There's no such thing as a 'true' model: the challenge of assessing face validity*

Vladimir Litvak, Amirhossein Jafarian, Peter Zeidman, Roni Tibon, Richard N. Henson, and Karl Friston

***Abstract*— To select among competing generative models of timeseries data, it is necessary to balance the goodness of fit (accuracy) and model complexity. Bayesian methods are a mathematically principled way to achieve this balance. However, when performing simulations – to assess the identifiability of models (face validation) – the best model identified by Bayesian model comparison might appear more complex than the model that actually generated the data. We illustrate this using dynamic causal models of human electrophysiological data, where models with multiple parameter modulations are selected as the best model, even if the true modulations are sparse. We explain this by the form of the complexity penalty, which is equivalent to weighted L2 norm. This phenomenon is an example of implicit prior biases that necessarily entail a complexity penalty.**

I. Introduction

When developing a new or improved mathematical model to explain empirical data, it is necessary to establish that the model and its parameters are *identifiable*, to the extent needed to distinguish different hypotheses about the causes of the data. This is typically assessed by simulating data using the candidate model and estimating the evidence for alternative models as explanations for the simulated data. Intuitively, the model that generated the data should have the strongest evidence. However, as we have found when developing models of neuroimaging data, this is not necessarily the case and the "true model" that generated the data is not necessarily the one assigned the strongest model evidence. This speaks to the adage that "all models are wrong but some are useful" [1]– even when it comes to simulations. In brief, the model used to generate some data is *not* the best explanation for those data, because there is always a simpler model that is more consistent with prior beliefs. Here, we explore this effect, which may be important when evaluating models.

The phenomena we describe are relevant to any inverse modelling problem; however, we will focus on the application of identifying functional brain circuitry from neuroimaging data, using Dynamic Causal Modelling (DCM). DCM is a Bayesian scheme for inferring physiologically meaningful parameters of neural systems from features of neurophysiological and neuroimaging data [2]. DCM entails the inversion of partially observed nonlinear dynamical systems to estimate biophysical parameters and their changes caused by experimental contexts – as well as to select the most plausible from several alternative models. DCM was first developed to provide insights about interactions between neuronal sources (effectivity connectivity) from functional magnetic resonance imaging (fMRI, [3]) data and subsequently further developed to investigate underlying generators of electrophysiological responses from EEG/MEG [4], [5], [6].

In Bayesian statistics, the goodness of models is measured by their marginal likelihood or log model evidence, $p(y|m)$ for data y and model m. This cannot typically be computed explicitly, so approximations are used. In DCM a lower bound on the evidence is defined called the free energy, equivalent to the Evidence Lower Bound (ELBO) in machine learning. This quantity is a useful score for the quality of the model, because it can be decomposed into the difference between the model likelihood (i.e., *accuracy*) and the Kullback-Leibler (KL) divergence between the prior and posterior densities of (unknown) model parameters (i.e., *complexity*). Parameter estimation in DCM is carried out through iterative optimization of the free energy, making it possible to optimally balance accuracy and complexity and thereby avoid overfitting [7]. The optimization scheme is a variational Bayesian (VB) procedure under the Laplace assumption (assuming a Gaussian form for probability densities functions). The free energy score of the estimated model in DCM is a statistical quantity that allows Bayesian model comparison (e.g., with a fixed effects [8] or random effects approach [9]).

Recently a novel post-hoc model selection approach known as Bayesian Model Reduction (BMR) was introduced [10], [11] for analytically deriving the evidence and parameters of sub-models of an estimated model. This can be used to rapidly assess the evidence for a small set of pre-defined, reduced models – or to automatically search over potentially thousands of candidate models – to find an optimal explanation for the data. The BMR approach may outperform separate VB estimations for each reduced model, because reduced models cannot be caught in different local optima.

How can we assess whether a particular DCM model or procedure is valid? The concept of model validity has been extensively explored in philosophical and statistical literature. When applying the notion of validity to DCM one can identify

*The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome (203147/Z/16/Z). The UK MEG community is supported by Medical Research Council grant MR/K005464/1.

VL, AJ, PZ and KF are with The Wellcome Centre for Human Neuroimaging, 12 Queen Square, London, UK. WC1N 3AR. (Corresponding author: Dr. Vladimir Litvak, phone: +44 20 3448 4370, fax: +44 20 7813 1420, e-mail: v.litvak@ucl.ac.uk).

RT and RNH are with MRC Cognition & Brain Sciences Unit, University of Cambridge, 15 Chaucer Rd, Cambridge, UK

three basic types: *(i) Face validity* – the ability of a DCM procedure to recover the model (and parameters) that generated the data, *(ii) Predictive validity* – the ability to obtain consistent results when applying DCM to multiple realizations of the same type of data; e.g., the inferences from one group predict the responses of another group of subjects drawn from the same population, *(iii) Construct validity* – the ability to obtain consistent results using different DCM procedures, different generative models or different measurements. Clearly, establishing construct validity will ensure that DCM results are applicable and useful in a wide variety of contexts. However, the first challenge is to establish face validity and predictive validity. Several previous works have addressed the predictive validity of DCM analyses for different procedures and data types [12]–[14].

The present paper deals with the problem of establishing the face validity of a DCM. This is a more difficult problem than may appear at first glance. The difficulty is that even when simulating data – with known parameters under a given model – the ensuing data can be emulated by a different set of parameters or different models. Therefore, Bayesian model comparison may not be able to distinguish the '*true*' *model* that generated the data from other models. Instead, it will favor *the least complex (simplest) model* (in the sense of the smallest complexity term) that can still fit the data features sufficiently accurately.

This notion of a *least complex model* may naïvely point to a model with fewer parameters or with the smallest changes with the same number of parameters. This suggests that model comparison is conservative, meaning that it might miss some aspects of how data are generated but the ones it does identify are true in the sense that they are necessary to explain the data. In this paper, we show that in DCM a model with small modulations of a large number of parameters can be less complex than a model with large modulations of a few parameters. This is just one example of the nature of Bayesian model comparison that should be considered when interpreting DCM results.

II. Methods

In this paper, we focus on DCM for evoked electromagnetic responses – and on the simplest problem of identifying the functional architecture using DCMs of single subject data or grand average responses over subjects. However, the same issue can arise when using other data features (e.g., spectral features) and other data modalities such as fMRI.

A. Data pre-processing

A publicly available group dataset of visual responses to familiar, unfamiliar and scrambled faces is used in this paper [15]. The dataset includes multiple neuroimaging modalities; however, in this paper, we only used EEG data collected from 16 subjects (7 female, age 26.4±2.9 years). The dataset was preprocessed in SPM12 (http://www.fil.ion.ucl.ac.uk/spm) as described in [16] and the evoked responses were averaged across subjects. The ensuing grand average was used as the observed data features in DCM for evoked responses. The Boundary Element Model [17] based on template head meshes and standard extended 10-20 electrode locations [18] was used to construct the forward model in DCM for EEG.

B. Model structure

Six regions including (left and right) bilateral primary visual cortex (V1) as well as the key areas known to be involved in processing of faces: bilateral occipital face area (OFA, [19]) and bilateral fusiform face area (FFA, [20]) were included in our DCM (Table 1 provides the MNI coordinates of these regions).

TABLE I. Coordinates of Model Sources

Source name	MNI coordinates (mm)		
	X	***Y***	***Z***
lV1	-12	-97	-1
rV1	12	-97	-1
lOFA	-39	-85	-13
rOFA	39	-79	-13
lFFA	-42	-55	-19
rFFA	42	-49	-19

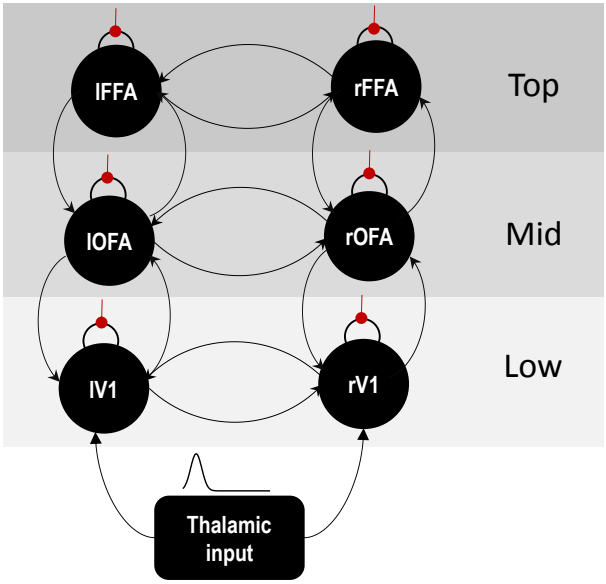

Figure 1. The architecture of DCM used for our analysis. See Table I for source coordinates.

We assumed that the hierarchical order of the network is in the form of V1→OFA→FFA, which is commonly used in literature [21]. We used the 'ERP' neural model in DCM, which is based

on the neural mass model formulation originally suggested by Jansen and Rit [22]. In this model, each source for node comprises three neuronal populations, each modelled as a single neural mass: pyramidal neurons, excitatory interneurons and inhibitory interneurons. The interconnections between different sources are defined by forward connections (targeting the excitatory interneurons), backward connections (targeting pyramidal neurons and inhibitory interneurons) and lateral connections (targeting all the three populations) [23]. In our model, the connections from V1 to OFA and from OFA to FFA were of the forward type, the reciprocal connections were of the backward type and the corresponding bilateral areas at each hierarchical level were linked by reciprocal lateral connections (See Fig. 1)

C. Between trial effects

We only modelled two of the three original experimental effects: unfamiliar and scrambled faces in the DCM input. This allowed us to focus on the well-known N170 response to faces [24], which is one of the strongest and most robust effects in cognitive neurophysiology. The aim of DCM – in this study – was to identify which hierarchical levels of the network increase their responsiveness to generate the N170. Following the established DCM procedures, we formed a model space (see Table II) comprising 7 models in total: 3 simple models with connectivity *modulations* at a single level (low – V1, mid – OFA and top – FFA), 3 models with modulations at two of the levels and one model with modulations at all the levels (the full model).

TABLE II. MODEL SPACE

Model number	Modulation levels
1	low
2	mid
3	top
4	low mid
5	low top
6	mid top
7	low mid top (full model)

D. Model inversion and comparison

We used both full VB inversion [7] and the BMR procedure [10] to invert the set of models described in the previous section. BMR is computationally efficient – and is not affected by the idiosyncratic local minima in the free energy landscape. However, for nonlinear models, the parameter estimates generated by BMR are generally not identical to those identified using VB due to the (Laplace) approximations BMR calls on. Therefore, we assessed the evidence for competing models using both VB and BMR and compared the results.

To perform BMR, the full model (model 7 in Table 1) was fitted to the empirical data using VB. The free energy and parameter estimates of the other six models, which are all reduced or sub-models of model 7, were then computed analytically from the posterior parameter estimates of the full model using BMR [10], [11]. Since all analyses were performed using the grand average (effectively a single subject) our model comparisons did not have to consider random effects at the between subject level.

E. Face validation using simulations

To address the question of whether the BMR procedure can identify the 'true' model, we simulated ERP data using the posterior parameters of the sub-models fitted to the data. The simulation was performed by integrating the model with the posterior expectations of the parameters (inferred by VB fitting to the original data) and adding random Gaussian fluctuations (sampled from the posterior noise covariance matrix estimated from the real data). This reproduces the amplitude and autocorrelation structure of the real noise. We then fitted the full model to each of the simulated ERPs and performed BMR with the 6 sub-models to ask whether the 'true' model could be recovered.

III. RESULTS

A. DCM results on real data

The results of the VB model inversion with the original grand average data are shown in Fig. 2. All the models fitted the N170 effect well. For both VB and BMR procedures, there is a clear tendency for models with modulations at more than one level to have higher evidence than the first three (simple) models. In both cases, the full model had the highest free energy by a large margin (a difference of 108 from the next best model for VB, 60 for BMR). From these log evidence differences – i.e. the log Bayes factors – the posterior probability for each model can be computed, under equal priors for each model. This equates to the full model having a posterior probability of very close to 1, with other models being close to zero.

Fig. 3 shows the posterior correlations matrix between the modulation parameters for the 6 sources, computed by normalizing the posterior covariance matrix which is produced by VB inversion of the full model. The pattern of correlations shows non-trivial conditional dependencies, particularly within a group of connectivity parameters from the left sources and the right V1.

B. Face validation with simulated data

We performed face validation for both the VB and BMR procedures with data generated from the three simple models. We tested whether each of the two procedures could identify the model that generated the data. The results of this analysis are

shown in Fig. 4. Only in one case ('true' model – model 2 with VB inversion), the generating model was correctly identified. In all other cases, the best model was one of the multi-level models. In two cases it was the full model. Note that if we limited our comparison to the simple models, we would have obtained perfect face validity in this case for both VB and BMR

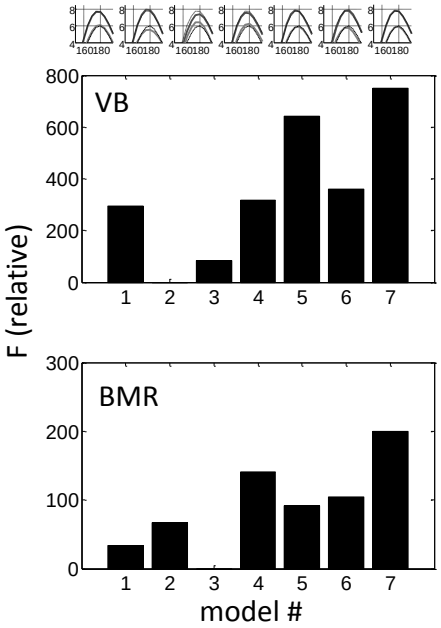

Figure 2. Model comparison for fitting DCM to the original data. Please refer to Table II for interpretation of the model numbers. The top bar chart shows the F values obtained from VB inversion. The plots above show the quality of the fits to the N170 effect for the 7 models. The dotted lines show the original data and the solid lines model fits. All the models fitted the effect well. The bottom bar chart shows the F values from BMR procedure with model 7 used as the full model. For both analyses, the full model had the highest evidence.

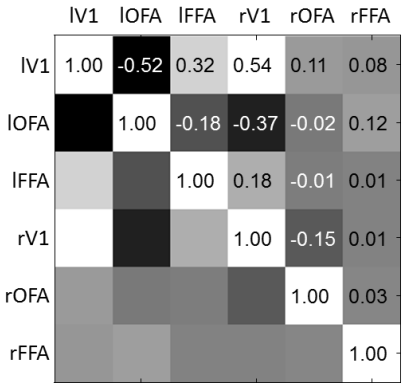

Figure 3. Posterior correlations for the modulation parameters of the full model (model 7) fitted to real data with VB. The numbers above diagonal show the corresponding values of the correlation coefficient. The group of sources including all the left sources and right V1 shows relatively high absolute correlations – 0.18 and above.

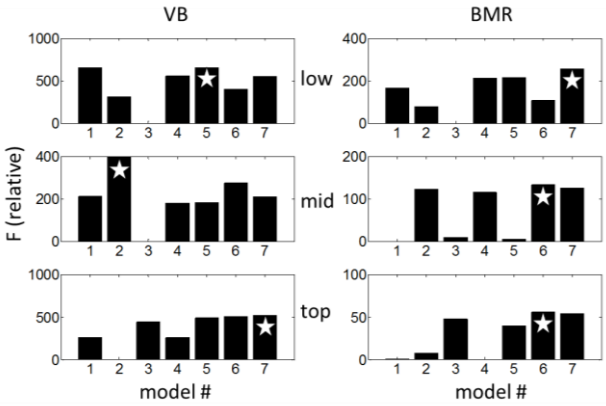

Figure 4. Model comparisons for the analysis of simulated data. The left column shows the results of the VB procedure and the right column those of BMR. Each row corresponds to a different true model (model 1 – top row, model 2 – middle row, model 3 – bottom row). The winning models are marked with stars.

IV. Discussion

A. Conditional dependencies.

A model can lack identifiability for various reasons. For example, the amount of data or the signal-to-noise ratio may be insufficient to confidently infer the presence or absence of various parameters. Additionally, conditional dependencies among parameters reduce identifiability, where more than one setting of parameter combinations can generate the same predicted data. In effect, the parameters with these sorts of dependencies cannot be estimated reliably. Conditional dependencies between estimated parameters (for a particular model and data) can be examined by looking at the posterior correlations. Such a matrix is calculated for the full model as shown in Fig 3, where one can see high conditional dependencies between four of the six modulation parameters.

In the DCM for fMRI literature, particular attention has been paid to minimization of conditional dependencies between the parameters of the neuronal model and between neuronal and hemodynamic parameters [25]. Correlations among hemodynamic parameters are usually not considered problematic, because these parameters do not show experimental effects.

In the case of DCM for M/EEG, the existence of conditional dependencies is closely related to the number of sources and their spatial proximity. MEG and EEG have low spatial resolution – and sources located within a few centimeters of each other can produce similar spatial patterns if their orientations are similar [26]. Therefore, any parameters that modulate the amplitude of spatially proximal sources will be conditionally dependent. In principle, it is possible to disambiguate some of their effects by their different signatures in terms of temporal aspects of the response [27] or the predicted responses of other sources in the network. However, multi-node DCMs have many degrees of freedom that could effectively

'absorb' this kind of differences predicting similar data for distinct parametric effects. Since, in the presence of realistic noise, there is no way to disambiguate these effects based on the goodness of fit (accuracy); one would expect that differences in model evidence are driven by model complexity, which in effect the least complex model could well not be the 'true' model.

B. Model complexity vs. parameter number

Under Laplace assumptions – that underlie both VB and BMR methods – the complexity penalty reduces to a measure of the deviation of the parameters from their prior values weighted by the respective prior precisions. Let us assume, there is a set of parameters with strong conditional dependencies for which prior precisions are the same (and they change in the same direction with similar effect). Which would be a simpler model: *(i)*, the one where only one parameter is modulated or *(ii)*, the one where all parameters change? The complexity penalty, in this case, would reduce to a scaled L2 norm of the parametric modulations. The L2 norm does not strongly penalize small deviations from zero but does penalize large deviations, in relation to a L1 norm [28]. This property well known in the M/EEG community from minimum norm solutions in source reconstruction; known to be spatially smeared rather than focal or sparse. These solutions use the L2 norm of source amplitudes as a regularization term [29]. Gaussian priors in VB make a similar kind of minimum norm assumption, which results in spreading of effects to all parameters that could improve the fit. This explains why full models or models with multiple modulations often win in DCM model comparison.

As seen in our simulation example, this could happen even for simulated data, where the ground truth is a sparse modulation. If we have a prior belief that differences in responses between our experimental conditions are caused by changes to a single connection, this belief must be explicitly expressed in the functional form of the complexity term. This is simple to implement with BMR, by using the number of allowable parameters as a prior (i.e., prior energy) over reduced models. In electromagnetic source reconstruction, this is known as a multiple sparse priors (MSP [30]).

The tendency for models with multiple modulations to have higher evidence – under L2 norm like complexity penalty – is just one example of what could be called implicit prior biases. Other examples, which we hope to address in our future work, include: *(i)* prior preferences for particular kinds of models induced by the model structure and *(ii)* biases induced by the relative strength of the effects of different parameters on the predicted data. Prior biases cannot be avoided as they reflect implicit prior assumptions about the generative model. However, one should be aware of them and ensure that they properly reflect prior beliefs about the generative process – in the same way as priors that are specified explicitly.

V. Summary and conclusions

In this paper, we suggest that the 'true' model – that is known to have generated some simulated data – is not necessarily the 'best' model of those data. If one finds this result for a particular model space or application, this may be useful information; by suggesting that the model space or data features are not fit for purpose for the hypotheses (i.e. models) being tested. We highlighted the issue of implicit bias towards non-sparse models, which stems from the use of Gaussian priors in DCM. This can be useful to bear in mind, when developing models or when designing a model space for empirical applications.

References

[1] G. E. P. Box, "Robustness in the Strategy of Scientific Model Building," *Robustness Stat.*, pp. 201–236, Jan. 1979.

[2] J. Daunizeau, O. David, and K. E. Stephan, "Dynamic causal modelling: A critical review of the biophysical and statistical foundations," *NeuroImage*, vol. 58, no. 2. pp. 312–322, 15-Sep-2011.

[3] K. J. Friston, L. Harrison, and W. Penny, "Dynamic causal modelling.," *Neuroimage*, vol. 19, no. 4, pp. 1273–302, Aug. 2003.

[4] O. David, S. J. Kiebel, L. M. Harrison, J. Mattout, J. M. Kilner, and K. J. Friston, "Dynamic causal modeling of evoked responses in EEG and MEG.," *Neuroimage*, vol. 30, no. 4, pp. 1255–72, May 2006.

[5] R. J. Moran, K. E. Stephan, T. Seidenbecher, H.-C. Pape, R. J. Dolan, and K. J. Friston, "Dynamic causal models of steady-state responses.," *Neuroimage*, vol. 44, no. 3, pp. 796–811, Feb. 2009.

[6] K. J. J. Friston, A. Bastos, V. Litvak, K. E. E. Stephan, P. Fries, and R. J. J. Moran, "DCM for complex-valued data: cross-spectra, coherence and phase-delays.," *Neuroimage*, vol. 59, no. 1, pp. 439–55, Jan. 2012.

[7] K. Friston, J. Mattout, N. Trujillo-Barreto, J. Ashburner, and W. Penny, "Variational free energy and the Laplace approximation," *Neuroimage*, vol. 34, no. 1, pp. 220–234, Jan. 2007.

[8] W. D. Penny, K. E. Stephan, A. Mechelli, and K. J. Friston, "Comparing dynamic causal models.," *Neuroimage*, vol. 22, no. 3, pp. 1157–72, Jul. 2004.

[9] K. E. Stephan, W. D. Penny, J. Daunizeau, R. J. Moran, and K. J. Friston, “Bayesian model selection for group studies.,” *Neuroimage*, vol. 46, no. 4, pp. 1004–17, Jul. 2009.

[10] K. J. Friston, V. Litvak, A. Oswal, A. Razi, K. E. Stephan, B. C. M. van Wijk, G. Ziegler, and P. Zeidman, “Bayesian model reduction and empirical Bayes for group (DCM) studies,” *Neuroimage*, vol. in press, 2015.

[11] K. Friston and W. Penny, “Post hoc Bayesian model selection,” *Neuroimage*, vol. 56, no. 4, pp. 2089–99, Jun. 2011.

[12] J. B. Rowe, L. E. Hughes, R. A. Barker, and A. M. Owen, “Dynamic causal modelling of effective connectivity from fMRI: Are results reproducible and sensitive to Parkinson’s disease and its treatment?,” *Neuroimage*, vol. 52, no. 3, pp. 1015–1026, Sep. 2010.

[13] D. Bernal-Casas, E. Balaguer-Ballester, M. F. Gerchen, S. Iglesias, H. Walter, A. Heinz, A. Meyer-Lindenberg, K. E. Stephan, and P. Kirsch, “Multi-site reproducibility of prefrontal–hippocampal connectivity estimates by stochastic DCM,” *Neuroimage*, vol. 82, pp. 555–563, Nov. 2013.

[14] V. Litvak, M. Garrido, P. Zeidman, and K. Friston, “Empirical bayes for group (Dcm) studies: A reproducibility study,” *Front. Hum. Neurosci.*, vol. 9, no. DEC, 2015.

[15] D. G. Wakeman and R. N. Henson, “A multi-subject, multi-modal human neuroimaging dataset,” *Sci. Data*, vol. 2, p. 150001, Jan. 2015.

[16] R. N. Henson, H. Abdulrahman, G. Flandin, and V. Litvak, “Multimodal integration of M/EEG and f/MRI data in SPM12,” *Front. Neurosci.*, 2019.

[17] C. Phillips, J. Mattout, and K. J. Friston, “Forward models for EEG,” in *Statistical Parametric Mapping*, Elsevier, 2007, pp. 352–366.

[18] R. Oostenveld and P. Praamstra, “The five percent electrode system for high-resolution EEG and ERP measurements,” *Clin. Neurophysiol.*, vol. 112, no. 4, pp. 713–719, Apr. 2001.

[19] D. Pitcher, V. Walsh, and B. Duchaine, “The role of the occipital face area in the cortical face perception network,” *Exp. Brain Res.*, vol. 209, no. 4, pp. 481–493, Apr. 2011.

[20] N. Kanwisher and G. Yovel, “The fusiform face area: a cortical region specialized for the perception of faces.,” *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 361, no. 1476, pp. 2109–28, Dec. 2006.

[21] F. Jiang, L. Dricot, J. Weber, G. Righi, M. J. Tarr, R. Goebel, and B. Rossion, “Face categorization in visual scenes may start in a higher order area of the right fusiform gyrus: evidence from dynamic visual stimulation in neuroimaging,” *J. Neurophysiol.*, vol. 106, no. 5, pp. 2720–2736, Nov. 2011.

[22] B. H. Jansen and V. G. Rit, “Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns.,” *Biol. Cybern.*, vol. 73, no. 4, pp. 357–66, Sep. 1995.

[23] O. David, L. Harrison, and K. J. Friston, “Modelling event-related responses in the brain.,” *Neuroimage*, vol. 25, no. 3, pp. 756–70, Apr. 2005.

[24] S. Bentin, T. Allison, A. Puce, E. Perez, and G. McCarthy, “Electrophysiological Studies of Face Perception in Humans.,” *J. Cogn. Neurosci.*, vol. 8, no. 6, pp. 551–565, Nov. 1996.

[25] K. E. Stephan, N. Weiskopf, P. M. Drysdale, P. A. Robinson, and K. J. Friston, “Comparing hemodynamic models with DCM.,” *Neuroimage*, vol. 38, no. 3, pp. 387–401, Nov. 2007.

[26] B. Lütkenhöner, “Magnetoencephalography and its Achilles’ heel.,” *J. Physiol. Paris*, vol. 97, no. 4–6, pp. 641–58, 2003.

[27] M. I. Garrido, J. M. Kilner, S. J. Kiebel, and K. J. Friston, “Evoked brain responses are generated by feedback loops.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 52, pp. 20961–6, Dec. 2007.

[28] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004, p. 716.

[29] M. S. Hämäläinen and R. J. Ilmoniemi, “Interpreting magnetic fields of the brain: minimum norm estimates.,” *Med. Biol. Eng. Comput.*, vol. 32, no. 1, pp. 35–42, Jan. 1994.

[30] K. Friston, L. Harrison, J. Daunizeau, S. Kiebel, C. Phillips, N. Trujillo-Barreto, R. Henson, G. Flandin, and J. Mattout, "Multiple sparse priors for the M/EEG inverse problem," *Neuroimage*, vol. 39, no. 3, pp. 1104–1120, Mar. 2008.