NeuroImage 75 (2013) 275–278

Contents lists available at SciVerse ScienceDirect

NeuroImage

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

Comments and Controversies

Model selection and gobbledygook: Response to Lohmann et al.

Karl Friston [a,*], Jean Daunizeau [b], Klaas Enno Stephan [a,b]

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

ARTICLE INFO

Article history:
Accepted 22 November 2011
Available online 1 December 2011

Keywords:
Model selection
Dynamic causal modelling
Model evidence
Bayesian
Inference

ABSTRACT

Lohmann et al. (this issue) make three unremarkable observations about model selection and use them to critique dynamic causal modelling—a Bayesian model selection procedure based on causal models of dynamical systems (Marreiros et al., 2010). In this response, we unpack their misconceptions and try to answer their questions.

Introduction

Lohmann et al. (this issue) make three observations about model selection to motivate a critique of dynamic causal modelling. We will not restrict our discussion to dynamic causal modelling *per se* (Marreiros et al., 2010) because the comments in Lohmann et al. pertain to all model selection or hypothesis testing procedures. Specifically, they note that the number of possible models one could entertain for any given data set can be very large. Second, they note that a model selected from a large number of models can have more evidence than a model selected from a subset of models. Finally, they note that models of multivariate data have to contend with heteroscedastic data (different noise levels). In what follows, we revisit these observations, clarify their implications and dismiss the critique of DCM offered by Lohmann et al.

Combinatorial explosions

Lohmann et al. note that the number of possible models (alternative hypotheses) that could be used to explain a particular data set can be very large. Their main point here is that a combinatorial explosion of model space is a problem, because it takes a finite amount of time to evaluate the evidence for each model. They then provide some quantitative illustrations of this problem based on an exhaustive search of model space using *post hoc* estimates of model evidence (or more strictly speaking the free energy bound on the log-evidence for dynamic causal models under the Laplace approximation).

The combinatorial explosion of model space is a problem if one is searching for the best model, because there is a possibility the best model might be missed if the model space is not big enough. However, this is only a problem if one is searching for the best model, which is not in fact the objective: Lohmann et al. seem to be conflating recent advances in DCM functionality (network discovery; Friston et al., 2011) with the use of DCM to characterise networks. In network discovery one can explore vast numbers of models. However, the rationale for including a large number of models is not to find the best model but to perform Bayesian model averaging (Hoeting et al., 1999) or Bayesian family comparison (Penny et al., 2010) in a way that properly accommodates uncertainty about the underlying model. Bayesian model averaging provides parameter estimates that are informed by the relative evidence for different models, while Bayesian family comparison enables inference about different model features. To assess the evidence for a network feature one usually partitions the model space into two subsets or families; for example, by comparing models with and without backward connexions. The resulting inference on model families is the usual end point of a search over large model spaces and is very different from selecting the best among millions of models. Crucially, if one evaluated the evidence for the best model, in relation to all others (Stephan et al., 2009), one would generally find that the relative evidence was very small (for large model spaces). This relates to the inevitable dilution of evidence over models (see Penny et al., 2010). Partitioning model space accommodates this dilution and prevents over-fitting at the

* Corresponding author at: Wellcome Trust Centre for Neuroimaging, Institute of Neurology, Queen Square, London, UK WC1N 3BG, UK.
E-mail address: k.friston@ucl.ac.uk (K. Friston).

1053-8119/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2011.11.064

level of models (we will see an example of this over-fitting from Lohmann et al. later). In short, in large model spaces, many models can be equally likely and any single model is generally uninteresting.

In general, being able to explore large model spaces does not mean that one should. The objective of scoring very large numbers of models is not to find the best model but to test hypotheses about different subsets (families) of models or perform Bayesian model averaging to make inferences about parameters within a family. To make meaningful inferences about one set of models, in relation to another, all the models should be equally plausible *a priori*. This means that the model space has to be chosen carefully because it entails prior assumptions about the models included. A common example of this is the careful choice of null and alternative hypotheses (models) in classical inference. DCM enables people to answer a question by evaluating the evidence for competing hypotheses; however, it does not furnish the question. The question posed in model selection is defined by the model or hypothesis space tested.

These considerations call into question what one means by 'network discovery'. Does it mean trying to discover the best network architecture or does it mean discovering the features of a network by evaluating all equally plausible models? The answer to this question may depend on the application. For example, if the objective is to characterise a network using measures from graph theory or to visualise network topography, one could use Bayesian model averaging to weight each coupling parameter in proportion to model evidence (e.g., to discover the weighted adjacency matrix). Alternatively, one may want to test a hypothesis by comparing two families with or without a particular connexion or class of connexions. This would involve pooling the evidence that had been discovered within each family. Given that the ability to score large model spaces in DCM is a fairly new development, one might anticipate developments along both of these lines.

On a technical note, the quantitative analysis of the time taken to score large model spaces in Lohmann et al. is wrong. It is based upon the assumption of an exhaustive search. In DCM large model spaces (with more than the 2^{16} models) are scored using a greedy search as described in Friston and Penny (2011). A technical description can be found in the appendix (spm_dcm_post_hoc: post hoc optimization). In this context, the number of models is constrained only by the number that can be represented in computer memory.

Model selection

In the second section, Lohmann et al. show that extending model space reveals new models that have more evidence than the best of the original models. This observation is in itself unremarkable; however, it is misinterpreted by Lohmann et al. For example, they say

"Of the top 400 models, 110 had no photic input into V1."

This (and a series of similar statements) creates the impression that the evidence for photic input to V1 is ambiguous and led Lohmann et al. to conclude "The most highly ranked models may be seen as neuroscientifically implausible." This is a false conclusion that follows from a failure to perform the appropriate model comparison. The proper way to assess the evidence for a photic input to V1 is to compare all models with and without photic input. This is an example of model family inference described above (Penny et al., 2010).[1] One would imagine, given that 290 of the top 400 models had a photic input to V1, there would be overwhelming evidence for this influence.

[1] This would normally entail adding the log evidences (or free energies) within (equally sized) subsets of models to ensure the difference is greater than three; in other words, the relative evidence is greater than $\exp(3) = 20$.

There is further confusion in this section, where it is suggested that "*F* values should be expanded to include non-uniform priors" so that

$$F = \ln p(y|m) + \ln p(m) \tag{1}$$

This equality is incorrect and reflects a misunderstanding of variational free energy. The free energy $F \approx \ln p(y|m)$ is (a bound approximation to) the log evidence. What Lohmann et al. were searching for was a way of combining the log evidence with log priors over models to provide the log posterior, where (ignoring constant terms that depend on the data)

$$\ln p(m|y) = F + \ln p(m) \tag{2}$$

In other words, the evidence $p(y|m)$ is not the probability of the model given the data; it is the probability of the data given the model. To convert the latter into the former one has to supply priors over models. Usually, these priors are implicit in the definition of the model space. Put simply, if one considers models with the same prior probability, then differences in log evidence (or free energy) become differences in the log of the probability of each model, given the data. A model that is implausible *a priori* can be included in the search space but then one has to specify its prior probability and use Eq. (2).

This is an important issue when evaluating model evidence and has broad implications for model selection in DCM. When using the evidence for one model (or family) in relation to others, one implicitly assumes that all models (or families) have the same prior probability. This means that the definition of model space (or families) is an implicit specification of prior beliefs about the plausibility of models (or families) considered. The motivation for including models with the same prior probability relates to the maximum entropy principle (Jaynes, 1957), which can be regarded as a generalisation of Laplace's principle of indifference. In the case of model selection, the prior with the greatest entropy is the prior in which all models are equally plausible.

One might ask if there is any way to optimise the priors that define models. The short answer is no. Model comparison is a procedure that provides answers to well posed questions that are cast in terms of alternative hypotheses or models—it does not tell you which models to consider. The longer answer is that it is possible parameterise the prior probability distribution over models and then optimise the parameters of the prior with respect to model evidence (Friston and Penny, 2011); however, there are still implicit priors on the parameters, which are usually assumed to be uninformative (i.e., to have maximum entropy).

True models

Lohmann et al. note that the concept of a 'true' model is quite elusive. Indeed, one could say there are no true models, unless one simulated data with a (known) model. However, Lohmann et al. then appear to assume that the true (known) model should always have the highest evidence. This is not necessarily the case: the data one simulates could have been produced by a simpler model, which would have higher evidence. This can occur when the true model has more degrees of freedom than the data, for example, in ill posed electromagnetic inverse problems. This is particularly true of DCM whether the true model has billions of neurons and parameters, which are not evidenced in the data. In short, a simpler (but equally accurate) explanation for data always has the greater evidence. In this sense, there is no true model (in the absence of simulated data); there is only a model with the highest evidence. This is the model that explains the data in an accurate and parsimonious way with minimal model complexity. Lohmann et al. also refer to the

notion of a 'plausible model'. Plausibility is another word for probability and can refer to a prior plausibility (or belief) or a posterior plausibility (after seeing data). In their Fig. 4 they show a model with high evidence and claim that it is implausible. But in what sense is it implausible? Presumably, they mean a low prior plausibility, in which case it should not have been included in the model space. If the model space contains *a priori* plausible and implausible models, it is necessary to compare posterior model probabilities, not model evidences; cf., Eq. (2).

Lohmann et al. specified the prior plausibility of models based on anatomical knowledge. In other settings, which ignore biological constraints, the model in Fig. 4 may have a high prior plausibility. In fact, it is much closer to the prior assumptions used in conventional mass-univariate analyses of fMRI data, in which every experimental factor has direct access to every node or voxel, and neuronal responses in any given region are uncoupled from those in others (see Fig. 1B). These considerations illustrate the danger of using terms like 'true' and 'plausible' without grounding them formally.

Model fit

The third critique takes us away from model comparison and addresses the distinction between accuracy and complexity. Model fit is a measure of the accuracy as represented by indices like the coefficient of determination R^2. This measures the proportion of variance explained by a model and is a common proxy for accuracy. As noted in Lohmann et al., the model with the highest evidence has a high accuracy and low complexity. This means that scoring models in terms of their accuracy alone is inappropriate (although things like the coefficient of determination can be useful when checking whether model optimisation has converged). Lohmann et al. "argue for an absolute goodness-of-fit measure" to score DCMs; however, the only relevant quantity for scoring a model is its evidence. This raises a key issue about the quality of the approximations to model evidence used to score models. These approximations usually rest on computationally expensive stochastic procedures, like Gibbs sampling or cross validation schemes, or analytic approximations, like variational free energy or the Bayesian information criterion. In the context of DCM for fMRI, variational free energy appears to be the best approximation (see Penny, 2012).

When examining the coupling between the lateral geniculate nucleus and striate cortex (LGN and V1), Lohmann et al. show rather convincingly that, over sessions and subjects, visual contrast selectively modulates the forward connexion from LGN to V1, in the context of reciprocal coupling. They then seem surprised to find that the amplitude of noise in the LGN is much greater, in relation to signal, than it is in V1. Lohmann et al. then use this unremarkable observation to motivate the importance of looking at the goodness of fit in different regions. They seem to erroneously conclude that because the signal-to-noise was higher in V1 than LGN, the ensuing "discrepancy in model fit between the two nodes" leads to "false inferences because of their lack of model fit." This is an elementary mistake—it is perfectly possible for different parts of the brain to have different levels of noise. There may be instances when the measurement noise in some regions is so high the data from these regions play no role in constraining parameter estimates. However, this does not mean that noisy regions (e.g., LGN) are not participating—if the LGN played no role in mediating distributed responses, then the model with contrast-dependent outputs from LGN would have had less evidence than a model without contrast-dependent outputs, not more. Important examples of regions or nodes whose dynamics are not informed by data are the hidden nodes in DCM for electromagnetic responses (David et al., 2011). In some instances, including hidden nodes can increase model evidence substantially.

Generally speaking, a model with a poor fit can have more evidence than a model with a good fit. For example, if you give a model pure measurement noise, a good model should properly identify that there is noise and no signal. This model will be better than a model that tries to (over) fit noisy fluctuations. It is important to note that in most cases of model inversion it is not just model parameters that are optimised but also the estimate of the amplitude of observation noise. In DCM, separate noise variance parameters are estimated for each node or region [these were not provided by Lohmann et al., although one would presume they are higher for LGN than V1]. These issues pertain to observation noise; however, there are also interesting questions about state-dependent changes in neuronal

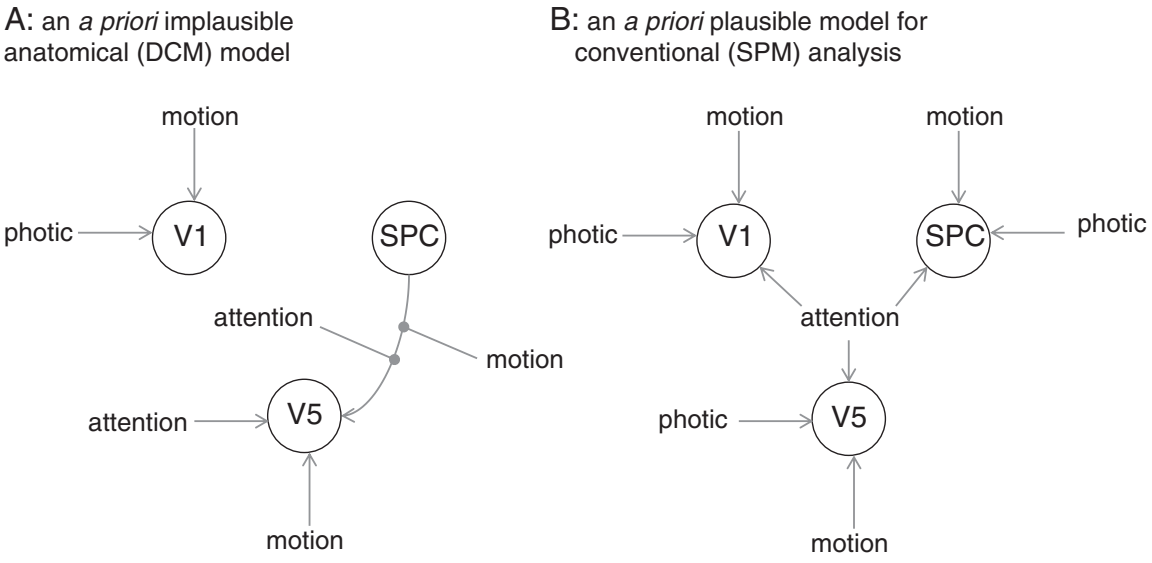

Fig. 1. Three node graphs depicting the 'implausible' architecture in Lohmann et al. (A) and the implicit architecture underlying conventional fMRI data analyses using general linear models (B), where all experimental factors can operate at every voxel and there are no connexions between voxels. In this instance, the input connexion strength is simply the conventional parameter estimate from a general linear convolution model.

fluctuations or system noise of the sort modelled in stochastic DCM. For example, the empirical results reported in Lohmann et al., LGN appear to show higher amplitude fluctuations when V1 is activated. This issue is pursued in an accompanying commentary by Breakspear.

Model validation

In their brief comments on model validation, Lohmann et al. note that spontaneous fluctuations make up a large proportion of signal variance in fMRI. While this is true, it is a little disingenuous to mention it, given their critique considered the estimation of endogenous fluctuations with stochastic DCM to be beyond "its scope." Lohmann et al. then assert that "in DCM model validation is primarily done by checking Bayes factors or model posteriors." This is incorrect: model selection is not model validation. Bayes factors (or differences in log evidence) are used to compare models, not validate them. In DCM, model validation proceeds in three stages. First, face validity is established by ensuring the model inversion does what it is supposed to do. This is the software testing described by Lohmann et al., but can only be done properly using simulated data. In other words, one generates simulated data and then tries to recover the known causes of those data using model optimisation. Note that Lohmann et al. tried to address face validity using empirical data, which is why they were unable to draw any definitive conclusions (because they did not know the 'true' model). Face validation is usually performed extensively for each new DCM, under a variety of model structures and noise assumptions (e.g., Friston et al., 2003, 2011; Stephan et al., 2008, 2009). The second phase involves construct validation. In other words, ensuring that one reaches similar conclusions using different constructs, such as different inversion schemes or non-Bayesian analyses like structural equation modelling (e.g., Penny et al., 2004). Finally, predictive validity is established in relation to independent data or knowledge. This usually involves a series of studies testing whether the model can predict some known or induced effect; e.g., the origin of an epileptic seizure (David et al., 2008), drug effects on ion channel function (Moran et al., 2011), or the presence of a remote lesion (Brodersen et al., 2011). The process of predictive validation can take several years, as different sorts of predictions are validated and confirmed.

Conclusion

Lohmann et al. conclude as follows:

> "In summary, we believe that DCM currently lacks convincing model validation methods, as well as a reliable model selection procedure, so that DCM models are based on insufficient evidence."

Given that dynamic causal modelling is a model selection procedure that identifies models with the greatest evidence, Lohmann et al. are effectively saying

> "In summary, we believe that model selection currently lacks convincing model validation methods, as well as a reliable model selection procedure, so that models with the greatest evidence are based on insufficient evidence."

Technically, this is gobbledygook. We have tried to address the questions (and gobbledygook) in Lohmann et al. and hope to have clarified some of the current tenets of model selection in DCM for people who have been wondering about these issues.

Acknowledgments

The Wellcome Trust funded this work. We would like to thank Michael Breakspear for invaluable guidance and two anonymous reviewers for helping make this response more constructive and didactic.

References

Brodersen, K.H., Schofield, T.M., Leff, A.P., Ong, C.S., Lomakina, E.I., Buhmann, J.M., Stephan, K.E., 2011. Generative embedding for model-based classification of fMRI data. PLoS Comput. Biol. 7 (6), e1002079 Jun.
David, O., Guillemain, I., Saillet, S., Reyt, S., Deransart, C., Segebarth, C., Depaulis, A., 2008. Identifying neural drivers with functional MRI: an electrophysiological validation. PLoS Biol. 6 (12), 2683–2697 Dec 23.
David, O., Maess, B., Eckstein, K., Friederici, A.D., 2011. Dynamic causal modeling of subcortical connectivity of language. J. Neurosci. 31 (7), 2712–2717 Feb 16.
Friston, K., Penny, W., 2011. Post hoc Bayesian model selection. NeuroImage 56 (4), 2089–2099 Jun 15.
Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. NeuroImage 19 (4), 1273–1302 Aug.
Friston, K.J., Li, B., Daunizeau, J., Stephan, K.E., 2011. Network discovery with DCM. NeuroImage 56 (3), 1202–1221 Jun 1.
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Stat. Sci. 14 (4), 382–417.
Jaynes, E.T., 1957. Information theory and statistical mechanics. Physical Review Series II 106 (no. 4), 620–630.
Marreiros, A.C., Friston, K.J., Stephan, K.E., 2010. Scholarpedia 5 (7), 9568.
Moran, R.J., Symmonds, M., Stephan, K.E., Friston, K.J., Dolan, R.J., 2011. An in vivo assay of synaptic function mediating human cognition. Curr. Biol. 21 (15), 1320–1325 Aug 9.
Penny, W.D., 2012. Comparing Dynamic Causal Models using AIC, BIC and Free Energy. NeuroImage 59 (1), 319–330 Jan 2.
Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004. Modelling functional integration: a comparison of structural equation and dynamic causal models. NeuroImage 23 (Suppl. 1), S264–S274.
Penny, W.D., Stephan, K.E., Daunizeau, J., Rosa, M.J., Friston, K.J., Schofield, T.M., Leff, A.P., 2010. Comparing families of dynamic causal models. PLoS Comput. Biol. 6 (3), e1000709 Mar 12.
Stephan, K.E., Kasper, L., Harrison, L.M., Daunizeau, J., den Ouden, H.E., Breakspear, M., Friston, K.J., 2008. Nonlinear dynamic causal models for fMRI. NeuroImage 42 (2), 649–662 Aug 15.
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 Jul 15.