

# Efficient Posterior Probability Mapping Using Savage-Dickey Ratios

William D. Penny\*, Gerard R. Ridgway

Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College, London, United Kingdom

## Abstract

Statistical Parametric Mapping (SPM) is the dominant paradigm for mass-univariate analysis of neuroimaging data. More recently, a Bayesian approach termed Posterior Probability Mapping (PPM) has been proposed as an alternative. PPM offers two advantages: (i) inferences can be made about effect size thus lending a precise physiological meaning to activated regions, (ii) regions can be declared inactive. This latter facility is most parsimoniously provided by PPMs based on Bayesian model comparisons. To date these comparisons have been implemented by an Independent Model Optimization (IMO) procedure which separately fits null and alternative models. This paper proposes a more computationally efficient procedure based on Savage-Dickey approximations to the Bayes factor, and Taylor-series approximations to the voxel-wise posterior covariance matrices. Simulations show the accuracy of this Savage-Dickey-Taylor (SDT) method to be comparable to that of IMO. Results on fMRI data show excellent agreement between SDT and IMO for second-level models, and reasonable agreement for first-level models. This Savage-Dickey test is a Bayesian analogue of the classical SPM-F and allows users to implement model comparison in a truly interactive manner.

**Citation:** Penny WD, Ridgway GR (2013) Efficient Posterior Probability Mapping Using Savage-Dickey Ratios. *PLoS ONE* 8(3): e59655. doi:10.1371/journal.pone.0059655

**Editor:** Kewei Chen, Banner Alzheimer's Institute, United States of America

**Received:** August 13, 2012; **Accepted:** February 19, 2013; **Published:** March 22, 2013

**Copyright:** © 2013 Penny, Ridgway. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This work was supported by the Wellcome Trust [grant number 091593/Z/10/Z] and the Medical Research Council [grant number MR/J014257/1]. The Wellcome Trust Centre for Neuroimaging is supported by core funding from the Wellcome Trust [grant number 091593/Z/10/Z]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: w.penny@ucl.ac.uk

## Introduction

Bayesian inference has been applied to the analysis of fMRI data in multiple domains, ranging from connectivity analysis [1–4], group analysis [5,6], haemodynamic modelling [7], spatial modelling [8], and state-space approaches [9,10]. Generically, the advantage of these Bayesian approaches is that they allow for seamless incorporation of prior knowledge and employ established procedures for parameter regularization and model selection. Bayesian methods have also been widely used in the MEG/EEG domain for tackling the problems of source reconstruction [11,12] and biologically informed connectivity analysis [13,14]. The development and application of Bayesian methods to neuroimaging is described in recent reviews [15,16]. The focus of this paper is a Bayesian method for the mass-univariate analysis of neuroimaging data, known as Posterior Probability Mapping (PPMs). Previously, PPMs have been proposed as a Bayesian alternative to Statistical Parametric Maps (SPMs) [17,18]. PPMs can be applied to several common neuroimaging modalities (fMRI, PET, MEG, EEG) and provide estimates of effect size that are informed by empirical priors.

PPMs address a key limitation of classical frequentist inference: while a small p-value allows rejection of the null hypothesis, a large p-value does not permit its acceptance. Informally, absence of evidence is not evidence of absence. Bayesian model comparison, on the other hand, can find either the null or alternative hypothesis more probable [19,20]. This enables imaging neuroscientists to infer that regions have not activated and so allows detection of double dissociations among brain regions and

cognitive processes. To date, this model comparison procedure has been implemented by estimating multiple models and computing the evidence for each, which is prohibitively time-consuming for investigating multiple hypotheses. This paper introduces a more computationally efficient method based on the Savage-Dickey ratio [21,22]. Before describing the method we review relevant concepts in Bayesian neuroimaging. Readers requiring a more comprehensive background to Bayesian inference are referred to standard texts [23,24].

## PPMs for Parameter Inference

PPMs are similar to SPMs in that they are also based on a mass univariate approach in which General Linear Models (GLMs) are fitted to data at each voxel [25]. They differ however in the statistical method used to estimate parameters and make inferences. Estimates of the GLM parameters, for example, are constrained using empirical priors.

Early work on Bayesian fMRI considered mass-univariate approaches to modeling spatial dependencies in the signal and noise. For example, Gossel et al. [8] proposed a separable spatio-temporal model where these spatial dependencies were characterized using Markov Random Field (MRF) priors. More recently, Woolrich et al. [18] described a Bayesian model of fMRI in which the noise process was characterized by separable or nonseparable spatio-temporal models. Both of these approaches used Markov Chain Monte Carlo (MCMC) to perform posterior inference, which is computationally expensive.

We have previously proposed a non-spatial PPM procedure employing global shrinkage priors which shrink parameter estimates toward zero [17]. We have additionally developed a PPM approach specifically for within-subject fMRI time series [26]. This allows users to specify either global shrinkage priors, or spatial priors based on Gauss-Markov Random Fields (GMRFs) which constrain effect sizes to be similar at nearby voxels. These models are particularly suited to within-subject fMRI, as the error correlations can be modelled using arbitrary-order voxel-specific autoregressive (AR) models. These AR models accurately describe the physiological noise processes in fMRI data [27]. Later work allows for spatial priors on the AR parameters [20] and the approach has been extended to incorporate spatial priors based on wavelets [28] and Gaussian processes [29].

For the above approaches, the result of the estimation is a posterior distribution of effect size at each voxel,  $p(c_i|y)$ , where  $c_i = c^T w_i$  is a linear combination or ‘contrast’ of the GLM parameters at the  $i$ th voxel,  $w_i$ . These voxel-wise posterior distributions or PPMs are visualised by specifying two thresholds – an effect size threshold,  $c_T$ , and a posterior probability threshold  $p_T$  – and plotting voxels for which  $p(c_i > c_T|y) > p_T$ . Depending on the software, what is actually plotted can be the posterior probability or the effect size itself. One may also have the option of plotting the log posterior odds,  $\log p/(1-p)$ , which improves the visualisation for voxels that have posterior probabilities close to unity.

Inferences based on PPMs thus allow researchers to be more specific as to the effects in which they are interested. For example, effect sizes less than 0.1% of the global mean may be deemed physiologically irrelevant (see also a related though less principled method to avoid declaring voxels with trivial effect sizes significant (in a frequentist sense) due to artefactually low variance [30]). An alternative perspective is that needing to specify an additional arbitrary threshold (the effect-size threshold) may be seen as a disadvantage of the method. This has motivated the development of PPMs for model inference.

### PPMs for Model Inference

We first distinguish between nested and non-nested model inference. In nested model inference, a ‘nested’ model can be formulated as a special case of a more general ‘full’ model. For example, nested models may be constructed by removing one or more explanatory variables from the full model. When models are not related in this way they are said to be non-nested. This will be the case if each model has its own unique set or subset of explanatory variables that are not found in the other model.

For non-nested model inference we can proceed by separately fitting the models of interest, computing the model evidence for each, and then plotting a map of the posterior model probability or log Bayes factor. This procedure, which we refer to as Independent Model Optimization (IMO), is straightforward because the evidence of a GLM can be computed exactly [31,32]. This is not the case for nonlinear models, such as the Dynamic Causal Models used in the study of brain connectivity [1].

This model inference approach has been applied in the context of within-subject models of fMRI time series [20], and allows one to compute a model evidence map; a map of (log) model evidence as a function of space. If one computes a model evidence map for each model of interest, and for each subject in a group, then one can make an inference at the population level as to which model is the most prevalent [33]. The method can accommodate any number of models (not just a null model and a single alternative). This approach has been used, for example, to show that in a forced-

choice decision task, anterior brain regions integrate contextual information over longer time periods than do posterior regions [34].

To show that a brain region does not activate requires a strong Bayes factor in favour of the null model over the alternative model for the data in that region. This inference requires the specification of a single parameter, namely what is meant by ‘strong’. Here we can refer to established scales of strengths of evidence [22,35] where, for example, a Bayes factor of at least 20 (or log Bayes factor of at least 3) corresponds to strong evidence. It is also possible to declare that a region does not activate using PPMs for parameter inference, but this requires specification of an additional parameter – the effect size threshold [17]. The model comparison approach is therefore more parsimonious.

Whilst PPMs based on model inference are a powerful paradigm for the analysis of fMRI time series, they are somewhat computationally demanding, because for every model comparison one wishes to make, it is necessary to fit all models over the spatial domain of interest, and compute the evidence for each. If one has a small region of interest this is less of an issue, but whole-brain analyses can require tens of minutes of fitting time for each model to be considered.

We now describe the special case of nested model comparison. Previously, we have proposed an analogue of the classical F-test, which instead uses a  $\chi^2$  test based on the posterior density [36]. The resulting test is conceptually rather unsatisfactory, however, as it implements a classical inference based on a Bayesian posterior density. This paper proposes replacing the  $\chi^2$  test with an inference based on the Savage-Dickey ratio. As we shall see, this new approach will also provide a computationally efficient method for non-nested model comparison. This extends recent work in brain connectivity analysis where we have proposed [37] and validated [38] a generalisation of the Savage-Dickey approach in the context of Dynamic Causal Modelling [39].

## Methods

This section first describes Bayesian model and parameter inference for the GLM. We then describe the statistical tests for nested and non-nested model comparison including the Savage-Dickey ratio. In our implementation of Posterior Probability Mapping (PPM) we do not store posterior covariance matrices as this would require a prohibitive amount of computer disk space. Instead, we store a small number of hyperparameters to reconstruct the covariance matrices using a Taylor series approximation. This additional step is described in a later subsection. We also show how PPMs can be derived for both first- and second-level models. In what follows  $N(x; \mu, \Sigma)$  denotes a multivariate Gaussian distribution with mean vector  $\mu$  and covariance matrix  $\Sigma$ , of which  $|\Sigma|$  is the determinant.

### Bayesian General Linear Model

We consider Bayesian inference for GLMs with data  $y$ , design matrix  $X$  and regression coefficients  $w$ . We assume a Gaussian prior over regression coefficients

$$p(w|m) = N(w; w_m, \Sigma_m) \quad (1)$$

where  $w_m$  and  $\Sigma_m$  are the prior mean and covariance for model  $m$ . In most applications to fMRI [17,26] the prior mean is set to zero, and the prior covariance is estimated using multiple time series over a spatial region. This is described in more detail below in the section on Empirical Bayes. The variable  $m$  symbolises the model assumptions. Different models are usually thought of as being

specified by having different design matrices. In GLMs a single parameter is associated with each column of the design matrix, therefore different models have different parameters. It is also possible to conceive of different models as having different priors, hence the notation above. For example, subspaces of the design matrix can be eliminated by setting the corresponding parameters to have zero prior mean and zero prior variance.

We also assume a Gaussian likelihood

$$p(y|w,m) = \mathbf{N}(y; Xw, \Sigma_y) \tag{2}$$

where  $\Sigma_y$  is the observation noise covariance matrix. Like the prior covariance, the noise covariance is typically estimated from the data, as described in the section on Empirical Bayes. Given a Gaussian prior and likelihood, the posterior over regression coefficients is also Gaussian [40]

$$p(w|y,m) = \mathbf{N}(w; w_N, \Sigma_N) \tag{3}$$

with posterior mean  $w_N$  and posterior covariance  $\Sigma_N$  given by

$$\begin{aligned} \Sigma_N^{-1} &= X^T \Sigma_y^{-1} X + \Sigma_m^{-1} \tag{4} \\ w_N &= \Sigma_N \left( X^T \Sigma_y^{-1} y + \Sigma_m^{-1} w_m \right) \end{aligned}$$

Bayesian inference over models is implemented by first computing the model evidence  $p(y|m)$ . If one has a prior distribution over a set of models,  $\mathcal{M}$ , this can be updated into a posterior distribution using Bayes rule and the model evidence

$$p(m|y) = \frac{p(y|m)p(m)}{\sum_{m' \in \mathcal{M}} p(y|m')p(m')} \tag{5}$$

For pairs of models with equal model priors,  $p(m_1) = p(m_2)$ , inference can be made based on the Bayes factor [22]. The Bayes factor for model  $m_1$  versus  $m_2$  is given by

$$BF_{12} = \frac{p(y|m_1)}{p(y|m_2)} \tag{6}$$

$$\log BF_{12} = \log p(y|m_1) - \log p(y|m_2)$$

For GLMs, assuming known  $\Sigma_m$  and  $\Sigma_y$ , the log model evidence,  $\log p(y|m)$ , can be computed as

$$\begin{aligned} L_m &= -\frac{1}{2} e_y^T \Sigma_y^{-1} e_y - \frac{1}{2} \log |\Sigma_y| - \frac{N}{2} \log 2\pi \tag{7} \\ &\quad - \frac{1}{2} e_w^T \Sigma_m^{-1} e_w - \frac{1}{2} \log |\Sigma_m| + \frac{1}{2} \log |\Sigma_N| \end{aligned}$$

where the ‘prediction errors’ are

$$\begin{aligned} e_y &= y - Xw_N \\ e_w &= w_N - w_m \end{aligned} \tag{8}$$

Unequal model priors are accommodated by making inferences using posterior odds ratios, instead of Bayes factors. The posterior odds is equal to the prior odds times the Bayes factor

$$\frac{p(m_1|y)}{p(m_2|y)} = \frac{p(y|m_1)p(m_1)}{p(y|m_2)p(m_2)} \tag{9}$$

Taking logs gives

$$\log \frac{p(m_1|y)}{p(m_2|y)} = \log BF_{12} + \log \frac{p(m_1)}{p(m_2)} \tag{10}$$

Thus if  $m_1$  is 100 times less likely a priori than  $m_2$ , the log posterior odds equals the log Bayes factor minus  $\log 0.01 = -4.6$ . Hence, unequal prior odds can be dealt with by a simple change to the decision threshold.

### Empirical Bayes

We first discuss the approach to second-level fMRI analysis which is described in [17]. This takes an Empirical Bayes approach which estimates parameters of the prior  $p(w)$  using data from all voxels in the search region. The prior mean is set to zero,  $w_m = 0$ , and the prior covariance is assumed diagonal  $\Sigma_m = \text{diag}(\alpha_m^{-1})$  with the  $k$ th element of  $\alpha_m$  denoting the prior precision of the  $k$ th parameter. The observation noise covariance matrix at the  $i$ th voxel, is then parameterised as

$$\Sigma_y(i) = \lambda_i V \tag{11}$$

where  $\lambda_i$  is a single voxel specific hyperparameter and  $V$  is a matrix which captures the global observation noise structure and has been estimated in a previous step. The hyperparameters  $\lambda_i$  and  $\alpha_m$  are then set to maximise the model evidence using an Empirical Bayes approach [17]. This optimisation does not require the model evidence itself to be computed.

For first-level models the approach is similar. The main difference is that  $\Sigma_y$  is set to accommodate voxel-wise Auto-Regressive (AR) noise processes of arbitrary order, so as to absorb aliased temporal fluctuations due for example to respiration and heartbeat. Here,  $\Sigma_y$  is parameterised using voxel-specific AR parameters. It is possible to set the AR model order to zero, in which case the likelihood reduces to that for the standard GLM. For the first-level models the priors can be either set as ‘global shrinkage priors’, which are identical to the second-level priors described above, or as spatial priors which encourage parameter estimates to be similar at nearby voxels [26]. All the hyperparameters are estimated together, along with the prior precisions using Empirical Bayes [26]. This paper is primarily concerned with evaluation of the Savage-Dickey approach for global shrinkage priors.

For the above Empirical Bayes approaches, the expression for the log model evidence in equation 7 should be augmented with penalty terms to accommodate the uncertainty in the estimation of the associated hyperparameters. These terms are provided, for example, in equations 8 and 13 in [20]. For the results in this paper the inclusion of these extra terms made little or no quantitative difference so, for ease of communication, the IMO results presented in this paper are based on equation 7.

### Nested Model Comparison

This section describes the Savage-Dickey approach for nested model comparison. If model  $m_1$  is nested within  $m_2$  where the models have common parameters  $w_1$  and  $m_2$  has additional parameters  $w_2$ , then the Bayes factor can be rewritten as follows. First, we write the evidence for model 2 given that  $w_2 = 0$

$$p(y|w_2=0, m_2) = \int p(y|w_2=0, w_1, m_2) p(w_1|m_2) dw_1 \quad (12)$$

Because we have a nested model the likelihood term  $p(y|w_2=0, w_1, m_2) = p(y|w_1, m_1)$ . This is the mathematical definition of a nested model. Second, if it is the case that the priors over the common parameters are the same for the two models,  $p(w_1|m_2) = p(w_1|m_1)$ , then we can write

$$p(y|w_2=0, m_2) = \int p(y|w_1, m_1) p(w_1|m_1) dw_1 = p(y|m_1) \quad (13)$$

Substituting into the Bayes factor (equation 6) gives

$$BF_{12} = \frac{p(y|w_2=0, m_2)}{p(y|m_2)} \quad (14)$$

Using Bayes rule over the posterior of  $m_2$  gives

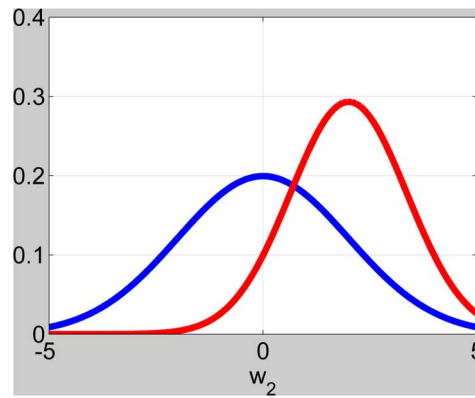
$$p(w_2=0|y, m_2) = \frac{p(y|w_2=0, m_2) p(w_2=0|m_2)}{p(y|m_2)} \quad (15)$$

We can therefore see that

$$BF_{12} = \frac{p(w_2=0|y, m_2)}{p(w_2=0|m_2)} \quad (16)$$

The formula makes intuitive sense and is known as the Savage-Dickey ratio [21]. If we believe it is more likely that parameters are zero after seeing the data than before, then  $BF_{12} > 1$  and we have evidence in favour of the nested model. Figure 1 illustrates the opposite case for a simple one-dimensional example. For nested model comparisons the Bayes factor can therefore be computed by fitting just the larger model. If the priors over the common parameters are not the same for two models then a correction factor, based on a sampling approach, can be computed [41].

The above procedure can be generalised to consider non-zero hypothesized values, and nested models defined as subspaces of full models. This is implemented using the usual approach of defining contrasts for linear models [42]. A single contrast vector, for example, can be used to specify a single hypothesis, whereas multiple contrast vectors combined into a matrix can be used to specify a compound hypothesis. For example, if  $C^T = [1, -1]$  then  $C^T w = 0$  specifies the single hypothesis that  $w_1 = w_2$ . Similarly, if  $C^T = [1, -1, 0; 0, 1, -1]$  then  $C^T w = 0$  specifies the compound hypothesis that  $w_1 = w_2$  and  $w_2 = w_3$ . This latter compound hypothesis is rejected if  $w_1 \neq w_2$  or  $w_2 \neq w_3$ . This type of contrast matrix is used, for example, in testing for main effects in factorial designs. More details on hypothesis testing in linear models can be found in standard textbooks [42].



**Figure 1.** The figure shows the prior density  $p(w_2|m_2)$  in blue and the posterior density  $p(w_2|m_2, y)$  in red. Here  $BF_{12} = 0.5$ , weakly favouring the more complex model  $m_2$ , since the parameter  $w_2$  is half as likely to be zero after seeing the data than before. doi:10.1371/journal.pone.0059655.g001

We now consider the use of contrasts for the case of Gaussian priors and posteriors. The Savage-Dickey approximation to the Bayes factor in favour of the alternative hypothesis (full model) over a particular null hypothesis (nested model  $i$ ) is given by

$$BF_{f,i} = \frac{p(C_i^T w = 0|m)}{p(C_i^T w = 0|y, m)} \quad (17)$$

where  $C_i$  is a contrast matrix and  $w$  are the regression coefficients. The Savage-Dickey ratio compares the probability density for the null hypothesis under the prior versus under the posterior. If it is a posteriori less likely then  $BF_{f,i}$  will be large, favouring the full model (as shown in Figure 1).

Given that the prior and posterior are both Gaussians this can be evaluated as

$$BF_{f,i} = \frac{N(0; \mu_0, S_0)}{N(0; \mu_N, S_N)} \quad (18)$$

where

$$\begin{aligned} \mu_0 &= C_i^T w_m \\ \mu_N &= C_i^T w_N \\ S_0 &= C_i^T \Sigma_m C_i \\ S_N &= C_i^T \Sigma_N C_i \end{aligned} \quad (19)$$

If the prior mean is  $w_m = 0$  (as it is for Bayesian GLMs implemented in SPM [43]) the log Bayes factor simplifies to

$$\log BF_{f,i} = \frac{1}{2} \mu_N^T S_N^{-1} \mu_N + \frac{1}{2} \log \frac{|S_N|}{|S_0|} \quad (20)$$

The Savage-Dickey ratio is exact if  $w_m$ ,  $\Sigma_m$  and  $\Sigma_y$  are identical for the two models being computed. Under these conditions it will give identical results to IMO. Most practical implementations of Bayesian inference for fMRI, however, set  $w_m = 0$  and use an Empirical Bayes procedure to estimate  $\Sigma_m$  and  $\Sigma_y$ . These parameters will therefore differ between models.

Consider, for example, the estimation of  $\Sigma_y$  when comparing a simple and complex model. If the simpler model is true then the error variances are likely to be very similar, whereas if the complex model is true then the error variances for the complex model are likely to be smaller. A redeeming feature of error variance estimation, however, is that these estimates are corrected for the degrees of freedom in the model. The effect of Empirical Bayes estimates is addressed empirically at the beginning of the results section.

### Non-nested Model Comparison

The previous section has shown how to compute  $\log BF_{f,i}$  which is the log Bayes factor of the full model with respect to the reduced model defined by contrast  $C_i$ . We can also consider a second contrast  $C_j$  and its associated term  $\log BF_{f,j}$ . Note that the contrasts  $C_i$  and  $C_j$  can define two separate subspaces of the full model, for example, by loading onto different sets of regressors in the design matrix. This means that model  $i$  need not be nested within model  $j$  or vice-versa. The only requirement is that both are nested within the full model  $f$ .

One can then combine the two log Bayes factors to get  $\log BF_{i,j}$  thus providing a procedure for the comparison of non-nested models. We have

$$BF_{ij} = \frac{p(y|m_i)}{p(y|m_j)} = \frac{p(y|m_i) p(y|m_f)}{p(y|m_j) p(y|m_f)} \quad (21)$$

Hence

$$\log BF_{ij} = \log BF_{f,i} - \log BF_{f,j} \quad (22)$$

This idea has been proposed in the Bayesian model selection literature [22] and has been employed [37] and validated [38] in the context of Dynamic Causal Models.

### Group Analysis

The implementation of non-nested model comparisons is based on the log Bayes factor images created as previously described. One can then compute differences in these, as indicated in equation 22, and enter these difference images into a group analysis. For nested model comparisons the log Bayes factor images, computed using equation 20, can also enter a group analysis in the same way.

To make model inferences regarding the population from which subjects were drawn one can use the same random-effects (RFX) model selection procedure as described previously [33]. Here the ‘random-effect’ is a discrete variable which indexes which model each subject uses. This presents an alternative to the standard group analysis which implements a random effects analysis over the parameters of a model. This RFX parameter inference procedure is described in standard references [25] and makes use of ‘second-level’ models.

RFX parameter inference looks for group effect sizes which are consistent in relation to the between-subject variability whereas RFX model comparison looks for the models which have the highest frequency in the population. If some subjects show strong negative and others strong positive effects then this could be detected with RFX model comparison but not with RFX parameter inference. Conversely, if there is a consistently signed but small effect RFX parameter inference may be more sensitive.

### Taylor Series Approximation

In our implementation of the above Bayesian estimation algorithms, the full voxel-wise posterior covariance matrices are not explicitly stored as this would require a prohibitive amount of disk space. For GLMs with  $k$  parameters each covariance matrix comprises  $k(k+1)/2$  real numbers. For brain images comprising  $v$  voxels this gives a total of  $S = vk(k+1)/2$  real numbers to store. For example, for  $k = 20$  and  $v = 50,000$  we have  $S = 1.05 \times 10^6$  or 210 images. Instead we store a small number of parameters that allow us to reconstruct these covariance matrices using a first-order Taylor series approximation. For example, in the ‘second-level’ PPM approach [17] the posterior covariance (4) at voxel  $i$  depends on  $\lambda_i$  via the noise covariance (11),

$$\Sigma_N^{-1}(\lambda_i) = X^T \Sigma_y^{-1} X + \Sigma_m^{-1} = X^T (\lambda_i V)^{-1} X + \Sigma_m^{-1} \quad (23)$$

where  $\lambda_i$  is a single voxel specific hyperparameter and  $V$  is a matrix which captures the observation noise structure and has been estimated in a previous step [17]. These hyperparameters  $\lambda_i$  are the same quantities referred to in the above section on Empirical Bayes. Viewed as a function of a continuous parameter  $\lambda$ ,  $\Sigma_N(\lambda)$  can be analytically differentiated, allowing the posterior covariance to be approximated using a first-order Taylor series

$$\Sigma_N(\lambda_i) = \Sigma_N(\bar{\lambda}) + \frac{d\Sigma_N}{d\lambda} \Big|_{\lambda=\bar{\lambda}} (\lambda_i - \bar{\lambda}) \quad (24)$$

where  $\bar{\lambda}$  is the mean hyperparameter averaged over the volume of interest. Thus we need only store a single voxel specific hyperparameter,  $\lambda_i$ ,  $\bar{\lambda}$  and the single Jacobian matrix  $\frac{d\Sigma_N}{d\lambda}$  evaluated at  $\bar{\lambda}$ . Thus for  $v$  voxels the total storage required is  $S = v + k(k+1)/2 + 1$ . This breaks down as  $v$  for the  $\lambda_i$ ,  $k(k+1)/2$  for the Jacobian and one for  $\bar{\lambda}$ . For our numerical example this gives  $S = 5.02 \times 10^4$  or between 1 and 2 images. This requires less storage by a factor of over 200.

A similar Taylor series approach is used for first-level models [44]. The fact that we will not be using the exact posterior distributions to compute the Savage-Dickey ratios in equation 20 will create an extra level of approximation in the computation of log Bayes factors. We therefore refer to the overall approach as the Savage-Dickey-Taylor (SDT) method.

### Summary

We have described the use of Savage-Dickey ratios initially for the case of nested model comparisons. This brings about a natural symmetry with classical inference based on SPMs. For SPMs there are two types of test. The SPM-t allows one to test for one-sided effects. The Bayesian analogue of the SPM-t is the PPM for parameter inference. The SPM-F allows one to test for two-sided effects for both uni-dimensional or multi-dimensional contrasts (the contrast matrix  $C_i$  has a single row, or multiple rows). The Bayesian analogue of this is the Savage-Dickey test for a nested model comparison.

We have also shown how the Savage-Dickey approach can be used for non-nested model comparison. Importantly, whether the comparison is nested or non-nested the computational saving is great, because we only need to estimate a single full model. To save storage space, practical implementations of these Bayesian algorithms reconstruct posterior parameter covariance matrices using a Taylor series approach. We therefore describe our overall approach as the Savage-Dickey-Taylor (SDT) method. In what follows we compare the proposed SDT method for model

inference with the previously proposed Independent Model Optimization (IMO) approach, which requires separate fitting of full and nested models.

### fMRI Data

We present first- and second-level analyses of data from an fMRI study of face processing. The data were collected to study neuronal responses to images of faces and are available from the SPM web site [43]. For a full description of this data set and similar analyses see [45]. Each face was presented twice and faces belonged to either familiar ('F') or unfamiliar ('N') people which gave rise to four conditions ('N1', 'N2', 'F1', 'F2'). For the first-level analyses hemodynamic responses were modelled with a single 'canonical' hemodynamic basis function [43]. Together with a constant column, this gives rise to a design matrix containing five columns which we refer to below as the 'standard' first-level model. We use this standard model to analyse data from a single subject.

The second-level analysis (RFX parameter inference) proceeds as follows. Data from 12 subjects were first analysed using 12 separate first-level models. These were not the standard model, as above, but treated all face presentations as a single event type. Responses were then modelled using a 12 time bin Finite Impulse Response (FIR) model as described in the Group analysis section of the SPM manual. Each time bin was 2 s wide thus covering a 24 s post-stimulus epoch. First-level contrasts were then used to produce summary statistic images for each time bin and for each subject. This resulted in 144 images which were used as data for the second-level models described in the results section below.

### Results

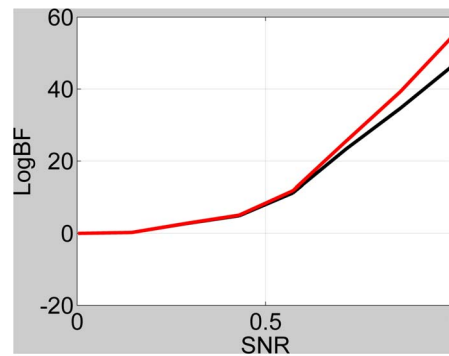
We first investigated the accuracy of the Savage-Dickey (SD) approximation using simulation studies to assess the effect of empirical estimation of observation noise and prior precision. We also assess the effect of the Taylor approximation. We then report the accuracy of SDT versus IMO on empirical first- and second-level fMRI data.

#### Observation Noise

As noted in the theory section, SD is exact if the likelihoods, and therefore the observation noise parameters, are the same between models. However, in practice the observation noise parameters are estimated from the data. Our simulations examined the effect of this estimation on the accuracy of the approximation.

We defined a 'reduced model' corresponding to the standard first-level model design described above. This has four regressors of interest, one for each of the four experimental conditions and an uninteresting constant column. We then defined a 'full model' which had these regressors, but in addition had two columns for parametric modulators. These modulators modelled responses as exponential functions of the lag between first and second presentations of face image  $i$ , in terms of the number of intervening faces. The exponential function was given by  $\exp(-l_i/10)$  where 10 denotes the chosen time constant (in units of number of faces presented).

We generated data sets with a range of signal-to-noise ratios (SNRs) similar to the simulations in [32]. Here SNR is defined as the ratio of signal standard deviation to noise standard deviation. Figure 2 shows the simulation results for the case of data generated from the full model, and Figure 3 for data generated from the reduced model. For the latter case, SD is almost exact as the noise estimates converge to the same values for full and reduced models. For the former case, SD becomes biased at high SNR because the



**Figure 2. Log Bayes factor versus SNR for full versus reduced model, when full model is true, for IMO approach (black line) and Savage-Dickey (red line).**

doi:10.1371/journal.pone.0059655.g002

observed noise is over-estimated for the reduced model due to the presence of unmodelled signals. However, this only occurs at very large values of log Bayes factor (favouring the full model) so is unlikely to have any practical effects on the resulting inference.

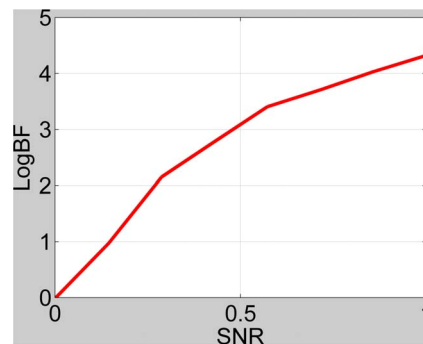
#### Prior Precisions

This simulation generates data from a design matrix that is similar to many second-level models. We use a design matrix  $X = I_K \otimes 1_N$  which models  $K$  effects using data from  $N$  subjects. This corresponds to a One-way ANOVA design with  $K$  levels. For the simulations we set  $K = 5$  and  $N = 20$ . We specify a prior over regression coefficients to have zero mean and precision  $\alpha = 30$  for each coefficient. The observation noise precision was set to  $\lambda = 1$ . We first draw the regression coefficients,  $w$  from this prior and produce data using  $y = Xw + e$  where  $e$  has zero mean and precision  $\lambda$ . We draw data at 1000 simulated voxels.

We then test for the effect of the first two regressors using the contrast

$$C^T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \quad (25)$$

The null model corresponding to this has design matrix  $X_n = XC_{null}$  where  $C_{null} = I_K - CC^+$  [42]. For the above contrast we have



**Figure 3. Log Bayes factor versus SNR for full versus reduced model, when reduced model is true, for IMO approach (black line) and Savage-Dickey (red line).** The lines overlap.

doi:10.1371/journal.pone.0059655.g003

$$C_{null}^T = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (26)$$

The SD log Bayes factor is computed using equation 20 using the true observation noise  $\lambda=1$ . Instead of using the true  $\alpha$ 's we use a modified set of alphas. We draw  $\alpha_k$  (for the  $k$ th regression coefficient) from a uniform distribution between plus and minus  $U\%$  percent of  $\alpha=30$ . This mimics the variability introduced by the Empirical Bayes estimates of the  $\alpha$ 's.

We then compute the IMO log Bayes factor by separately computing the model evidence for the full and null models. Again, we use the true observation noise  $\lambda=1$  but use a modified set of alphas. Here, the alphas for the full model are the same as for the SD simulation above. But the alphas for the null model are adjusted using the same uniform sampling approach to produce a different set of  $\alpha_k$ 's. This reflects the fact that the Empirical Bayes IMO approach uses two different sets of alphas; one for the full model and one for the null model.

We repeat the above procedure for four levels of variability in the prior precisions;  $U=0\%$ ,  $17\%$ ,  $33\%$  or  $50\%$ . Figure 4 shows SD versus IMO estimates of the log Bayes factor for these four different levels. For all modifications of prior precisions, larger log Bayes factors are accurately approximated. There is, however, increasing levels of disagreement at the lower range. The most noticeable feature is a 'bottoming-out', most clearly observable for the  $50\%$  condition. This occurs because the IMO estimate is a function of two sets of  $\alpha$ 's (full and null model) whereas the SD estimate is only a function of one set of  $\alpha$ 's (full model).

For null prior precisions which are smaller than full prior precisions, the IMO estimate is more negative - hence the dots left of the red line in Figure 4. Null prior precisions larger than full prior precisions produce dots to the right of the red line. Similar results have been obtained (not shown) when using contrasts testing for additive or differential effects.

We repeated the above procedure but this time using the standard first-level fMRI design matrix. An observation noise precision of  $\lambda=5$ , which is representative of values estimated from event-related fMRI data (see below), was set to be the same for both models. The results are shown in Figure 5 for a contrast testing for a differential effect. Again, we observe a bottoming-out effect. Further simulations showed that the bottoming-out effect could be produced for first- or second-level designs, and for subset, differential or additive contrasts. This effect could be alleviated by setting the observation noise precision to a sufficiently high level. To summarise, SD and IMO agree well for moderately positive IMO log BFs. But for negative IMO log Bayes factors, the discrepancy becomes commensurately larger for decreasing observation noise precision and increasing heterogeneity of the prior precision estimates.

Finally, we compare SD and IMO estimates to the true log Bayes factors. In these simulations, regression coefficients were sampled from distributions with known prior precision ( $\alpha=30$ , as above) and this value was used to compute the true log Bayes factor. IMO estimates were based on full prior precisions and null prior precisions that were both modified by a maximum proportion  $z$ . The SD estimates were based on the modified full prior precisions. Bayes factors were then computed for 1000 data sets and produced the results in Figure 6. Here we can see that SD and IMO produce different patterns of errors in their estimation, with SD showing a degree of bias and IMO showing a degree of

variance. We then computed the Root Mean Squared Error (RMSE) in estimating the log Bayes factor for the above results. This procedure was repeated 100 times. For  $U=17\%$ ,  $33\%$  and  $50\%$  the RMSE's are 0.07, 0.14 and 0.24 for SD and 0.07, 0.15 and 0.25 for IMO. There is therefore very little difference in the average accuracy of the estimates.

### Taylor Approximation

We repeated the 'first-level' fMRI simulations described in the above section on observation noise. But this time we hold the noise precision fixed and look at the effect of approximating the posterior density using the Taylor series approximation. We used empirical values of observation noise levels from 2000 voxels of first level fMRI data taken from slice  $z=-11mm$  (see below). These ranged from  $\lambda=0.05$  to  $\lambda=4.5$ . We compared the log Bayes factors as estimated using SDT versus SD over 2000 simulated voxels and found excellent agreement. The SDT estimates were within 0.00007%, 0.00009% and 0.00022% of the SD values, for AR model orders of 1, 2 and 3 respectively. Plots of SDT and SD versus SNR (not shown) are visually indistinguishable.

### First-level fMRI

We first fitted the first-level models using the '1st level' Bayesian estimation algorithm described in [20] using a 'global' prior. We additionally constrained the analysis to within brain voxels using an explicit mask (the brainmask.nii image in SPM's apriori directory). Model fitting took 6 minutes on a high-end desktop PC (dual 3.2 GHz Intel Xeon CPUs, 12 GB Ram, 64-bit Windows Vista).

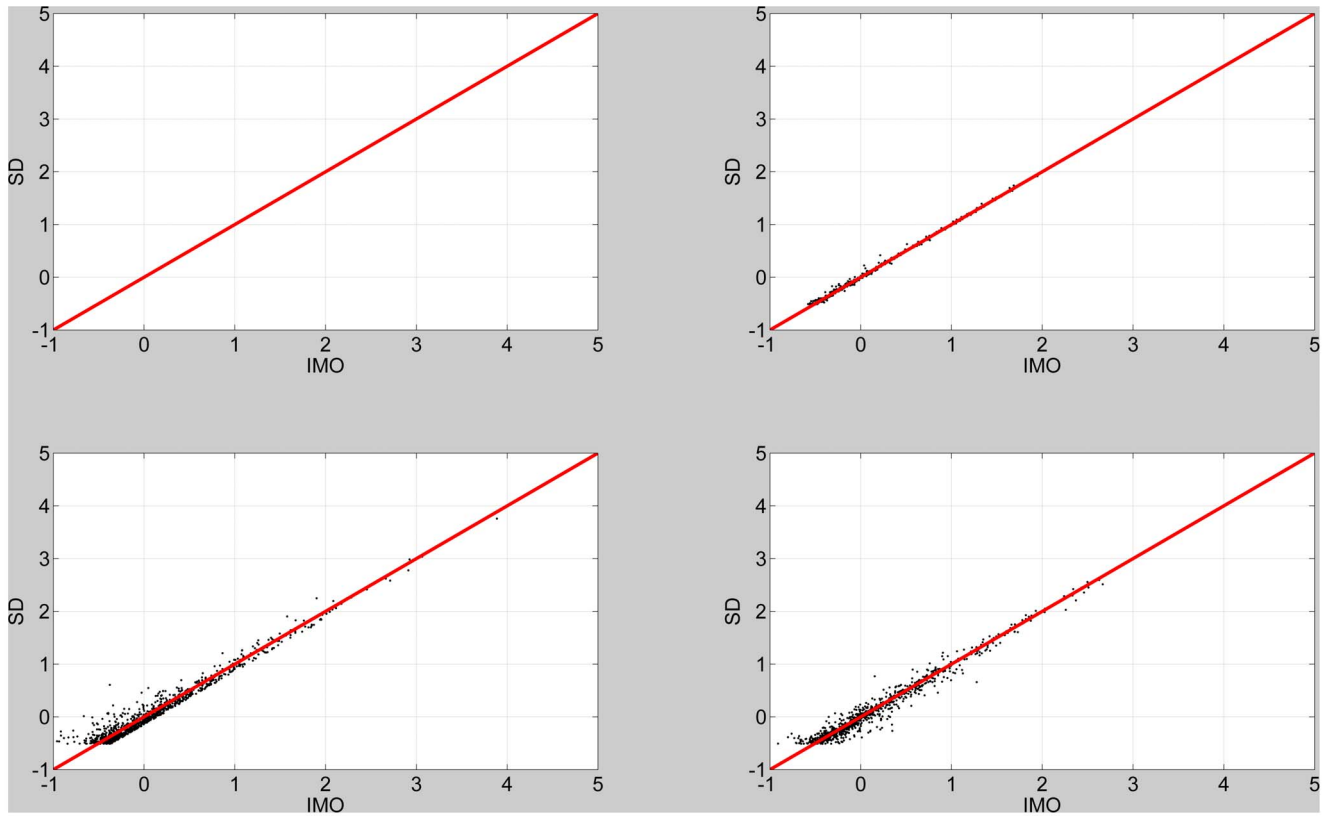
We used the SDT approach to compute Bayes factors testing for responses to non-familiar images (the fifth column of zeroes in the contrast relates to the uninteresting constant column in the design matrix, and is often not explicitly included when defining contrasts in the SPM software)

$$C_2^T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \quad (27)$$

The log Bayes factor at each voxel was computed using equation 20. We also computed log BF maps using the IMO approach by fitting two models. First, we fitted the standard model and computed its log evidence,  $L_1$ , at each voxel using equation 7. Second, we fitted a reduced model which did not model responses to non-familiar faces. Thus, the reduced model has three regressors whereas the standard model has five. We then computed the log evidence map  $L_2$ . The log BF map testing for responses to non-familiar faces is  $L_1-L_2$ . The models were estimated using the '1st level' Bayesian estimation algorithm described in [20] using a 'global' prior. Model fitting took 14 minutes for the standard model and 12 minutes for the reduced model. Each estimation took longer for the IMO approach because the model evidence was computed at each voxel.

Figure 7 (top panel) plots SDT versus IMO log Bayes factors for voxels in slice  $z=-11mm$ . This shows good agreement, except at large values of IMO log Bayes factor. The overall correlation is  $r=0.993$ . The plot shows a similar effect to that observed in Figure 2, suggesting that the discrepancy may be due to inconsistent estimates of observation noise precision. We then repeated the above analysis but with the contrast now testing for the main effect of familiarity.





**Figure 4. Second-level design. Savage-Dickey log Bayes factor versus IMO log Bayes factor for four levels of variability in prior precisions: 0% (top left), 17% (top right), 33% (bottom left) and 50% (bottom right).** The red line denotes equality.  
doi:10.1371/journal.pone.0059655.g004

$$C_2^T = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \end{bmatrix} \quad (28)$$

This tests for differences between responses to familiar versus unfamiliar faces, collapsed across repetition. Figure 7 (bottom panel) plots SDT versus IMO log Bayes factors for voxels in slice  $z = -11 \text{ mm}$ . This shows poor agreement over a large range of IMO log Bayes factor values. The overall correlation is  $r = 0.707$ . The plot shows a similar effect to that observed in Figure 5, suggesting that the discrepancy may be due to inconsistent estimates of prior precision in the context of large observation noise.

### Second-level fMRI

We fitted a second-level model to the FIR summary statistic images as described earlier using the global shrinkage prior approach [17]. This was a one-way ANOVA model with a single time-bin factor. We then used SDT to compute the log Bayes factors for comparing the standard model to a nested model which did not include responses in the 3 time bins from 6–12 s. This was implemented using equation 20 and the appropriate contrast (an identity matrix over columns 4, 5 and 6). We then estimated this log Bayes factor using the IMO approach by separately fitting the standard and reduced models and computing the model evidences using equation 7. Figure 8 (top panel) plots the log Bayes factors for SDT versus IMO approaches for voxels in the  $z = -11 \text{ mm}$  slice. This shows a very strong correlation between the measures

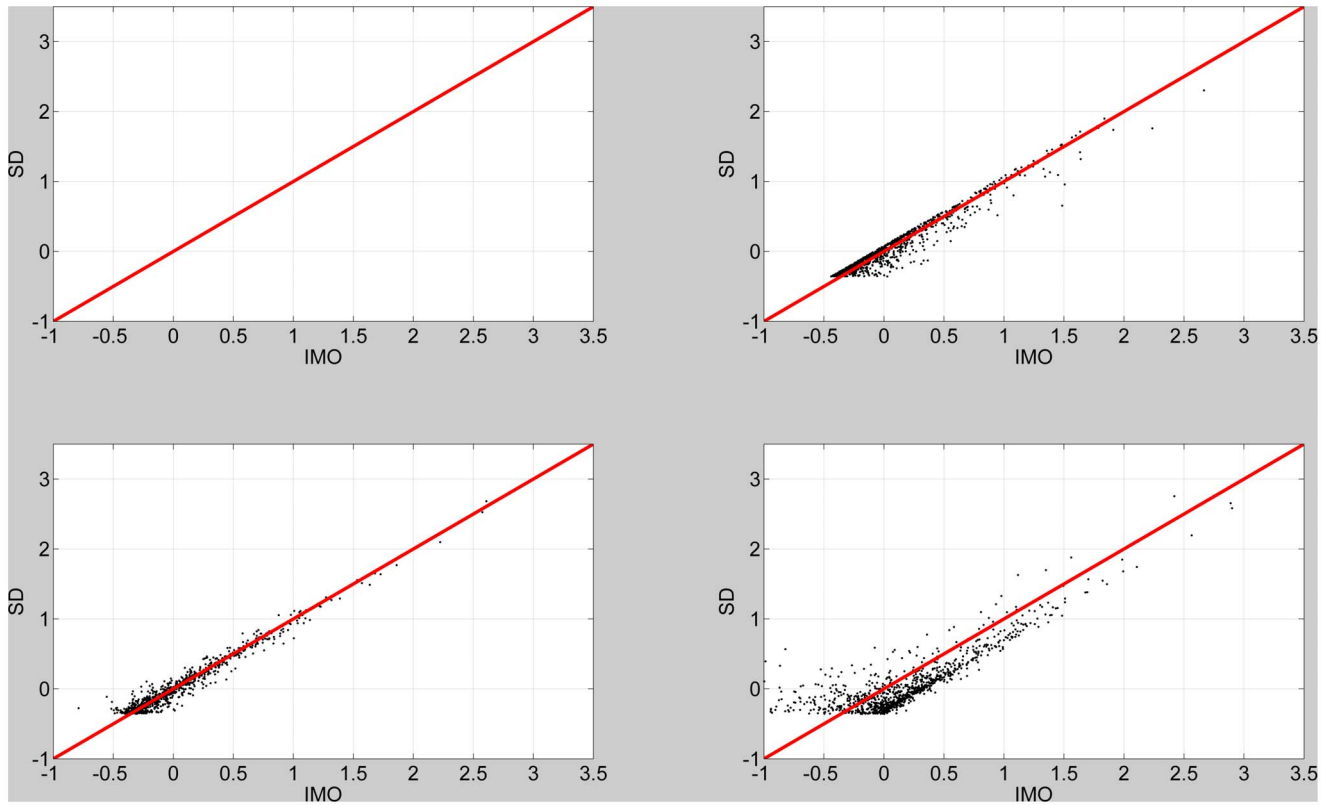
( $r = 0.994$ ). Our decision to look for late responses, in the 6–12 s window, is rather arbitrary but we note that similarly good agreements between SDT and IMO were found for other time windows.

We also implemented a non-nested model comparison to find where in the brain BOLD responses were better explained by a 4 to 6 s bin model versus a 6 to 8 s bin model. This model comparison looks at the relative amounts of variance explained by the different models, and is not the same as a contrast testing for a difference in the mean response in each bin. This test was first implemented using the SDT approach by specifying the two contrasts and subtracting the resulting log Bayes factor images using equation 22. This was then compared to the IMO approach where we separately computed the evidence for each model. We then plotted the log Bayes factors for SDT versus IMO approaches in Figure 8 (bottom panel). This figure is for voxels in the  $z = -11 \text{ mm}$  slice. This shows a very strong correlation between the measures ( $r = 0.999$ ). Similarly good agreements were found over a range of time bin comparisons.

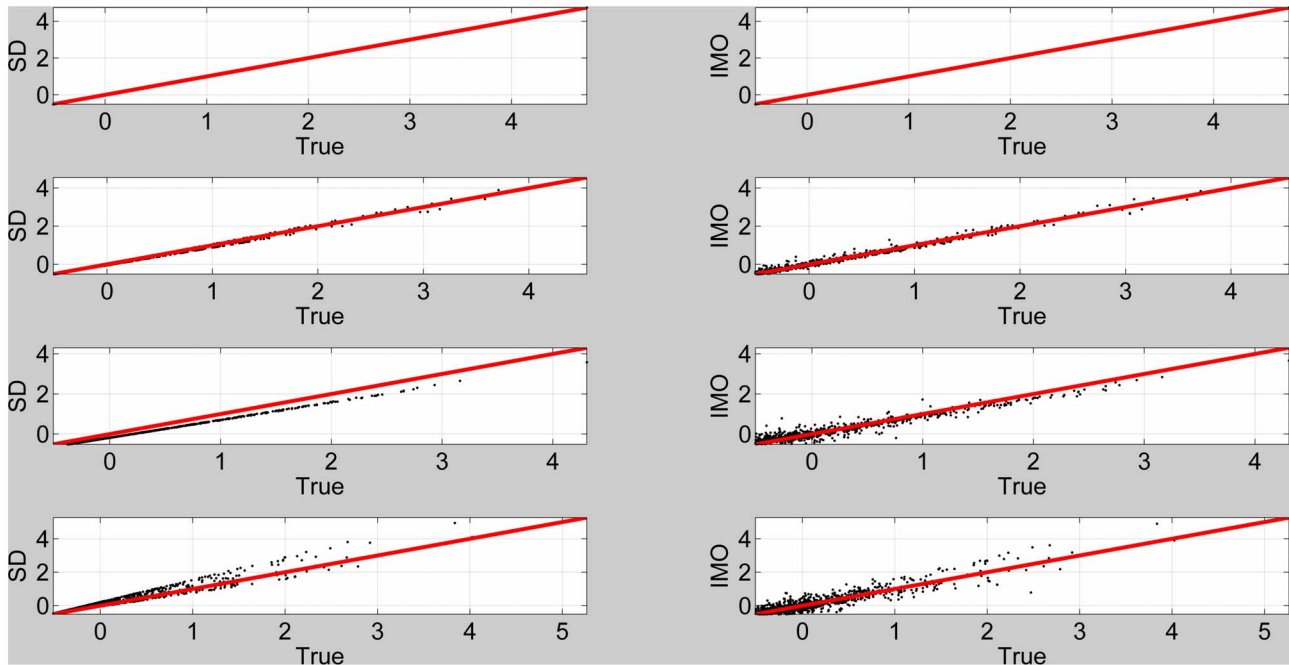
### Discussion

Statistical Parametric Mapping (SPM) has become the dominant paradigm for mass-univariate analysis of neuroimaging data. This paper has examined an alternative Posterior Probability Mapping (PPM) approach which offers two advantages (i) inferences can be made about effect size, thus lending a precise physiological meaning to activated regions, (ii) regions can be declared inactive. This latter facility is most parsimoniously provided by PPMs based on Bayesian model comparisons.

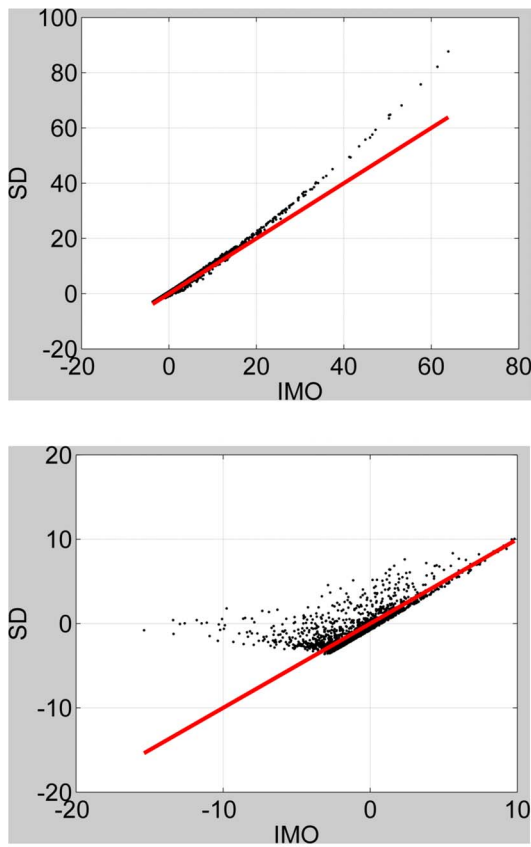




**Figure 5. First-level design. Savage-Dickey log Bayes factor versus IMO log Bayes factor for four levels of variability in prior precisions: 0% (top left), 17% (top right), 33% (bottom left) and 50% (bottom right).** The red line denotes equality.  
doi:10.1371/journal.pone.0059655.g005



**Figure 6. Left Panel: Savage-Dickey log Bayes factor versus true log Bayes factors for four levels of variability in prior precisions.** Right Panel: IMO log Bayes factor versus true log Bayes factors for four levels of variability in prior precisions. These are 0% (first row), 17% (second row), 33% (third row) and 50% (last row).  
doi:10.1371/journal.pone.0059655.g006



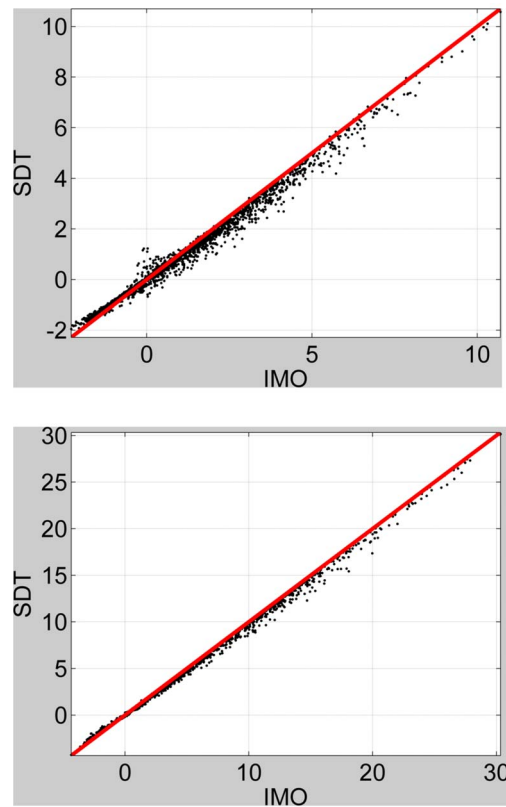
**Figure 7. First-level models.** Top Panel: log Bayes factor for SDT versus IMO approaches testing for any response to non-familiar faces. The red line denotes equality. Bottom Panel: log Bayes factor for SDT versus IMO approaches testing for main effect of familiarity  
doi:10.1371/journal.pone.0059655.g007

Previously, these comparisons have been implemented by an Independent Model Optimization (IMO) procedure which separately fits null and alternative models. In this paper we have proposed a more computationally efficient method based on Savage-Dickey approximations to the Bayes factor and Taylor series approximations to the voxel-wise posterior covariance matrices.

The IMO approach is more time consuming both due to the time taken to estimate the models and the user's time taken to set up the relevant design matrices. The Savage-Dickey-Taylor (SDT) approach is quicker on both counts and allows the user to explore the model space in a truly interactive way which is analogous to the use of F-contrasts in classical inference. Simulations show that the accuracy of the SDT method is comparable to that of the IMO method. Results on fMRI data show a correlation between SDT and IMO estimates, that is consistently high for second-level data, but is only moderately high for first-level data.

Our current Empirical Bayes implementation for estimating prior precisions works slice-by-slice for first-level data, due to computational constraints, but over the whole volume for second-level data. This has the effect of rendering the estimates of prior precisions more variable for the first than the second-level. The results in this paper suggest we revisit this implementation. Until these first-level estimates have been re-implemented we recommend that SDT only be used at the second level.

In general, the SDT approach would be suitable for all neuroimaging modalities. However, in this paper we have only



**Figure 8. Second-level models.** Top Panel: log Bayes factor for SDT versus IMO approaches testing for response in 6 to 12 s time bins. The red line denotes equality. Bottom Panel: log Bayes factor for SDT versus IMO approaches testing for responses that are better explained by a 4 to 6 s model than a 6 to 8 s model.  
doi:10.1371/journal.pone.0059655.g008

implemented it for the case of global shrinkage priors; these are appropriate for fMRI because the null hypothesis is of no activity on average [17]. For PET and M/EEG, when processed so that the data features represent activation (or, more generally, differences between conditions, whose expectation is zero under the null hypothesis) the methods presented here are similarly appropriate.

However, some modalities have imaging data that would not be zero under the null, such as voxel-based morphometry (VBM), whose voxel-wise data represent local tissue volumes [46] or forms of PET with a single image per subject that does not represent a difference between conditions, for example amyloid imaging [47]. For these data, shrinkage of the voxel-wise parameter estimates towards a *non-zero* overall mean should be appropriate and straightforward. We will therefore examine the use of SDT for these non-zero mean priors in a future publication. This future work will also extend SDT to work with spatial priors [20]. Both of these extensions are mathematically straightforward but beyond the scope of the current paper.

### A Software Implementation

Many of the algorithms referred to in this paper are available in the SPM software package which is available from <http://www.fil.ion.ucl.ac.uk/spm/>. The PPM procedure employing global shrinkage priors which shrink estimated parameters towards zero [17] can be accessed in the user interface of SPM by choosing '2nd-level' fMRI or M/EEG models and selecting the Bayesian option. The PPM approach for the analysis of within-subject fMRI

time series [26] can be accessed in the user interface of SPM by choosing ‘1st level’ fMRI models and selecting the Bayesian option.

## References

1. Friston K, Harrison L, Penny W (2003) Dynamic Causal Modelling. *NeuroImage* 19: 1273–1302.
2. Harrison L, Penny W, Friston K (2003) Multivariate autoregressive modelling of fMRI time series. *NeuroImage* 19: 1477–1491.
3. Bowman F, Caffo B, Bassett S, Kilts C (2008) A Bayesian hierarchical framework for spatial modelling of fMRI data. *NeuroImage* 39: 146–156.
4. Zhang L, Agravat S, Derado G, Chen S, McIntosh B, et al. (2012) BSMac: A MATLAB toolbox implementing a Bayesian spatial model for brain activation and connectivity. *Journal of Neuroscience Methods* 204: 133–143.
5. Friston KJ, Stephan KE, Lund TE, Morcom A, Kiebel S (2005) Mixed-effects and fMRI studies. *Neuroimage* 24: 244–252.
6. Woolrich MW, Behrens TEJ, Beckmann CF, Jenkinson M, Smith SM (2004) Multilevel linear modelling for fMRI group analysis using Bayesian inference. *Neuroimage* 21: 1732–1747.
7. Woolrich M, Behrens T, Smith S (2004) Constrained linear basis sets for HRF modelling using Variational Bayes. *NeuroImage* 21: 1748–1761.
8. Gossel C, Auer D, Fahrmeir L (2001) Bayesian spatiotemporal inference in functional magnetic resonance imaging. *Biometrics* 57: 554–562.
9. Penny W, Ghahramani Z, Friston K (2005) Bilinear Dynamical Systems. *Phil Trans R Soc B* 360: 983–993.
10. Smith JF, Pillai A, Chen K, Horwitz B (2011) Effective connectivity modeling for fMRI: Six issues and possible solutions using linear dynamic systems. *Front Syst Neurosci* 5: 104.
11. Friston K, Harrison L, Daunizeau J, Kiebel S, Phillips C, et al. (2008) Multiple sparse priors for the M/EEG inverse problem. *NeuroImage* 39: 1104–1120.
12. Wipf D, Nagarajan S (2009) A unified Bayesian framework for MEG/EEG source imaging. *Neuroimage* 44: 947–966.
13. David O, Kiebel S, Harrison L, Mattout J, Kilner J, et al. (2006) Dynamic causal modelling of evoked responses in EEG and MEG. *NeuroImage* 30: 1255–1272.
14. Valdes-Sosa PA, Roebroeck A, Daunizeau J, Friston K (2011) Effective connectivity: influence, causality and biophysical modeling. *Neuroimage* 58: 339–361.
15. Woolrich M (2012) Bayesian inference in fMRI. *Neuroimage* 62: 801–810.
16. Litvak V, Mattout J, Kiebel S, Phillips C, Henson R, et al. (2011) EEG and MEG data analysis in SPM8. *Comput Intell Neurosci* 2011: 852961.
17. Friston K, Penny W (2003) Posterior probability maps and SPMs. *NeuroImage* 19: 1240–1249.
18. Woolrich M, Jenkinson M, Brady M, Smith S (2004) Fully Bayesian spatiotemporal modeling of fMRI data. *IEEE Trans Med Imaging* 23: 213–231.
19. Dienes Z (2011) Bayesian versus orthodox statistics: which side are you on? *Perspectives on Psychological Science* 6: 274–290.
20. Penny WD, Kilner J, Blankenburg F (2007) Bayesian comparison of spatially regularised general linear models. *Human Brain Mapping* 28: 275–293.
21. Dickey J (1971) The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics* 42: 204–223.
22. Kass R, Raftery A (1995) Bayes factors. *Journal of the American Statistical Association* 90: 773–795.
23. Gelman A, Carlin J, Stern H, Rubin D (1995) *Bayesian Data Analysis*. Chapman and Hall, Boca Raton.
24. Mackay D (2003) *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge.
25. Friston K, Ashburner J, Kiebel S, Nichols T, Penny W, editors (2007) *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press.
26. Penny W, Trujillo-Barreto N, Friston K (2005) Bayesian fMRI time series analysis with spatial priors. *Neuroimage* 24: 350–362.
27. Penny W, Kiebel S, Friston K (2003) Variational Bayesian Inference for fMRI time series. *NeuroImage* 19: 727–741.
28. Flandin G, Penny W (2007) Bayesian fMRI data analysis with sparse spatial basis function priors. *Neuroimage* 34: 1108–1125.
29. Harrison LM, Penny W, Daunizeau J, Friston KJ (2008) Diffusion-based spatial priors for functional magnetic resonance images. *Neuroimage* 41: 408–423.
30. Ridgway GR, Litvak V, Flandin G, Friston KJ, Penny WD (2012) The problem of low variance voxels in statistical parametric mapping; a new hat avoids a ‘haircut’. *Neuroimage* 59: 2131–2141.
31. Bishop CM (1995) *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford.
32. Penny WD (2012) Comparing dynamic causal models using AIC, BIC and free energy. *Neuroimage* 59: 319–330.
33. Rosa M, Bestmann S, Harrison L, Penny W (2010) Bayesian model selection maps for group studies. *Neuroimage* 49: 217–224.
34. Harrison L, Bestmann S, Rosa M, Penny W, Green G (2011) Time scales of representation in the human brain: weighing past information to predict future events. *Front Hum Neurosci* 5: 37.
35. Penny W, Stephan K, Mechelli A, Friston K (2004) Comparing Dynamic Causal Models. *NeuroImage* 22: 1157–1172.
36. Penny W, Flandin G (2005) Bayesian analysis of fMRI data with spatial priors. In: *Proceedings of the Joint Statistical Meeting (JSM)*. American Statistical Association.
37. Friston K, Penny W (2011) Post hoc Bayesian model selection. *Neuroimage* 56: 2089–2099.
38. Rosa MJ, Friston K, Penny W (2012) Post-hoc selection of dynamic causal models. *J Neurosci Methods* 208: 66–78.
39. Friston KJ, Harrison L, Penny W (2003) Dynamic causal modelling. *Neuroimage* 19: 1273–1302.
40. Bishop C (2006) *Pattern Recognition and Machine Learning*. Springer.
41. Verdinelli I, Wasserman L (1995) Computing Bayes factors using a generalisation of the Savage-Dickey density ratio. *Journal of the American Statistical Association* 90: 614–618.
42. Christensen R (2002) *Plane answers to complex questions: the theory of linear models*. Springer-Verlag, New York, US.
43. Friston K, Ashburner J, Kiebel S, Nichols T, Penny W, editors (2007) *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press. Software and data available from <http://www.fil.ion.ucl.ac.uk/spm/>. Accessed 2013 Feb 25.
44. Penny W, Trujillo-Barreto N, Flandin G (2005) Bayesian analysis of single-subject fMRI: SPM implementation. Technical report, Wellcome Department of Imaging Neuroscience. URL <http://www.fil.ion.ucl.ac.uk/spm/doc/papers/vb3.pdf>. Accessed 2013 Feb 25.
45. Henson R, Shallice T, Gorno-Tempini M, Dolan R (2002) Face repetition effects in implicit and explicit memory tests as measured by fMRI. *Cerebral Cortex* 12: 178–186.
46. Ashburner J, Friston K (2000) Voxel-based morphometry – the methods. *Neuroimage* 11: 805–821.
47. Herholz K, Ebmeier K (2011) Clinical amyloid imaging in Alzheimer’s disease. *The Lancet Neurology* 10: 667–670.

## Author Contributions

Conceived and designed the experiments: WP GR. Performed the experiments: WP. Analyzed the data: WP. Contributed reagents/materials/analysis tools: WP. Wrote the paper: WP GR.