NeuroImage 141 (2016) 469–489

Contents lists available at ScienceDirect

NeuroImage

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

How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection

Joram Soch[a,f,*], John-Dylan Haynes[a,b,c,d,e,f], Carsten Allefeld[a,b]

[a] Bernstein Center for Computational Neuroscience, Berlin, Germany
[b] Berlin Center for Advanced Neuroimaging, Berlin, Germany
[c] Berlin School of Mind and Brain, Berlin, Germany
[d] Excellence Cluster NeuroCure, Charité-Universitätsmedizin Berlin, Germany
[e] Department of Neurology, Charité-Universitätsmedizin Berlin, Germany
[f] Department of Psychology, Humboldt-Universität zu Berlin, Germany

ARTICLE INFO

Article history:
Received 25 February 2016
Accepted 24 July 2016
Available online 28 July 2016

Keywords:
fMRI-based neuroimaging
mass-univariate GLM
model misspecifcation
underfitting versus overfitting
cross-validation
Bayesian model selection

ABSTRACT

Voxel-wise general linear models (GLMs) are a standard approach for analyzing functional magnetic resonance imaging (fMRI) data. An advantage of GLMs is that they are flexible and can be adapted to the requirements of many different data sets. However, the specification of first-level GLMs leaves the researcher with many degrees of freedom which is problematic given recent efforts to ensure robust and reproducible fMRI data analysis. Formal model comparisons that allow a systematic assessment of GLMs are only rarely performed. On the one hand, too simple models may underfit data and leave real effects undiscovered. On the other hand, too complex models might overfit data and also reduce statistical power. Here we present a systematic approach termed cross-validated Bayesian model selection (cvBMS) that allows to decide which GLM best describes a given fMRI data set. Importantly, our approach allows for non-nested model comparison, i.e. comparing more than two models that do not just differ by adding one or more regressors. It also allows for spatially heterogeneous modelling, i.e. using different models for different parts of the brain. We validate our method using simulated data and demonstrate potential applications to empirical data. The increased use of model comparison and model selection should increase the reliability of GLM results and reproducibility of fMRI studies.

Introduction

Over the last 20 years, the *general linear model* (GLM; Friston et al., 1994; Holmes and Friston, 1998) has been the main workhorse in the analysis of functional neuroimaging data. Though many alternatives are available – e.g. network modelling approaches (Smith et al., 2011) such as dynamic causal modelling (Friston et al., 2003), multivariate pattern analysis (Haxby, 2012) or representational similarity analysis (Kriegeskorte et al., 2008) –, the GLM approach remains the most important analysis method for task-based experiments using *functional magnetic resonance imaging* (fMRI), among other reasons because it is well understood and highly flexible.

However, this flexibility has its drawbacks as well. Using the GLM, a researcher has to make a lot of modelling decisions – which processes to consider relevant, which experimental events to model and in which detail, which hemodynamic basis set to use etc. –, many of which strongly influence the sensitivity of the analysis. This situation tempts the researcher to run not just one analysis based on *a priori* information about the paradigm and previous experience or general recommendations (Josephs and Henson, 1999; Grinband et al., 2008), but to try many different models until good results are obtained (Leek and Peng, 2015). This strategy, a form of so-called *p-hacking* (Simonsohn et al., 2014), may render the false positive rate much higher than indicated by the nominal significance level (Simmons et al., 2011) and is one of the reasons for recent doubt about the *reproducibility* of neuroimaging studies (Pernet and Poline, 2015; Glatard et al., 2015). On the other hand, the obvious solution – to decide for a single analysis pipeline in advance and refrain from *post-hoc* modifications, a procedure referred to as *pre-registration* (Boekel et al., 2015) – often leads to a suboptimal model which may fail to detect a real effect.

In this paper, we propose a different approach: to select an optimal model from a number of candidate models, but without recourse to the statistical significance of a test based on this model. Instead, we apply *Bayesian model selection* (BMS; MacKay, 2003; Bishop, 2007) to

* Corresponding author at: BCCN Berlin, Philippstraße 13, Haus 6, 10115 Berlin, Germany.
E-mail address: joram.soch@bccn-berlin.de (M. Soch).

http://dx.doi.org/10.1016/j.neuroimage.2016.07.047
1053-8119/

general linear models. Note that we do not propose to replace null hypothesis tests with Bayesian procedures, but only suggest to select the optimal model via BMS and then apply a standard GLM-based hypothesis test (*t*-test or *F*-test).

In the following section, we examine the problem of *model misspecification* in more detail, review relevant literature and motivate our own approach which uses cross-validation (Arlot and Celisse, 2010) on the first level to utilize empirically informed priors and employs random-effects model selection (Stephan et al., 2009) on the second level to account for possibly different optimal models in different subjects. We also give a short overview of the method's output for non-technical readers. The Theory section derives the mathematical details of our approach which is applied to an empirical data set in Validation: empirical data. Finally, we validate our method in two simulated model spaces in Validation: simulated data and discuss several of its implications.

Background

Model misspecification

For almost every psychological paradigm that can be implemented as an fMRI experiment, there exist several plausible ways to model it in a GLM (Monti, 2011; Carp, 2012). Typical questions that occur in setting up the analysis include: Should one include this regressor of no interest or not? Does this cognitive process represent the baseline or must it be considered activation? What is the duration of the underlying neuronal process? Should one model cue stimuli, button presses, error trials or feedback screens? How should the hemodynamic response be modelled? etc. If decisions are made that do not satisfy the data, this leads to a misspecified model.

Two basic forms of mismodelling are *underfitting* and *overfitting* (Guyon and Yao, 1999; see also MacKay, 2003; Bishop, 2007). In the first case, a necessary regressor is omitted from the model which reduces the model's descriptive power with respect to the given data (increasing the in-sample error). For example, the subject's reaction times (RT) in task-based fMRI experiments can account for a considerable amount of signal variance (Grinband et al., 2011; Yeung et al., 2011) so that omitting an RT regressor from the GLM would cause underfitting. In the second case, unnecessary regressors are included in the model, which are then just fitted to noise and thereby reduce the model's descriptive power with respect to other data of the same kind (increasing the generalization error). For example, global signal regression (GSR) in resting-state fMRI might induce spurious anti-correlations (Murphy et al., 2009; Chai et al., 2012) so that including it in the model would cause overfitting. It is also possible that a relevant effect is modelled, but the regressor does not have the optimal shape, e.g. because the assumed hemodynamic response function (HRF) does not reflect the actual hemodynamic response (Monti, 2011). Since this can be seen as omitting the correct regressor but including a wrong one, in this case overfitting and underfitting occur at the same time. From the perspective of a hypothesis test based on the respective model, underfitting increases the residual error while overfitting decreases the error degrees of freedom which is why in consequence both reduce the sensitivity of the test, i.e. its statistical power.

These theoretical considerations are supported by empirical evidence of potential underfitting (Lieberman and Cunningham, 2009), potential overfitting (Eklund et al., 2012) and potential p-hacking (Vul et al., 2009). It has also been demonstrated that activations detected using GLM-based fMRI analysis substantially depend on the model which is employed and its particular assumptions (Carp, 2012). Since these are only rarely scrutinized during GLM specification, mismodelling can be considered a serious problem in fMRI data analysis (Monti, 2011).

An alternative way to describe under- and overfitting is through the concepts of *accuracy* and *complexity*. An underfitted model does not have sufficient accuracy (too large in-sample error or residual variance), an overfitted model is too complex (has more parameters than can be well estimated from the given data). In the following, we will mainly use this latter terminology.

Previous measures against model misspecification

In practice, when faced with the question whether a particular regressor should be included in a model, sometimes a null hypothesis test is performed to make this decision (see e.g. Henson et al., 2001). A statistical parametric map for the corresponding parameter is computed at the group level, and the regressor is included if there are significant effects in brain areas that are thought to be relevant for the paradigm. There are several disadvantages to this approach: First, it is restricted to nested model comparison (Penny, 2012), i.e. choosing between two models that only differ by including (or not including) a regressor (or a set of regressors). Second, it only measures the gain in model accuracy through better fit (lower residual variance), but not model complexity and how well it generalizes to other data (Oaksford, 2002) which is only indirectly accounted for via e.g. numerator degrees of freedom in the F-statistic. Third, it also fails to detect the relevance of regressors that have a non-significant effect at the group level, but still contribute to model fit on the level of individual subjects (Rigoux et al., 2014).

Beyond this naïve approach, the problem of model misspecification in fMRI analysis has been recognized by researchers and a number of strategies have been developed to deal with potential mismodelling. Kherif and Loh have proposed algorithms for model optimization, but they only allow inference pertaining to nested model comparisons (Kherif et al., 2002) and condition regressor timing (Loh et al., 2008). Razavi et al. (2003) investigated model accuracy based on the goodness of fit, but do not consider model complexity. Luo and Nichols (2003) have supplied a range of frequentist methods for the diagnosis of a given model and exploratory data analysis. Measures against overfitting were proposed for certain situations, e.g. noise-model selection (Penny et al., 2007b) and hemodynamic response modelling (Kay et al., 2008a). In addition, there are some recommendations regarding the temporal specification of experimental conditions (Josephs and Henson, 1999; Yarkoni et al., 2009), orthogonalisation of regressors (Mumford et al., 2015) and modelling of non-white noise (Lund et al., 2006). However, none of these methods and recommendations provide a general remedy for all possible forms of model misspecification and they are not based on systematic model comparison.

General linear model selection for fMRI data analysis

An approach that avoids these shortcomings is Bayesian model selection (BMS). One reason for this is that it uses the Bayesian *log model evidence* (LME) as a measure of model quality, a measure which accounts for both model accuracy and model complexity, so that maximizing it avoids both underfitting and overfitting (MacKay, 2003; Bishop, 2007). Moreover, it attaches a model quality to any GLM and therefore enables *non-nested model comparison*, i.e. comparing more than two models that do not just differ by one or more regressors but in arbitrary ways (Penny, 2012).

The concept of Bayesian model selection is not new to neuroimaging. In dynamic causal modelling (DCM; Friston et al., 2003; Stephan et al., 2008), the free energy, a variational approximation to the log model evidence (Friston et al., 2007) is an established measure to quantify model quality (Penny et al., 2004). For the GLM, the past has seen *Variational Bayesian* (Penny et al., 2003, 2005, 2007a) and *Empirical Bayesian* (Friston et al., 2008; Penny and Ridgway, 2013) methods to calculate the log model evidence. In this work, we introduce a number of methodological improvements which address

shortcomings of these approaches and increase the usability of BMS across GLMs for fMRI.

First, a practical problem for the use of BMS is that it requires prior distributions on the model parameters. Using too wide priors will exaggerate the complexity of models which include more regressors while too narrow priors will underestimate it. Optimally, priors should be informed by previous experience about typical parameter values (Stephan, 2010). However, such numeric information is normally not included in neuroimaging papers and concrete values strongly depend, among others, on scanner technology (Triantafyllou et al., 2005). We here solve this problem by using part of the available data to obtain an informative prior and another part to calculate the log model evidence, leading to the *cross-validated log model evidence* (cvLME). BMS based on this quantity is referred to as *cross-validated Bayesian model selection* (cvBMS).

Second, as opposed to previous Bayesian analyses (Penny et al., 2003, 2005, 2007a), we use a simple, but reasonable model structure which exactly corresponds to the model underlying standard first-level fMRI data analyses (Friston et al., 1994) as performed e.g. by Statistical Parametric Mapping (SPM). This enables an *Analytical Bayesian* computation for the Bayesian GLM with normal-gamma priors (GLM-NG) which makes our method both fast and precise.

The cvLME for the GLM-NG is combined with a voxel-wise implementation of *random-effects Bayesian model selection* (RFX BMS; Stephan et al., 2009; Rosa et al., 2010; Penny et al., 2010; Rigoux et al., 2014), a common procedure for second-level model inference that accounts for the optimal model to vary across voxels as well as subjects. This allows for *spatially heterogeneous modelling*, i.e. using different models for different areas of the brain (Razavi et al., 2003).

Application of the method

On the first level, once a model is specified and estimated for a given subject, cvBMS calculates a cross-validated version of the log model evidence (cvLME) for each voxel in the brain. This results in a cvLME map which quantifies the model's performance in different parts of the brain (see Fig. 1A). The cvLME is a relative measure of model quality and its absolute value has no direct interpretation.

For a single subject, two or more models are compared by converting the log model evidences jointly into *posterior model probabilities*. If the models fall into one or more model families, where models of a family have a particular modelling decision in common, log model evidences can be first aggregated into log family evidences which are then converted into *posterior family probabilities*. The optimal model or model family is the one with the largest posterior probability.

On the second level, cvBMS accounts for the fact that different models may be optimal in different subjects. Using the first-level log model (or family) evidences for all models (or families) in all subjects, it estimates how frequently each model is optimal in the population of subjects. This results in brain maps of *estimated model frequencies* (see Fig. 1B) or, alternatively, maps of the posterior probability that a given model is more frequent than all others, the so-called *exceedance probability*. The optimal model or model family is the one with the largest estimated frequency.

Just as log model evidences, posterior probabilities and model frequencies are voxel-wise and have the form of brain maps. Selecting models on a voxel-by-voxel basis results in a *selected-model map* (SMM) which can then be used to restrict GLM-based analyses to those voxels where the corresponding model is optimal.

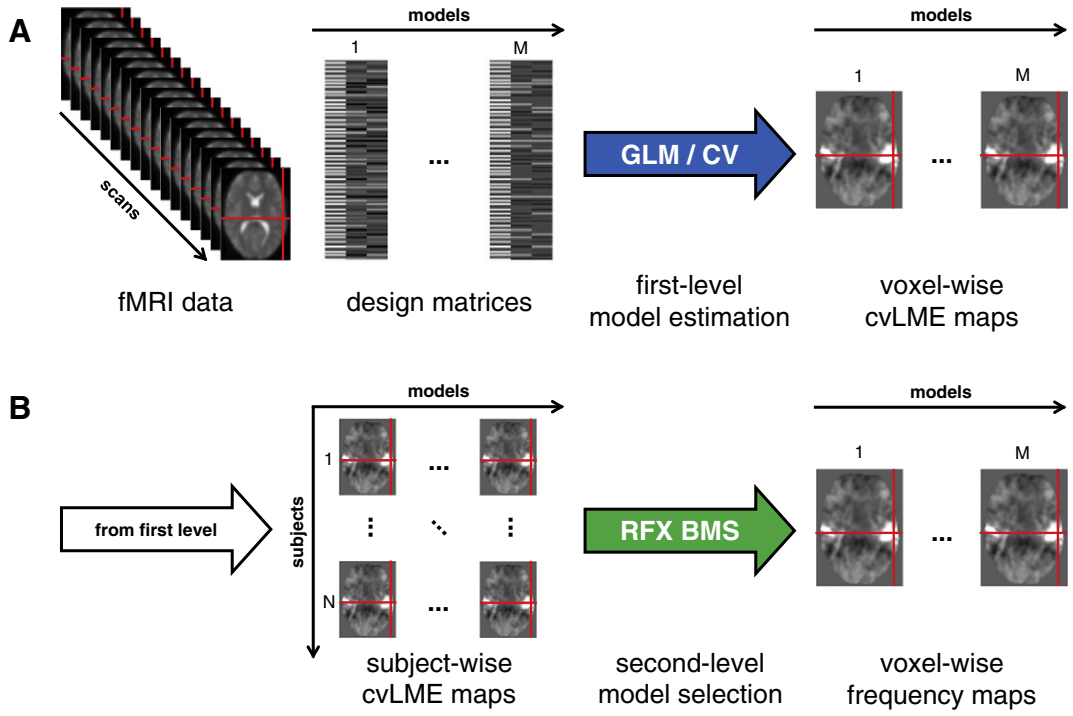

Fig. 1. Model selection for general linear models in fMRI data analysis. This figure summarizes our approach of cross-validated Bayesian model selection (cvBMS). All calculations are performed voxel-wise, an exemplary voxel is highlighted using red crosshairs. (A) First level: All fMRI data from one subject and design matrices for each model enter general linear model estimation (GLM) with cross-validation across sessions (CV) which produces voxel-wise cross-validated log model evidence (cvLME) maps for each model on which population inference is based. (B) Second level: The cvLME images from all subjects (N = number of subjects) and all models (M = number of models) enter random-effects Bayesian model selection (RFX BMS) which produces voxel-wise frequency maps on which model decisions are based. In each voxel, the model with the highest frequency in the population is selected for data analysis which constitutes a selected-model map (SMM). Source: Parts of this figure are adapted from SPM course material (Stephan, 2010).

Theory

In this section, we describe in detail the mathematical derivation of our proposed method. For the non-technical reader mainly interested in the practical use of the method, we recommend to directly proceed to Validation: empirical data demonstrating the application of the method to empirical data.

The general linear model

As linear models, GLMs for fMRI (Friston et al., 1994; Kiebel and Holmes, 2011) assume an additive relationship between experimental conditions and the fMRI BOLD signal, i.e. a linear summation of expected hemodynamic responses into the measured hemodynamic signal. Consequently, in the GLM, a single voxel's fMRI data (y) are modelled as a linear combination (β) of experimental factors and potential confounds (X), where errors (ε) are assumed to be normally distributed around zero and to have a known covariance structure (V), but unknown variance factor (σ^2):

$$y = X\beta + \varepsilon, \;\; \varepsilon \sim \mathrm{N}\left(0, \sigma^2 V\right) \tag{1}$$

In this equation, X is an $n \times p$ matrix called the "design matrix" and V is an $n \times n$ matrix called a "correlation matrix" where n is the number of data points and p is the number of regressors. In standard analysis packages like Statistical Parametric Mapping (SPM) (Ashburner et al., 2013), V is typically estimated from the signal's temporal auto-correlations across all voxels using a Restricted Maximum Likelihood (ReML) approach (Friston et al., 2002a,b). In contrast to that, X has to be constructed by the user which is essential to the specification of a GLM (Monti, 2011). As there are often many possible models for a given psychological paradigm, GLMs for fMRI need to be subjected to model selection in order to avoid model misspecification.

The general linear model in Eq. (1) implicitly defines the following likelihood function:

$$p\left(y \left| \beta, \sigma^2 \right.\right) = \mathrm{N}\left(y; X\beta, \sigma^2 V\right) \tag{2}$$

Classical estimation of the GLM proceeds by maximizing the likelihood function $p(y|\beta, \sigma^2)$ with respect to the unknown parameter β to find the optimal parameter estimates. Under temporal independence, i.e. if $V = I_n$, this Maximum Likelihood (ML) solution is obtained by Ordinary Least Squares (OLS) estimation (Bishop, 2007, eq. 3.15).

If V is not equal to the identity matrix I_n, i.e. errors ε are not assumed independent and identically distributed (i.i.d.), Weighted Least Squares (WLS) estimation is employed which gives rise to Gauss-Markov (GM) estimates (Koch, 2007, eq. 4.29). This WLS solution is equivalent to multiplying data y and design X with a whitening matrix $W = V^{-1/2}$ and then performing OLS estimation.

Statistical inference proceeds by multiplying the parameter estimates $\hat{\beta}$ with a contrast vector or matrix c and thresholding statistical parametric maps (SPM) to perform a t-test or an F-test on the parameter estimates (Stephan, 2010).

Classical inference for the GLM using the ReML approach, WLS estimation, contrasts and SPMs of t-values or F-values is implemented in standard analysis packages such as SPM where it is applied to each voxel indepedently which is usually called a "voxel-wise" or "mass-univariate" analysis of fMRI data.

Bayesian inference for the GLM

Classical and Bayesian estimation agree in their use of the likelihood given by Eq. (2). For mathematical convenience, we rewrite the likelihood function in terms of an $n \times n$ precision matrix $P = V^{-1}$ and the inverse residual variance $\tau = 1/\sigma^2$:

$$p\left(y \left| \beta, \tau \right.\right) = \mathrm{N}\left(y; X\beta, (\tau P)^{-1}\right) \tag{3}$$

Other than classical inference, Bayesian inference requires prior distributions on the model parameters. The straightforward choice of prior distribution for the general linear model with unknown regression coefficients β and inverse residual variance τ is the conjugate prior relative to this likelihood function (Koch, 2007, ch. 4.3.2; Bishop, 2007, ch. 3.3), the normal-gamma distribution (Koch, 2007, eq. 2.212; Bishop, 2007, eq. B.52)

$$\begin{aligned} p\left(\beta \left| \tau \right.\right) &= \mathrm{N}\left(\beta; \mu_0, (\tau\Lambda_0)^{-1}\right) \\ p(\tau) &= \mathrm{Gam}\left(\tau; a_0, b_0\right) \end{aligned} \tag{4}$$

where μ_0 and Λ_0 are the prior mean and the prior precision of β and a_0 and b_0 are the prior shape and rate parameters for τ (Gelman et al., 2013, p. 68).

Bayes' theorem implies that the posterior is proportional to the product of likelihood and prior. As we show in Appendix A, the likelihood function from Eq. (3) and the parameter prior from Eq. (4) result in the following posterior distribution on the model parameters

$$\begin{aligned} p\left(\beta \left| \tau, y \right.\right) &= \mathrm{N}\left(\beta; \mu_n, (\tau\Lambda_n)^{-1}\right) \\ p\left(\tau \left| y \right.\right) &= \mathrm{Gam}\left(\tau; a_n, b_n\right) \end{aligned} \tag{5}$$

where the posterior parameters are given by (Koch, 2007, eq. 4.159)

$$\begin{aligned} \mu_n &= \Lambda_n^{-1}\left(X^T P y + \Lambda_0 \mu_0\right) \\ \Lambda_n &= X^T P X + \Lambda_0 \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}\left(y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n\right) \end{aligned} \tag{6}$$

By calculating a posterior distribution, we update our belief about the model parameters regarding their location and precision. Upon model estimation, this posterior distribution can be used for statistical inference about the unknown parameters β and τ.

The Bayesian model evidence

Consider Bayesian inference on data y using model m with parameters θ. In this case, Bayes' theorem is a statement about the posterior density (Gelman et al., 2013, eq. 1.1):

$$p\left(\theta \left| y, m \right.\right) = \frac{p\left(y \left| \theta, m \right.\right) \; p\left(\theta \left| m \right.\right)}{p\left(y \left| m \right.\right)} \tag{7}$$

The denominator $p(y|m)$ acts as a normalization constant on the posterior density $p(\theta|y, m)$ and is given by (Gelman et al., 2013, eq. 1.3)

$$p\left(y \left| m \right.\right) = \int p\left(y \left| \theta, m \right.\right) \; p\left(\theta \left| m \right.\right) \; \mathrm{d}\theta \tag{8}$$

This is the probability of the data given only the model, independent of any particular parameter values. It is also called "marginal likelihood" or "model evidence" and can act as a model quality criterion in Bayesian inference, because parameters are integrated out of the likelihood (Penny, 2012). For computational reasons, only

the logarithmized or log model evidence (LME) $\mathrm{L}(m) = \log p(y|m)$ is of interest in most cases.

As we show in Appendix B, the LME for the GLM-NG is given by

$$\begin{aligned}\mathrm{L}(m) = &\frac{1}{2}\log|P| - \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Lambda_0| - \frac{1}{2}\log|\Lambda_n| \\ &+ \log\Gamma(a_n) - \log\Gamma(a_0) + a_0\log b_0 - a_n\log b_n\end{aligned} \tag{9}$$

where the posterior parameters are given by Eq. (6).

Accuracy and complexity

Recall Bayes' theorem for model-based inference on unknown parameters θ given by Eq. (7). Rearranging this equation, the model evidence can also be written as

$$p(y|m) = \frac{p(y|\theta,m)\ p(\theta|m)}{p(\theta|y,m)} \tag{10}$$

Logarithmizing both sides of the equation and taking the expectation with respect to the posterior density over model parameters θ gives (Penny et al., 2007a, eq. 8)

$$\mathrm{L}(m) = \int p(\theta|y,m)\ \log p(y|\theta,m)\ \mathrm{d}\theta - \int p(\theta|y,m)\ \log\frac{p(\theta|y,m)}{p(\theta|m)}\ \mathrm{d}\theta \tag{11}$$

Using this reformulation, the LME as a model quality measure can be naturally decomposed into an accuracy term, the posterior expected log likelihood, and a complexity penalty, the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between the posterior and the prior distribution (Friston et al., 2007; Penny et al., 2007a)

$$\begin{aligned}\mathrm{L}(m) &= \mathrm{Acc}(m) - \mathrm{Com}(m) \\ \mathrm{Acc}(m) &= \langle\log p(y|\theta,m)\rangle_{p(\theta|y,m)} \\ \mathrm{Com}(m) &= \mathrm{KL}\left[p(\theta|y,m)\,||\,p(\theta|m)\right]\end{aligned} \tag{12}$$

This reflects the capability of the LME to select models that achieve the best balance between accuracy and complexity, i.e. models that explain the observations sufficiently well (high accuracy) without employing too many principles (low complexity). Model accuracy and complexity also relate to the well-known bias-variance trade-off: When model complexity grows, this decreases an estimator's bias, but increases its variance, leading to lower in-sample error, but higher generalization error (Hastie et al., 2001).

The log model evidence is superior to other model quality criteria like Akaike's information criterion (AIC) (Akaike, 1974) or Bayesian information criterion (BIC) (Schwarz, 1978) in two ways. First, in the accuracy term, it does not use frequentist point estimates like the ML, but accounts for the possible uncertainty about model parameters θ. Second, in the complexity penalty, different parameters receive different complexities which accounts for the fact that model parameters might be correlated (Penny, 2012) or might introduce a different degree of flexibility into the model.

Cross-validated log model evidence

These benefits of the Bayesian model evidence come at the cost of having to specify prior distributions on the model parameters (Gelman, 2008). However, in fMRI data analysis with new or modified experimental paradigms and different types of scanners or scanning protocols, such prior information is usually not at hand. Non-informative priors can be used in this absence of knowledge, but they are usually improper and let the model evidence diverge. However, improper priors can lead to proper posteriors. We therefore use a non-informative prior for estimating the model from the data of all but one session and then employ the resulting posterior distribution as an informative prior, allowing us to compute the model evidence for the left-out session.[1]

In more detail, we perform S-fold cross-validation where S is the number of sessions, and priors in each session are calculated as posteriors from all other sessions. Using non-informative (improper) priors for the combined data $y_{\bar{s}}$ from $S-1$ sessions $\bar{s}$ (e.g. $\bar{s} = \{1,2,3,4\}$), we obtain combined posteriors $p(\theta|y_{\bar{s}})$. Using these posteriors as informative (proper) priors in the remaining session s (i.e. $s = 5$) and based on the common assumption of statistical independence between sessions, we obtain an out-of-sample log model evidence (oosLME) as follows:

$$\log p(y_s|m) = \log\int p(y_s|\theta,m)\ p(\theta|y_{\bar{s}},m)\ \mathrm{d}\theta \tag{13}$$

This procedure embodies the Bayesian maxim that "today's posterior can be tomorrow's prior" (Stephan, 2010) and is similar to the log predictive density score (LPDS) approach (Villani et al., 2009; Li et al., 2010). The cross-validated log model evidence (cvLME) is then given by

$$\log p(y|m) = \sum_{i=1}^{S}\log p(y_i|m) \tag{14}$$

As non-informative priors for cross-validation in the context of GLMs for fMRI, we use a normal-gamma distribution given in Eq. (4) with the prior parameters

$$\mu_0 = 0_p,\ \Lambda_0 = 0_{pp} \quad \text{and} \quad a_0 = 0,\ b_0 = 0 \tag{15}$$

which yields an infinitely wide and therefore flat Gaussian for the regression coefficients β (Friston et al., 2002b) and Jeffreys prior for the inverse residual variance τ (Jeffreys, 1946). As one can see from Eq. (6), these (improper) priors are non-informative in the sense that only the data remain to influence the (proper) posteriors.

First-level Bayesian model inference

If several models m of the same data y can be specified, the model evidence can be used for model inference. In the simplest case of only two models, log Bayes factors (LBF) are used for model comparison (Koch, 2007, eq. 3.67):

$$\mathrm{LBF}_{i,j} = \log p(y|m_i) - \log p(y|m_j) \tag{16}$$

Evidence in favor of one model is usually considered "strong" when the LBF exceeds three, meaning that seeing the data y is about $\exp(3) \approx 20$ times more likely under model m_i than m_j (Kass and Raftery, 1995). In the case of two or more models, posterior model probabilities can be calculated using Bayes' theorem (Penny et al., 2010):

$$p(m|y) = \frac{p(y|m)\ p(m)}{\sum_{i=1}^{M} p(y|m_i)\ p(m_i)} \tag{17}$$

One can also consider a model family f, i.e. a set of models m that share some characteristic (Stephan et al., 2009), e.g. certain

[1] For single-session fMRI data, we suggest split-half cross-validation which we have successfully tested on an SPM template data set (Ashburner et al., 2013; Henson et al., 2002; Soch et al., 2014).

parameters like a set of regressors in the case of the GLM. Using the law of marginal probability, a family evidence can be calculated from the evidences of the models belonging to this family as follows:

$$p\,(y|f) = \sum_{m \in f} p\,(y|m)\;\; p\,(m|f) \tag{18}$$

If models or families are to be compared in several subjects, it is advisable to use an explicit population proportion model which will be described in the next section.

Second-level Bayesian model selection

On the second level, between-subject variance in individual model preferences can be accounted for by a hierarchical model in which first-level model evidences $p(y_i|m_j)$ (model j applied to subject i) serve as the second-level likelihood function $p(y|m)$. Models m are then assumed to follow a multinomial distribution with model frequencies r and a Dirichlet distribution with concentration parameters α is used as the prior distribution for r.

This extension of Pólya's urn model (Mahmoud, 2008), also called random-effects Bayesian model selection (RFX BMS), has been introduced and validated (Stephan et al., 2009), extended and refined (Penny et al., 2010) and revisited (Rigoux et al., 2014). It is widely applied (Stephan et al., 2010) in dynamic causal modelling (DCM) and was also implemented for voxel-wise model inferences (Rosa et al., 2010).

An iterative estimation algorithm has been developed to infer a posterior distribution over model frequencies $p(r|y)$ from prior concentration parameters α_0 (Stephan et al., 2009). This procedure and the model are given in detail in Appendix D. In this work, we apply RFX BMS to cvLMEs from the first level and use a uniform prior over r. Then, the posterior distribution over r informs us about the proportion of the population whose data are best explained by each of the models.

Decision-theoretic model choice

In DCM, it is common practice (see e.g. Deserno et al., 2012) to report group-level model selection results using exceedance probabilities (EP), i.e. the posterior probability of a model being more frequent than any other model (Stephan et al., 2009; Penny et al., 2010). Additionally, RFX BMS also gives rise to estimated model frequencies in the form of expected frequencies (EF), i.e. the posterior means $\langle r \rangle$, or likeliest frequencies (LF), i.e. the posterior modes $\hat{r}_{\mathrm{MAP}}$. When a uniform prior is used, LFs directly quantify the proportion of subjects in the sample in which a particular model is optimal.

In model selection, we are not so much interested in inference about the model frequencies r, but rather in an optimal decision regarding the models m, i.e. which model to use to analyze a group of subjects. As we show in Appendix E, this optimal model choice can be finessed using Bayesian decision theory (BDT) and is achieved by choosing the model with the largest estimated frequency (EF or LF) in an RFX BMS.

Selected-model maps (SMM) are therefore defined as binary maps indicating where a particular model has the largest frequency. In this paper, we use SMMs with an additional cluster threshold of 10 voxels for all whole-brain results and use LFs when we make quantitative statements about evidence in favor of each model.

Validation: empirical data

We demonstrate and validate the cvBMS approach using empirical fMRI data acquired in a sample of 22 subjects using orientation pop-out processing (Bogler et al., 2013).

In this experiment, the screen showed a 3 × 7 array of homogeneous bars oriented either 0°, 45°, 90° or 135° relative to the vertical axis (see Fig. 2A). This background stimulation changed every second and during each trial, one target bar on the left and one target bar on the right were independently rotated either 0°, 30°, 60° or 90° relative to the rest of the stimulus display (see Fig. 2B). Those trials of orientation contrast (OC) lasted 4 s and were alternated with inter-trial intervals of 7, 10 or 13 s.

Below the grid of bars in central position, a square was presented that opened up to the left or to the right side for 1 s every 2 s. Subjects were asked to indicate via button press whether the square opened up to the same side or the opposite side relative to the last opening.

This study also included a localizer paradigm in which a single bar identical to one of the target bars from the main experiment (see Fig. 2B) was presented in blocks. These blocks were separated by intervals of no stimulation.

fMRI data was preprocessed using SPM8, Revision 5236 per 04/02/2013 (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/). Functional MRI scans were corrected for acquisition time delay (slice timing) and head motion (spatial realignment), normalized to MNI space and smoothed with a Gaussian kernel with an FWHM of 6 × 6 × 6 mm.

For the localizer data, a GLM with two regressors, left and right visual stimulation, was estimated. At the group level, two paired t-tests (left vs. right and vice versa) were performed to define regions of interest (ROI) that were found to be responsive to oriented bars in the left and right hemisphere (Bogler et al., 2013). Statistical inference was performed using family-wise error (FWE) correction, a significance level of $p \leq 0.05$ and an extent threshold of $k = 10$. The resulting localizer masks were used to look for model preferences in particular regions.

1st model space: technical model parameters

In the first model space, we begin with the most common approach and build a model with one onset regressor per experimental condition, leading to 16 regressors per session.

We include motion parameters from spatial realignment, we activate temporal filtering using a high-pass filter at $T = 128$ s and we choose temporal non-sphericity correction using a first-order auto-regressive AR(1) model. The feature that varies over models is the number of hemodynamic response function (HRF) kernels. In the first model, box-car stimulus functions are convolved with a canonical hemodynamic response function (cHRF) as implemented in SPM. In the second model (cHRF + 1), we additionally include the temporal derivative of the cHRF. In the third model (cHRF + 2), we include the temporal and dispersion derivative of the cHRF (Henson et al., 2002).

Using group-level model selection and subsequent generation of selected-model maps, we find that the cHRF does not underfit and is sufficient in almost all parts of the brain (see Fig. 3A). There are regions in cerebral cortex, potentially belonging to the dorsal attention network (DAN) (Fox et al., 2005, 2006), that require cHRF+1, but no voxels where cHRF + 2 was selected. Since the temporal derivative can account for differences in the latency of the peak response (Henson et al., 2001), this could indicate that the HRF peaks earlier or later in DAN regions. As the dispersion derivative can capture differences in the duration of the peak response (Henson et al., 2001), this suggests that the duration of the peak response does not vary significantly across voxels.

2nd model space: modelling experimental factors

For the second model space, we use the winning model with a canonical HRF and no HRF derivatives as a starting point and focus on different ways of modelling the neural processes underlying the

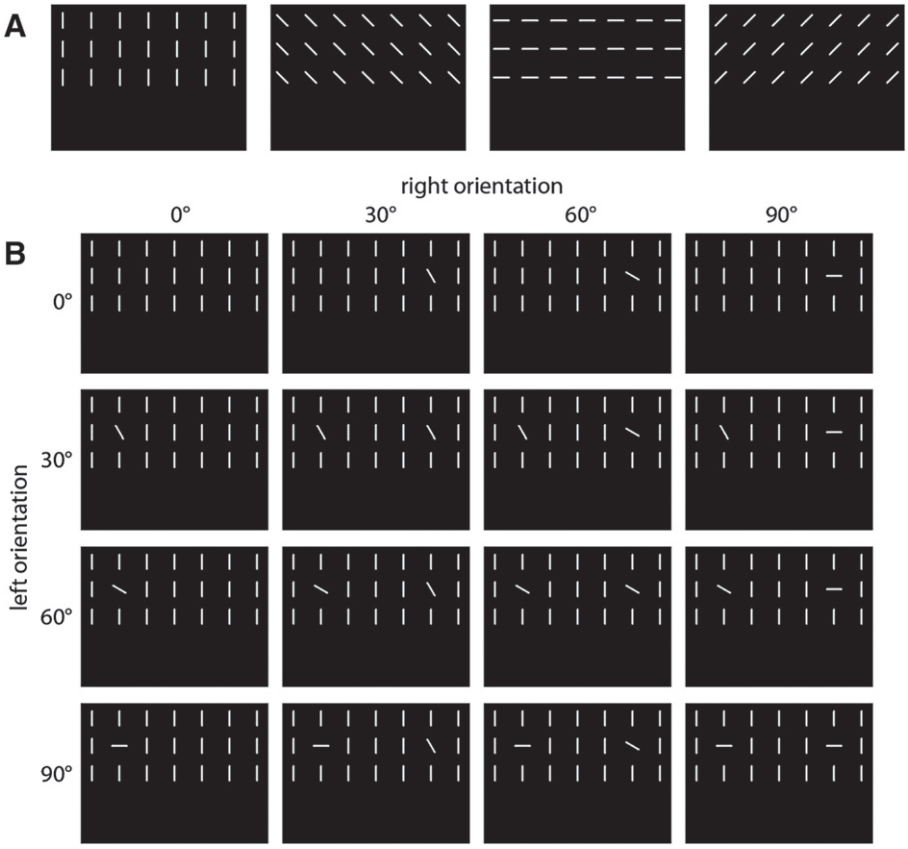

Fig. 2. Experimental paradigm for orientation pop-out processing. This figure describes the psychological paradigm used in our empirical validation data set. (A) A 3×7 background display of homogeneously oriented bars rotated by 0°, 45°, 90° or 135° was alternated every second. (B) During each four-second trial, one bar on the left and one bar on the right were randomly rotated 0°, 30°, 60° or 90° counter-clockwise relative to the background display. This is referred to as left and right orientation contrast (OC). Source: This figure is adapted from the original study (Bogler et al., 2013).

perception of orientation contrast (OC). The model introduced above (GLM I) considers the experiment a factorial design with two factors (left vs. right OC) having four levels (0°, 30°, 60°, 90°) each. This results in $4 \times 4 = 16$ possible combinations or experimental conditions modelled by 16 onset regressors convolved with the canonical HRF.

The second model (GLM II) puts all trials from all conditions into one HRF-convolved regressor and encodes orientation contrast using a parametric modulator (PM) that is given as

$$\mathrm{PM} = \frac{\mathrm{deg}}{90^\circ}$$

with $\mathrm{deg} = (0^\circ, 30^\circ, 60^\circ, 90^\circ)$, resulting in $\mathrm{PM} = \left(0, \frac{1}{2}, \frac{1}{3}, 1\right)$, such that the parametric modulator is proportional to orientation contrast. There was one PM for each factor of the design, i.e. one PM for left OC and one PM for right OC.

In behavioral pre-tests for the original study, it was observed that reaction time is not linear in orientation contrast. Instead, the reaction time when being presented with the respective OC and asked to detect where OC occured saturates at lower levels for higher OC (Bogler et al., 2013). Thus, a further model was implemented.

The third model (GLM III) is similar to the second model, but uses a nonlinear transformation of orientation contrast to calculate a parametric modulator (PM) that is given by

$$\mathrm{PM} = \frac{1}{\frac{\mathrm{deg}}{30^\circ} + 1}$$

with $\mathrm{deg} = (0^\circ, 30^\circ, 60^\circ, 90^\circ)$, resulting in $\mathrm{PM} = \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}\right)$, such that the parametric modulator is proportional to the expected average reaction time. This OC model is intended to be a psychologically more plausible description of the neural processing underlying OC perception.

This model space is equivalent to the set of three models already used in the original publication (Bogler et al., 2013). We will refer to GLM I as the "categorical model", GLM II as the "linear-parametric model" and GLM III as the "nonlinear-parametric model". In all models, background stimulation was treated as implicit baseline so that only target onsets were explicitly modelled. Note that these models are not nested in each other and therefore significance testing of additional regressors is not applicable for model selection. Furthermore, we cannot just use the best model in each single subject as this would make second-level analysis impossible due to different model parameters.

Accuracy and complexity

First, just consider GLM I and GLM II. These two models encode exactly the same information, namely the values of the factors "left OC" and "right OC" at each point in time. However, they represent this information in different ways: Whereas the parametric model assumes a parametric shape of the measured signal, the categorical model allows for a greater flexibility of activation patterns across experimental conditions. This means that every signal that can be identified using GLM II can also be detected using GLM I, but not vice versa.

We performed group-level model selection on the cross-validated log model evidences (cvLME) to find voxels where GLM II has a higher model frequency than GLM I. Due to the specific assumptions in GLM II and the higher flexibility of GLM I, we hypothesized that GLM I and GLM II might be equal in terms of model accuracy

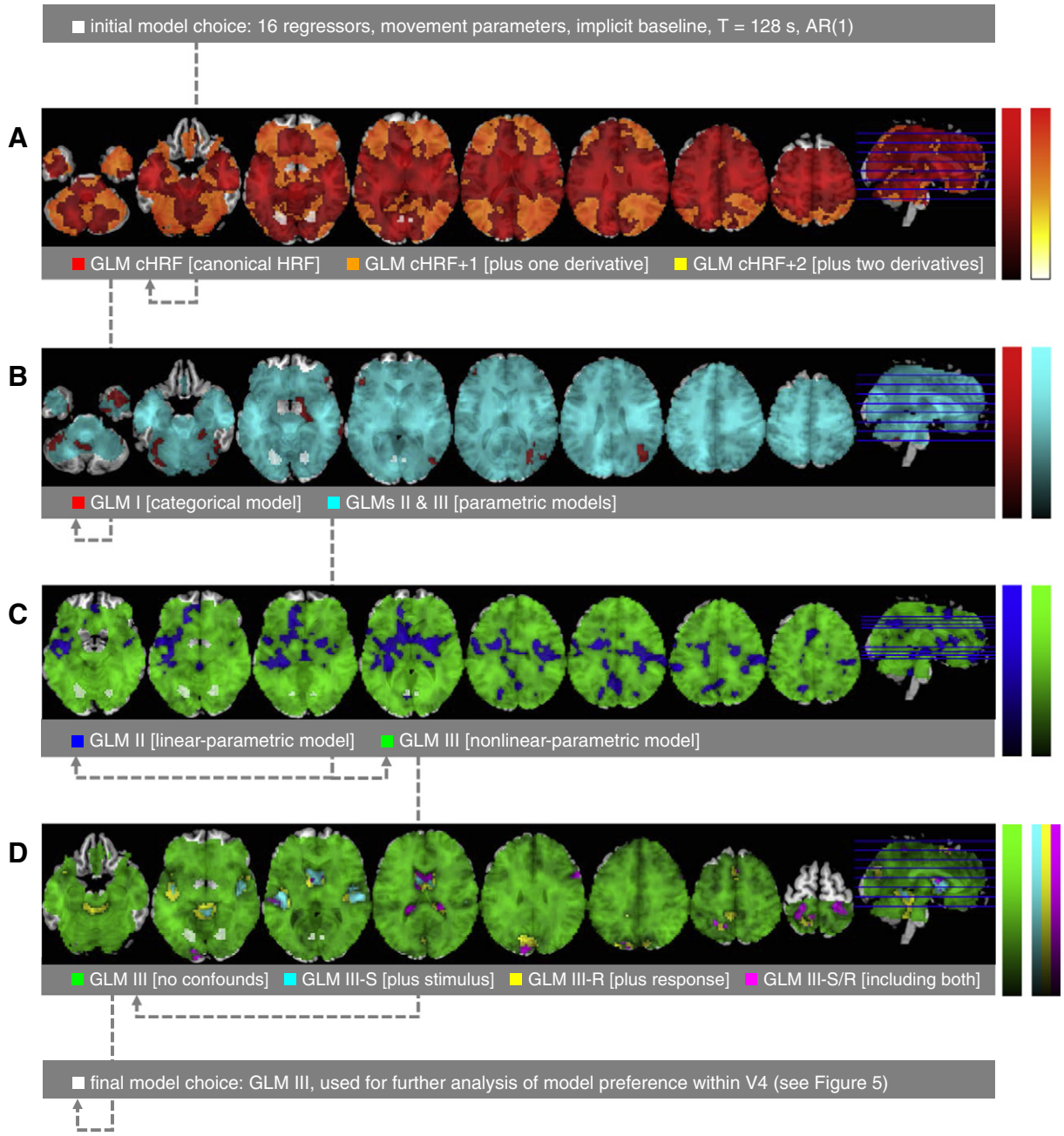

Fig. 3. Sequential model selection for orientation pop-out processing. This figure shows selected-model maps (SMM) obtained from group-level model selection performed in several steps (see Validation: empirical data) and displays the likeliest frequency (LF) for the selected model in each voxel. Color scales range from 0 to 1, except for the cyan, yellow and violet bars in panel D which range from 0 to 0.5. The white area in each section highlights the orientation-sensitive V4 (more ventral) and anterior parts of V1 (more dorsal) that were obtained from analysis of the localizer paradigm. Note that different slices are displayed in each panel in order to make model preferences perfectly visible. (A) 1st model space: different hemodynamic basis sets were compared; the canonical hemodynamic response function (cHRF) alone (red) came out best. (B) 2nd model space, family selection: using the cHRF, the categorical model was compared against the family of parametric models; parametric models (cyan) perform better. (C) 2nd model space, model selection: within the parametric models, a linear and nonlinear parametric modulator were compared; the nonlinear-parametric model (green) is superior. (D) 3rd model space: using the nonlinear-parametric model, confound modelling was investigated; confound modelling was only required in regions not of interest to the task (cyan, yellow, violet). Exact voxel counts for each model comparison are given in Table 1.

and the difference would primarily come from on a complexity advantage of GLM II over GLM I. Within left V4, which is known to be sensitive to orientation contrast (Burrows and Moore, 2009; Bogler et al., 2013), we identified the peak voxel (likeliest frequency (LF) = 86.29%, exceedance probability (EP) = 99.97%, [x y z] = [−18, −76, −8] mm) and extracted cvLME as well as model accuracy and model complexity from this voxel for each subject (see Fig. 4). For this purpose, accuracy and complexity were calculated based on theorem (12) and using Eqs. (C.2) and (C.4) derived in Appendix C.

From the differences in cvLME, accuracy and complexity, it can be seen that model accuracy alone weakly favors GLM I (see Fig. 4A), because it better explains the signal, but model complexity alone strongly favors GLM II (see Fig. 4B), because it uses fewer variables to do so. Taken together, the log model evidence clearly favors GLM II (see Fig. 4C). We therefore conclude that, given the data have a functional form such that the measured signal is continuously related to the levels of an experimental factor, there is no need for categorizing the data by these levels or there might even be the danger

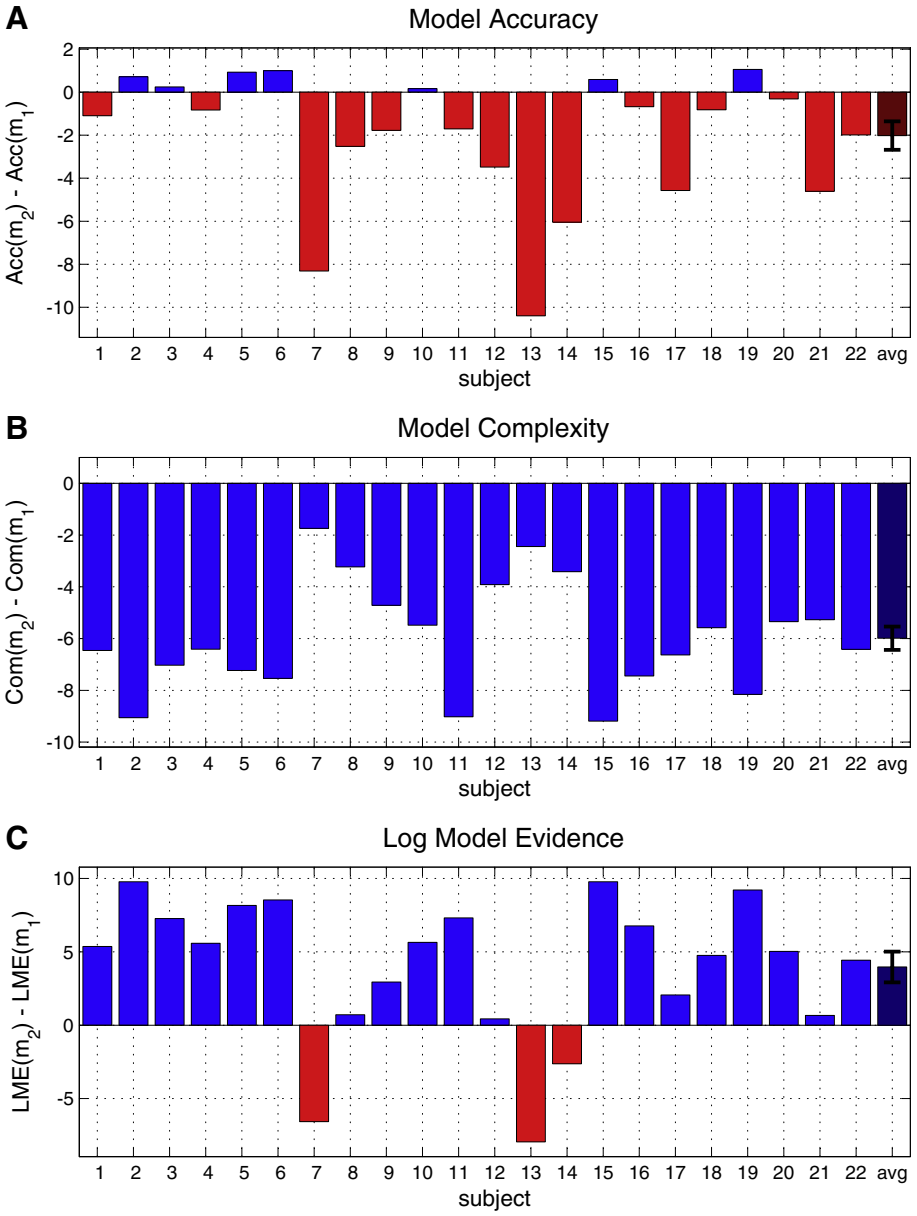

Fig. 4. Accuracy and complexity for categorical and parametric model. This figure illustrates the trade-off between model accuracy and model complexity in the log model evidence. A group-level model selection was performed between the categorical (GLM I) and the linear-parametric (GLM II) model (see Accuracy and Complexity). From the peak voxel in favor of GLM II within left V4, we extracted model accuracy (Acc), model complexity (Com) as well as log model evidence (LME) and calculated the difference of GLM I (m_1) relative to GLM II (m_2) indicating an advantage (higher accuracy, lower complexity, greater LME) for either GLM I (red) or GLM II (blue). (A) Regarding model accuracy, there is a slight advantage for the categorical model due to its higher flexibility. (B) However, model complexity is consistently lower for the parametric model due to its fewer regressors. (C) Since LME = Acc – Com (and therefore ΔLME = ΔAcc – ΔCom), this has the consequence that the log model evidence most often favors GLM II. This demonstrates that the complexity penalty is able to prevent overfitting that would occur from looking at the accuracy term alone.

to overfit data using a categorical model regarding this factor. Using cvBMS, this overfitting can be detected and effectively eliminated.

Model family comparison

Next, consider all the three models. It is clear that GLM II and GLM III are more similar to each other than each of them is similar to GLM I. When a model space is not balanced and some models are more similar to each other than others, more similar models share evidence between them. This effect is called "model dilution" and can lead to false conclusions about model preferences (Penny et al., 2010).

In order to avoid model dilution, we divided the model space into two model families: a family of parametric models consisting of GLM II and GLM III and a "family" of categorical models consisting only of GLM I. We calculated a log family evidence for the parametric models using Eq. (18) and compared it to the log model evidence of the categorical model using group-level model selection.

We find that there is overwhelming evidence for parametric processing in the entire brain and especially in the regions that we were most interested in, namely parts of V1 and the OC-sensitive V4 (see Fig. 3B). Only in some voxels, mostly located in right parietal cortex, the categorical model is better.

Why are the parametric models also superior in voxels not related to the paradigm? If there is no signal, model selection prefers simpler models. This means, although there is no signal that can be attributed to orientation pop-out processing in large parts of the cortex and

white matter, we might still develop a model preference in these voxels. If none of the models yields a useful description of the observed variations, it is only reasonable to work with the least complex model using the fewest regressors. Since the parametric models only have 3 regressors, they are to be preferred over the categorical model with 16 regressors, leading to a very strong model quality difference in white matter regions.

Individual model selection

Last, consider only GLM II and GLM III. From the previous "between-family" analysis, these two parametric models remain as candidates for the optimal model of orientation pop-out processing as operationalized by the present paradigm. We therefore performed a "within-family" analysis and submitted their log model evidences to another group-level model selection.

We find that, within the family of parametric models, there is overwhelming evidence for the nonlinear-parametric model in the entire brain and in particular the localizer regions V1 and V4 (see Fig. 3C). Voxels attributed to the linear-parametric model do not show a clear pattern, but seem to be restricted to white-matter parts of the brain.

Why is there a difference between the models in voxels not related to the paradigm? Model complexity is not just the number of parameters. This means, although both models have the same number of regressors and a similar structure with one onset regressor and two parametric modulators, they can still receive different complexity penalties which apparently has been the case here. The reason for this is that the complexity measure in the log model evidence is a measure of the informational content of the model and also depends on the data through the posterior distribution.

3rd model space: modelling confounding variables

So far, we have identified the optimal modelling of the variance components related to the processes of interest, namely orientation pop-out processing. In addition to this, there are processes of no interest, particularly the control fixation task employed in this experiment. To know whether these processes of no interest should be modelled or not is important, because how they are modelled might influence parameter estimates and statistical inferences for the processes of interest. This is what we are trying to find out using the third model space.

We use the winning model from the previous analysis (GLM III) as the null model for this model space. To this basic model, certain regressors are either added or not added, resulting in a model space spanned by binary axes: First, we can model the background stimulation (B), so far treated as an implicit baseline, using four regressors for the levels of orientation (0°, 45°, 90° or 135°). Second, we can model the fixation stimulus (S) shown on the screen, using two regressors for the square either opening to the left or to the right side. Third, we can model the fixation response (R) given by the subject, using two regressors for one-back responses and responses indicating the opposite.

If present, all regressors are timed according to the experiment: Background stimulations change every second, fixation stimuli occur every 2 s and are active for 1 s and fixation responses are locked to the time of the button press with a duration of zero. Since there are three features to be additionally modelled (background, stimulus, response) and there are two options for each feature (modelling or not modelling it), there are $2^3 = 8$ models in this model space with GLM III being the simplest and GLM III-B/S/R being the most complex model.

In summary, we find that none of these modifications is required for appropriate modelling of this data set (see Fig. 3D). In almost all parts of the brain, including V1 and V4, GLM III is selected as the best model of neural activity. GLMs including background stimulation are not selected in any voxel, suggesting that background stimulation in fact plays the role of an implicit baseline. Among the models without background stimulation, the control-task fixation stimulus (S) is hard to disentangle from the co-occurring fixation response (R), so that GLM III-S, GLM III-R and GLM III-S/R form clusters in motor cortex (due to button presses during the fixation response), auditory cortex (due to rhythmicity of the fixation stimulus) and ventricles (due to motion energy induced by motor activity).

The observed model preferences are thus consistent with the functional neuroanatomy of these regions. We suggest that modelling confounding variables is not required here, because (i) the changes that are captured are faster than the acquisition frequency implied by the TR = 2400 ms (background stimulation: T = 1 s; fixation stimulus: T = 2 s; fixation response: T ≈ 2 s), (ii) the additional regressors are highly correlated with the onset regressor describing the task and (iii) processes of interest and of no interest do not engage the same regions of the brain.

4th model space: lateralization of visual perception

Finally, we would like to demonstrate that model selection cannot only be used for methodological control of data analysis, but also to decide between competing hypotheses about neural processing. Imagine we would want to know which part of the visual field is processed in which part of the brain.[2] Formally, this corresponds to specifying models that describe different hemifields and comparing them in different hemispheres. For this purpose, we take the winning model from the last two model spaces, the nonlinear-parametric GLM III, and specify two new models: The first one only models OC on the left side (GLM III-l) and the second one only models OC on the right side (GLM III-r). This was achieved by just removing the opposite PM regressor in each case.

We present model selection results using histograms of likeliest frequencies (LF) across voxels from left and right V4, as defined by the localizer analysis. As can be seen, modelling only right OC makes a better model in left V4 (see Fig. 5A) and modelling only left OC is performing better in right V4 (see Fig. 5B) meaning that orientation pop-out processing is contra-lateralized (see Fig. 5C). This is consistent with the fact that contents of the left visual hemifield are neurally processed in the right visual cortex and vice versa (Van Essen et al., 1992) and shows that model selection is capable of detecting anatomically plausible lateralization effects.

For comparison purposes, we have also performed all nested model comparisons (1st, 3rd and 4th model space) using conventional GLM techniques, i.e. omnibus *F*-tests on the additional regressors. These results are provided in the supplementary online material (see Supplementary Fig. S1 and Table S1). Note that this type of classical model selection cannot be employed for non-nested model comparison (2nd model space).

Validation: simulated data

We additionally validate the cvBMS approach using simulated fMRI data. We place these simulations after the empirical validation, because we want to generate realistic fMRI signals and thus use one of the model spaces that were explained in the previous section and already applied to empirical fMRI data. We perform two simulations and investigate certain properties of the cvLME criterion in order to demonstrate that the cvBMS approach is an appropriate tool for model selection. We only focus on the performance of the cvLME on the first level and not on the validity of RFX BMS at the second

[2] Note that this model space represents a sanity-check analysis by which we just try to recover known model preferences, because we already know the answer.

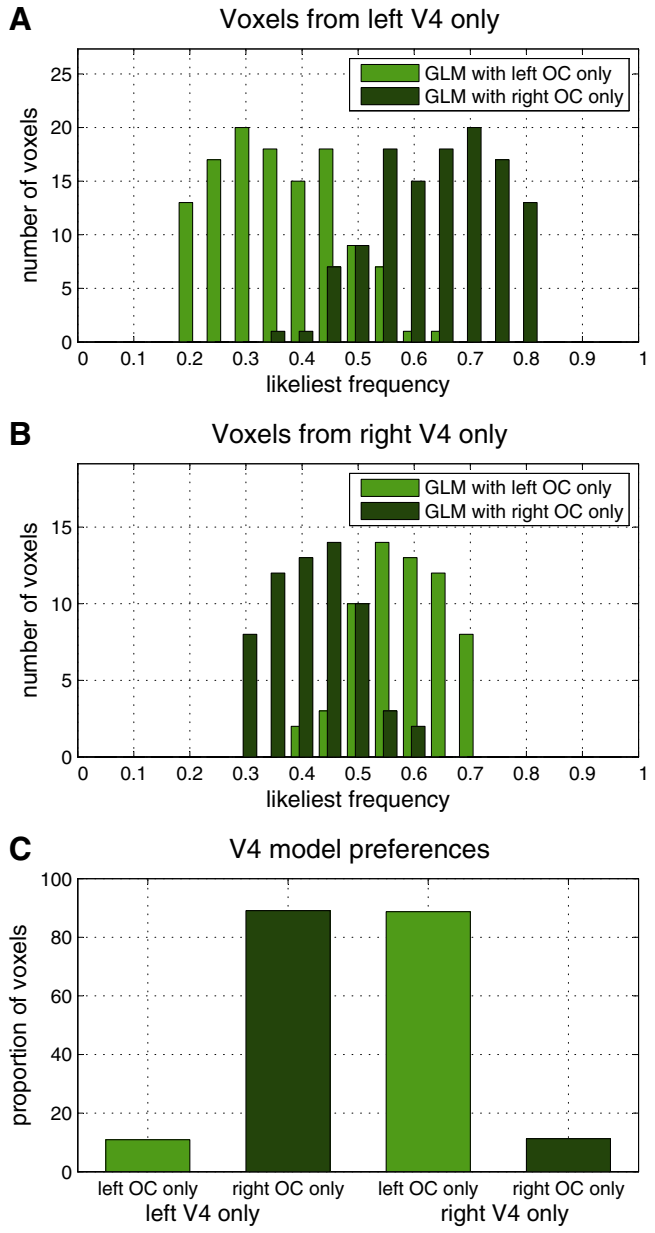

Fig. 5. Confirming lateralization of visual cortex via model selection. This figure illustrates the application of model selection to answer questions about functional neuroanatomy. A group-level model selection was performed between two models of orientation contrast (see 4th model space: lateralization of visual perception). We analyze likeliest frequencies (LF), which describe the proportion of subjects in which a particular model is optimal, and look at the performance of two models (including left OC or right OC) in two regions (left V4 and right V4). (A), (B) Histograms of LFs across the voxels from V4 as obtained from the localizer analysis. Clearly, the GLM with right OC only performs better in voxels from left V4 only and the other way around. (C) From the LF histograms, we calculated the proportion of voxels in which the two models were selected based on having the higher LF. The resulting plot demonstrates the interaction effect between side of the visual field and side of the visual cortex which is consistent with the contra-lateral processing in visual perception.

level since this latter part of the technique has already been validated (Stephan et al., 2009; Penny et al., 2010).

First, as model selection serves to infer the most likely model given the data, it should be capable of identifying true models – something which we do not know when analyzing empirical data, but which we can control in simulation-based validation. We therefore investigate the capability of the cvLME to identify the GLM that certain data were generated with. We analyze this while varying between-session variance of the true parameter values, because the cvLME assumes stationarity of model parameters across recording sessions. This means that model selection performance is expected to decrease as non-stationarity across sessions increases. For this simulation, we use the three non-nested models from the 2nd model space that represent different ways of how experimental factors can be described.

Second, the reason to perform model selection is not only to eventually use the generating model of an observed process, but also to estimate parameter values more precisely and to perform statistical tests more accurately. We therefore investigate the impact of model selection on mean squared errors of parameter estimates as well as sensitivity and specificity of statistical tests on model parameters. We analyze this in the context of correlated regressors, since high correlation of an additional regressor with existing regressors is often taken as an indication not to include it in order to preserve design orthogonality. In this simulation, we use a nested model difference motivated by GLM II and manipulate the correlation of one additional regressor with two regressors already present in the model.

1st simulation: influence of between-session variance

Methods

The model space consists of a categorical model (GLM I) which models all conditions of the 4×4 design using 16 regressors, a linear-parametric model (GLM II) which puts all trials into one regressor and encodes orientation contrast on the left and right visual hemifield as two parametric modulators and a nonlinear-parametric model (GLM III) in which the parametric modulators are a nonlinear function of orientation contrast. For the purpose of this simulation, movement parameters and implicit baseline were removed from the design matrix in each model (see Fig. 6A).

From the first model, we extract the covariance matrix V as computed by SPM during model estimation in order to induce the same temporal auto-correlation structure into all data. For each model, we extract the design matrix X as generated by SPM during model specification in order to simulate realistic fMRI signals. Both X and V describe 5 sessions of fMRI recording. Synthetic data is then generated as follows.

First, true regression coefficients are drawn using the relation

$$\beta_{ij} = x_j + y_{ij}$$
$$x_j \sim \mathrm{N}\left(0, \sigma_v^2\right)$$
$$y_{ij} \sim \mathrm{N}\left(0, \sigma_s^2\right)$$

where i and j index session and parameter respectively, σ_v^2 represents the voxel-to-voxel variance and σ_s^2 represents the session-to-session variance. In other words, we are working with regionally specific effects which are zero mean across all realizations (Friston et al., 2002a) and spherical covariance matrices on the regression coefficients (Penny, 2012).

The sampled parameters imply a certain true signal $X\beta$ and induce a certain signal variance $\mathrm{var}(X\beta)$. A residual or scan-to-scan variance σ^2 is chosen such that $\mathrm{var}(X\beta)/\sigma^2 = \mathrm{SNR}$ for a desired signal-to-noise ratio (SNR).

Second, simulated data is generated by sampling zero-mean Gaussian observation noise according to the relation $\varepsilon \sim \mathrm{N}(0, \sigma^2 V)$ and then adding the random noise to the true signal to get a measured signal $y = X\beta + \varepsilon$.

In our simulation, we set $\sigma_v^2 = 2.5$ and $\mathrm{SNR} = 4.5$ (which approximately correspond to the median values 2.67 and 4.57 observed in

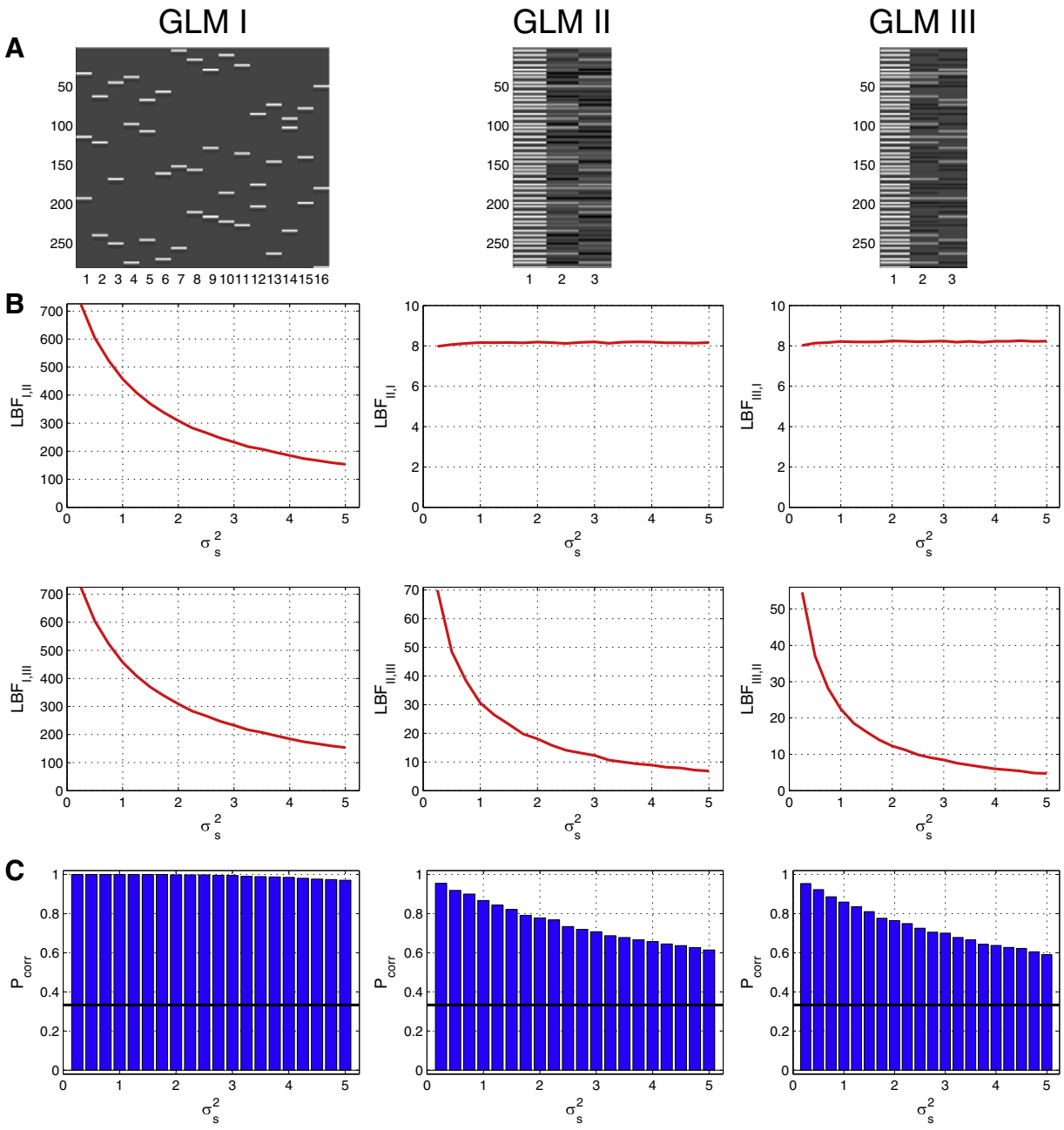

Fig. 6. Simulation performance of non-nested model selection. (A) Design matrices of the three models that were used in this simulation. Each column of the figure represents simulations in which the model at the top was the true model. (B) Average log Bayes factors (LBF) in favor of the true model, i.e. against the other two false models, as a function of between-session variance. (C) Proportion of correct model selections, i.e. successful identifications of the true model by maximum log model evidence (LME), as a function of between-session variance.

the SPM template data set[3]) while between-session variance σ_s^2 is varied from 0.25 to 5 in steps of 0.25.

For each level of σ_s^2, $N = 10{,}000$ data sets are generated using a true model which results in $N_{\text{sim}} = N \times 20 = 200{,}000$ simulations per model. Then, for each data set from each model, we calculate the cvLME using each model which gives rise to $N_{\text{sim}} \times 3 \times 3 = 1{,}800{,}000$ cvLMEs. Performance of cvLME is evaluated based on average model differences and the ability to recognize the true model.

[3] These data were first published as a study on repetition priming (Henson et al., 2002), previously used for model comparison (Penny et al., 2007a) and analyzed according to the SPM8 Manual (Ashburner et al., 2013).

Results

The results from this simulation are shown in Fig. 6. Average log Bayes factors (LBF) in favor of the true model and proportions of correct model selections (P_{corr}) are plotted as a function of the between-session variance σ_s^2. Note that σ_s^2 is proportional to the squared coefficient of variation of the regression coefficients across sessions $\text{CV}^2 = (\sigma_s / \langle|\beta_j|\rangle)^2$ and therefore can be seen as a measure of non-stationarity.

First, GLM I can be detected as the generating model with an overwhelming precision of 99.18 % (see Fig. 6C, left). This is due to the fact that this model differs strongly from the other two models which in turn only differ from each other in small details. Additionally, GLM II and GLM III make very special assumptions about the

signal as a function of orientation contrast while GLM I can account for more diverse activation patterns across experimental conditions.

Second, GLM II and GLM III can be reliably distinguished from each other, but less reliably selected against the other model. The reason for this is that true signals generated using GLM II or GLM III can also be identified by GLM I due to its higher flexibility. This means that GLM I can only be ruled out based on a complexity disadvantage. Still, detecting GLM II and GLM III as generating models works at acceptable accuracies of 75.06 % and 73.77 % (see Fig. 6C, middle and right).

To our knowledge, the session-to-session variance (here: $\sigma_s^2 = 0.25, \ldots, 5$) is almost always lower than the voxel-to-voxel variance (here: $\sigma_v^2 = 2.5$) in real fMRI data which is why we expect that first-level model detection accuracy will be higher than 80 % even for model differences hard to detect like GLM II/III vs. GLM I. With second-level model selection, an additional degree of accuracy will be achieved.

2nd simulation: influence of regressor correlation

Methods

We want to systematically study how model selection performance is influenced by shared variance between a model's regressors and another regressor that might be added or not added to the model. To this end, we take the design matrix of the linear-parametric model (GLM II, null model) and use its parametric modulators to create a more complex model (GLM II+, full model). Ignoring movement parameters and the implicit baseline, GLM II has three regressors (see Fig. 7A): an onset regressor encoding orientation contrast events (x_1), a parametric modulator for left OC (x_2) and a parametric modulator for right OC (x_3). Then, GLM II+ has an additional artificial regressor (x_4) that is given by

$$x_4 = \cos(\alpha)\, x_m + \sin(\alpha)\, x_o$$
$$x_m = \frac{1}{2}(x_2 + x_3)$$
$$x_2 \perp x_o \perp x_3$$

where x_m is the mean of x_2 and x_3 and x_o is a regressor that is orthogonal to x_2 and x_3.[4] In other words, we form x_4 as an average of a highly correlated regressor x_m and a completely uncorrelated regressor x_o where α determines the weighting and thereby controls how much variance x_4 shares with x_2 and x_3. Note that x_2 and x_3 as well as x_m and x_o are normalized to unit vectors in order to avoid scaling effects.

If α is very high, x_4 will be dominated by the orthogonal regressor so that its correlation with x_2 and x_3 will be very low. In the case that $\alpha = 0^\circ$, x_4 has a correlation of around 0.7 with x_2 and x_3 and a correlation of 1 with $(x_2 + x_3)$. Importantly, we will not investigate this extreme case, because the design matrix of GLM II+ is colinear in this situation, making it a redundant model that is to be avoided a priori. Instead, we vary α from 5° to 90° in steps of 5° and analyze different properties of the selected model.

Like in the first simulation, the design matrix X and the covariance matrix V are extracted from the estimated SPM models and describe 5 sessions of fMRI recording. Synthetic data is then generated using the same procedure as before, again using $\sigma_v^2 = 2.5$, but with a fixed $\sigma_s^2 = 1.25$ which induces a ratio of between-voxel to between-session variance of 2 and a coefficient of variation for the regression coefficients of around 0.88. Together with the very subtle model difference of a single covariate regressor, this relatively low effect consistency was intended to make model comparison as hard as possible.

For each level of α, we generate the corresponding x_4, and for each of these x_4, $N = 10{,}000$ data sets are generated resulting in $N_{\text{sim}} = N \times 20 = 200{,}000$ simulations per model (SNR = 0.1). In another $N_{\text{sim}} = 200{,}000$ simulations for each model (SNR = 10), the true values of the common model parameters x_2 and x_3 are set to zero in order to investigate specificity (when true values are zero) and sensitivity (when true values are not zero) of statistical tests later on. For all these simulations, cvLMEs are calculated and LBFs are used to define the selected and the rejected model in each simulation.

Results

The results from this simulation are shown in Fig. 7. The correlation of x_4 with x_2 and x_3 is shown as a function of regressor angle α. Moreover, mean squared errors (MSE) for the common model parameters x_2 and x_3 as well as sensitivity (true positive rate, TPR) and specificity (true negative rate, TNR) of an omnibus F-test on the parameters x_2 and x_3 are plotted against α.

The two models have three common parameters. The MSE describes the expected squared deviations of estimated parameters from their true values across simulations. When α is low and correlation is high, β_2 and β_3 are estimated closer to their true values when the selected model is used compared to when the rejected model is used (see Fig. 7C). Intuitively, this effect disappears with higher α. This confirms that even if the additional regressor is not of interest in itself, selecting the correct model still improves the parameter estimation for regressors of interest.

The capability of selected models to more precisely estimate parameters of interest also carries over to the quality of statistical test based on these parameter estimates. The sensitivity or TPR is the probability that the null hypothesis is rejected, given that it is false, and the specificity or TNR is the probability that the null hypothesis is not rejected, given that it is true. In our case, null hypothesis and alternative hypothesis are given by

$$H_0 : x_2 = x_3 = 0 \quad \text{and} \quad H_1 : x_2 \neq 0 \text{ or } x_3 \neq 0$$

which means that we are testing whether there is an effect of orientation contrast for at least one side of the visual field. We have simulations using both models, GLM II and GLM II+, in which H_1 is true (when parameters are drawn from distributions), as well as simulations in which H_0 is true (when parameters are set to zero deliberately).

The simulation shows that sensitivity is reduced when the null model is true and the full model is used (see Fig. 7D, left) and specificity is reduced when the full model is true and null model is used (see Fig. 7D, right). The fact that the specificity suffers massively when the more complex model is true and the rejected model is used means that the F-test can be heavily invalidated by a large false positive rate (FPR) when models omit variables that would be required for sufficient explanation of the data. In contrast, the specificity does not decrease very much when the simpler model is true and the rejected model is used in our example. This is because the full model also contains all the regressors that are in the null model.

The problem of adding non-orthogonal regressors to a design matrix is well known among practitioners of GLM-based fMRI analysis (Andrade et al., 1999) and design orthogonality is commonly used to assess the amount of covariation between regressors (Ashburner et al., 2013, ch. 28). As is evident from the plots, differences in MSE as well as sensitivity and specificity disappear when α

[4] In this implementation, x_o is chosen from the orthonormal basis of the design matrix calculated in MATLAB via `null([x1 x2 x3 xc]')` where x_c is a constant regressor.

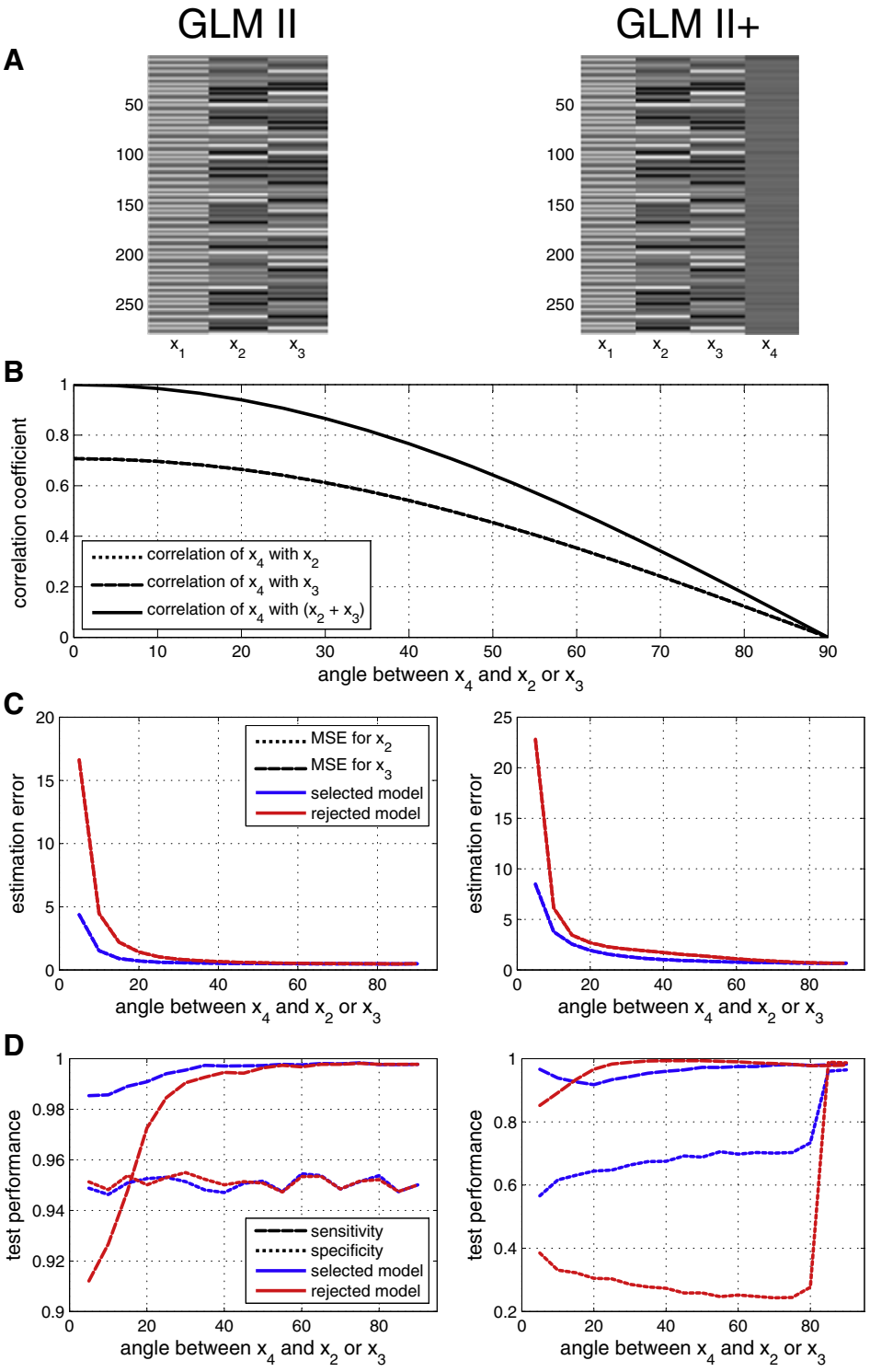

Fig. 7. Simulation performance of nested model comparison. (A) Design matrices of the two models that were used in this simulation. Each column of the figure represents simulations in which the model at the top was the true model. (B) The additional regressor of GLM II+ (x_4) was manipulated so that it showed different degrees of correlation with the two other regressors of interest (x_2 and x_3). (C) Mean-squared errors (MSE) for the common parameters x_2 (dotted) and x_3 (dashed) when analyzing the data using the selected model (blue) or the rejected model (red) as a function of their vector angle with x_4. (D) Sensitivity (dashed) and specificity (dotted) of an omnibus *F*-test on the common parameters x_2 and x_3 when analyzing the data using the selected model (blue) or the rejected model (red) as a function of their vector angle with x_4.

is high, such that correlation with x_4 becomes low. We therefore argue that regressors can be harmlessly added to the design matrix, if their correlation with the regressors already present is negligible. However, if they share considerable variance with other regressors, model selection is required to keep statistical inference valid.

Discussion

We have introduced a model selection approach for constraining analysis approaches when analyzing *functional magnetic resonance imaging* (fMRI) data using *general linear models* (GLMs). We have

demonstrated that *cross-validated Bayesian model selection* (cvBMS) serves its intended purpose and that it is useful in practice. Usage of model assessment and model selection effectively removes uncertainty from GLM-based fMRI analysis, reduces model misspecification and thereby enhances the methodological quality of functional neuroimaging studies (Friston, 2009).

Model selection against first-level mismodelling

In functional neuroimaging, mismodelling can lead to suboptimal signal explanation and false statistical inferences. If just one model is estimated and no other modelling approaches are considered, this can cause underfitting or overfitting so that researchers may fail to reject null hypotheses that are false (*false negatives*). If many models are tested and model choice is made by looking at significant results, this encourages p-hacking which may lead researchers to reject null hypotheses that are true (*false positives*). Using model selection, we can solve this problem. Model selection encourages multiple model estimation in order to avoid mismodelling, but bases the final model choice on objective criteria to avoid p-hacking. In our case, these objective criteria are how well a model fits the data (*model accuracy*), as a protection against underfitting, and how well it generalizes to new data (negative *model complexity*), as a protection against overfitting.

The advantages of our approach are manifold: It represents a substantial advance over occasionally performed informal model selection using classical significance tests. It properly quantifies *accuracy and complexity*, it can handle *arbitrary model differences* and it delivers *voxel-wise optimal models*. A disadvantage of the method is that cvLME maps have to be calculated for each model separately without the possibility of only estimating one global model and then comparing within this model (Kherif et al., 2002; Penny and Ridgway, 2013). This is however alleviated by the technique's computational efficiency and the modularity facilitated by the model-wise evidence calculation. Lastly, the possibility of using different GLMs for different parts of the brain can be easily implemented by masking group analyses with the *selected-model map* belonging to the model that the group analyses were performed with. This means that second-level model estimation is performed in all voxels, but second-level statistical inference is restricted to voxels where the corresponding first-level GLM is optimal.

The approach of cross-validated model selection thereby complements existing cross-validated pre-processing validation techniques addressing the issue of nuisance variable regression (Kay et al., 2013; Churchill et al., 2015). We envisage that also a lot of methodological problems of model building that have been discussed in the past, e.g. relating to the optimal hemodynamic response function (Lindquist et al., 2009; Steffener et al., 2010) or optimal reaction time modelling (Todd et al., 2013; Woolgar et al., 2014; Soch et al., 2015), can be approached using the proposed method. Beyond that, competing scientific theories, e.g. relating to prediction error processing (Knutson et al., 2001; Abler et al., 2006) or response conflict and time-on-task (Grinband et al., 2011; Yeung et al., 2011), might be tested against each other. Furthermore, subject-level measures of model quality such as posterior model probabilities (see First-level Bayesian model inference) may allow to uncover individual variation in stimulus processing, e.g. in the context of visual receptive field mapping (Thirion et al., 2006; Kay et al., 2008b).

In the remainder of this section, we discuss issues relating to the theoretical aspects of our model quality measure as well as its practical application to empirical data and directions for future research.

Cross-validation and Bayesian model selection

In principle, cross-validation in a Bayesian context does not differ from cross-validation in a classical context. Parameters of interest are estimated from one (part of the) data set in order to test generalization of a model to another (part of the) data set. In fact, the cross-validated log model evidence (cvLME) is just the Bayesian analogue to the sum of out-of-sample log-likelihoods (oosLL) in frequentist statistics, called the cross-validated log-likelihood (cvLL):

$$\mathrm{cvLL} = \sum_{i=1}^{S} \log\ p\left(y_i|\hat{\theta}\left(\cup_{j\neq i}\ y_j\right)\right)$$

$$\mathrm{cvLME} = \sum_{i=1}^{S} \log \int p\left(y_i|\theta\right)\ p\left(\theta|\cup_{j\neq i}\ y_j\right)\ \mathrm{d}\theta$$

Here, S denotes the number of sessions indicating S-fold cross-validation and the union symbol $\cup$ signifies that parameters are estimated from several sessions as ML estimates $\hat{\theta}(y)$ for cvLL or as posterior distributions $p(\theta|y)$ for cvLME. Whereas cvLL uses point estimates and the likelihood function, cvLME accounts for the whole uncertainty about model parameters by using the marginal likelihood.

Cross-validation requires that data sets are drawn from the same source or population. This means that, in application to fMRI, the above formulas are based on the assumption of independence between recording sessions, mathematically implying that likelihood functions and marginal likelihoods factorize over parts of the data. Furthermore, cross-validating the log model evidence requires a certain degree of stationarity across sessions. Using simulations, we have investigated the influence of non-stationarity in the form of between-session variance on model parameters and found that the cvLME can accommodate moderate stationarity violations.

Usually, cross-validation is considered to be a measure against overfitting. Since we are controlling for overfitting (and underfitting) already using the model evidence itself, our use of cross-validation has three other reasons: First, we wanted to avoid having to specify prior distributions on the model parameters. Second, we wanted to avoid approximations that originate from methods estimating hyperparameters like Empirical Bayes (Friston et al., 2008; Penny and Ridgway, 2013) or methods employing hyperpriors like Variational Bayes (Penny et al., 2003, 2005, 2007a). Third, and most importantly, we wanted to use the whole potential of the complexity penalty in the log model evidence along the following rationale.

As we have seen earlier (see Accuracy and Complexity), the Bayesian complexity is identical to a Kullback-Leibler (KL) divergence between the posterior distribution and the prior distribution over model parameters. If a model is more likely than another, i.e. better describing the underlying causes of the measured signal, it will estimate the parameters more uniformly across sessions, because it exactly and only incorporates the required parameters. This means that priors obtained via cross-validation from all except one session will be closer to the posteriors in the remaining session. Consequently, for such models, the KL divergence between the posterior distribution and a prior obtained from independent data will be smaller on average and they will receive a higher log model evidence. On these grounds, one could even say that the complexity penalty in the log model evidence is never really used properly unless empirical priors are employed.

Pragmatics of model selection for methodological control

There is no statistical inference without a statistical model. This is the reason why model selection is different from scientific inference. While we might not declare a model difference significant, we will still make a decision which model we use for data analysis, because in the end we have to use one model to make any inference at all. This highlights that a different rationale is required in model selection, because it is a decision-theoretic, not an inferential problem. We have shown that our approach can be framed in the logic of decision

Table 1
Sequential model selection for orientation pop-out processing. This table breaks down model selection results for the four model spaces reported earlier (see Figs. 3 and 5). For each model space, the number of voxels in which each model is selected as the optimal model by cvBMS is given for the two regions of interest, anterior parts of V1 and the orientation-sensitive V4, as obtained from the localizer paradigm and separated by cranial hemisphere, as well as at the whole-brain level.

	whole-brain	V1		V4	
		left	right	left	right
number of voxels	53265	16	18	119	62
1st model space: technical model parameters					
GLM cHRF	37215	16	18	58	26
GLM cHRF + 1	16050	0	0	61	36
GLM cHRF + 2	0	0	0	0	0
2nd model space: modelling experimental factors (family selection)					
GLM I	1194	0	0	0	0
GLMs II & III	52071	16	18	119	62
2nd model space: modelling experimental factors (model selection)					
GLM II	4910	3	0	0	0
GLM III	48355	13	18	119	62
3rd model space: modelling confounding variables					
GLM III	51050	16	18	119	62
GLM III-S	647	0	0	0	0
GLM III-R	1074	0	0	0	0
GLM III-S/R	494	0	0	0	0
4th model space: lateralization of visual perception					
GLM III-l	23006	0	1	13	55
GLM III-r	30259	16	17	106	7

theory and leads to a very simple decision rule. As a consequence, when for example comparing two models, the threshold for a model to be selected will not be an estimated model frequency of say 0.95, but 0.5 in order to force a decision for one of the two models in any possible situation.

Another problem is model dilution. Like any comparison of more than two models, RFX BMS with three or more models is susceptible to the problem of model dilution (Penny et al., 2010). Model dilution occurs when some models are more similar to each other than to others and evidential support for this type of model is shared between them so that a sub-optimal model is selected as the winning model (Penny et al., 2010, p. 5). For this reason, a model space has to be balanced regarding all its axes in order to avoid model dilution. For example, our 3rd model space was balanced, because it had three axes (background stimulation, fixation stimulus, fixation response) with two levels each (modelling or not modelling it) and all $2^3 = 8$ possible combinations were evaluated. In contrast, our 2nd model space with one categorical model and two parametric models was not balanced which is why we first did a family selection and then performed model selection within the winning family.

A third issue is sequential model selection. While the separation of different modelling decisions into consecutive model spaces with the winning model at one stage being the starting point for the next stage enhances the practical usability of our method, it is based on the assumption that model space dimensions from different stages do not interact with each other. For example, it is possible (though not very likely) that the parametric models would have performed better with two HRF derivatives (GLM II/III, cHRF + 2) which was not tested, because the categorical model came out best with just the canonical HRF (GLM I, cHRF). In other words, when we use a model in configuration A_1 (rather than A_2) and compare model configurations B_1 and B_2, we believe that the preference regarding B_1 vs. B_2 will be the same with a model in configuration A_2 (just like in A_1). If we do not believe this, sequential model selection is invalid and the full model space ($A_1/B_1, A_1/B_2, A_2/B_1, A_2/B_2$) has to be analyzed.

Across-voxel model selection for fMRI data analysis

Usually, in fMRI, signals from tens of thousands of voxels are acquired which makes computational complexity an important issue. We restricted the method to simple, but reasonable assumptions about model structure corresponding to established first-level data analyses, and we were able to derive analytical solutions for the general linear model with normal-gamma priors (GLM-NG) which allows whole-brain model comparison for large model spaces and many subjects in affordable time.[5]

The mass-univariate nature of standard fMRI data analysis also raises the question of across-voxel inference for model selection. Across-voxel model selection is the task of selecting an optimal model for a set of voxels. While the best model in the full brain was always more or less clear (see Fig. 3 and Table 1), there has also been variation in the optimal model across voxels. On the one hand, different voxels might require different analysis strategies since the corresponding brain area may be involved in the cognitive task in different ways. On the other hand, it could be useful to model a particular brain region uniformly using the same model in all voxels or to know which model is best in the whole brain. Overall, we see four ways to deal with this topic.

First, one can regard across-voxel inference as not sensible at the whole-brain level and just mask second-level inference based on a specific first-level model with the respective selected-model map, as proposed in this paper. This enables spatially heterogeneous modelling and ensures that inference is safe at each voxel.

[5] We tested the speed of our method by calculating voxel-wise cvLME for GLM III in our empirical data set. With a time consumption of 02:18 min, cvLME calculation was almost as fast as SPM's classical (ReML) model estimation (01:33 min) and clearly outperformed computation of a log model evidence map using SPM's (Variational) Bayesian GLM with standard settings [AR order of 3, UGL spatial priors] (40:06 min) and with most parsimonious settings [AR order of 0, no spatial prior information] (15:16 min). All computations were performed using MATLAB R2013b on a 64-bit Windows 7 PC with 16 GB RAM and four hyperthreaded Intel i7 CPU kernels working at 3.40 GHz.

Second, a model explicitly accounting for the covariance of model preference across voxels could be specified. This can happen at the first level by building multivariate GLMs on multiple time series y (Allefeld and Haynes, 2014) or at the second level by building a spatial model over model frequencies r (Rosa et al., 2010, p. 224).

Third, instead of operating on the parameter estimates from the same first-level model, second-level analysis could be calculated from the estimates of different GLMs, depending on which model is optimal in each single subject. Using subject-wise selected-model maps, this could be easily implemented.

Fourth, Bayesian model averaging (BMA) on the common model parameters might be applied (Penny et al., 2010) before entering first-level estimates into second-level analyses. Using the posterior model probabilities as weighting factors, this would account for the whole uncertainty about modelling approaches.

The last two approaches make across-voxel inference unnecessary, because there would be no decision for a particular model at each voxel to be made anymore, but are also restricted to model spaces in which the regressors of interest are present in all models. The question of across-voxel model selection and the potential of BMA across GLMs for fMRI could therefore be a promising direction for future research.

Software note

An implementation of voxel-wise cross-validated Bayesian model selection (cvBMS) compatible with SPM8 and SPM12 can be obtained from the corresponding author.

Acknowledgments

This work was supported by the Bernstein Computational Neuroscience Program of the German Federal Ministry of Education and Research (BMBF grant 01GQ1001C), the Research Training Group "Sensory Computation in Neural Systems" (GRK 1589/1-2), the Collaborative Research Center "Volition and Cognitive Control: Mechanisms, Modulations, Dysfunctions" (SFB 940/1) and the German Research Foundation (DFG grants EXC 257 and KFO 247).

Joram Soch received a Humboldt Research Track Scholarship and receives an Elsa Neumann Scholarship from the State of Berlin.

The authors would like to thank Torsten Wüstenberg for providing ideas to this project and Carsten Bogler for providing access to the fMRI data set.

The authors have no conflict of interest, financial or otherwise, to declare.

Appendix A. Parameter estimation for the Bayesian GLM

For the Bayesian GLM, we write the likelihood function as:

$$\begin{aligned} p\left(y|\beta,\tau\right) &= \mathrm{N}\left(y; X\beta, (\tau P)^{-1}\right) \\ &= \sqrt{\frac{|\tau P|}{(2\pi)^n}} \exp\left[-\frac{\tau}{2}(y-X\beta)^T P\,(y-X\beta)\right] \end{aligned} \tag{A.1}$$

As a prior distribution, we use the normal-gamma distribution:

$$\begin{aligned} p(\beta|\tau) &= \mathrm{N}\left(\beta; \mu_0, (\tau\Lambda_0)^{-1}\right) = \sqrt{\frac{|\tau\Lambda_0|}{(2\pi)^p}} \exp\left[-\frac{\tau}{2}(\beta-\mu_0)^T\Lambda_0(\beta-\mu_0)\right] \\ p(\tau) &= \mathrm{Gam}\left(\tau; a_0, b_0\right) = \frac{b_0^{a_0}}{\Gamma(a_0)}\,\tau^{a_0-1}\exp[-b_0\tau] \end{aligned} \tag{A.2}$$

As every Bayesian analysis, Bayesian inference for the GLM is based on Bayes' theorem:

$$p(\theta|y,m) = \frac{p\left(y|\theta,m\right)\; p\left(\theta|m\right)}{p(y|m)} \tag{A.3}$$

For the Bayesian general linear model with normal-gamma priors (GLM-NG), this reads:

$$p(\beta,\tau|y) = \frac{p\left(y|\beta,\tau\right)\; p\left(\beta,\tau\right)}{p(y)} \tag{A.4}$$

Since $p(y)$ is just a normalization factor, the equation becomes a proportionality:

$$p\left(\beta,\tau|y\right) \propto p\left(y|\beta,\tau\right)\; p\left(\beta,\tau\right) \tag{A.5}$$

Writing this out for the GLM-NG yields:

$$\begin{aligned} p\left(\beta,\tau|y\right) \propto p\left(y|\beta,\tau\right)\, p\left(\beta,\tau\right) &= p\left(y|\beta,\tau\right)\, p\left(\beta|\tau\right)\, p(\tau) \\ &= \sqrt{\frac{|\tau P|}{(2\pi)^n}}\; \exp\left[-\frac{\tau}{2}(y-X\beta)^T P\,(y-X\beta)\right] \\ &\cdot \sqrt{\frac{|\tau\Lambda_0|}{(2\pi)^p}}\; \exp\left[-\frac{\tau}{2}(\beta-\mu_0)^T\Lambda_0\,(\beta-\mu_0)\right] \\ &\cdot \frac{b_0^{a_0}}{\Gamma(a_0)}\,\tau^{a_0-1}\exp\left[-b_0\tau\right] \end{aligned} \tag{A.6}$$

Collecting identical variables gives:

$$\begin{aligned} p\left(\beta,\tau|y\right) \propto\; & \sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}}|P||\Lambda_0|}\; \frac{b_0^{a_0}}{\Gamma(a_0)}\,\tau^{a_0-1}\exp\left[-b_0\tau\right] \\ & \cdot \exp\left[-\frac{\tau}{2}\left((y-X\beta)^T P(y-X\beta) + (\beta-\mu_0)^T\Lambda_0\,(\beta-\mu_0)\right)\right] \end{aligned} \tag{A.7}$$

Completing the square over β gives:

$$\begin{aligned} p\left(\beta,\tau|y\right) \propto\; & \sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}}\,|P||\Lambda_0|}\; \frac{{b_0}^{a_0}}{\Gamma(a_0)}\,\tau^{a_0-1}\exp\left[-b_0\tau\right] \\ & \cdot \exp\left[-\frac{\tau}{2}\left((\beta-\mu_n)^T\Lambda_n\,(\beta-\mu_n) + \left(y^T P y + \mu_0^T\Lambda_0\mu_0 - \mu_n^T\Lambda_n\mu_n\right)\right)\right] \end{aligned} \tag{A.8}$$

From this, we can isolate the posterior distribution over β:

$$\begin{aligned} p\left(\beta|\tau,y\right) &= \mathrm{N}\left(\beta; \mu_n, (\tau\Lambda_n)^{-1}\right) \\ \mu_n &= \Lambda_n^{-1}\left(X^T P y + \Lambda_0\mu_0\right) \\ \Lambda_n &= X^T P X + \Lambda_0 \end{aligned} \tag{A.9}$$

From the remaining term, we can isolate the posterior distribution over τ:

$$\begin{aligned} p\left(\tau|y\right) &= \mathrm{Gam}(\tau; a_n, b_n) \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}\left(y^T P y + \mu_0^T\Lambda_0\mu_0 - \mu_n^T\Lambda_n\mu_n\right) \end{aligned} \tag{A.10}$$

Appendix B. Log model evidence for the Bayesian GLM

According to the law of marginal probability, the model evidence is given by

$$p\left(y|m\right) = \int p\left(y|\theta,m\right)\ p\left(\theta|m\right)\ \mathrm{d}\theta \tag{B.1}$$

Writing this out for the GLM-NG yields:

$$p\left(y|m\right) = \iint p\left(y|\beta,\tau\right)\ p\left(\beta|\tau\right)\ p\left(\tau\right)\ \mathrm{d}\beta\ \mathrm{d}\tau \tag{B.2}$$

At first, we only focus on the integrand

$$p\left(y,\beta,\tau\right) = p\left(y|\beta,\tau\right)\ p\left(\beta|\tau\right)\ p\left(\tau\right) \tag{B.3}$$

We have already evaluated this term as

$$\begin{aligned} p\left(y,\beta,\tau\right) = & \sqrt{\frac{\tau^n|P|}{(2\pi)^n}}\sqrt{\frac{\tau^p|\Lambda_0|}{(2\pi)^p}}\ \frac{b_0^{a_0}}{\Gamma(a_0)}\ \tau^{a_0-1}\exp\left[-b_0\tau\right] \\ & \cdot\exp\left[-\frac{\tau}{2}\left(\left(\beta-\mu_n\right)^T\Lambda_n\left(\beta-\mu_n\right)+\left(y^TPy+\mu_0^T\Lambda_0\mu_0-\mu_n^T\Lambda_n\mu_n\right)\right)\right] \end{aligned} \tag{B.4}$$

Using the posterior distribution over β, we can rewrite this as

$$\begin{aligned} p\left(y,\beta,\tau\right) = & \sqrt{\frac{\tau^n|P|}{(2\pi)^n}}\sqrt{\frac{\tau^p|\Lambda_0|}{(2\pi)^p}}\sqrt{\frac{(2\pi)^p}{\tau^p\,|\Lambda_n|}}\ \frac{b_0^{a_0}}{\Gamma(a_0)}\ \tau^{a_0-1}\exp\left[-b_0\tau\right] \\ & \cdot\mathrm{N}\left(\beta;\mu_n,\left(\tau\Lambda_n\right)^{-1}\right)\exp\left[-\frac{\tau}{2}\left(y^TPy+\mu_0^T\Lambda_0\mu_0-\mu_n^T\Lambda_n\mu_n\right)\right] \end{aligned} \tag{B.5}$$

Now, β can be integrated out easily:

$$\begin{aligned} \int p\left(y,\beta,\tau\right)\ \mathrm{d}\beta = & \sqrt{\frac{\tau^n|P|}{(2\pi)^n}}\sqrt{\frac{|\Lambda_0|}{|\Lambda_n|}}\ \frac{b_0^{a_0}}{\Gamma(a_0)}\ \tau^{a_0-1}\exp\left[-b_0\tau\right] \\ & \cdot\exp\left[-\frac{\tau}{2}\left(y^TPy+\mu_0^T\Lambda_0\mu_0-\mu_n^T\Lambda_n\mu_n\right)\right] \end{aligned} \tag{B.6}$$

Using the posterior distribution over τ, we can rewrite this as

$$\int p\left(y,\beta,\tau\right)\ \mathrm{d}\beta = \sqrt{\frac{|P|}{(2\pi)^n}}\sqrt{\frac{|\Lambda_0|}{|\Lambda_n|}}\ \frac{b_0^{a_0}}{\Gamma(a_0)}\ \frac{\Gamma(a_n)}{b_n^{a_n}}\ \mathrm{Gam}\left(\tau;a_n,b_n\right) \tag{B.7}$$

Finally, τ can also be integrated out:

$$\iint p\left(y,\beta,\tau\right)\ \mathrm{d}\beta\ \mathrm{d}\tau = \sqrt{\frac{|P|}{(2\pi)^n}}\sqrt{\frac{|\Lambda_0|}{|\Lambda_n|}}\ \frac{\Gamma(a_n)}{\Gamma(a_0)}\ \frac{b_0^{a_0}}{b_n^{a_n}} = p\left(y|m\right) \tag{B.8}$$

Thus, the log model evidence is given by

$$\begin{aligned} \log p\left(y|m\right) = & \frac{1}{2}\log|P| - \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Lambda_0| - \frac{1}{2}\log|\Lambda_n| \\ & + \log\Gamma(a_n) - \log\Gamma(a_0) + a_0\log b_0 - a_n\log b_n \end{aligned} \tag{B.9}$$

Appendix C. Accuracy/Complexity for the Bayesian GLM

The Bayesian model accuracy is defined as

$$\mathrm{Acc}(m) = \left\langle\log p\left(y|\theta,m\right)\right\rangle_{p(\theta|y,m)} = \int p(\theta|y,m)\ \log p(y|\theta,m)\ \mathrm{d}\theta \tag{C.1}$$

For the GLM-NG, it evaluates to

$$\begin{aligned} \mathrm{Acc}(m) = & -\frac{1}{2}\frac{a_n}{b_n}\left[\left(y-X\mu_n\right)^TP\left(y-X\mu_n\right)\right]+\frac{1}{2}\log|P|-\frac{n}{2}\log(2\pi) \\ & +\frac{n}{2}\left(\psi\left(a_n\right)-\log(b_n)\right)-\frac{1}{2}\mathrm{tr}\left(X^TPX\Lambda_n^{-1}\right) \end{aligned} \tag{C.2}$$

The Bayesian model complexity is defined as

$$\mathrm{Com}(m) = \mathrm{KL}\left[p(\theta|y,m)||p(\theta|m)\right] = \int p\left(\theta|y,m\right)\ \log\frac{p\left(\theta|y,m\right)}{p\left(\theta|m\right)}\ \mathrm{d}\theta \tag{C.3}$$

For the GLM-NG, it evaluates to

$$\begin{aligned} \mathrm{Com}(m) = & +\frac{1}{2}\frac{a_n}{b_n}\left[\left(\mu_0-\mu_n\right)^T\Lambda_0\left(\mu_0-\mu_n\right)-2\left(b_n-b_0\right)\right] \\ & -\frac{1}{2}\log\frac{|\Lambda_0|}{|\Lambda_n|}+\frac{1}{2}\mathrm{tr}\left(\Lambda_0\Lambda_n^{-1}\right)-\frac{p}{2}+a_0\log\frac{b_n}{b_0} \\ & -\log\frac{\Gamma(a_n)}{\Gamma(a_0)}+\left(a_n-a_0\right)\psi(a_n) \end{aligned} \tag{C.4}$$

Appendix D. Second-level Bayesian model selection

We consider model evidences $p(y_i|m_j)$ for the data from subjects $i = 1,\ldots,N$ analyzed using models $j = 1,\ldots,M$. In this setting, model frequencies can be inferred using a hierarchical population proportion model defined by the following probability densities:

$$\begin{aligned} p\left(y|m\right) &= \prod_{i=1}^{N}p\left(y_i|m_i\right) = \prod_{i=1}^{N}\prod_{j=1}^{M}p\left(y_i|e_j\right)^{m_{ij}} \\ p(m|r) &= \prod_{i=1}^{N}\mathrm{Mult}(m_i;1,r) = \prod_{i=1}^{N}\prod_{j=1}^{M}r_j^{m_{ij}} \\ p\left(r|\alpha\right) &= \mathrm{Dir}\left(r;\alpha\right) = \frac{\Gamma\left(\sum_{j=1}^{M}\alpha_j\right)}{\prod_{j=1}^{M}\Gamma\left(\alpha_j\right)}\prod_{j=1}^{M}r_j^{\alpha_j-1} \end{aligned} \tag{D.1}$$

Accounting for different subjects being best explained by different models, i.e. allowing the variable model to be a random effect, a Variational Bayesian (VB) iterative estimation algorithm has been developed (Stephan et al., 2009) to infer a posterior distribution over

model frequencies $p(r|y)$ from prior concentration parameters α_0 as follows:

$$\begin{aligned}
&\alpha = \alpha_0 \\
&\text{until convergence} \\
&\quad u_{ij} = \exp\left[\log p\left(y_i|e_j\right) + \psi\left(\alpha_j\right) - \psi\left(\sum_{j=1}^{M} \alpha_j\right)\right] \\
&\quad \beta_j = \sum_{i=1}^{N} \frac{u_{ij}}{u_i} \quad \text{where} \quad u_i = \sum_{j=1}^{M} u_{ij} \\
&\quad \alpha = \alpha_0 + \beta \\
&\text{end} \\
&p(r|y) = \mathrm{Dir}\left(r; \alpha\right)
\end{aligned} \tag{D.2}$$

This algorithm is best understood as distributing subjects on models where each subject can be divided onto the models depending on its relative model evidences. Therefore, when the uniform prior $\alpha_0 = [1, \ldots, 1]$ is used, the posterior $\alpha_j - 1$ can be regarded as the effective number of subjects "using" model j (Stephan et al., 2009).

Appendix E. Decision-theoretic model choice

We define the 1-0 utility function on the variable m

$$U\left(m, \hat{m}\right) = \left[m = \hat{m}\right] \tag{E.1}$$

and the benefit function as the posterior expected utility

$$B(\hat{m}) = \mathrm{E}_{r|y}\left[\mathrm{E}_{m|r}\left[U(m, \hat{m})\right]\right] \tag{E.2}$$

where $p(m|r)$ and $p(r|y)$ come from second-level RFX BMS. Since the utility function is only one when the true model is correctly identified, the benefit function becomes

$$B\left(\hat{m}\right) = \int p\left(\hat{m}|r\right)\ p(r|y)\ \mathrm{d}r \tag{E.3}$$

With the probability densities (D.1), this evaluates to

$$B(\hat{m}) = \frac{\Gamma\left(\sum_{j=1}^{M} \alpha_j\right)}{\prod_{j=1}^{M} \Gamma\left(\alpha_j\right)} \frac{\prod_{j=1}^{M} \Gamma\left(\alpha_j + \sum_{i=1}^{N} \hat{m}_{ij}\right)}{\Gamma\left(\sum_{j=1}^{M}\left(\alpha_j + \sum_{i=1}^{N} \hat{m}_{ij}\right)\right)} \tag{E.4}$$

With the estimation algorithm (D.2), this evaluates to

$$B(\hat{m}) = \frac{\Gamma(M+N)}{\Gamma(M+2N)} \prod_{j=1}^{M} \prod_{k=0}^{\sum \hat{m}_{ij}-1} \left(\alpha_j + k\right) \tag{E.5}$$

The optimal decision regarding models m is the one that maximizes the benefit:

$$m^* = \arg\max_{\hat{m}} B(\hat{m}) \tag{E.6}$$

Remember that m is an $N \times M$ matrix. Therefore, in principle, a different model could be chosen for each subject. However, if the same model k has to be used in all subjects which generally is the case in fMRI data analysis, the benefit function becomes

$$B(k) \propto \prod_{i=1}^{N} \left(\alpha_k + i - 1\right) \tag{E.7}$$

With this correction on the benefit function, the optimal decision is given by

$$m^* = 1_N \otimes e_k, \quad \text{with} \quad k = \arg\max_j \alpha_j \tag{E.8}$$

This means that choosing models with respect to the largest posterior alphas or, equivalently, the largest estimated frequency, in an RFX BMS maximizes the expected utility or benefit function and is the recommended choice according to Bayesian decision theory (BDT).

Appendix F. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2016.07.047.

References

Abler, B., Walter, H., Erk, S., Kammerer, H., Spitzer, M., 2006. Prediction error as a linear function of reward probability is coded in human nucleus accumbens. NeuroImage 31, 790–795. http://linkinghub.elsevier.com/retrieve/pii/S1053811906000188. http://dx.doi.org/10.1016/j.neuroimage.2006.01.001.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19, 716–723. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1100705. http://dx.doi.org/10.1109/TAC.1974.1100705.

Allefeld, C., Haynes, J.D., 2014. Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. NeuroImage 89, 345–357. http://linkinghub.elsevier.com/retrieve/pii/S1053811913011920. http://dx.doi.org/10.1016/j.neuroimage.2013.11.043.

Andrade, A., Paradis, A.L., Rouquette, S., Poline, J.B., 1999. Ambiguous results in functional neuroimaging data analysis due to covariate correlation. NeuroImage 10, 483–486. http://www.sciencedirect.com/science/article/pii/S1053811999904792. http://dx.doi.org/10.1006/nimg.1999.0479.

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Stat. Surveys 4, 40–79. http://projecteuclid.org/euclid.ssu/1268143839. http://dx.doi.org/10.1214/09-SS054.

Ashburner, J., Friston, K., Penny, W., Stephan, K.E., et al. 2013. SPM8 Manual. http://www.fil.ion.ucl.ac.uk/spm/doc/spm8_manual.pdf.

Bishop, C.M., 2007. Pattern Recognition and Machine Learning. 1st ed., Springer, New York.

Boekel, W., Wagenmakers, E.J., Belay, L., Verhagen, J., Brown, S., Forstmann, B.U., 2015. A purely confirmatory replication study of structural brain-behavior correlations. Cortex 66, 115–133. http://linkinghub.elsevier.com/retrieve/pii/S0010945215000155. http://dx.doi.org/10.1016/j.cortex.2014.11.019.

Bogler, C., Bode, S., Haynes, J.D., 2013. Orientation pop-out processing in human visual cortex. NeuroImage 81, 73–80. http://linkinghub.elsevier.com/retrieve/pii/S105381191300534X. http://dx.doi.org/10.1016/j.neuroimage.2013.05.040.

Burrows, B.E., Moore, T., 2009. Influence and limitations of popout in the selection of salient visual stimuli by area V4 neurons. J. Neurosci. 29, 15169–15177. http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3710-09.2009. http://dx.doi.org/10.1523/JNEUROSCI.3710-09.2009.

Carp, J., 2012. On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments. Front. Neurosci. 6, A149. http://journal.frontiersin.org/article/10.3389/fnins.2012.00149/abstract. http://dx.doi.org/10.3389/fnins.2012.00149.

Chai, X.J., Castañón, A.N., Öngür, D., Whitfield-Gabrieli, S., 2012. Anticorrelations in resting state networks without global signal regression. NeuroImage 59, 1420–1428. http://www.sciencedirect.com/science/article/pii/S1053811911009657. http://dx.doi.org/10.1016/j.neuroimage.2011.08.048.

Churchill, N.W., Spring, R., Afshin-Pour, B., Dong, F., Strother, S.C., 2015. An automated, adaptive framework for optimizing preprocessing pipelines in task-based functional MRI. PLOS ONE 10, e0131520. http://dx.plos.org/10.1371/journal.pone.0131520. http://dx.doi.org/10.1371/journal.pone.0131520.

Deserno, L., Sterzer, P., Wustenberg, T., Heinz, A., Schlagenhauf, F., 2012. Reduced prefrontal-parietal effective connectivity and working memory deficits in schizophrenia. J. Neurosci. 32, 12–20. http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3405-11.2012. http://dx.doi.org/10.1523/JNEUROSCI.3405-11.2012.

Eklund, A., Andersson, M., Josephson, C., Johannesson, M., Knutsson, H., 2012. Does parametric fMRI analysis with SPM yield valid results?—an empirical study of 1484 rest datasets. NeuroImage 61, 565–578. http://linkinghub.elsevier.com/retrieve/pii/S1053811912003825. http://dx.doi.org/10.1016/j.neuroimage.2012.03.093.

Fox, M.D., Corbetta, M., Snyder, A.Z., Vincent, J.L., Raichle, M.E., 2006. Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems.

Proc. Natl. Acad. Sci. 103, 10046–10051. http://www.pnas.org/cgi/doi/10.1073/pnas.0604187103. http://dx.doi.org/10.1073/pnas.0604187103.
Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., Raichle, M.E., 2005, Jul. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proc. Natl. Acad. Sci. 102, 9673–9678. http://www.pnas.org/cgi/doi/10.1073/pnas.0504136102. http://dx.doi.org/10.1073/pnas.0504136102.
Friston, K., Chu, C., Mourão-Miranda, J., Hulme, O., Rees, G., Penny, W., Ashburner, J., 2008. Bayesian decoding of brain images. NeuroImage 39, 181–205. http://linkinghub.elsevier.com/retrieve/pii/S1053811907007203. http://dx.doi.org/10.1016/j.neuroimage.2007.08.013.
Friston, K., Glaser, D., Henson, R., Kiebel, S., Phillips, C., Ashburner, J., 2002b. Classical and Bayesian inference in neuroimaging, part II: applications. NeuroImage 16, 484–512. http://linkinghub.elsevier.com/retrieve/pii/S1053811902910918. http://dx.doi.org/10.1006/nimg.2002.1091.
Friston, K., Harrison, L., Penny, W., 2003, Aug. Dynamic causal modelling. NeuroImage 19 (4), 1273–1302. http://dx.doi.org/10.1016/S1053-8119(03)00202-7.
Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W., 2007. Variational free energy and the Laplace approximation. NeuroImage 34, 220–234. http://linkinghub.elsevier.com/retrieve/pii/S1053811906008822. http://dx.doi.org/10.1016/j.neuroimage.2006.08.035.
Friston, K., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J., 2002a. Classical and Bayesian inference in neuroimaging, part I: theory. NeuroImage 16, 465–483. http://linkinghub.elsevier.com/retrieve/pii/S1053811902910906. http://dx.doi.org/10.1006/nimg.2002.1090.
Friston, K.J., 2009. Modalities, modes, and models in functional neuroimaging. Science 326, 399–403. http://www.sciencemag.org/cgi/doi/10.1126/science.1174521. http://dx.doi.org/10.1126/science.1174521.
Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.J., 1994. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. 2, 189–210. http://doi.wiley.com/10.1002/hbm.460020402. http://dx.doi.org/10.1002/hbm.460020402.
Gelman, A., 2008. Objections to Bayesian statistics. Bayesian Anal. 3, 445–450. http://ba.stat.cmu.edu/journal/2008/vol03/issue03/gelman.pdf. http://dx.doi.org/10.1214/08-BA318.
Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis. 3rd ed., Chapman and Hall/CRC, Boca Raton.
Glatard, T., Lewis, L.B., Ferreira da Silva, R., Adalat, R., Beck, N., Lepage, C., Rioux, P., Rousseau, M.E., Sherif, T., Deelman, E., Khalili-Mahani, N., Evans, A.C., 2015. Reproducibility of neuroimaging analyses across operating systems. Front. Neuroinform. 9, A12. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4408913/. http://dx.doi.org/10.3389/fninf.2015.00012.
Grinband, J., Savitskaya, J., Wager, T.D., Teichert, T., Ferrera, V.P., Hirsch, J., 2011. The dorsal medial frontal cortex is sensitive to time on task, not response conflict or error likelihood. NeuroImage 57, 303–311. http://www.sciencedirect.com/science/article/pii/S1053811910016101. http://dx.doi.org/10.1016/j.neuroimage.2010.12.027.
Grinband, J., Wager, T.D., Lindquist, M., Ferrera, V.P., Hirsch, J., 2008. Detection of time-varying signals in event-related fMRI designs. NeuroImage 43, 509–520. http://linkinghub.elsevier.com/retrieve/pii/S1053811908009075. http://dx.doi.org/10.1016/j.neuroimage.2008.07.065.
Guyon, X., Yao, J. f., 1999. On the underfitting and overfitting sets of models chosen by order selection criteria. J. Multivar. Anal. 70, 221–249. http://linkinghub.elsevier.com/retrieve/pii/S0047259X99918286. http://dx.doi.org/10.1006/jmva.1999.1828.
Hastie, T., Friedman, J., Tibshirani, R., 2001. Model assessment and selection. The Elements of Statistical Learning. Springer New York, New York, NY, pp. 193–224. http://link.springer.com/10.1007/978-0-387-21606-5_7.
Haxby, J.V., 2012. Multivariate pattern analysis of fMRI: the early beginnings. NeuroImage 62, 852–855.
Henson, R., Rugg, M.D., Friston, K.J., 2001. The choice of basis functions in event-related fMRI. NeuroImage 13, 149. http://www.fil.ion.ucl.ac.uk/spm/download/data/rfx-multiple/hbm-fir.pdf.
Henson, R.N.A., Shallice, T., Gorno-Tempini, M.L., Dolan, R.J., 2002. Face repetition effects in implicit and explicit memory tests as measured by fMRI. Cereb. Cortex 12, 178–186. http://www.cercor.oxfordjournals.org/cgi/doi/10.1093/cercor/12.2.178. http://dx.doi.org/10.1093/cercor/12.2.178.
Holmes, A., Friston, K., 1998. Generalisability, random effects & population inference. NeuroImage 7, S754.
Jeffreys, H., 1946. An invariant form for the prior probability in estimation problems. Proc. R. Soc. A Math. Phys. Eng. Sci. 186, 453–461. http://rspa.royalsocietypublishing.org/cgi/doi/10.1098/rspa.1946.0056. http://dx.doi.org/10.1098/rspa.1946.0056.
Josephs, O., Henson, R.N.A., 1999. Event-related functional magnetic resonance imaging: modelling, inference and optimization. Philos. Trans. R. Soc. B: Biol. Sci. 354, 1215–1228. http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.1999.0475. http://dx.doi.org/10.1098/rstb.1999.0475.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795. http://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572. http://dx.doi.org/10.1080/01621459.1995.10476572.
Kay, K.N., David, S.V., Prenger, R.J., Hansen, K.A., Gallant, J.L., 2008a. Modeling low-frequency fluctuation and hemodynamic response timecourse in event-related fMRI. Hum. Brain Mapp. 29, 142–156. http://doi.wiley.com/10.1002/hbm.20379. http://dx.doi.org/10.1002/hbm.20379.
Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L., 2008b. Identifying natural images from human brain activity. Nature 452, 352–355. http://www.nature.com/doifinder/10.1038/nature06713. http://dx.doi.org/10.1038/nature06713.
Kay, K.N., Rokem, A., Winawer, J., Dougherty, R.F., Wandell, B.A., 2013. GLMdenoise: a fast, automated technique for denoising task-based fMRI data. Front. Neurosci. 7, http://journal.frontiersin.org/article/10.3389/fnins.2013.00247/abstract. http://dx.doi.org/10.3389/fnins.2013.00247.
Kherif, F., Poline, J.B., Flandin, G., Benali, H., Simon, O., Dehaene, S., Worsley, K.J., 2002. Multivariate model specification for fMRI data. NeuroImage 16, 1068–1083. http://linkinghub.elsevier.com/retrieve/pii/S1053811902910943. http://dx.doi.org/10.1006/nimg.2002.1094.
Kiebel, S., Holmes, A., 2011. The general linear model. Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press., pp. 101–125.
Knutson, B., Adams, C.M., Fong, G.W., Hommer, D., 2001. Anticipation of increasing monetary reward selectively recruits nucleus accumbens. J. Neurosci. 21, RC159. http://www.jneurosci.org/content/21/16/RC159.
Koch, K.R., 2007. Introduction to Bayesian Statistics. 2nd ed., Springer, Berlin; New York.
Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2, A4.
Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22, 79–86. http://www.jstor.org/stable/2236703.
Leek, J.T., Peng, R.D., 2015. Statistics: *P* values are just the tip of the iceberg. Nature 520, http://www.nature.com/doifinder/10.1038/520612a. http://dx.doi.org/10.1038/520612a. 612–612.
Li, F., Villani, M., Kohn, R., 2010. Flexible modeling of conditional distributions using smooth mixtures of asymmetric student *t* densities. J. Stat. Plann. Infer. 140, 3638–3654. http://linkinghub.elsevier.com/retrieve/pii/S0378375810002119. http://dx.doi.org/10.1016/j.jspi.2010.04.031.
Lieberman, M.D., Cunningham, W.A., 2009. Type I and Type II error concerns in fMRI research: re-balancing the scale. Soc. Cogn. Affect. Neurosci. 4, 423–428. http://scan.oxfordjournals.org/cgi/doi/10.1093/scan/nsp052. http://dx.doi.org/10.1093/scan/nsp052.
Lindquist, M.A., Meng Loh, J., Atlas, L.Y., Wager, T.D., 2009. Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. NeuroImage 45, S187–S198. http://linkinghub.elsevier.com/retrieve/pii/S1053811908012056. http://dx.doi.org/10.1016/j.neuroimage.2008.10.065.
Loh, J.M., Lindquist, M.A., Wager, T.D., 2008. Residual analysis for detecting mis-modeling in fMRI. Stat. Sin. 18, 1421. http://www3.stat.sinica.edu.tw/statistica/password.asp?vol=18&num=4&art=10.
Lund, T.E., Madsen, K.H., Sidaros, K., Luo, W.L., Nichols, T.E., 2006. Non-white noise in fMRI: does modelling have an impact? NeuroImage 29, 54–66. http://linkinghub.elsevier.com/retrieve/pii/S105381190500501X. http://dx.doi.org/10.1016/j.neuroimage.2005.07.005.
Luo, W.L., Nichols, T.E., 2003. Diagnosis and exploration of massively univariate neuroimaging models. NeuroImage 19, 1014–1032. http://linkinghub.elsevier.com/retrieve/pii/S1053811903001496. http://dx.doi.org/10.1016/S1053-8119(03)00149-6.
MacKay, D.J.C., 2003. Information Theory, Inference and Learning Algorithms. Cambridge University Press, Cambridge; New York.
Mahmoud, H., 2008. Pólya Urn Models. 1st ed., Chapman and Hall/CRC, Boca Raton.
Monti, M., 2011. Statistical analysis of fMRI time-series: a critical review of the GLM approach. Front. Hum. Neurosci. 5, A28. http://journal.frontiersin.org/article/10.3389/fnhum.2011.00028/abstract. http://dx.doi.org/10.3389/fnhum.2011.00028.
Mumford, J.A., Poline, J.B., Poldrack, R.A., 2015. Orthogonalization of regressors in fMRI models. PLOS ONE 10, e0126255. http://dx.plos.org/10.1371/journal.pone.0126255. http://dx.doi.org/10.1371/journal.pone.0126255.
Murphy, K., Birn, R.M., Handwerker, D.A., Jones, T.B., Bandettini, P.A., 2009. The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? NeuroImage 44, 893–905. http://linkinghub.elsevier.com/retrieve/pii/S1053811908010264. http://dx.doi.org/10.1016/j.neuroimage.2008.09.036.
Oaksford, M., 2002. How does it fit? Trends Cogn. Sci. 6, 412–413. http://linkinghub.elsevier.com/retrieve/pii/S136466130201999X. http://dx.doi.org/10.1016/S1364-6613(02)01999-X.
Penny, W., 2012. Comparing dynamic causal models using AIC, BIC and free energy. NeuroImage 59, 319–330. http://linkinghub.elsevier.com/retrieve/pii/S1053811911008160. http://dx.doi.org/10.1016/j.neuroimage.2011.07.039.
Penny, W., Flandin, G., Trujillo-Barreto, N., 2007a. Bayesian comparison of spatially regularised general linear models. Hum. Brain Mapp. 28, 275–293. http://onlinelibrary.wiley.com/doi/10.1002/hbm.20327/abstract. http://dx.doi.org/10.1002/hbm.20327.
Penny, W., Kiebel, S., Friston, K., 2003. Variational Bayesian inference for fMRI time series. NeuroImage 19, 727–741. http://linkinghub.elsevier.com/retrieve/pii/S1053811903000715. http://dx.doi.org/10.1016/S1053-8119(03)00071-5.
Penny, W., Kilner, J., Blankenburg, F., 2007b. Robust Bayesian general linear models. NeuroImage 36, 661–671. http://linkinghub.elsevier.com/retrieve/pii/S1053811907000869. http://dx.doi.org/10.1016/j.neuroimage.2007.01.058.
Penny, W., Stephan, K., Mechelli, A., Friston, K., 2004. Comparing dynamic causal models. NeuroImage 22, 1157–1172. http://linkinghub.elsevier.com/retrieve/pii/S1053811904001648. http://dx.doi.org/10.1016/j.neuroimage.2004.03.026.
Penny, W.D., Ridgway, G.R., 2013. Efficient posterior probability mapping using Savage-Dickey ratios. PLoS ONE 8, e59655. http://dx.plos.org/10.1371/journal.pone.0059655. http://dx.doi.org/10.1371/journal.pone.0059655.
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, e1000709. http://dx.plos.org/10.1371/journal.pcbi.1000709. http://dx.doi.org/10.1371/journal.pcbi.1000709.

Penny, W.D., Trujillo-Barreto, N.J., Friston, K.J., 2005. Bayesian fMRI time series analysis with spatial priors. NeuroImage 24, 350–362. http://linkinghub.elsevier.com/retrieve/pii/S1053811904004975. http://dx.doi.org/10.1016/j.neuroimage.2004.08.034.

Pernet, C., Poline, J.B., 2015. Improving functional magnetic resonance imaging reproducibility. GigaScience 4, A15. http://www.gigasciencejournal.com/content/4/1/15. http://dx.doi.org/10.1186/s13742-015-0055-8.

Razavi, M., Grabowski, T.J., Vispoel, W.P., Monahan, P., Mehta, S., Eaton, B., Bolinger, L., 2003. Model assessment and model building in fMRI. Hum. Brain Mapp. 20, 227–238. http://doi.wiley.com/10.1002/hbm.10141. http://dx.doi.org/10.1002/hbm.10141.

Rigoux, L., Stephan, K., Friston, K., Daunizeau, J., 2014. Bayesian model selection for group studies – revisited. NeuroImage 84, 971–985. http://linkinghub.elsevier.com/retrieve/pii/S1053811913009300. http://dx.doi.org/10.1016/j.neuroimage.2013.08.065.

Rosa, M., Bestmann, S., Harrison, L., Penny, W., 2010. Bayesian model selection maps for group studies. NeuroImage 49, 217–224. http://linkinghub.elsevier.com/retrieve/pii/S105381190900963X. http://dx.doi.org/10.1016/j.neuroimage.2009.08.051.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464. http://projecteuclid.org/euclid.aos/1176344136. http://dx.doi.org/10.1214/aos/1176344136.

Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22, 1359–1366. http://pss.sagepub.com/lookup/doi/10.1177/0956797611417632. http://dx.doi.org/10.1177/0956797611417632.

Simonsohn, U., Nelson, L.D., Simmons, J.P., 2014. P-curve: a key to the file-drawer. J. Exp. Psychol. Gen. 143, 534.

Smith, S.M., Miller, K.L., Salimi-Khorshidi, G., Webster, M., Beckmann, C.F., Nichols, T.E., Ramsey, J.D., Woolrich, M.W., 2011. Network modelling methods for FMRI. NeuroImage 54, 875–891. http://linkinghub.elsevier.com/retrieve/pii/S1053811910011602. http://dx.doi.org/10.1016/j.neuroimage.2010.08.063.

Soch, J., Allefeld, C., Haynes, J.D., 2014. Solving the problem of overfitting in neuroimaging? Use of voxel-wise model comparison to test design parameters in first-level fMRI data analysis. F1000Research. http://f1000research.com/posters/1096034.

Soch, J., Allefeld, C., Haynes, J.D., 2015. Solving the problem of overfitting in neuroimaging? Cross-validated Bayesian model selection for methodological control in fMRI data analysis. F1000Research. http://dx.doi.org/10.7490/f1000research.1000161.1.

Steffener, J., Tabert, M., Reuben, A., Stern, Y., 2010. Investigating hemodynamic response variability at the group level using basis functions. NeuroImage 49, 2113–2122. http://linkinghub.elsevier.com/retrieve/pii/S1053811909011987. http://dx.doi.org/10.1016/j.neuroimage.2009.11.014.

Stephan, K., Penny, W., Moran, R., den Ouden, H., Daunizeau, J., Friston, K., 2010. Ten simple rules for dynamic causal modeling. NeuroImage 49, 3099–3109. http://linkinghub.elsevier.com/retrieve/pii/S1053811909011999. http://dx.doi.org/10.1016/j.neuroimage.2009.11.015.

Stephan, K.E., 2010. Methods & models for fMRI data analysis in neuroeconomics. http://www.socialbehavior.uzh.ch/teaching/methodsspring10.html.

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, 649–662. http://linkinghub.elsevier.com/retrieve/pii/S1053811908005983. http://dx.doi.org/10.1016/j.neuroimage.2008.04.262.

Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J., 2009. Bayesian model selection for group studies. NeuroImage 46, 1004–1017. http://linkinghub.elsevier.com/retrieve/pii/S1053811909002638. http://dx.doi.org/10.1016/j.neuroimage.2009.03.025.

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.B., Lebihan, D., Dehaene, S., 2006. Inverse retinotopy: inferring the visual content of images from brain activation patterns. NeuroImage 33, 1104–1116. http://linkinghub.elsevier.com/retrieve/pii/S1053811906007373. http://dx.doi.org/10.1016/j.neuroimage.2006.06.062.

Todd, M.T., Nystrom, L.E., Cohen, J.D., 2013. Confounds in multivariate pattern analysis: theory and rule representation case study. NeuroImage 77, 157–165. http://linkinghub.elsevier.com/retrieve/pii/S1053811913002887. http://dx.doi.org/10.1016/j.neuroimage.2013.03.039.

Triantafyllou, C., Hoge, R., Krueger, G., Wiggins, C., Potthast, A., Wiggins, G., Wald, L., 2005. Comparison of physiological noise at 1.5 T, 3 T and 7 T and optimization of fMRI acquisition parameters. NeuroImage 26, 243–250. http://linkinghub.elsevier.com/retrieve/pii/S1053811905000339. http://dx.doi.org/10.1016/j.neuroimage.2005.01.007.

Van Essen, D., Anderson, C., Felleman, D., 1992. Information processing in the primate visual system: an integrated systems perspective. Science 255, 419–423. http://www.sciencemag.org/cgi/doi/10.1126/science.1734518. http://dx.doi.org/10.1126/science.1734518.

Villani, M., Kohn, R., Giordani, P., 2009. Regression density estimation using smooth adaptive Gaussian mixtures. J. Econ. 153, 155–173. http://linkinghub.elsevier.com/retrieve/pii/S0304407609001419. http://dx.doi.org/10.1016/j.jeconom.2009.05.004.

Vul, E., Harris, C., Winkielman, P., Pashler, H., 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspect. Psychol. Sci. 4, 274–290. http://pps.sagepub.com/lookup/doi/10.1111/j.1745-6924.2009.01125.x. http://dx.doi.org/10.1111/j.1745-6924.2009.01125.x.

Woolgar, A., Golland, P., Bode, S., 2014. Coping with confounds in multivoxel pattern analysis: what should we do about reaction time differences? A comment on Todd, Nystrom & Cohen 2013. NeuroImage 98, 506–512. http://linkinghub.elsevier.com/retrieve/pii/S1053811914003395. http://dx.doi.org/10.1016/j.neuroimage.2014.04.059.

Yarkoni, T., Barch, D.M., Gray, J.R., Conturo, T.E., Braver, T.S., 2009. BOLD correlates of trial-by-trial reaction time variability in gray and white matter: a multi-study fMRI analysis. PLoS ONE 4, e4257. http://dx.plos.org/10.1371/journal.pone.0004257. http://dx.doi.org/10.1371/journal.pone.0004257.

Yeung, N., Cohen, J.D., Botvinick, M.M., 2011. Errors of interpretation and modeling: a reply to Grinband et al., NeuroImage 57, 316–319. http://linkinghub.elsevier.com/retrieve/pii/S1053811911004289. http://dx.doi.org/10.1016/j.neuroimage.2011.04.029.