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:

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

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:

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

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:

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

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.: $H_1 \xrightarrow{\alpha_0 \to \infty} H_0$. 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:

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

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:

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