MACS – a new SPM toolbox for model assessment, comparison and selection

— *Research Article* submitted to *Journal of Neuroscience Methods* on 03/11/2017 —

Joram Soch[a,c,•], Carsten Allefeld[a,b]

[a] Bernstein Center for Computational Neuroscience, Berlin, Germany
[b] Berlin Center for Advanced Neuroimaging, Berlin, Germany
[c] 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

Background: In cognitive neuroscience, functional magnetic resonance imaging (fMRI) data are widely analyzed using general linear models (GLMs). However, model quality of GLMs for fMRI is rarely assessed, in part due to the lack of formal measures for statistical model inference.
New Method: We introduce a new SPM toolbox for model assessment, comparison and selection (MACS) of GLMs applied to fMRI data. MACS includes classical, information-theoretic and Bayesian methods of model assessment previously applied to GLMs for fMRI as well as recent methodological developments of model selection and model averaging in fMRI data analysis.
Results: The toolbox – which is freely available from GitHub – directly builds on the Statistical Parametric Mapping (SPM) software package and is easy-to-use, general-purpose, modular, readable and extendable. We validate the toolbox by reproducing model selection and model averaging results from earlier publications.
Comparison with Existing Methods: A previous toolbox for model diagnosis in fMRI has been discontinued and other approaches to model comparison between GLMs have not been translated into reusable computational resources in the past.
Conclusions: Increased attention on model quality will lead to lower false-positive rates in cognitive neuroscience and increased application of the MACS toolbox will increase the reliability of GLM results and reproducibility of fMRI studies.

Keywords

fMRI-based neuroimaging, mass-univariate GLM, SPM toolbox, analysis pipelines, model assessment, model comparison, model selection, model averaging

Toolbox DOI

DOI 10.5281/zenodo.845404

Contents

1 Introduction **1**

2 Theory **2**
2.1 Model assessment 2
2.1.1 Goodness-of-fit measures 2
2.1.2 Classical information criteria 4
2.1.3 Bayesian log model evidence 5
2.2 Model comparison 6
2.3 Model selection 7

3 Methods **8**
3.1 Three layers of implementation 8
3.2 Two modes of batch processing 9

4 Results **10**
4.1 Model assessment: repetition priming 10
4.2 Model selection: orientation pop-out 11
4.3 Model averaging: conflict adaptation 12

5 Discussion **13**

6 Software Note **14**

7 Acknowledgements **14**

Highlights

- Functional MRI (fMRI) data are analyzed using general linear models (GLMs).
- Model quality is rarely assessed or controlled when employing GLMs for fMRI.
- We introduce an SPM toolbox for model assessment, comparison and selection (MACS).
- The structure of the MACS toolbox facilitates diverse analysis pipelines.
- The toolbox is validated by reproducing published model selection results.

1 Introduction

In functional magnetic resonance imaging (fMRI), general linear models (GLMs) are routinely applied to relate measured hemodynamic signals with controlled experimental variables in order to make statements about the localization of cognitive functions in brain regions (Friston et al., 1994; Holmes and Friston, 1998). Like any other statistical inference, GLM analyses are dependent of the particular choice of the model (Andrade et al., 1999; Carp, 2012) and model quality can have a profound effect on sensitivity and specificity of statistical tests (Razavi et al., 2003; Monti, 2011).

Model assessment (Akaike, 1974; Schwarz, 1978), model comparison and selection (Kass and Raftery, 1995) as well as model averaging (Hoeting et al., 1999) are established techniques to control for model quality by inferring the most likely model given the data and by conditioning data analysis on the optimal model for each data set. While model assessment (Friston et al., 2007), comparison (Penny et al., 2004; Penny, 2012), selection (Stephan et al., 2009) and averaging (Penny et al., 2010) have become standard practice in dynamic causal modelling (DCM) for fMRI (Friston et al., 2003), such techniques are almost not used in the application of GLMs for fMRI (Friston et al., 1994).

Previous attempts to control for model quality in GLMs for fMRI include statistical tests for goodness of fit (Razavi et al., 2003) and the application of Akaike's or Bayesian information criterion for activation detection (Seghouane and Ong, 2010) or theory selection (Gläscher and O'Doherty, 2010). Additionally, voxel-wise Bayesian model assessment (Penny et al., 2003, 2005, 2007) and random-effects Bayesian model selection (Rosa et al., 2010) have been included in the popular software package *Statistical Parametric Mapping* (SPM), but are only rarely used due to low visibility, high analytical complexity and interpretational difficulty. Finally, a toolbox for frequentist model diagnosis and exploratory data analysis called *SPMd* ("d" for "diagnostic") has been released for SPM (Luo and Nichols, 2003), but was discontinued several years ago[1] (Nichols, 2013).

With this work, we provide a new SPM toolbox for **m**odel **a**ssessment, **c**omparison and **s**election termed *MACS* (pronounced as "Max"). Purposes of this software package are (i) to unify classical, information-theoretic and Bayesian measures of model quality previously applied to fMRI data, (ii) to provide a common pipeline for our previously developed methods of cross-validated Bayesian model selection (cvBMS; Soch et al., 2016) and averaging (cvBMA; Soch et al., 2017) that so far existed as seperate toolkits and (iii) to build a framework in which all these model quality measures can be conveniently applied to GLMs estimated in SPM by directly building on `SPM.mat` files.

The rest of the paper falls into three parts. First, we give a technical overview of the features included in the MACS toolbox (see Section 2). Among others, we have implemented goodness-of-fit (GoF) measures, classical information criteria (ICs) and the Bayesian log model evidence (LME). Second, we explain the structure of the MACS toolbox and how it facilitates diverse and user-specific analyses (see Section 3). Most importantly, the embedding of MACS into SPM's batch systems accelerates and simplies fMRI analyses. Finally, we present two exemplary applications illustrating cvBMS and cvBMA as our latest methodological advances for fMRI (see Section 4) and discuss our work.

[1] Tom Nichols gave a "Requiem for SPMd" on Slide 3 of his talk "Building Confidence in fMRI Results with Model Diagnosis" in the workshop "How Not to Analyze Your Data: A Skeptical Introduction to Modeling Methods" at the 2013 Annual Meeting of the Organization for Human Brain Mapping.

2 Theory

In the MACS toolbox, methods of model inference are broadly categorized into the three categories *model assessment* (MA), *model comparison* (MC) and *model selection* (MS). This section gives a technical introduction of these different techniques which are also summarized in Table 1.

2.1 Model assessment

Model assessment (MA) is defined as any method that quantifies the quality of an individual model by a single real number. MA methods can again be sub-divided into three groups of tools: (i) *goodness-of-fit measures* (GoF), (ii) *classical information criteria* (ICs) and (iii) *Bayesian log model evidence* (LME).
In all of the following, we consider a generative model m with model parameters θ describing measured data y with estimated parameters $\hat{\theta}$ giving predicted data $\hat{y}$, with number of data points n and number of model parameters p. In the context of first-level fMRI data analysis, m is typically a *general linear model* (GLM), i.e. univariate linear regression analysis; θ includes the *regression coefficients* β and the *residual variance* σ^2; y is a time course of the *blood-oxygenation level dependent* (BOLD) signal measured via fMRI; $\hat{y}$ is the fitted BOLD signal; n is the number of fMRI scans and p is the number of GLM regressors[2].

2.1.1 Goodness-of-fit measures

The purpose of GoF measures is to get a first impression of how well a model – in this case: a GLM – describes a given set of data – in this case: voxel-wise fMRI signals. One obvious thing to look at would be the *residual variance* of the GLM calculated as

$$\begin{aligned} \hat{\sigma}^2_{\mathrm{ML}} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ \hat{\sigma}^2_{\mathrm{ub}} &= \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \end{aligned} \tag{1}$$

where $\hat{\sigma}^2_{\mathrm{ML}}$ is the maximum-likelihood estimator and $\hat{\sigma}^2_{\mathrm{ub}}$ is the unbiased estimator of that quantity. In SPM, the residual variance is calculated somewhat differently as

$$\hat{\sigma}^2_{\mathrm{SPM}} = \frac{1}{\mathrm{tr}(RV)}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

where R is the residual forming matrix and V is the error correlation matrix of the corresponding GLM which accounts for the loss of effective residual degrees of freedom (erdf) by whitening and filtering the data. Generally, the lower the residual variance, the better model predictions and measured data agree with each other. However, as the

[2]Note that p is the number of GLM regressors in GoF measures, but this number plus one in classical ICs, because information criteria account for the total number of model parameters which include the residual variance σ^2 in addition to the regression coefficients β.

residual variance does not have an upper bound, it is more useful to look at the *coefficient of determination* (Razavi et al., 2003, eq. 2)

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = \frac{\mathrm{ESS}}{\mathrm{TSS}} \tag{3}$$

where RSS is the residual sum of squares $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, ESS is the explained sum of squares $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ and TSS is the total sum of squares $\sum_{i=1}^{n}(y_i - \bar{y})^2$, such that TSS = ESS + RSS. Generally, the closer R^2 gets to 1, the better the model captures the variation. However, as the coefficient of determination always increases when another variable is included in the model, an improved measure is given by the *adjusted coefficient of determination* (Razavi et al., 2003, eq. 4)

$$R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{RSS}/\mathrm{df}_r}{\mathrm{TSS}/\mathrm{df}_t} \tag{4}$$

where the residual degress of freedom $\mathrm{df}_r = n - p$ and the total degrees of freedom $\mathrm{df}_t = n - 1$ are used to adjust the sums of squares to account for additional model parameters. For this reason, the adjusted R^2 can be seen as a precursor to classical information criteria (see next section).
Note that, whereas R^2 always falls between 0 and 1, R^2_{adj} can also become negative. Both R^2 and R^2_{adj} are also related to the F-statistic in an F-test of a GLM against only the constant regressor which is given by

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/\mathrm{df}_n}{\mathrm{RSS}/\mathrm{df}_d} \tag{5}$$

where the numerator degress of freedom $\mathrm{df}_n = p - 1$ and the denominator degrees of freedom $\mathrm{df}_d = n - p$ are the parameters of the F-distribution under the null hypothesis that the signal only consists of a mean μ.
Goodness-of-fit measures also include the *signal-to-noise ratio* (SNR):

$$\begin{aligned} \mathrm{SNR}_{\mathrm{mf}} &= \frac{|\hat{\mu}|}{\hat{\sigma}}, \quad \hat{\mu} = \bar{y} \\ \mathrm{SNR}_{\mathrm{mb}} &= \frac{\mathrm{var}(X\hat{\beta})}{\hat{\sigma}^2} \ . \end{aligned} \tag{6}$$

Here, $\mathrm{SNR}_{\mathrm{mf}}$ is the model-free SNR which is calculated as the inverse of the *coefficient of variation* (CoV) and quantifies the ratio of signal average to signal variation (Welvaert and Rosseel, 2013, def. 1). Signals that are very low or very variable will receive a lower SNR given this measure. In contrast, $\mathrm{SNR}_{\mathrm{mb}}$ corresponds to a model-based SNR which divides the variance of the explained signal by the variance of the unexplained signal (Welvaert and Rosseel, 2013, def. 5).
Note that, if the measured signal y and the predicted signal $\hat{y}$ have the same mean, $\mathrm{var}(X\hat{\beta}) = \mathrm{var}(\hat{y})$ is proportional to the explained sum of squares and $\mathrm{SNR}_{\mathrm{mb}} = \mathrm{ESS}/\mathrm{RSS} = R^2/(1-R^2)$. Crucially, with a change of the GLM, the model-based SNR will also change while the model-free SNR will not.
In our toolbox, all these measures are implemented via the routine `MA_inspect_GoF` and can be exported as whole-brain maps or explored in a voxel-wise manner (see Figure 4A).

2.1.2 Classical information criteria

Classical information criteria (ICs), in the terminology of the MACS toolbox, are model quality measures that only depend on the *maximum log-likelihood* (MLL), as a measure of model accuracy, as well as n and p, giving rise to some measure of model complexity. The first of these ICs was *Akaike's information criterion* (Akaike, 1974)

$$\mathrm{AIC} = -2 \log p(y|\hat{\theta}, m) + 2k \tag{7}$$

where the maximum log-likelihood $\mathrm{MLL} = \log p(y|\hat{\theta}, m)$ and $k = p + 1$ is the number of model parameters (see Fn. 2). The *corrected AIC*, an improved version of the AIC for finite-sample calculations, has been derived as (Hurvich and Tsai, 1989)

$$\mathrm{AIC_c} = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1} \ . \tag{8}$$

Historically, the second IC was the *Bayesian information criterion* (Schwarz, 1978)

$$\mathrm{BIC} = -2 \log p(y|\hat{\theta}, m) + k \log n \tag{9}$$

which means that AIC and BIC are equal for sample sizes around $n = e^2 \approx 7.39$. Although the name suggests that the BIC rests on Bayesian inference, it does not and is only theoretically connected to Bayesian statistics by being an infinite-sample approximation to the Bayesian log model evidence (see next section).
In contrast to that, the *deviance information criterion* is a "real" Bayesian information criterion, because it requires the specification of prior distributions on the model parameters and posterior distributions are used to calculate expected likelihoods. Technically, the DIC therefore depends on more than just MLL, n and p. Formally, the DIC is given by (Spiegelhalter et al., 2002)

$$\mathrm{DIC} = -2 \log p(y| \langle\theta\rangle_{\theta|y}, m) + 2p_D \tag{10}$$

where the effective number of parameters is

$$p_D = \langle D(\theta)\rangle_{\theta|y} - D(\langle\theta\rangle_{\theta|y}) \tag{11}$$

and $D(\theta)$ is the log-likelihood function on the deviance scale

$$D(\theta) = -2 \log p(y|\theta, m) \ , \tag{12}$$

such that $\langle D(\theta)\rangle_{\theta|y}$ is proportional to the *posterior expected log-likelihood* (PLL) and $D(\langle\theta\rangle_{\theta|y})$ is proportional to the *log-likelihood at the posterior expectation* (LLP). In these equations, the operator $\langle\cdot\rangle_{\theta|y}$ denotes an expectation across the posterior distribution (Woolrich et al., 2004) which naturally depends on the prior distribution. However, the DIC is very robust with respect to the choice of the prior and we have therefore implemented it with a non-informative prior (Soch et al., 2016, eq. 15).
In our toolbox, all these criteria are implemented via the routine `MA_classic_ICs` which can also be accessed through the corresponding batch editor job (see Figure 2).

2.1.3 Bayesian log model evidence

An alternative to those classical information criteria are Bayesian measures of model quality based on the marginal log-likelihood, also called the *log model evidence* (LME) which is given by (Gelman et al., 2013, eq. 1.3; Bishop, 2007, eq. 1.45)

$$\mathrm{LME}(m) = \log p(y|m) = \log \int p(y|\theta, m)\, p(\theta|m)\, \mathrm{d}\theta \ . \tag{13}$$

Note that, while classical ICs operate on the deviance scale (see eq. 12), the LME operates on the log-likelihood scale (see eq. 13), so that thresholds commonly applied to LME differences (Kass and Raftery, 1995) must be multiplied by -2 when applying them to classical ICs. Such thresholds include log Bayes factors of 0, 1, 3 or 5 corresponding to Bayes factors of around 1, 3, 20 or 150 (see next section).
The LME has several advantages over classical information criteria including (i) its automatic penalty for additional model parameters (Penny, 2012), (ii) its natural decomposition into model accuracy and model complexity (Penny et al., 2007) and (iii) its consideration of the whole uncertainty about parameter estimates (Friston et al., 2007). The LME however has the disadvantage that it depends on a prior distribution $p(\theta|m)$ over model parameters and the choice of the prior can have a profound impact on resulting model comparisons. In recent work on GLMs for fMRI, we have therefore suggested the *out-of-sample log model evidence* (Soch et al., 2016, eq. 13)

$$\mathrm{oosLME}_i(m) = \log \int p(y_i|\theta, m)\, p(\theta| \cup_{j \neq i}\ y_j, m)\, \mathrm{d}\theta \tag{14}$$

and the *cross-validated log model evidence* (Soch et al., 2016, eq. 14)

$$\mathrm{cvLME}(m) = \sum_{i=1}^{S} \log \int p(y_i|\theta, m)\, p(\theta| \cup_{j \neq i}\ y_j, m)\, \mathrm{d}\theta \tag{15}$$

where S is the number of data subsets – in our case: fMRI recording sessions. These can be seen as Bayesian measures of model quality, but avoid having to specify prior distributions on the model parameters. Instead, the prior distribution is obtained by training the model on independent data of the same kind.
Given that LME, cvLME or oosLME have been calculated and models m can be sensibly grouped into model families f, *log family evidences* (LFE) can be easily calculated using the law of marginal probability as (Soch et al., 2016, eq. 18)

$$\mathrm{LFE}(f) = \log p(y|f) = \log \sum_{m \in f} p(y|m)\, p(m|f) \tag{16}$$

where $p(y|m)$ is the exponential of LME, cvLME or oosLME and $p(m|f)$ is a within-family prior, usually set to a discrete uniform distribution, embodying the assumption that within each family, all the models are equally likely *a priori*.
In our toolbox, the cvLME is the default model selection criterion and can be calculated via the routine `MA_cvLME_multi` or the corresponding batch editor job (see Figure 2).

2.2 Model comparison

Model comparison (MC) is defined as any method that quantifies the quality difference between two models by a single real number. Most trivially, one can for example calculate the *difference between two information criteria* (Penny et al., 2004, eq. 23)

$$\Delta\mathrm{IC}_{ij} = \mathrm{IC}(m_i) - \mathrm{IC}(m_j) \tag{17}$$

which can be performed for AIC, BIC etc. and compared with particular scales. When calculating this difference between two LMEs, the result is called the logarithm of a Bayes factor or a *log Bayes factor* (Penny et al., 2004, eq. 22)

$$\begin{aligned} \mathrm{BF}_{ij} &= \frac{p(y|m_i)}{p(y|m_j)} \\ \mathrm{LBF}_{ij} &= \mathrm{LME}(m_i) - \mathrm{LME}(m_j) \end{aligned} \tag{18}$$

which can be applied to LME, cvLME or oosLME. When analyzing a group of subjects in an experimental study and assuming independence across subjects, subject-level LMEs can be summed up into a *group-level LME* (Penny et al., 2010, eq. 6)

$$\begin{aligned} p(Y|m) &= \prod_{i=1}^{N} p(y_i|m) \\ \log p(Y|m) &= \sum_{i=1}^{N} \log p(y_i|m) \end{aligned} \tag{19}$$

where $Y = \{y_1, \ldots, y_N\}$ are the group data and N is the number of subjects. Comparing two models across several subjects then gives rise to a group Bayes factor or *log group Bayes factor* which is given by (Stephan et al., 2007, eq. 20)

$$\begin{aligned} \mathrm{GBF}_{ij} &= \frac{p(Y|m_i)}{p(Y|m_j)} \\ \mathrm{LGBF}_{ij} &= \log p(Y|m_i) - \log p(Y|m_j) \ . \end{aligned} \tag{20}$$

Importantly, this LGBF approach assumes the same model for each subject and therefore corresponds to a fixed-effects (FFX) analysis. If this assumption cannot be justified, a random-effects (RFX) model has to be used (see next section).
Finally, in the case that just two models are considered and a uniform prior across models is assumed, (log) Bayes factors can be directly transformed into *posterior model probabilities* using the simple relation (Penny and Ridgway, 2013, eq. 10)

$$p(m_i|y) = \frac{\mathrm{BF}_{ij}}{\mathrm{BF}_{ij} + 1} = \frac{\exp(\mathrm{LBF}_{ij})}{\exp(\mathrm{LBF}_{ij}) + 1} \tag{21}$$

and additionally

$$p(m_j|y) = 1 - p(m_i|y) \ . \tag{22}$$

2.3 Model selection

Model selection (MS) is defined as any method that chooses one particular model from a set of candidate models based on some objective criterion. For example, *fixed-effects Bayesian model selection* (FFX BMS) generalizes the idea of posterior model probabilities to more than two models, such that (Penny et al., 2010, eq. 7)

$$p(m_i|Y) = \frac{p(Y|m_i)\,p(m_i)}{\sum_{j=1}^{M} p(Y|m_j)\,p(m_j)} \tag{23}$$

where M is the number of models. In contrast to that, *random-effects Bayesian model selection* (RFX BMS) allows for the possibility that different subjects have different optimal models. In each subject, the generating model is assumed to come from a multinomial distribution with a Dirichlet as the prior distribution over multinomial frequencies. However, the models are not observed directly, but only indirectly via the model evidences from first-level model assessment (Stephan et al., 2009, eqs. 3-5):

$$\begin{aligned} p(y|m) &= \prod_{i=1}^{N} p(y_i|m_i) = \prod_{i=1}^{N}\prod_{j=1}^{M} p(y_i|e_j)^{m_{ij}} \\ p(m|r) &= \prod_{i=1}^{N} \mathrm{Mult}(m_i; 1, r) = \prod_{i=1}^{N}\prod_{j=1}^{M} r_j^{m_{ij}} \\ p(r|\alpha) &= \mathrm{Dir}(r;\alpha) = \frac{\Gamma(\sum_{j=1}^{M}\alpha_j)}{\prod_{j=1}^{M}\Gamma(\alpha_j)} \prod_{j=1}^{M} r_j^{\alpha_j - 1} \end{aligned} \tag{24}$$

Together, these probability distribution define a second-level hierarchical population proportion model that can be inverted using either Variational Bayes (VB) (Stephan et al., 2009, eq. 14) or Gibbs Sampling (GS) (Penny et al., 2010, eqs. 13-17) techniques.
Finally, *Bayesian model averaging* (BMA) is also grouped under model selection here as it implicitly consists in, based on the set of candidate models, fitting a larger model which effectively marginalizes over modelling approaches. In our implementation, averaged model parameter estimates are calculated as (Soch et al., 2017, eq. 8)

$$\hat{\beta}_{\mathrm{BMA}} = \sum_{i=1}^{M} \hat{\beta}_i \cdot p(m_i|y) \tag{25}$$

where $\hat{\beta}_i$ is the i-th model's parameter estimate for a particular regressor in the GLM and $p(m_i|y)$ is the i-th model's posterior probability calculated from cvLMEs.
Whereas *model selection* decides for one particular model that maximizes a certain function such as log model evidence, posterior model probability or group-level model frequency, *model averaging* combines several models using their posterior probabilities calculated from model evidences. Given that finite data never provide absolute certainty about the modelling approach, this approach allows to account for modelling uncertainty by utilizing parameter estimates from all models in the model space instead of discarding the valuable information from all but one model.

3 Methods

The MACS toolbox is written in MATLAB and designed to work as an extension for Statistical Parametric Mapping (SPM), Version 8 oder 12 (SPM8/12). This section describes how three layers of implementation make the toolbox accessible for different users and how two modes of batch processing make the toolbox useful in different contexts.

3.1 Three layers of implementation

An overview of the toolbox functions is given in Figure 1. This figure shows the three types of functions included in the toolbox: (i) mathematical functions, (ii) interface functions and (iii) batch functions. *Mathematical functions* (see Figure 1C) directly implement the statistical analyses (such as `ME_GLM_NG` for estimating a Bayesian GLM) required for toolbox features. *Interface functions* (see Figure 1B) represent toolbox features (such as `MA_cvLME_multi` for calculating the voxel-wise cvLME) and execute mathematical functions. *Batch functions* (see Figure 1A) take information from SPM batch editor jobs (such as `batch_MA_cvLME_auto` for performing cvLME assessment) and use it to call interface functions.
Regular toolbox users will never access any of these functions directly in MATLAB, but will use toolbox features through SPM's batch editor only. This can be achieved by installing the MACS toolbox and opening the SPM batch editor (see Section 6 and Figure 3). Users can then fill out regular batch jobs like in SPM and directly run them from the batch editor. Batch job templates only use few options most of which can often be left at their default values. Examples are given in Figures 3 and 4:

```
- Figure 3 : batch for cross-validated Bayesian model selection
- Figure 4A: output from goodness-of-fit inspection for a GLM
- Figure 4B: output from high-dimensional data visualization
```

Intermediate toolbox users may also use the batch editor, but will additionally call the interface functions via MATLAB's command-line interface. This could for example happen in a pipeline script which is intended to process many models applied to many subjects. Note however that those pipeline-style analyses can also be implemented using the regular batch system. An exemplary command-line call could look like this:

```
>> load C:\fMRI\SUB_01\GLM_A\SPM.mat    % load GLM estimated in SPM
>> MA_cvLME_multi(SPM)                  % calculate voxel-wise cvLME
>> MA_classic_ICs(SPM)                  % calculate voxel-wise AIC/BIC
```

Advanced toolbox users and developers can also rely on the batch editor and interface functions, but will additionally call mathematical functions to perform their own customized analyses. This could for example happen when one is interested in Bayesian GLM parameter estimates or when the cvLME is to be calculated for a large number of models in a small number of voxels. This type of usage can also lead to new methods developments. An extract from a possible script could read like this:

```
% load model specification and prepare model estimation
load C:\fMRI\SUB_01\GLM_A\SPM.mat    % load GLM estimated in SPM
Y = MA_load_data(SPM);               % load voxel-wise fMRI data
X = SPM.xX.X;                        % get design matrix
V = SPM.xVi.V;                       % get covariance matrix
P = spm_inv(V);                      % get precision matrix
[n, p] = size(X);                    % get model dimensions
m0 = zeros(p,1);                     % set prior beta mean
L0 = zeros(p,p);                     % set prior beta covariance

% estimate general linear model with normal-gamma priors
[mn, Ln, an, bn] = ME_GLM_NG(Y, X, P, m0, L0, 0, 0);
```

3.2 Two modes of batch processing

An overview of the toolbox modules is given in Figure 2. This figure shows the two modes of batch processing that the toolbox allows: (i) automatic processing and (ii) manual processing. In short, automatic processing requires specifying the model space only once whereas manual processing requires that log model evidence images are entered for each model and each subject in every processing step.

Automatic processing consists in creating a model space file called `MS.mat` that contains references to `SPM.mat` files belonging to several GLMs for several subjects. Given that the model space has been specified, cvLME model assessment, Bayesian model comparison, model selection, model averaging and other modules can be implemented by a simple reference to the model space file which automatically applies the analyses to all subjects. The automatric procesing mode is the preferred one as it allows larger analyses in shorter time and also avoids unnecessary analysis errors.

Manual processing means that GLM folders or cvLME maps are not referenced automatically using a model space file, but manually by reference to their filepaths. This is useful when for example cvLME model assessment is prefered to be performed in one batch with GLM model estimation or when LME images are processed, e.g. spatially smoothed or recalculated into LFE images, before further steps such as model comparison, model selection or model averaging are performed.

Like the SPM batch system, the MACS toolbox uses *dependencies between modules*. Dependencies are references to files generated in earlier processing steps that are to be used in later processing steps. For example, an `SPM.mat` file generated during model estimation may be used for cvLME calculation; or a cvLME map generated during model assessment may be used for group-level BMS; or a `BMS.mat` file generated during model selection may be used for generation of selected-model maps (SMM); etc. The MACS toolbox encourages the use of dependencies and they are most conveniently implemented in the automatic processing mode.

4 Results

The MACS toolbox was primarily designed to summarize tools for *cross-validated Bayesian model selection* (cvBMS) and *cross-validated Bayesian model averaging* (cvBMA). In this section, we propose pipelines and analyze examples to demonstrate how these techniques can be effortlessly implemented with MACS. Before that, we describe analyses illustrating how to apply goodness-of-fit measures and classical information criteria to first-level GLMs in SPM.

4.1 Model assessment: repetition priming

To demonstrate first-level model assessment, we reanalyze a single-subject, single-session data set from a study on repetition priming (Henson et al., 2002) that is used as example data in the SPM manual (Ashburner et al., 2016, ch. 31). In this experiment, subjects were viewing faces which either belonged to famous or non-famous people ("familiarity") and each face was presented twice ("repetition"). This implies a 2×2 experimental design with 4 experimental conditions. This demonstration builds on data preprocessed according to the SPM12 manual (Ashburner et al., 2016, sec. 31.1) and addresses model assessment for the "categorical" model (Ashburner et al., 2016, sec. 31.2) and the "parametric" model (Ashburner et al., 2016, sec. 31.3).

First, we want to get an impression of where in the brain the categorical model performs better and where it performs worse. We use the MACS module for goodness-of-fit inspection to assess signal-to-noise ratio (SNR) and coefficient of determination (R^2) across the whole brain. We find that in visual cortex, especially fusiform face area (FFA), the model achieves the highest variance explanation which is consistent with the facial stimuli of the experiment (see Figure 4A).

Second, we want to find out where the model fit is supported by which model parameters. We use the MACS module for high-dimensional data visualization to explore the estimated regression coefficients from the main experimental conditions in a voxel-wise fashion. For orientation purposes, we overlay this query with an R^2 image obtained in the first step. We find that in one region of parietal cortex, possibly the superior parietal gyrus (SPG), there is a negative effect of both experimental factors, familiarity and repetition (see Figure 4B).

Finally, we want to compare the categorical against the parametric GLM. We use the MACS module for classical information criteria (ICs) to compute the voxel-wise Bayesian information criterion (BIC) for both models. Computing the information difference (ΔIC) between the two models is then implemented using the log Bayes factor (LBF) module. Model comparison at an ad-hoc threshold of $\Delta\text{IC} > 10$ in favor of the parametric model gives rise to a large cluster in parietal cortex, close to the postcentral gyrus (PCG), that was also reported when performing an F-test on the additional regressors in the parametric model[3] (Ashburner et al., 2016, fig. 31.17).

[3]Note that (i) we of course advocate to perform this model comparison using Bayesian model quality measures such as the cvLME rather than information criteria like BIC and (ii) the model comparison performed here is, strictly speaking, not equivalent to the F-test in the SPM manual, because the parametric model also excludes regressors that are included in the categorical model, namely experimental condition regressors convolved with the hemodynamic derivatives.

4.2 Model selection: orientation pop-out

To demonstrate the cvBMS pipeline, we reanalyze data from an earlier study (Bogler et al., 2013) that has previously been used for the cvBMS paper (Soch et al., 2016). In this data set, we compared three models: a categorical model (GLM I) that describes the 4×4 experimental design using 16 regressors, one for each experimental condition; a linear-parametric model (GLM II) that uses one onset regressor indicating stimulation phases (orientation pop-out) and two parametric modulators encoding levels of the two factors (left and right orientation contrast); and a nonlinear-parametric model (GLM III) that uses nonlinear parametric regressors, motivated by behavioral pre-tests (Bogler et al., 2013; Soch et al., 2016). In this demonstration, we will build on estimated first-level GLMs and only describe the steps of model assessment and model selection.

First, in order to avoid model dilution, we want to compare the family of categorical models (consisting of GLM I only) against the family of parametric models (consisting of GLMs II & III). To this end, we specify a job in which we define a model space that consists of all 3 models applied to 22 subjects. Using dependencies on this model space information, automatic calculation of cvLME maps and LFE maps is performed (see Figure 5A). Following model family assessment, group-level BMS has to be specified manually, because log *family* evidences are not associated to particular *models* – which would be necessary for the automatic processing mode –, but only exist as individual images. After group-level BMS, selected-model maps are generated (see Figure 5A). This analysis leads to the results depicted in Figure 3B of the cvBMS paper.

Second, within the winning family of parametric models, we want to compare the linear-parametric (GLM II) to the nonlinear-parametric (GLM III) model. To this end, we define a new model space consisting only of these 2 models applied to 22 subjects. As cvLME maps have already been generated in the first step, model assessment does not have to be performed in this second step. Using dependencies on model space information, automatic group-level BMS and generation of SMMs are performed (see Figure 5A). This analysis leads to the results depicted in Figure 3C of the cvBMS paper.

Third, using the winning nonlinear-parametric model, we want to investigate the effect of removing the parametric modulator either for left or for right orientation contrast (OC), i.e. the effect of visual hemifield on model preference. This leads to a model describing only left OC (GLM III-l) and a model describing only right OC (GLM III-r). Again, we define a model space consisting of these 2 models applied to 22 subjects. As these are new models, cvLME maps have to be calculated (automatically). Finally, group-level BMS and selected-model maps are added to the pipeline (see Figure 5A). From the SMMs, we extract the proportion of voxels in left and right V4, respectively, with preferences for left or right OC, respectively, according to the group-level model selection. This analysis leads to the results depicted in Figure 5C of the cvBMS paper.

4.3 Model averaging: conflict adaptation

To demonstrate the cvBMA pipeline, we reanalyze data from another study (Meyer and Haynes, in prep.) that has previously been used for the cvBMA paper (Soch et al., 2017). For this data set, we estimated four models: all models include regressors for main blocks and delay phases in a conflict adaptation paradigm (Meyer and Haynes, in prep.); one model includes no additional processes (GLM 1); one model additionally describes the first trials following preparatory delays (GLM 2); one model additionally describes the cue phases preceding preparatory delays (GLM 3); and one model includes both additional processes, cue phases and first trials (GLM 4). Critically, cue phases and first trials are temporally correlated to preparatory delays, such that model choice influences parameter estimates for the delays phases. However, as each model includes regressors for the delay phases, we can use Bayesian model averaging to remove this modelling uncertainty (Soch et al., 2017). In this demonstration, we will build on estimated first-level GLMs and only describe the steps of model assessment and model averaging.

First, following model estimation, we want to calculate averaged estimates for the common model parameters in these four GLMs. To this end, we specify a job in which we define a model space that consists of all 4 models applied to 24 subjects. Using dependencies on this model space information, automatic calculation of cvLME maps and automatic group-level BMA are performed (see Figure 5B). Additionally, we perform automatic group-level BMS and generate selected-model maps. This analysis leads to the results reported in Table 2 and Figure 4A of the cvBMA paper. Group-level SMMs can be used for masking second-level analyses – an alternative to BMA that is not included in this demonstration, but discussed in the publication (Soch et al., 2017).

Second, following model averaging, we want to perform second-level analysis, but based on the averaged model parameter estimates calculated in the first step instead of the parameter estimates from a particular GLM as in standard processing pipelines. To this end, we use SPM's factorial design specification to set up a paired t-test between parameter estimates belonging to the two delay phase regressors (see Figure 5B). Subsequent model estimation, contrast calculation and statistical inference lead to the results reported in Table 1 and Figure 4A of the cvBMA paper. As expected, a main effect between switch and stay delays phases and, in more detail, a positive effect of switch over stay delay phases can be observed in left primary motor cortex (Soch et al., 2017).

SPM batch editor job files from all demonstrations, classical model assessment as well as cvBMS and cvBMA, are included as mat-files in the MACS toolbox repository. They can be used for educational purposes illustrating suggested pipelines or as analysis templates underlying the user's applications. Additionally, time consumption of creating and executing these demonstrations are reported in Table 2. In summary, when using the automatic processing mode, complex analyses, such as whole cvBMS and cvBMA pipelines, can be quickly implemented, because almost all analysis steps can simply reference the model space module which usually is the only time-consuming operation.

5 Discussion

In functional magnetic resonance imaging (fMRI), model quality of general linear models (GLMs) is rarely assessed (Razavi et al., 2003) which may have contributed to the current reproducibility crisis in cognitive neuroscience (Pernet and Poline, 2015). We have introduced an SPM toolbox for model assessment, comparison and selection (MACS) of GLMs for fMRI. The MACS toolbox implements classical, information-theoretic and Bayesian measures of model quality and allows to guide first- and second-level fMRI analyses by identifying optimal modelling approaches (Soch et al., 2016, 2017). Increased application of the MACS toolbox should increase the methodological quality of statistical analyses in fMRI research.

The key features of our toolbox are (i) *simplicity* – the toolbox is very easy to use and directly operates on `SPM.mat` files from GLM estimation in SPM; (ii) *generality* – model assessment and selection are not restricted to special model comparisons, but operate on arbitrary design matrices; (iii) *modularity* – toolbox features can be combined into unique processing batches or even appended to existing SPM infrastructure; (iv) *readability* – toolbox code is clearly written for developers and includes extensive comments explaining mathematical background as well as implementational details[4]; and (v) *extendability* – the toolbox can be easily augmented with other methods of model assessment, e.g. when one wants to employ Variational Bayesian (Penny et al., 2007) or Empirical Bayesian (Penny and Ridgway, 2013) versions of the log model evidence instead of our cvLME approach (Soch et al., 2016).

Version 1.0 of the MACS toolbox has been uploaded to GitHub in May 2017 (DOI: 10.5281/zenodo.845404). The developers intend to immediately commit bug fixes to this repository (see Section 6). Future developments of the MACS toolbox may include group-level statistical tests for goodness-of-fit measures (Razavi et al., 2003) and information criteria (Vuong, 1989) as well as improved voxel-wise visualization for second-level model selection results.

[4]Information on any toolbox function (see Figure 1) can be obtained by simply typing `help MACS_fct_name` (e.g. `help MA_cvLME_multi`) into the MATLAB command window after successful MACS toolbox installation (see Section 6). A list of functions with descriptions can be generated using the built-in script `function_index` (resulting in `function_index.xls`).

6 Software Note

The latest version of the MACS toolbox can be downloaded from GitHub (https://github.com/JoramSoch/MACS/releases) and has to be placed as a sub-directory "MACS" into the SPM toolbox folder. Upon starting SPM, batch modules for toolbox features can be assessed by clicking "SPM → Tools → MACS Toolbox" in the SPM batch editor (see Figure 3). MACS is optimized for SPM12, but also compatible with SPM8.
With the adoption of MACS, software for cross-validated Bayesian model selection (cvBMS; https://github.com/JoramSoch/cvBMS) and cross-validated Bayesian model averaging (cvBMA; https://github.com/JoramSoch/cvBMA) has become obsolete and the user is highly encouraged to replace their old pipelines with the new toolbox.

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 an Elsa Neumann Scholarship from the State of Berlin.
The authors would like to thank Carsten Bogler and Achim Meyer for acquiring the fMRI data sets used in this publication.
The authors have no conflict of interest, financial or otherwise, to declare.

Figures

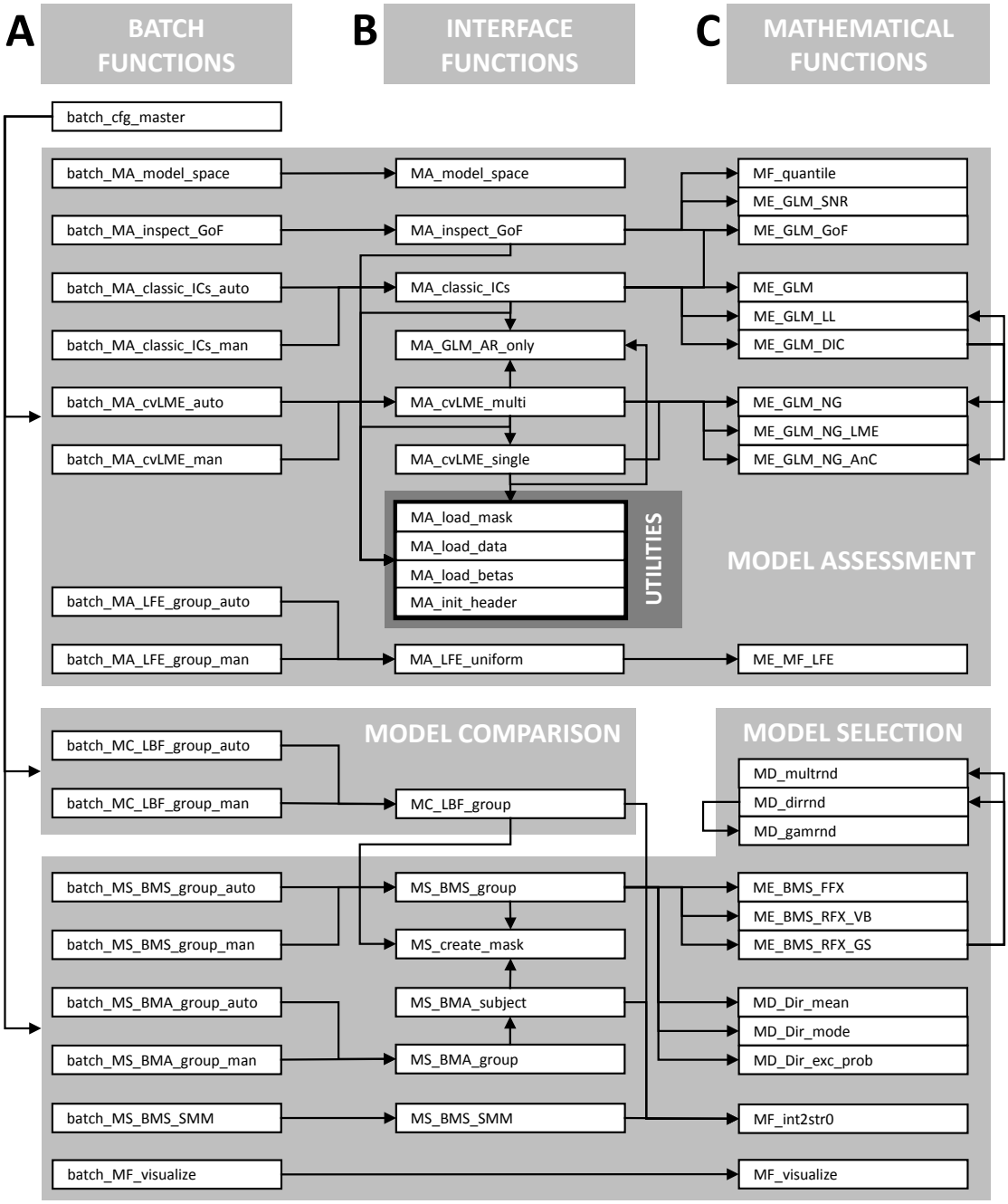

Figure 1. Architecture of the MACS toolbox. This figure is for developers only and lists all functions from the toolbox. A black arrow indicates that one function calls the other. The horizontal dimension corresponds to the type of function (see Section 3.1): Generally, (A) batch functions call (B) interface functions which in turn call (C) mathematical functions. The vertical dimension corresponds to the type of feature (see Section 2): Interface functions consist of tools for model assessment (function prefix "MA"), model comparison (MC) and model selection (MS); mathematical functions consist of tools for model estimation (ME), many distributions (MD) and more functions (MF).

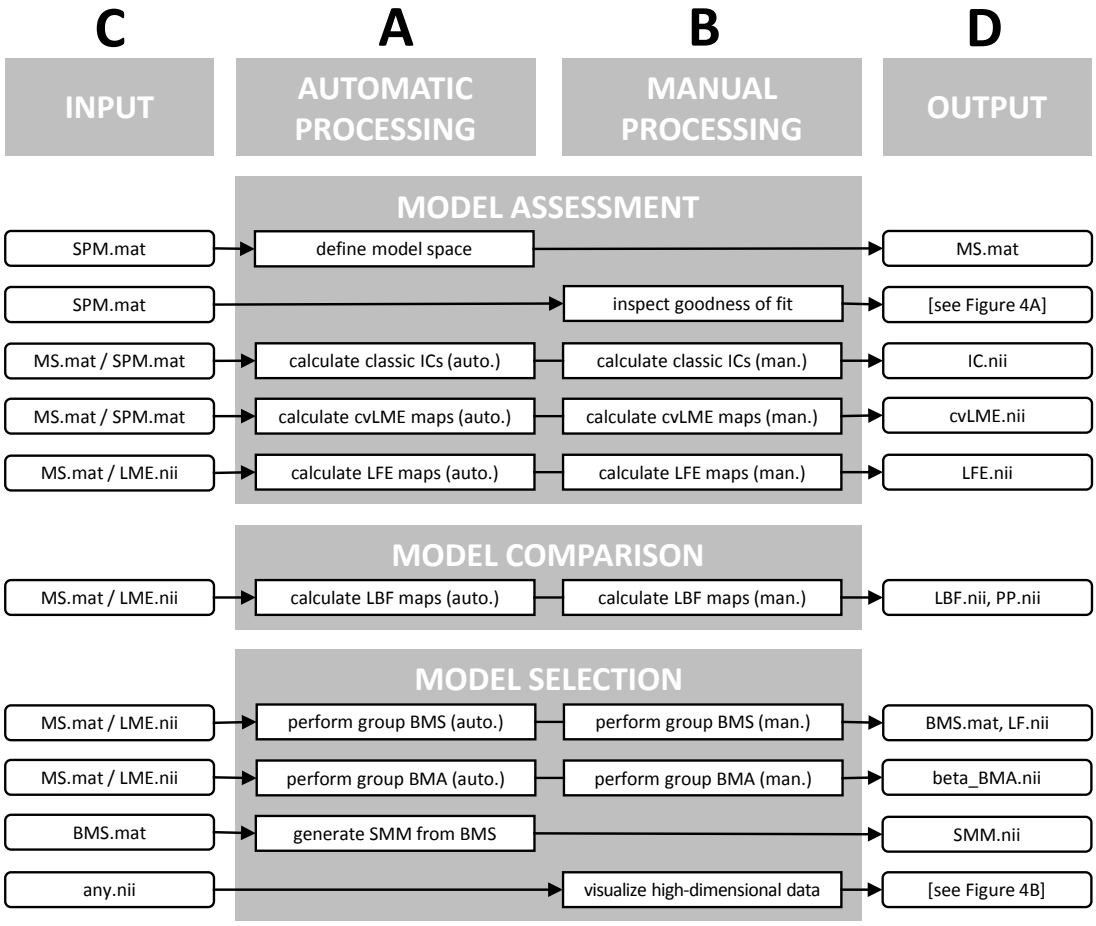

Figure 2. Batch modules in the MACS toolbox. This figure lists batch modules from the toolbox that are accessible via the SPM batch editor. Modules either belong to (A) the automatic processing stream or (B) the manual processing stream (see Section 3.2): automatic processing is based on a model space file (`MS.mat`) whereas manual processing is based on manual input of GLM info (`SPM.mat`) or cvLME maps (`LME.nii`). Modules take (C) particular input files and result in (D) particular output files: MATLAB files contain structured information (such as GLM info) and NIfTI files are brain images (such as cvLME maps). For abbreviations, confer the list of features (see Section 2 and Table 1).

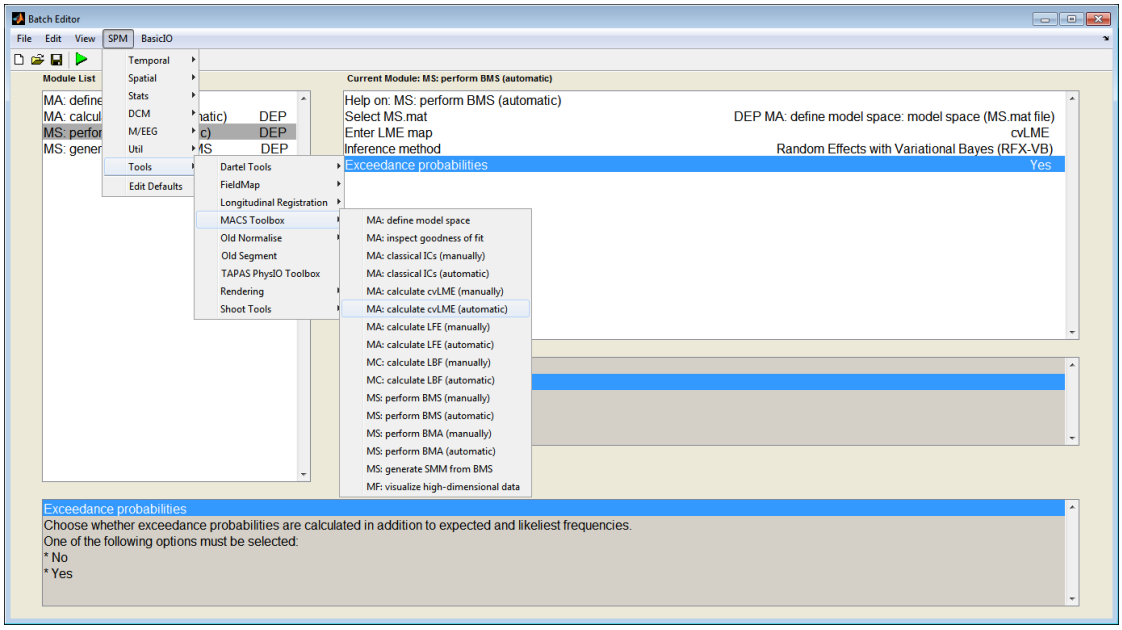

Figure 3. Exemplary batch for Bayesian model selection. This figure shows the SPM batch editor with a typical batch for cross-validated Bayesian model selection (cvBMS) including an active module for group-level Bayesian model selection (BMS). The unfolded menu also hints at where to find features from the MACS toolbox upon successful integration into the SPM software, with the module for automatic cvLME model assessment being highlighted by mouse-over.

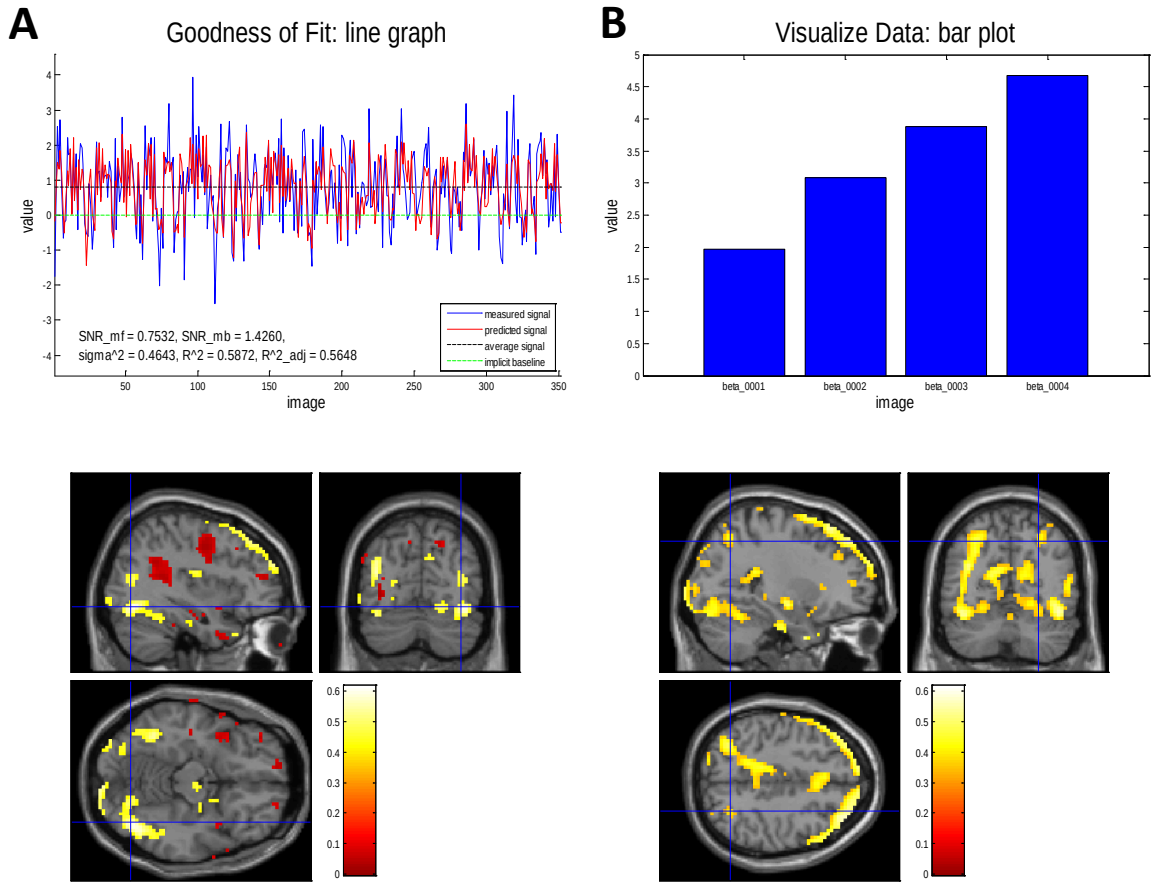

Figure 4. Demonstrations of voxel-wise visualization. This figure shows SPM graphics windows resulting from voxel-wise visualization tools included in the MACS toolbox. (A) Using the module "inspect goodness of fit", measured and predicted fMRI signal are shown alongside goodness-of-fit measures (see Section 2.1.1) which can be browsed for every in-mask voxel included in a particular first-level analysis (see Ashburner et al., 2016, sec. 31.2). The lower panel highlights voxels with the top 5% (yellow, good fit) and the bottom 5% (red, bad fit) in terms of coefficient of determination (R^2) for this GLM. (B) Using the module "visualize high-dimensional data", several data points in each brain region can be displayed and browsed in a voxel-wise fashion. The lower panel highlights voxels in which the R^2 of a GLM (the same as in A) is larger than $0.\bar{3}$. The upper panel bar plots beta estimates from the four experimental condition regressors of this model, indicating a negative effect of both two-level factors in this voxel (see Ashburner et al., 2016, fig. 31.1). Note that this feature may also be used to display voxel-wise posterior probabilities or model frequencies in large model spaces for comparing model preferences in different regions of interest.

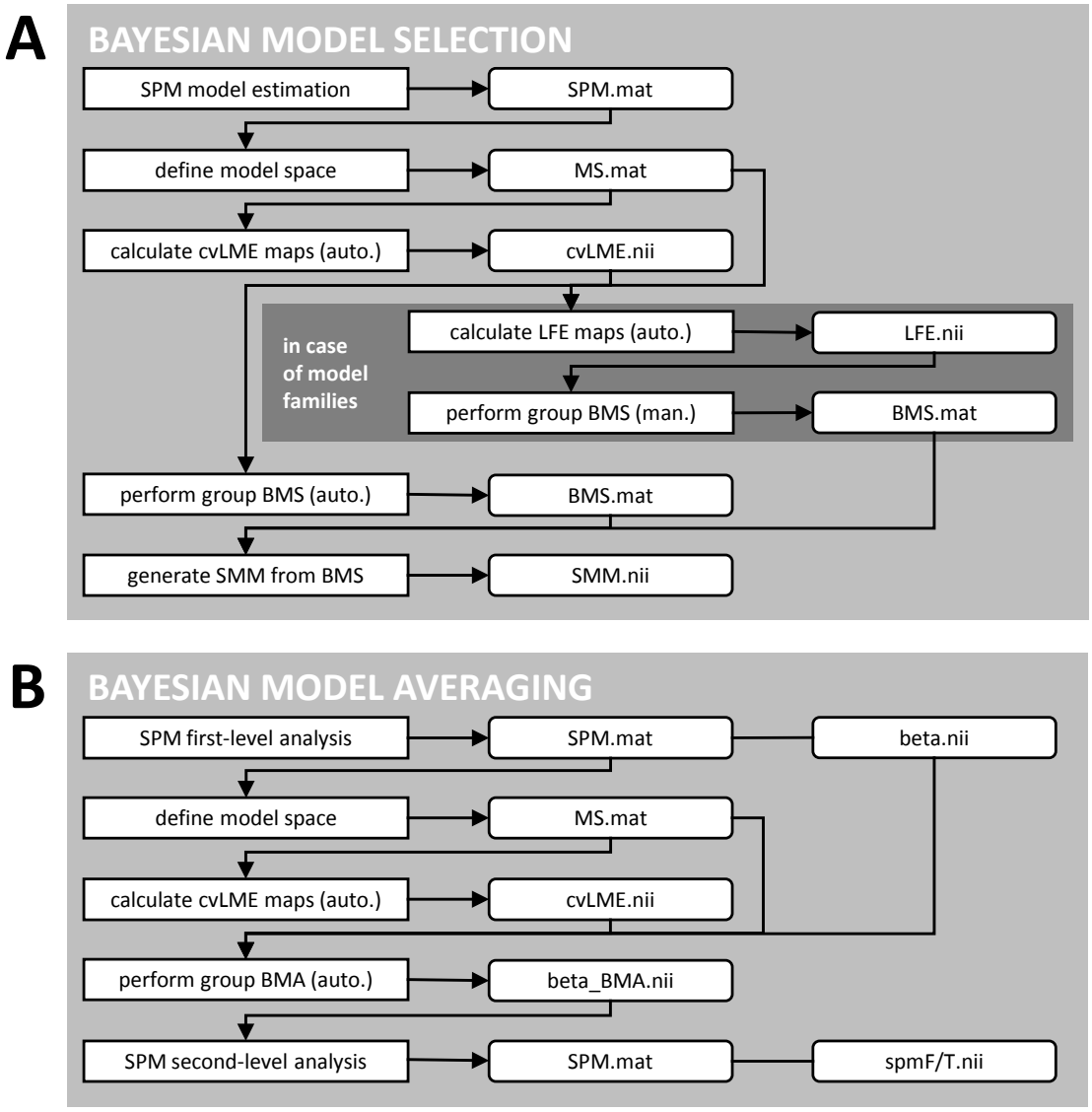

Figure 5. Suggested pipelines for the MACS toolbox. This figure shows processing pipelines underlying the examples reported in this paper (see Section 4). (A) For Bayesian model selection (BMS), SPM model estimation should be followed by definition of the model space, automatic calculation of cvLME maps, automatic group-level BMS and generation of selected-model maps (SMM) from BMS results. In the case that individual models can be grouped into model families, automatic calculation of LFE maps would be followed by manual group-level BMS. (B) For Bayesian model averaging (BMA), SPM first-level analysis should be followed by definition of the model space, automatic calculation of cvLME maps, automatic group-level BMA which invokes subject-level BMAs (see Figure 1) and SPM second-level analysis based on averaged parameters.

Tables

Abbreviation	Feature	Equation	Reference
		Model assessment: goodness-of-fit measures	
R^2	coefficient of determination	$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$	Razavi et al., 2003
R^2_{adj}	coefficient of determination (adjusted)	$R^2_{\mathrm{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p}$	Razavi et al., 2003
$\mathrm{SNR}_{\mathrm{mf}}$	signal-to-noise ratio (model-free)	$\mathrm{SNR} = \frac{\lvert\hat{\mu}\rvert}{\hat{\sigma}},\ \hat{\mu} = \bar{y}$	Welvaert and Rosseel, 2013
$\mathrm{SNR}_{\mathrm{mb}}$	signal-to-noise ratio (model-based)	$\mathrm{SNR} = \frac{\mathrm{var}(X\hat{\beta})}{\hat{\sigma}^2}$	Welvaert and Rosseel, 2013
		Model assessment: classical information criteria	
AIC	Akaike's information criterion	$\mathrm{AIC} = -2\log p(y\lvert\hat{\theta}, m) + 2k$	Penny et al., 2004
$\mathrm{AIC_c}$	Akaike's information criterion (corrected)	$\mathrm{AIC_c} = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1}$	Penny, 2012
BIC	Bayesian information criterion	$\mathrm{BIC} = -2\log p(y\lvert\hat{\theta}, m) + k\log n$	Penny et al., 2004
DIC	Deviance information criterion	$\mathrm{DIC} = -2\log p(y\lvert\,\langle\theta\rangle_{\theta\lvert y}, m) + 2p_D$	Woolrich et al., 2004
		Model assessment: Bayesian log model evidence	
LME	log model evidence	$p(y\lvert m) = \int p(y\lvert\theta, m)\,p(\theta\lvert m)\,\mathrm{d}\theta$	Penny et al., 2007
oosLME	out-of-sample log model evidence	$\mathrm{oosLME}_i = \log\int p(y_i\lvert\theta, m)\,p(\theta\lvert\cup_{j\neq i}\ y_j, m)\,\mathrm{d}\theta$	Soch et al., 2016
cvLME	cross-validated log model evidence	$\mathrm{cvLME} = \sum_{i=1}^{S}\mathrm{oosLME}_i$	Soch et al., 2016
LFE	log family evidence	$p(y\lvert f) = \sum_{m\in f} p(y\lvert m)\,p(m\lvert f)$	Soch et al., 2016
		Model comparison	
ΔIC	information difference	$\Delta\mathrm{IC}_{ij} = \mathrm{IC}(m_i) - \mathrm{IC}(m_j)$	Penny et al., 2004
LBF	log Bayes factor	$\mathrm{LBF}_{ij} = \mathrm{LME}(m_i) - \mathrm{LME}(m_j)$	Penny et al., 2004
LGBF	log group Bayes factor	$\mathrm{LGBF}_{ij} = \log p(Y\lvert m_i) - \log p(Y\lvert m_j)$	Stephan et al., 2007
PP	posterior model probability	$p(m_i\lvert y) = \frac{\exp(\mathrm{LBF}_{ij})}{\exp(\mathrm{LBF}_{ij})+1}$	Penny and Ridgway, 2013
		Model selection	
FFX BMS	fixed-effects Bayesian model selection	$p(m_i\lvert Y) = \frac{p(Y\lvert m_i)\,p(m_i)}{\sum_{j=1}^{M} p(Y\lvert m_j)\,p(m_j)}$	Penny et al., 2010
RFX BMS (VB)	random-effects Bayesian model selection (Variational Bayes)	see eq. 14 in reference	Stephan et al., 2009
RFX BMS (GS)	random-effects Bayesian model selection (Gibbs Sampling)	see eqs. 13-17 in reference	Penny et al., 2010
BMA	Bayesian model averaging	$\hat{\beta}_{\mathrm{BMA}} = \sum_{i=1}^{M}\hat{\beta}_i \cdot p(m_i\lvert y)$	Soch et al., 2017

Table 1. Features in the MACS toolbox. This table summarizes methods of model inference in MACS. References correspond to exemplary applications in the fMRI literature. For the respective statistical literature, confer the list of features (see Section 2).

Modules	**time needed to create batch**	**time needed to execute batch**
Model assessment: repetition priming		
Step 1: goodness of fit of categorical model	1 model	
inspect GoF	< 1 min	< 00:01 h
visualize HD data	1 min	< 00:01 h
Step 2: categorical model vs. parametric model	2 models	
model space, ICs (auto.), LBF (auto.)	2 min	< 00:01 h
Model selection: orientation pop-out		
Step 1: categorical GLM vs. parametric GLMs	3 models / 2 families	
model space, cvLME (auto.), LFE (auto.)	15 min	02:38 h
BMS (man.), SMM from BMS	15 min	< 00:01 h
Step 2: linear-parametric vs. nonlinear-parametric GLM	2 models	
model space, BMS (auto.), SMM from BMS	3 min	< 00:02 h
Step 3: nonlinear-parametric GLM with left vs. right OC	2 models	
model space, cvLME (auto.), BMS (auto.), SMM from BMS	8 min	01:43 h
Model averaging: conflict adaptation		
Step 1: averaged model parameter estimates	4 models	
model space, cvLME (auto.), BMA (auto.), BMS (auto.), SMM	25 min	07:05 h
Step 2: second-level statistical analysis	1 (averaged) model	
factorial design, model estimation, contrast manager (SPM)	10 min	< 00:01 h

Table 2. Demonstrations using the MACS toolbox. This table lists processing steps used in first-level model assessment or belonging to cvBMS and cvBMA pipelines (see Section 4). The second and third column give times required for specifying and running each batch in the SPM batch editor. The person that created the batch modules had 7 years of experience with SPM and was listening to Disc 1 of The Beatles' *White Album* while operating the MACS toolbox.

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., 2016. SPM12 Manual. URL: http://www.fil.ion.ucl.ac.uk/spm/doc/spm12_manual.pdf.

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

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.

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.

Friston, K., Harrison, L., Penny, W., 2003. Dynamic causal modelling. NeuroImage 19, 1273–1302. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811903002027, doi: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. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811906008822, doi:10.1016/j.neuroimage.2006.08.035.

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.

Gläscher, J.P., O'Doherty, J.P., 2010. Model-based approaches to neuroimaging: combining reinforcement learning theory with fMRI data. Wiley Interdisciplinary Reviews: Cognitive Science , n/a–n/aURL: http://doi.wiley.com/10.1002/wcs.57, doi:10.1002/wcs.57.

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.

Hurvich, C.M., Tsai, C.L., 1989. Regression and Time Series Model Selection in Small Samples. Biometrika 76, 297. URL: http://www.jstor.org/stable/2336663?origin=crossref, doi:10.2307/2336663.

Kass, R.E., Raftery, A.E., 1995. Bayes Factors. Journal of the American Statistical Association 90, 773–795. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572, doi:10.1080/01621459.1995.10476572.

Luo, W.L., Nichols, T.E., 2003. Diagnosis and exploration of massively univariate neuroimaging models. NeuroImage 19, 1014–1032. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811903001496, doi:10.1016/S1053-8119(03)00149-6.

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.

Nichols, T.E., 2013. Building Confidence in fMRI Results with Model Diagnosis. URL: https://www.pathlms.com/ohbm/courses/250/sections/2156/slide_presentations/17067.

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., Kiebel, S., Friston, K., 2003. Variational Bayesian inference for fMRI time series. NeuroImage 19, 727–741. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811903000715, doi:10.1016/S1053-8119(03)00071-5.

Penny, W., Stephan, K., Mechelli, A., Friston, K., 2004. Comparing dynamic causal models. NeuroImage 22, 1157–1172. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811904001648, doi: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. URL: http://dx.plos.org/10.1371/journal.pone.0059655, doi: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 Computational Biology 6, e1000709. URL: http://dx.plos.org/10.1371/journal.pcbi.1000709, doi: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. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811904004975, doi:10.1016/j.neuroimage.2004.08.034.

Pernet, C., Poline, J.B., 2015. Improving functional magnetic resonance imaging reproducibility. GigaScience 4. URL: http://www.gigasciencejournal.com/content/4/1/15, doi: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. Human Brain Mapping 20, 227–238. URL: http://doi.wiley.com/10.1002/hbm.10141, doi:10.1002/hbm.10141.

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

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.

Seghouane, A.K., Ong, J.L., 2010. A Bayesian model selection approach to fMRI activation detection, in: 2010 IEEE International Conference on Image Processing, IEEE. pp. 4401–4404. URL: http://ieeexplore.ieee.org/document/5653354/, doi:10.1109/ICIP.2010.5653354.

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.

Soch, J., Meyer, A.P., Haynes, J.D., Allefeld, C., 2017. How to improve parameter estimates in GLM-based fMRI data analysis: cross-validated Bayesian model averaging. bioRxiv URL: http://biorxiv.org/lookup/doi/10.1101/095778, doi:10.1101/095778.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64, 583–639. URL: http://doi.wiley.com/10.1111/1467-9868.00353, doi:10.1111/1467-9868.00353.

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. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811909002638, doi:10.1016/j.neuroimage.2009.03.025.

Stephan, K.E., Weiskopf, N., Drysdale, P.M., Robinson, P.A., Friston, K.J., 2007. Comparing hemodynamic models with DCM. NeuroImage 38, 387–401. URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811907006489, doi:10.1016/j.neuroimage.2007.07.040.

Vuong, Q.H., 1989. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica 57, 307. URL: http://www.jstor.org/stable/1912557?origin=crossref, doi:10.2307/1912557.

Welvaert, M., Rosseel, Y., 2013. On the Definition of Signal-To-Noise Ratio and Contrast-To-Noise Ratio for fMRI Data. PLoS ONE 8, e77089. URL: http://dx.plos.org/10.1371/journal.pone.0077089, doi:10.1371/journal.pone.0077089.

Woolrich, M., Jenkinson, M., Brady, J., Smith, S., 2004. Fully Bayesian Spatio-Temporal Modeling of FMRI Data. IEEE Transactions on Medical Imaging 23, 213–231. URL: http://ieeexplore.ieee.org/document/1263611/, doi:10.1109/TMI.2003.823065.