begin{aligned} \alpha_k &= \alpha_0 + \sum_{i=1}^{n} z_{ik} \\ z_{ik} &= \frac{\exp(L_{ik} + \psi(\alpha_k))}{\sum_{k'=1}^{K}\exp(L_{ik'} + \psi(\alpha_{k'}))} \end{aligned} \tag{A21}$$

for $k = 1,\ldots, K$ and $i = 1,\ldots, n$. Inserting Eq. (A21) (after convergence of the VB algorithm) yields the lower bound to the model evidence $F_1 = \log p(y|H_1)$. The Bayesian omnibus risk can now be simply evaluated as follows: $BOR = \frac{1}{1+\exp(F_1 - F_0)}$.

Appendix 3. Dynamic Causal Modeling (DCM)

In addition to localizing brain regions that encode specific sensory, motor or cognitive processes, neuroimaging data is nowadays further exploited to understand how information is transmitted through brain networks. Addressing such questions is the *raison d'être* of Dynamic Causal Modeling (DCM), which allows for the formal (Bayesian) statistical analysis of large-scale network connectivity based upon realistic biophysical models of brain responses. In brief, a set of differential equations describe how neuronal populations interact and respond to external perturbation (e.g., sensory stimulation). These state equations are augmented with an observation model that maps the dynamics of the hidden neuronal states (such as average membrane depolarization within neuronal ensembles) to neuroimaging data time series (such as EEG[5] of fMRI[6]). Important DCM parameters are, for example, connection strengths and their modulation. Note that the latter usually embody the question of interest — which is usually framed in terms of plasticity (change in connectivity) induced by drugs, lesions or task demands.

Note that the impact of model parameters on the data is nonlinear and obscured by measurement noise. This is why DCM relies upon variational approaches to approximate Bayesian inference (Friston et al., 2007b), which are informed about the (a priori) likely values of model parameters. In essence, the variational Bayesian (VB) scheme recovers the approximate posterior density $q(\theta) \approx p(\theta|y,m)$ under the Laplace approximation (Friston et al., 2007b), given the synthetic fMRI data time series. This density can be used to define parameter estimates ($\hat{\theta} = E[\theta|y,m]$), which can then enter a second-level analysis (cf. the Random effects BMS and classical random effects analysis of parameter estimates section). It also provides a free energy (bound) approximation to the log model evidence $\log p(y|m)$:

$$\begin{aligned} F &= I(\mu) + \frac{1}{2}\ln|\Sigma| + \frac{p}{2}\ln 2\pi \\ \Sigma &= -\left[\frac{\partial^2 I}{\partial\theta^2}\Big|_{\mu}\right]^{-1} \\ I(\theta) &= \ln p(y|\theta,m) + \ln p(\theta|m) \end{aligned} \tag{A22}$$

where $I(\theta)$ is the log joint density over data y and parameters θ under the generative model m, p is the number of parameters, and μ and Σ are the main and variance of the approximate Gaussian posterior density $q(\theta) = N(\mu,\Sigma)$. Here, θ includes neural parameters (e.g., network connectivity strengths), as well as other parameters that control, for example, the shape of the hemodynamic response function. Priors $p(\theta|m)$

[5] ElectroEncephaloGraphy.
[6] Functional Magnetic Resonance Imaging.

on model parameters can be found in Daunizeau et al. (2012). Note that Bayesian model selection rests on the VB free energy approximation to the model evidence (cf. Eq. (A22)). In this work, DCM simulations and model inversions used the VBA-toolbox[7] (http://code.google.com/p/mbb-vb-toolbox/).

References

Abramowitz, M., Stegun, I.A., 1968. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Dover, New York (Eds.).

Akaike, H., 1973. Information measures and model selection. Bull. Int. Stat. Inst. 50, 277–290.

Boly, M., Garrido, M.I., Gosseries, O., Bruno, M.A., Boveroux, P., Schnakers, C., Massimini, M., Litvak, V., Laureys, S., Friston, K., 2011. Preserved feedforward but impaired top–down processes in the vegetative state. Science 332, 858–862.

Casella, G., Berger, R.L., 2001. Statistical Inference, 2nd edition. Duxbury Press.

Chen, C.C., Henson, R.N., Stephan, K.E., Kilner, J.M., Friston, K.J., 2009. Forward and backward connections in the brain: a DCM study of functional asymmetries. Neuroimage 45, 453–462.

Daunizeau, J., David, O., Stephan, K.E., 2011a. Dynamic causal modeling: a critical review of the biophysical and statistical foundations. Neuroimage 58 (2), 312–322.

Daunizeau, J., Preuschoff, K., Friston, K.J., Stephan, K.E., 2011b. Optimizing experimental design for comparing models of brain function. PLoS Comp. Biol. 7 (11), e1002280.

Daunizeau, J., Stephan, K.E., Friston, K.J., 2012. Stochastic dynamic causal modelling of fMRI data: should we care about neural noise? Neuroimage 62, 464–481.

Dan Romik, 2000. Stirling's approximation for n!: the ultimate short proof? Am. Math. Mon. 107 (6), 556–557.

Daw, N., Gershman, S.J., Seymour, B., Dayan, P., Dolan, R.J., 2011. Model-based influences on humans' choices and striatal prediction errors. Neuron 69, 1204–1215.

den Ouden, H.E.M., Daunizeau, J., Roiser, J., Friston, K.J., Stephan, K.E., 2010. Striatal prediction error modulates cortical coupling. J. Neurosci. 30, 3210–3219.

Ethofer, T., Anders, S., Erb, M., Herbert, C., Wiethoff, S., Kissler, J., Gord, W., Wildgruber, D., 2006. Cerebral pathways in processing of affective prosody: a dynamic causal modeling study. Neuroimage 30 (2), 580–587.

Fienberg, S.E., 2006. When did Bayesian inference become 'Bayesian'? Bayesian Anal. 1, 1–40.

Fleming, S., Thomas, C.L., Dolan, R.J., 2010. Overcoming status quo bias in the human brain. PNAS 107, 6005–6009.

Friston, K.J., Harrison, L., Penny, W.D., 2003. Dynamic causal modelling. Neuroimage 19, 1273–1302.

Friston, K.J., Stephan, K.E., Lund, T.E., Morcom, A., Kiebel, S.J., 2005. Mixed-effects and fMRI studies. Neuroimage 24, 244–252.

Friston, K.J., Ashburner, J.T., Kiebel, S.J., Nichols, T.E., Penny, W.D. (Eds.), 2007a. Statistical Parametric Mapping. Academic Press.

Friston, K.J., Mattout, J., Trujillo-Barreto, Ashburner J., Peeny, W., 2007b. Variational free energy and the Laplace approximation. Neuroimage 34, 220–234.

Garrido, M.I., Kilner, J.M., Kiebel, S.J., Friston, K.J., 2007. Evoked brain responses are generated by feedback loops. Proc. Nat Acad. Sci. 104, 20961–20966.

Holmes, A.P., Friston, K.J., 1998. Generalisability, random effects and population inference. Neuroimage 7, S754.

Johnson, N.L., Kotz, S., 1977. In: Wiley, John (Ed.), Urn Models and Their Application.

Kass, R., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90 (430), 773–795.

Leff, A., Schofield, T., Stephan, K.E., Crinion, J.T., Friston, K.J., Price, C.J., 2008. The cortical dynamics of intelligible speech. J. Neurosci. 28 (49), 13209–13215.

Madigan, D., Raftery, A.E., Volinsky, C., Hoeting, J., 1996. Bayesian model averaging. AAAI Workshop Integrat. Mult. Learn. Mod. 77–83.

Penny, W.D., 2012. Comparing dynamic causal models using AIC, BIC and free energy. Neuroimage 59 (1), 319–330.

Penny, W.D., Holmes, A.P., Friston, K.J., 2007. Hierarchical models. In: Friston, K.J., Ashburner, J., Kiebel, S., Nichols, T., Penny, W. (Eds.), Statistical Parametric Mapping. Academic Press.

Penny, W., Joao, M., Flandin, G., Daunizeau, J., Stephan, K.E., Friston, K.J., Schofield, T., Leff, A.P., 2010. Comparing families of dynamic causal models. PLoS Comp. Biol. 6 (3), e1000709.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464.

Stephan, K.E., Weiskopf, N., Drysdale, P.M., Robinson, P.A., Friston, K.J., 2007. Comparing hemodynamic models with DCM. Neuroimage 38, 387–401.

Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J., 2009. Bayesian model selection for group studies. Neuroimage 46 (4), 1004–1017.

Tricomi, E., Rangel, A., Camerer, C.F., O'Dohery, J., 2010. Neural evidence for inequality-averse social preferences. Nature 463, 1089–1091.

van Leeuwen, T.M., den Ouden, H.E., Hagoort, P., 2011. Effective connectivity determines the nature of subjective experience in grapheme-color synesthesia. J. Neurosci. 31 (27), 9879–9884.

Vuong, Q., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57 (2), 307–333.

[7] For our current focus (namely: group-BMS), this particular software implementation can be considered identical to that of the academic freeware Statistical Parametric Mapping (www.fil.ion.ucl.ac.uk/spm/).