How to improve parameter estimates in GLM-based fMRI data analysis: cross-validated Bayesian model averaging

— Preprint, submitted to *NeuroImage* on 18/11/2016 —

Joram Soch[a,f,•], Achim Meyer[a],
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

• Corresponding author: joram.soch@bccn-berlin.de.
BCCN Berlin, Philippstraße 13, Haus 6, 10115 Berlin, Germany.

Abstract

In functional magnetic resonance imaging (fMRI), model quality of general linear models (GLMs) for first-level analysis is rarely assessed. In recent work (Soch et al., 2016: "How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection", NeuroImage, vol. 141, pp. 469-489; DOI: 10.1016/j.neuroimage.2016.07.047), we have introduced cross-validated Bayesian model selection (cvBMS) to infer the best model for a group of subjects and use it to guide second-level analysis. While this is the optimal approach given that the same GLM has to be used for all subjects, there is a much more efficient procedure when model selection only addresses nuisance variables and regressors of interest are included in all candidate models. In this work, we propose cross-validated Bayesian model averaging (cvBMA) to improve parameter estimates for these regressors of interest by combining information from all models using their posterior probabilities. This is particularly useful as different models can lead to different conclusions regarding experimental effects and the most complex model is not necessarily the best choice. We find that cvBMS can prevent not detecting established effects and that cvBMA can be more sensitive to experimental effects than just using even the best model in each subject.

Keywords

fMRI-based neuroimaging, mass-univariate GLM, nuisance variables, correlated regressors, cross-validation, Bayesian model averaging.

1 Introduction

In *functional magnetic resonance imaging* (fMRI), data are most commonly analyzed using *general linear models* (GLMs) which construct a relation between psychologically defined conditions and the measured hemodynamic signal (Friston et al., 1994; Holmes and Friston, 1998). This allows to infer significant effects of cognitive states on brain activation, based on certain assumptions about the measured signal. As different GLMs can lead to different conclusions regarding experimental effects (Andrade et al., 1999; Carp, 2012), proper model assessment and model comparison is critical for statistically valid fMRI data analysis (Razavi et al., 2003; Monti, 2011).

In previous work, we have proposed *cross-validated Bayesian model selection* (cvBMS) to identify voxel-wise optimal models at the group level and then restrict group-level analysis to the best model in each voxel which avoids underfitting and overfitting in GLM-based fMRI data analysis (Soch et al., 2016). Importantly, cvBMS detects the model which is optimal in the majority of subjects, but not necessarily in all of them. Therefore, while this approach is powerful in a lot of cases and optimal in a decision-theoretic sense, it is not necessary in other cases and more appropriate alternatives exist.

The critical question here is whether *regressors of interest* are contained in all models to be compared. By "regressors of interest", we refer to those predictors whose estimates enter second-level analysis after first-level estimation. If these regressors are not part of all models in the model space, e.g. because the models differ by a categorical vs. parametric description of the experiment using completely different predictors (Bogler et al., 2013), one must employ the same model in all subjects in order to perform a sensible group analysis. In this case, cvBMS is the method of choice.

However, if regressors of interest are contained in all models, e.g. because the models only differ by *nuisance regressors* describing processes of no interest (Meyer and Haynes, in prep.), each model provides estimates for the parameters going into group analysis. In this case, cvBMS might unnecessarily lead one to use a model that is optimal in most subjects, but still sub-optimal in a lot of them. One could therefore speculate about performing second-level analysis on parameter estimates from different first-level models, depending on which model is optimal in each subject and voxel, which could be easily implemented using subject-wise selected-model maps (Soch et al., 2016).

In the present work, we take on this suggestion and motivate a model averaging approach (Hoeting et al., 1999), more precisely a form of *Bayesian model averaging* (BMA). In fMRI data analysis, BMA has been described for dynamic causal models (Penny et al., 2010), but not so far for general linear models (Penny et al., 2007). In BMA, estimates of the *same* parameter from *different* models are combined with the models' posterior probabilities (PP) and give rise to *averaged* parameter values which are more precise than *individual* models' estimates. Here, we calculate PPs from *out-of-sample log model evidences* (oosLME) and refer to this as *cross-validated Bayesian model averaging* (cvBMA).

The rest of this paper falls into three parts. In Section 2, we describe the mathematical details of cvBMA for GLMs in fMRI, resting on both classical and Bayesian inference for the GLM. In Section 3, we apply cvBMA to simulated data and show that it leads to parameter estimates which are closer to their true values than estimates from the best model in each simulation. In Section 4, we apply cvBMA to empirical data before we discuss our results. Again, we show that cvBMA can be more sensitive than just using even the best model in each subject. Moreover, we find that cvBMS can prevent not detecting established effects when using only one model.

2 Theory

2.1 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):

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

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 ist 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 autocorrelations across all voxels using a Restricted Maximum Likelihood (ReML) approach (Friston et al., 2002b, a). In contrast to that, X has to be set by the user. Especially if regressors are correlated with each other, there can be doubt about which model to use. The general linear model implicitly defines the following likelihood function:

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

GLMs are typically inverted by applying maximum likelihood (ML) estimation to equation (2). This leads to ordinary least squares (OLS) estimates (Bishop, 2007, eq. 3.15) if $V = I_n$, i.e. under temporal independence, or weighted least squares (WLS) estimates (Koch, 2007, eq. 4.29) if $V \neq I_n$, i.e. when errors ε are not assumed independent and identically distributed (i.i.d.).
Based on these ML estimates, statistical tests can be performed to investigate brain activity during different experimental conditions. These tests however strongly depend on the design matrix of the underlying model (Carp, 2012). When events overlap in time or are closely spaced temporally, convolution with the hemodynamic response function (Henson et al., 2001) will lead to positive correlation between the corresponding regressors. This influences parameter estimates for regressors of interest which in turn influences statistical tests and can change non-significant to significant or vice versa.
For mathematical convenience, we will rewrite the likelihood function as

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

In this equation, $P = V^{-1}$ is a $n \times n$ precision matrix and $\tau = 1/\sigma^2$ is the inverse residual variance (Koch, 2007, eq. 4.116). For Bayesian inference, it is advantageous to use the conjugate prior relative to equation (3). We have described this model, the general linear model with normal-gamma priors (GLM-NG), earlier and derived posterior distributions on the model parameters (Soch et al., 2016, eqs. 6) as well as the log model evidence for model comparison (Soch et al., 2016, eqs. 9).
In the following, we will introduce this model quality criterion (Section 2.2) and show how it can give rise to averaged model parameters (Section 2.3).

2.2 The log 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):

$$p(\theta|y,m) = \frac{p(y|\theta,m)\,p(\theta|m)}{p(y|m)} \tag{4}$$

Here, $p(y|\theta,m)$ is the likelihood function, $p(\theta|m)$ is the prior distribution and the posterior distribution $p(\theta|y,m)$ is given as the normalized product of likelihood and prior. The denominator $p(y|m)$ on the right-hand side acts as a normalization constant on the posterior density $p(\theta|y,m)$ and is given by (Gelman et al., 2013, eq. 1.3)

$$p(y|m) = \int p(y|\theta,m)\,p(\theta|m)\,\mathrm{d}\theta \tag{5}$$

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 (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. For the GLM-NG, we have derived the posterior distribution (Soch et al., 2016, eq. 6) and log model evidence (Soch et al., 2016, eq. 9) in earlier work on model selection for GLMs.

The LME is a reliable model selection criterion as it (i) automatically penalizes for additional model parameters by integrating them out of the likelihood (Penny, 2012), (ii) can be naturally decomposed into model accuracy and model complexity (Penny et al., 2007) and (iii) accounts for the whole uncertainty about parameter estimates (Gelman et al., 2013) instead of using point estimates like classical information criteria such as AIC (Akaike, 1974) and BIC (Schwarz, 1978).

The LME however also requires prior distributions on the model parameters and typically diverges with ML-style flat priors. As multi-session fMRI data provides a natural basis for the application of cross-validation (CV), we therefore suggested to use the LME in conjunction with CV (Soch et al., 2015) in order to avoid the necessity to specify prior distributions which are usually hard to come up with in fMRI research.

In this procedure, a posterior distribution is estimated from all sessions $j \neq i$ using non-informative prior distributions and then used as an informative prior distribution on the remaining session i to calculate the LME for this session (Soch et al., 2016). This is referred to as the out-of-sample log model evidence (oosLME):

$$\log p(y_i|m) = \log \int p(y_i|\theta,m)\,p(\theta|\cup_{j\neq i}\ y_j,m)\,\mathrm{d}\theta \tag{6}$$

Here, we will use these oosLMEs for session-wise parameter inference, but one can also calculate a cross-validated log model evidence (cvLME) as

$$\log p(y|m) = \sum_{i=1}^{S} \log p(y_i|m) \tag{7}$$

The cvLME has been validated using extensive simulations (Soch et al., 2016) and is insensitive to the number of folds into which the data are partitioned for cross-validation (Soch and Allefeld, in prep.).

2.3 Bayesian model averaging

As log model evidences (LME) represent conditional probabilities, they can be used to calculate posterior probabilities (PP) using Bayes' theorem (Penny et al., 2010, eq. 7):

$$p(m_i|y) = \frac{p(y|m_i)\,p(m_i)}{\sum_{j=1}^{M} p(y|m_j)\,p(m_j)} \tag{8}$$

Here, M is the number of models, $p(y|m_i)$ is the i-th model evidence where the exponentiated oosLME has to be plugged in and $p(m_i)$ is the i-th prior probability which is usually set to $p(m_i) = 1/M$ making all models equally likely *a priori*. In this latter case, posterior probabilities are obtained as normalized exponentiated log model evidences. Conceptually, this approach uses the first-level model evidences as the second-level likelihood function to make probabilistic statements about the model space.
After model assessment using LMEs, the PPs can be used to calculate averaged parameter estimates by performing Bayesian model averaging (BMA) (Penny et al., 2010, eq. 27):

$$\hat{\beta}_{\mathrm{BMA}} = \sum_{i=1}^{M} \hat{\beta}_i \cdot p(m_i|y) \tag{9}$$

Here, $\hat{\beta}_i$ is the i-th model's parameter estimate for a specific regressor and $p(m_i|y)$ is the i-th model's posterior probability. Formally, BMA estimates can be seen as the parameter estimates of a larger model in which the variable "model" is marginalized out using the law of marginal probability. Note that, when one model is highly favored by the LME with a PP close to one, BMA is equivalent to just selecting this model's parameter estimate. However, BMA automatically generalizes to cases where LMEs are less clear.
Please also note that all these analyses are session-wise and therefore the oosLMEs are used. In the case of single-session fMRI data, we suggest to use split-half cross-validation (Soch et al., 2014) and use the cvLME for model averaging, as there is only one estimate for each regressor in each voxel in that case.
Critically, the regressor or regressors for which the averaged parameter values are calculated must be contained in all models of the model space. For this reason, BMA is particularly interesting when having identical regressors of interest, but varying regressors of no interest potentially correlated to the regressors of interest, which is often the case in fMRI data analysis due to different possible nuisance variables (see Figure 1A). We will investigate such cases in simulation settings (see Section 3) as well as with empirical data (see Section 4).
For the application of BMA to GLMs for fMRI, we use voxel-wise maximum-likelihood parameter estimates from SPM's 1st level analysis (Ashburner et al., 2013) and voxel-wise out-of-sample log model evidences as described earlier (Soch et al., 2016) (see Figure 1B). Finally, voxel-wise BMA estimates are calculated according to equation (9) (see Figure 1C). This is referred to as cross-validated Bayesian model averaging (cvBMA).

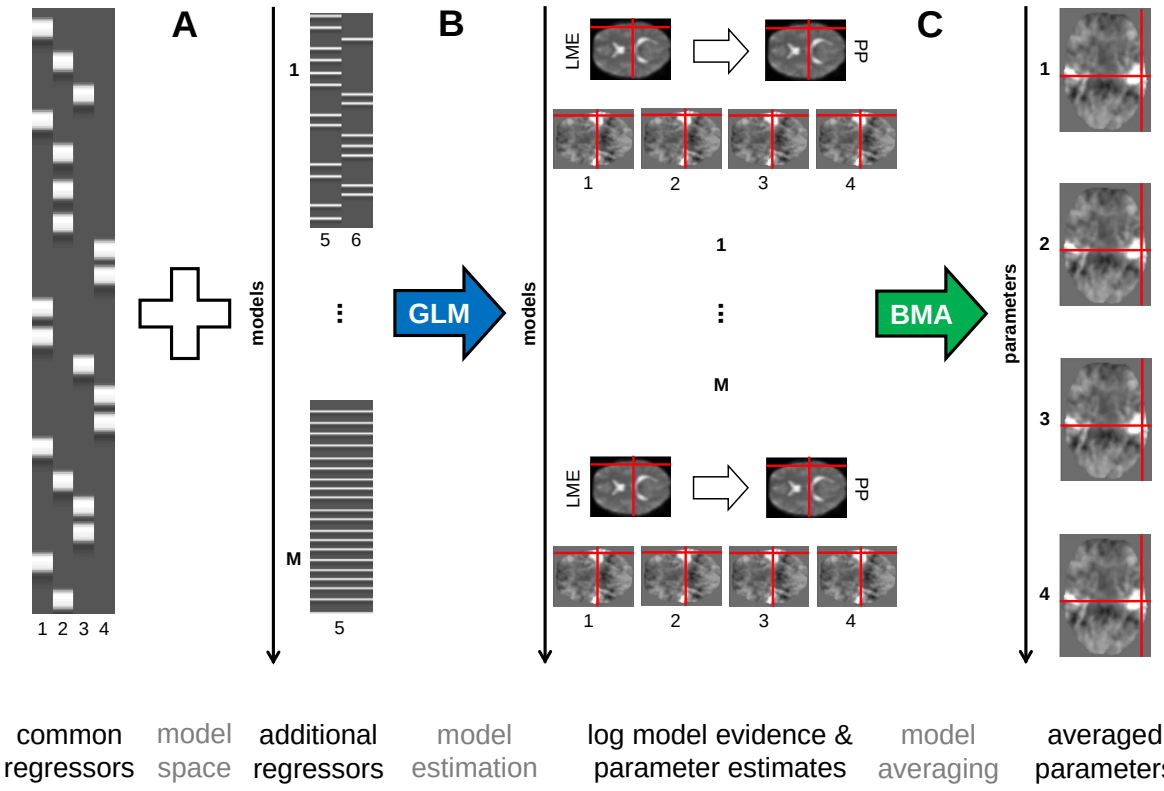

Figure 1. Model averaging for general linear models in fMRI data analysis. This figure summarizes our approach of cross-validated Bayesian model averaging (cvBMA). All calculations are performed voxel-wise, an exemplary voxel is highlighted using red crosshairs. (A) A model space is constructed by adding varying additional regressors (e.g. describing cues and feedback) to a set of common regressors (e.g. describing targets) which are included in all models. (B) Model estimation proceeds by classical GLM inversion, resulting in maximum likelihood (ML) parameter estimates (1-4), and by Bayesian GLM inversion with cross-validation (CV), resulting in maps of out-of-sample log model evidences (LME) from which posterior probabilities (PP) can be calculated. (C) Model averaging proceeds by weighting different models' estimates for the same regressor with the corresponding PP to obtain BMA estimates on which second-level inference can be performed. Parts of this figure are adapted from SPM course material (Stephan, 2010).

3 Simulation

3.1 Methods

We test the cvBMA approach using simulated data and specifically investigate the impact of regressor correlation on various parameter estimation methods.
To this end, we imagine three different regressors: a "target" regressor specifying event onsets for a condition of interest (x_1), a "cue" regressor with event onsets before targets (x_2) and a "feedback" regressor with event onsets after targets (x_3). Critically, these binary indicator regressors have close temporal proximity leading to non-orthogonality, with their overlap being modulated from 0 to 4 volumes (see Figure 2A).
Based on these regressors, we define four different models: one consisting of only the target regressor (m_1), one having the target regressor with either the cue regressor (m_2) or the feedback regressor (m_3) and one containing all three regressors (m_4). In the full model m_4, this leads to covariation of targets x_1 with cues x_2 and feedbacks x_3 where correlation increases and angle decreases with overlap (see Figure 2A).
For each level of overlap, $N_{\text{sim}} = 10{,}000$ simulations are performed as follows. First, a true model is randomly drawn from $M = \{m_1, m_2, m_3, m_4\}$. For each model and overlap, design matrices X for $S = 5$ sessions were generated before simulation. Each session consisted of $n = 200$ data points containing 9 trials with a duration of 6 scans and the respective overlap of target with cue and feedback.
Second, true regression coefficients are drawn using the relation

$$\begin{aligned} \beta_{ij} &= x_j + y_{ij} \\ x_j &\sim \mathrm{N}(0, \sigma_v^2) \\ y_{ij} &\sim \mathrm{N}(0, \sigma_s^2) \end{aligned}$$

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.
Third, simulated data are generated by sampling zero-mean Gaussian observation noise $\varepsilon \sim \mathrm{N}(0, \sigma^2 I_n)$ and then adding the random noise to the true signal to get a measured signal $y = X\beta + \varepsilon$ where the 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).
In our simulations, 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 the SPM template data set[1]) while between-session variance σ_s^2 is set to 1.25 implying a ratio of within-session variance to between-session variance of 2 (Soch et al., 2016).
Finally, as the target regressor was included in all design matrices, the model parameter corresponding to target presentation (β_1) was estimated using all four models and models were quantified using the out-of-sample log model evidence (oosLME). Then, we compared three parameter estimates for β_1: the one obtained using the best model (maximal cvLME), using Bayesian model averaging (BMA estimates) and using the true model (used to generate the data).

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

3.2 Results

The impact of covariation between regressors on parameter estimates in the general linear model (GLM) using ordinary least squares (OLS) is best captured using the inner product of these regressors. The normalized inner product of two vectors is equal to the cosine of their angle. With zero overlap, regressors are orthogonal, i.e. their angle is 90°. With an overlap of four data points, we observed an angle of 48.2° in our simulations (see Figure 2A), implying a certain degree of non-orthogonality, but not collinearity between target and cue or feedback regressors.

The precision of parameter estimates can be described by the mean squared error (MSE), i.e. squared differences between true and estimated parameter values, averaged across simulations. We were interested in MSE(β_1), because the parameter estimate for the target regressor was able to be confounded by the cue and feedback regressors, depending on whether they were part of the true model or not.

First and trivially, in the case of zero overlap, the MSE for this parameter is the same for all models compared, because additional regressors do not change parameter estimates if they are orthogonal to the regressor of interest (see Figure 2B, left panel). Second and also not surprisingly, when there is overlap between regressors, the MSE is smallest when the true model is used (see Figure 2B, blue bars), because the model used to generate the data leads to the most precise parameter estimates. Third and most importantly, while the true model outperforms BMA estimates, BMA estimates outperform the best model (see Figure 2B, green/red), because BMA accounts for the uncertainty over models and does not just select from the model with maximal LME.

This demonstrates that Bayesian model averaging, i.e. weighting parameter estimates according to the models' posterior probabilities, can be better than using the best model, i.e. taking parameter estimates from the model with maximal posterior probability. Although BMA is worse than using the true model, it is the optimal approach for empirical data, because the true model is usually unknown in such cases. These simulations are therefore a first indication for employing BMA in first-level fMRI data analysis. A second indication will be provided by the analysis of empirical data.

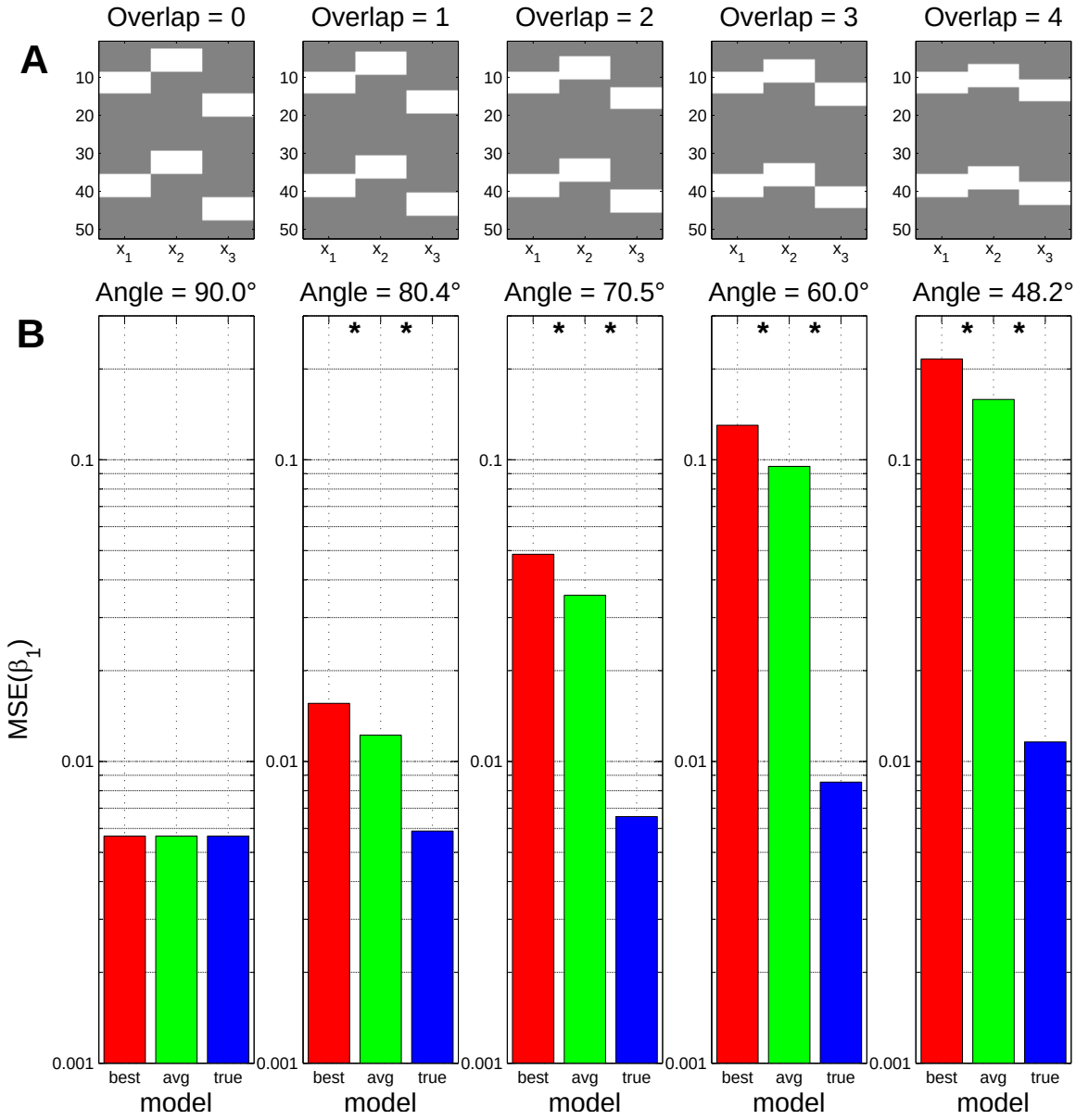

Figure 2. Simulation performance of cross-validated Bayesian model averaging. This figure demonstrates that Bayesian model averaging (BMA) can be superior to always using parameters from the best model as identified by the maximal log model evidence (LME). (A) Different degrees of overlap between a target regressor (x_1) and preceding cues (x_2) as well as subsequent feedback (x_3) regressor are simulated where an overlap given in volumes implies a certain angle between x_1 and x_2 as well as x_1 and x_3 in degrees. (B) Across all simulations, the mean-squared error (MSE) for the target regressor was calculated when using the best model (maximal cvLME, red), the averaged model (BMA estimates, green) or the true model (used to generate the data, blue) which is unknown for empirical data, but known in simulation settings. The higher the overlap gets, the higher the MSE becomes for all three estimation techniques (note that the y-axes are log scales). However, BMA outperforms the best model and the true model outperforms BMA at all overlaps greater than zero. Except from the left-most panel, all differences are significant at a significance level of $p \leq 0.05$.

4 Application

4.1 Methods

We test the cvBMA approach using empirical data from a conflict adaptation paradigm (Meyer and Haynes, in prep.) and specifically investigate the capability of model averaging to identify experimental effects that would be undetectable by individual models.
The experimental paradigm (see Figure 3) was an Eriksen flanker task (Eriksen and Eriksen, 1974) combined with a response rule switch (Bode and Haynes, 2009) giving a 2×2 factorial design, the two factors being conflict (congruent vs. incongruent) and task set (response rule 1 vs. 2). Arrows were arranged vertically on the screen. Difference in conflict was operationalized via congruent or incongruent flanking arrows. Subjects were requested to indicate the direction of the target arrow via right-hand button press. Task set difference was operationalized by a response button switch (Rule 1: up $\rightarrow$ left & down $\rightarrow$ right; Rule 2: up $\rightarrow$ right & down $\rightarrow$ left).
Stimuli were presented in 6 sessions with 20 blocks of 10 trials. Each trial consisted of 1 second presentation and 1 second fixation. Each block lasted $10 \times 2 = 20$ seconds, was preceded by a 500 ms auditory cue signalling the response rule as well as a 5.5 second delay phase to prepare for the coming block and was succeded by a 1 second feedback phase indicating how many blocks would still follow. The feedback phase occurred between two blocks and separated two intervals with durations exponentially distributed around 3 seconds. Thus, the total duration of one block was around 33 seconds. In total, there were 5 blocks per condition in each fMRI session lasting 666 seconds.
fMRI data were preprocessed using SPM12, Revision 6225 per 01/10/2014 (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 $6 \times 6 \times 6$ mm.
For first-level analysis, we categorized each block as containing congruent or incongruent stimuli and by whether the response rule switched or stayed relative to the preceding block.[2] This lead to four categories of blocks: congruent-stay, congruent-switch, incongruent-stay, incongruent-switch. For each subject and each session, a GLM was specified including four regressors modelling these four types of blocks and two regressors modelling delay phases for switch blocks and for stay blocks.
In second-level analysis, we were interested in the different neural activity in the switch-delays preceding the application of a new response rule and the stay-delays preceding the application of the same response rule as before. To again induce correlation between regressors, we introduce two variable model space features with regressors overlapping with the delay phase regressors and therefore influencing their estimates.
First, the 500 ms auditory cues before the 5.5 second delay phases were modelled by two extra regressors, also separated by the switch-stay difference. This model feature was motivated by the fact these stimulations indeed require two different cognitive processes, namely auditory perception for the cues and executive planning for the delays.
Second, the first trials of stimulation blocks were additionally modelled by four extra regressors, separated exactly like the block regressors. This model feature was motivated by the fact that the first trial of each block could demand a restart cost which was also

[2]The first block was categorized as a switch block, because a new rule had to be applied.

observable in the distribution of reaction times, i.e. a significantly higher reaction time in the first trial compared to later trials (Meyer and Haynes, in prep.).
Taken together, this resulted in four possible models for first-level data (see Table 2): all of them with blocks and delays being modelled; one with only cue phases being additionally modelled, one with only first trials being additionally modelled, one with both and one without both. Across all subjects and sessions, average correlation between delays and cue phases was 0.63 and average correlation between delays and first trials was 0.46. Like in our simulation study, the goal was to investigate the properties of statistical inference being performed with individual models, using the best model as identified by maximal cvLME and using model averaging based on the oosLMEs.

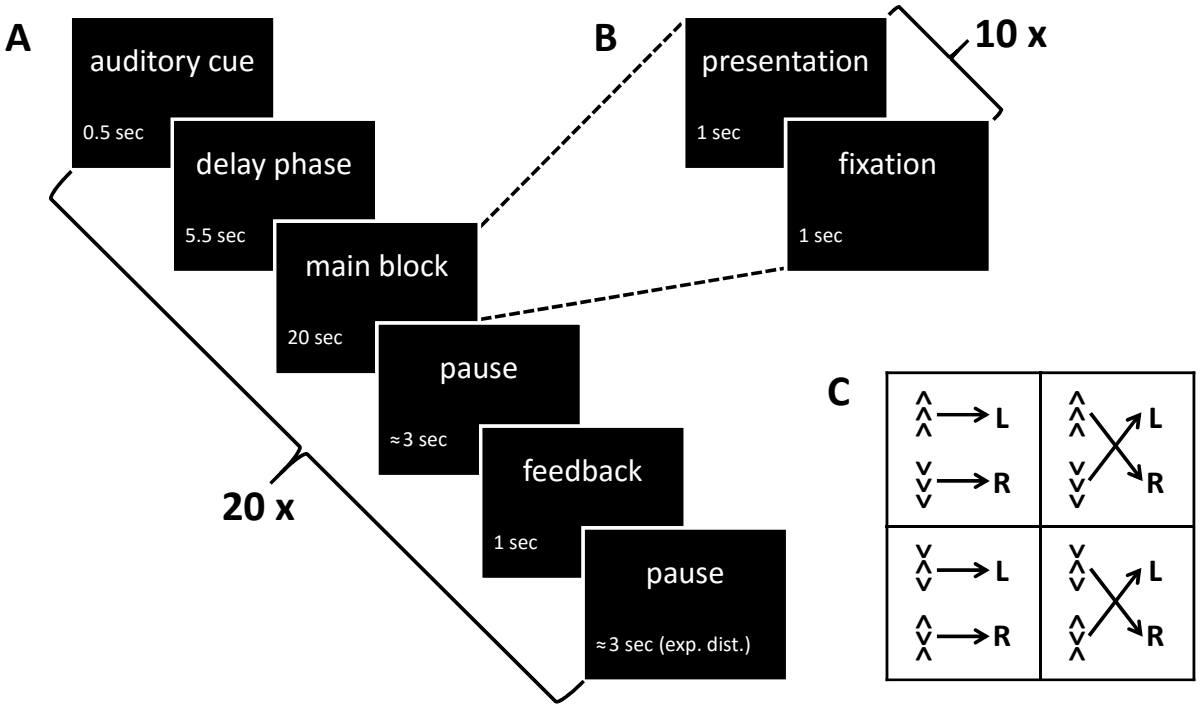

Figure 3. Experimental design of the conflict adaptation paradigm. This figure describes the experiment underlying the data set used for empirical validation of our method. (A) Sequence of events and exact timing during each of the 20 blocks per session. (B) Sequence of events and exact timing during one of the 10 trials per block. (C) Experimental conditions: The paradigm was a 2×2 design with conflict (congruent vs. incongruent) and task set (response rule 1 vs. 2) being the two factors. Abbreviations: sec = seconds; exp. dist. = exponentially distributed; L/R = left/right button.

4.2 Results

On the second level, we focused on the delay phases and looked for a main effect of stay vs. switch blocks. We hypothesized that delay phases after cues indicating a switch of the response rule might elicit preparatory processes that lead to higher motor cortex activity in the left hemisphere (participants responded with their right hand) when being compared to delay phases preceding blocks with the same response rule as before (Brass and Cramon, 2002). In fact, this main effect later turned out to be a positive effect of switch over stay blocks (see Table 1, right-hand side).

First, we tried to identify this effect using the four first-level models as such. We performed second-level analysis using the summary-statistic approach (Holmes and Friston, 1998) and were able to detect a main effect of stay vs. switch in left primary motor cortex using all models except the one modelling both, cue phases and first trials (see Table 1, upper section). The effect was not significant at the cluster level under correction for family-wise errors (FWE) when using the model with cue phases but without first trials, indicating that modelling the cue phase had a higher impact on parameter estimates for the delay phase due to their shared variance and higher correlation, making the difference between stay and switch blocks insignificant.

Next, we performed cross-validated Bayesian model selection (cvBMS) to identify the group-level optimal model in each voxel (Soch et al., 2015) and observed that all models are optimal in at least some voxels of the left precentral gyrus (see Table 2). We used this information to generate selected-model maps (SMM) indicating for each model in which voxels it is optimal and masked second-level analyses using these SMMs in order to restrict statistical inferences to those voxels where the corresponding model is best explaining the data at the group level. This approach, as suggested in previous work (Soch et al., 2016), lead to the effect only being detected by the models not accounting for cue phases (see Table 1, middle section), again suggesting that modelling them had the greater influence on delay phase significance. This demonstrates that cvBMS can prevent us from overfitting and not detecting established effects when using just one model. Notably, the most complex model including both, cue phases and first trials, would not have been the best choice here.

Last, we compared two estimation methods not being based on individual models: using parameter estimates from the subject-level optimal models as identified by maximal cvLME and using cross-validated Bayesian model averaging (cvBMA) as developed in the present work. For both methods, voxel-wise out-of-sample log model evidences (oosLME) were calculated for each model in each subject. For the best-model approach, first-level parameter estimates in each voxel were taken from the model having the highest cvLME in this voxel and then subjected to second-level analysis. For the model averaging approach, first-level parameter estimates in each voxel were weighted according to posterior probabilities calculated from the models' oosLMEs in this voxel (see Figure 1) and the averaged parameters were subjected to second-level analysis. The best-model approach is equivalent to setting the best model's posterior probability to one and then performing Bayesian model averaging. We observe that the effect in question can be identified using both methods, but is only FWE-significant at the cluster level when using cvBMA and not when using the subject-wise best model. Like in the individual model including first trials and not including cue phases, the main effect of stay vs. switch blocks in left pri-

mary motor cortex was also the global maximum on the respective contrast (see Table 1, lower section). Interestingly, cvBMA does not only calculate the weighted average of the models' parameter estimates, but also seems to make a compromise regarding the spatial location of the main effect when compared to statistical inferences based on the individual models (see Figure 4).

Like our simulation study, this empirical example therefore indicates that performing cvBMA can be superior to using parameter estimates from the subject-wise best models. Using cvBMA, we can improve parameter estimates for regressors of interest by drawing information from a variety of models instead of just relying on one particular model. This is possible, because regressors of interest are included in all models, so that each model provides parameter estimates for them and because the models differ in how well they are supported by the measured data, as quantified by their posterior probabilties. As demonstrated in simulation, by factoring in our uncertainty in this way, parameter estimates move closer to their true values which in turn increases the sensitivity for experimental effects.

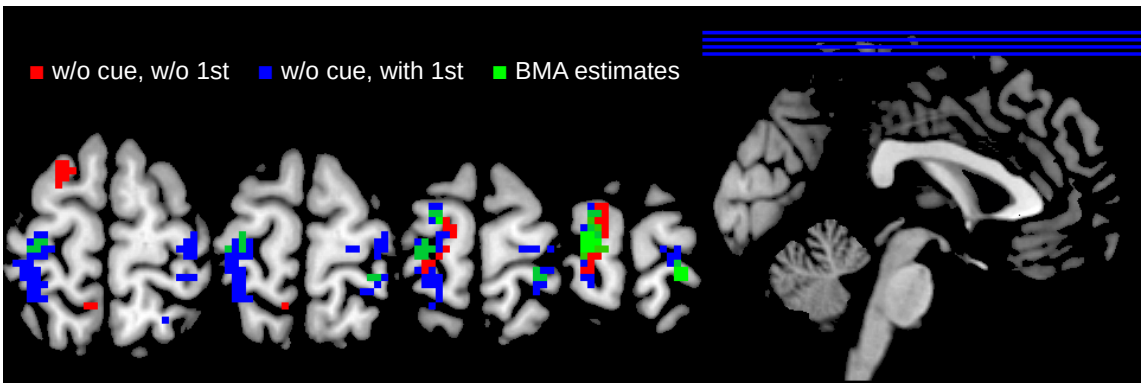

Figure 4. Empirical example for cross-validated Bayesian model averaging. This figure illustrates how model averaging achieves a compromise between the different models' parameter estimates. Colored voxels indicate a significant main effect of stay vs. switch blocks observed when using a model without cue phases and first trials (red), a model without cue phases but with first trials (blue), both masked with their group-level selected-model map (SMM) obtained through cross-validated Bayesian model selection (cvBMS), as well as observed in a second-level analysis on the averaged parameters (green) obatined via Bayesian model averaging (BMA). The BMA seems to not only average parameter values, but to also make a compromise regarding spatial location with respect to the blue (more ventral-posterior) and red (more dorsal-anterior) clusters.

First-level model used for second-level analysis	Main effect "stay ≠ switch"							Negative effect "stay < switch"						
	peak-level				cluster-level		set-level	peak-level				cluster-level		set-level
	x	y	z	F	k	p_{FWE}	GM?	x	y	z	t	k	p_{FWE}	GM?
Second-level analyses with unconstrained whole-brain statistical inference (unc., p ≤ 0.001, k = 10)														
w/o cue phase, w/o 1st trials	-15	-13	77	45.98	106	<0.001	no	-15	-13	77	6.78	180	<0.001	no
w/o cue phase, with 1st trials	-15	-13	77	43.60	175	<0.001	no	-15	-13	77	6.60	239	<0.001	no
with cue phase, w/o 1st trials	-18	-25	77	20.76	14	0.401	no	-18	-25	77	4.56	21	0.237	no
with cue phase, with 1st trials	no significant effect in left primary motor cortex (AAL 001)							no significant effect in left primary motor cortex (AAL 001)						
Second-level analyses masked with selected-model maps (cross-validated Bayesian model selection, cvBMS)														
w/o cue phase, w/o 1st trials	-15	-13	77	45.98	13	0.513	no	-15	-13	77	6.78	15	0.501	no
	-18	-28	77	36.62	16	0.367	no	-18	-28	77	6.05	18	0.381	no
	-18	8	68	23.72	10	0.691	no	-18	8	68	4.87	14	0.547	no
w/o cue phase, with 1st trials	-30	-40	71	42.87	100	<0.001	yes	-30	-40	71	6.55	135	<0.001	yes
								-42	-16	53	4.84	14	0.533	no
								-24	-55	59	3.86	10	0.739	no
with cue phase, w/o 1st trials	no significant effect in left primary motor cortex (AAL 001)							no significant effect in left primary motor cortex (AAL 001)						
with cue phase, with 1st trials	no significant effect in left primary motor cortex (AAL 001)							no significant effect in left primary motor cortex (AAL 001)						
Second-level analyses based on averaged parameters (cross-validated Bayesian model averaging, cvBMA)														
averaged model parameter estimates	-18	-22	77	29.08	35	0.015	yes	-18	-22	77	5.39	50	0.007	yes
best model's parameter estimates	-15	-22	77	34.93	10	0.316	yes	-15	-22	77	5.91	23	0.033	yes

Table 1. Empirical example for cross-validated Bayesian model averaging. In each row, peak-level, cluster-level and set-level statistics are given for an F-test of the main effect between stay and switch blocks as well as a t-test of the negative effect of stay against switch blocks. The upper section of the table summarizes unconstrained whole-brain statistical inference using the four models. The effect in question can be detected using three models, but not the most complex one. The middle section of the table summarizes second-level analyses that were masked using selected-model maps (SMM) from cross-validated Bayesian model selection (cvBMS). The effect in question can be detected using the two models that do not include cue phase regressors. The lower section of the table summarizes second-level analysis based on averaged parameters (BMA estimates) and the subject-wise best model's parameter estimates (maximal cvLME). The effect in question is cluster-level significant using BMA, but not with the approach using only the best model. Abbreviations: x, y, z = MNI coordinates; F/t = F-/t-statistic, k = cluster size; p_{FWE} = family-wise error-corrected p-value; GM = global maximum.

First-level model	whole-brain	precentral gyrus	
		left	right
number of voxels	46334	855	792
Number of voxels on selected-model maps from cvBMS			
w/o cue phase, w/o 1st trials	23548	140	258
w/o cue phase, with 1st trials	9609	252	328
with cue phase, w/o 1st trials	1510	14	0
with cue phase, with 1st trials	11667	449	206

Table 2. Selected models from cross-validated Bayesian model selection. This table summarizes model selection results for the four models used in the empirical example (see Figures 3 and 4). For each model, the number of voxels in which it is selected as the optimal model by cross-validated Bayesian model selection (cvBMS) is given for the left and right precentral gyrus, taken from the Automated Anatomical Labeling (AAL) atlas (AAL 001 and AAL 002) and putatively including the primary motor cortices, as well as at the whole-brain level.

5 Discussion

We have introduced a model averaging approach for optimizing parameter estimates when analyzing *functional magnetic resonance imaging* (fMRI) data using *general linear models* (GLMs). We have demonstrated that *cross-validated Bayesian model averaging* (cvBMA) serves its intended purpose and that it is useful in practice. As nuisance variables and correlated regressors are common topics in fMRI data analysis, usage of this technique reduces model misspecification and thereby enhances the methodological quality of functional neuroimaging studies (Friston, 2009).

Often, psychological paradigms combined with fMRI use trials or blocks with multiple phases (e.g. cue – delay – target – feedback, see Meyer and Haynes, in prep.), so that the basic model setup (the target regressors) is fixed, but there is uncertainty about which processes of no interest (cues, delays, feedback) should be included into the model (Andrade et al., 1999). Especially in, but not restricted to these cases of correlated regressors (Mumford et al., 2015), cvBMA has its greatest potential which is why our simulated data and the empirical examples were constructed like this.

Typically, if one is unsure about the optimal analysis approach in such a situation, just one model is estimated or, even worse, a lot of models are estimated and model selection is made by looking at significant effects (Soch et al., 2016). Here, model averaging provides a simple way to avoid such biases. It encourages multiple model estimation in order to avoid mismodelling, but calculates weighted parameter estimates by combining the models in order to avoid subjective model selection.

Using simulated data, we were able to show that averaged model parameter estimates have a smaller mean squared error than even the best model's parameter estimates. Using empirical data, we demonstrated the trivial fact that different GLMs can lead to the same effect being either significant or insignificant. Interestingly, we found that the most complex GLM is not always the best, speaking against the fMRI practitioner's maxim that the design matrix should "embody all available knowledge about experimentally controlled factors and potential confounds" (Stephan, 2010), though it should still be applied in the absence of any knowledge about model quality.

Although the most complex model was not optimal in this case, our previously suggested approach of *cross-validated Bayesian model selection* (cvBMS) and subsequent masking of second-level analyses with selected-model maps (SMM) was able to protect against not detecting an established experimental effect which additionally validates this technique (Soch et al., 2016). Moreover, *cross-validated Bayesian model averaging* (cvBMA) was found to be more sensitive to experimental effects than simply extracting parameter estimates from the best model in each subject which again highlights its applicability in situations of uncertainty about modelling processes of no interest.

All in all, we therefore see cvBMA as a complement to the recently developed cvBMS. While cvBMS is the optimal approach when parameters of interest are not identical across the model space, e.g. because one part of the models uses a categorical and another part uses a parametric description of the paradigm (Bogler et al., 2013), cvBMA is the more appropriate analysis when regressors of interest are the same in all models (Meyer and Haynes, in prep.), such that their estimates can be averaged across models and taken to second-level analysis for sensible population inference.

6 Statement of Transparency

For an analytical technique that is directed against mismodelling, overfitting, p-hacking and fishing expeditions, it is important that the validation of this technique itself is not based on fishy analyses. We therefore briefly describe which exploratory analyses have led to the results reported in this work.
For simulation validation, just one simulation was performed and it was planned in exactly the way it is reported in this work – in part, because it is based on a simulation already used in earlier work (Soch et al., 2016).
For empirical validation, three model spaces were investigated. In the first model space, nuisance regressors were not correlated to regressors of interest, so that model averaging did not outperform statistical inference based on the individual models, similar to our simulation in which the regressor overlap was zero (see Figure 2B, left panel). In the second model space, nuisance regressors were correlated to regressors of interest, but statistical inference addressed a difference – effect of response rule, not stay vs. switch blocks – that putatively only elicits multivariate effects in fMRI (Bode and Haynes, 2009). The third model space is the one reported in this work. We believe that each model space was motivated by lessons learned from the previous one and we want to emphasize that the discovery-style exploratory analysis that we performed led us to establish in which context our method should actually be applied.

7 Acknowledgements

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

8 Software Note

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

References

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

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. URL: http://www.sciencedirect.com/science/article/pii/S1053811999904792, doi:10.1006/nimg.1999.0479.

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

Bishop, C.M., 2007. Pattern Recognition and Machine Learning. 1st ed. 2006. corr. 2nd printing 2011 ed., Springer, New York.

Bode, S., Haynes, J.D., 2009. Decoding sequential stages of task preparation in the human brain. NeuroImage 45, 606–613. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811908012226, doi:10.1016/j.neuroimage.2008.11.031.

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

Brass, M., Cramon, D.Y., 2002. The Role of the Frontal Cortex in Task Preparation. Cerebral Cortex 12, 908–914. URL: http://www.cercor.oupjournals.org/cgi/doi/10.1093/cercor/12.9.908, doi:10.1093/cercor/12.9.908.

Carp, J., 2012. On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments. Frontiers in Neuroscience 6. URL: http://journal.frontiersin.org/article/10.3389/fnins.2012.00149/abstract, doi:10.3389/fnins.2012.00149.

Eriksen, B.A., Eriksen, C.W., 1974. Effects of noise letters upon the identification of a target letter in a nonsearch task. Perception & Psychophysics 16, 143–149. URL: http://www.springerlink.com/index/10.3758/BF03203267, doi:10.3758/BF03203267.

Friston, K., Glaser, D., Henson, R., Kiebel, S., Phillips, C., Ashburner, J., 2002a. Classical and Bayesian Inference in Neuroimaging: Applications. NeuroImage 16, 484–512. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811902910918, doi:10.1006/nimg.2002.1091.

Friston, K., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J., 2002b. Classical and Bayesian Inference in Neuroimaging: Theory. NeuroImage 16, 465–483. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811902910906, doi:10.1006/nimg.2002.1090.

Friston, K.J., 2009. Modalities, Modes, and Models in Functional Neuroimaging. Science 326, 399–403. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1174521, doi: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. Human Brain Mapping 2, 189–210. URL: http://doi.wiley.com/10.1002/hbm.460020402, doi:10.1002/hbm.460020402.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis. 3rd edition ed., Chapman and Hall/CRC, Boca Raton.

Henson, R., Rugg, M.D., Friston, K.J., 2001. The choice of basis functions in event-related fMRI. NeuroImage 13, 149–149. URL: http://www.fil.ion.ucl.ac.uk/spm/data/face_rfx/pdf/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. Cerebral Cortex 12, 178–186. URL: http://www.cercor.oxfordjournals.org/cgi/doi/10.1093/cercor/12.2.178, doi:10.1093/cercor/12.2.178.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian Model Averaging: A Tutorial. Statistical Science 14, 382–401. URL: http://www.jstor.org/stable/2676803.

Holmes, A., Friston, K., 1998. Generalisability, random effects & population inference. NeuroImage 7, S754.

Kiebel, S., Holmes, A., 2011. The General Linear Model, in: Statistical Parametric Mapping: The Analysis of Functional Brain Images: The Analysis of Functional Brain Images. Academic Press, pp. 101–125.

Koch, K.R., 2007. Introduction to Bayesian Statistics. 2nd, updated and enlarged ed. 2007 edition ed., Springer, Berlin ; New York.

Meyer, A., Haynes, J.D., in prep. Decoding behavioral adaptation under stimulus conflict .

Monti, M., 2011. Statistical Analysis of fMRI Time-Series: A Critical Review of the GLM Approach. Frontiers in Human Neuroscience 5. URL: http://journal.frontiersin.org/article/10.3389/fnhum.2011.00028/abstract, doi: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. URL: http://dx.plos.org/10.1371/journal.pone.0126255, doi:10.1371/journal.pone.0126255.

Penny, W., 2012. Comparing Dynamic Causal Models using AIC, BIC and Free Energy. NeuroImage 59, 319–330. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811911008160, doi:10.1016/j.neuroimage.2011.07.039.

Penny, W., Flandin, G., Trujillo-Barreto, N., 2007. Bayesian comparison of spatially regularised general linear models. Human Brain Mapping 28, 275–293. URL: http://onlinelibrary.wiley.com/doi/10.1002/hbm.20327/abstract, doi:10.1002/hbm.20327.

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 Computational Biology 6, e1000709. URL: http://dx.plos.org/10.1371/journal.pcbi.1000709, doi:10.1371/journal.pcbi.1000709.

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

Schwarz, G., 1978. Estimating the Dimension of a Model. The Annals of Statistics 6, 461–464. URL: http://projecteuclid.org/euclid.aos/1176344136, doi:10.1214/aos/1176344136.

Soch, J., Allefeld, C., in prep. The cross-validated log model evidence: a new model selection criterion .

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, in: F1000Research. URL: http://f1000research.com/posters/1096034, doi:10.7490/f1000research.1096034.1.

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, in: F1000Research. URL: http://dx.doi.org/10.7490/f1000research.1000161.1, doi:10.7490/f1000research.1000161.1.

Soch, J., Haynes, J.D., Allefeld, C., 2016. How to avoid mismodelling in GLM-based fMRI data analysis: Cross-validated Bayesian model selection. NeuroImage URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811916303615, doi:10.1016/j.neuroimage.2016.07.047.

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