



# Variational Bayesian Parameter Estimation Techniques for the General Linear Model

Ludger Starke<sup>1</sup> and Dirk Ostwald<sup>1,2,3\*</sup>

<sup>1</sup> Arbeitsbereich Computational Cognitive Neuroscience, Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany, <sup>2</sup> Center for Cognitive Neuroscience Berlin, Freie Universität Berlin, Berlin, Germany, <sup>3</sup> Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

## OPEN ACCESS

### Edited by:

John Ashburner,  
UCL Institute of Neurology,  
United Kingdom

### Reviewed by:

Lester Melie-Garcia,  
Centre Hospitalier Universitaire  
Vaudois (CHUV), Switzerland  
Mark Rowland,  
University of Cambridge,  
United Kingdom

### \*Correspondence:

Dirk Ostwald  
dirk.ostwald@fu-berlin.de

### Specialty section:

This article was submitted to  
Brain Imaging Methods,  
a section of the journal  
Frontiers in Neuroscience

**Received:** 06 July 2017

**Accepted:** 24 August 2017

**Published:** 15 September 2017

### Citation:

Starke L and Ostwald D (2017)  
Variational Bayesian Parameter  
Estimation Techniques for the General  
Linear Model.  
*Front. Neurosci.* 11:504.  
doi: 10.3389/fnins.2017.00504

Variational Bayes (VB), variational maximum likelihood (VML), restricted maximum likelihood (ReML), and maximum likelihood (ML) are cornerstone parametric statistical estimation techniques in the analysis of functional neuroimaging data. However, the theoretical underpinnings of these model parameter estimation techniques are rarely covered in introductory statistical texts. Because of the widespread practical use of VB, VML, ReML, and ML in the neuroimaging community, we reasoned that a theoretical treatment of their relationships and their application in a basic modeling scenario may be helpful for both neuroimaging novices and practitioners alike. In this technical study, we thus revisit the conceptual and formal underpinnings of VB, VML, ReML, and ML and provide a detailed account of their mathematical relationships and implementational details. We further apply VB, VML, ReML, and ML to the general linear model (GLM) with non-spherical error covariance as commonly encountered in the first-level analysis of fMRI data. To this end, we explicitly derive the corresponding free energy objective functions and ensuing iterative algorithms. Finally, in the applied part of our study, we evaluate the parameter and model recovery properties of VB, VML, ReML, and ML, first in an exemplary setting and then in the analysis of experimental fMRI data acquired from a single participant under visual stimulation.

**Keywords:** variational Bayes, general linear model (GLM), fMRI neuroimaging, restricted maximum likelihood estimation, covariance estimation, data analysis, machine learning

## 1. INTRODUCTION

Variational Bayes (VB), variational maximum likelihood (VML) (also known as expectation-maximization), restricted maximum likelihood (ReML), and maximum likelihood (ML) are cornerstone parametric statistical estimation techniques in the analysis of functional neuroimaging data. In the SPM software environment (<http://www.fil.ion.ucl.ac.uk/spm/>), one of the most commonly used software packages in the neuroimaging community, variants of these estimation techniques have been implemented for a wide range of data models (Penny et al., 2011; Ashburner, 2012). For fMRI data, these models vary from mass-univariate general linear and auto-regressive models (e.g., Friston et al., 1994, 2002a,b; Penny et al., 2003), over multivariate decoding models (e.g., Friston et al., 2008a), to dynamic causal models (e.g., Friston et al., 2003; Marreiros et al., 2008; Stephan et al., 2008). For M/EEG data, these models range from channel-space general linear models (e.g., Kiebel and Friston, 2004a,b), over dipole and distributed source reconstruction

models (e.g., Friston et al., 2008b; Kiebel et al., 2008; Litvak and Friston, 2008), to a large family of dynamic causal models (e.g., David et al., 2006; Chen et al., 2008; Moran et al., 2009; Pinotsis et al., 2012; Ostwald and Starke, 2016).

Because VB, VML, ReML, and ML determine the scientific inferences drawn from empirical data in any of the above mentioned modeling frameworks, they are of immense importance for the neuroimaging practitioner. However, the theoretical underpinnings of these estimation techniques are rarely covered in introductory statistical texts and the technical literature relating to these techniques is rather evolved. Because of their widespread use within the neuroimaging community, we reasoned that a theoretical treatment of these techniques in a familiar model scenario may be helpful for both neuroimaging novices, who would like to learn about some of the standard statistical estimation techniques employed in the field, and for neuroimaging practitioners, who would like to further explore the foundations of these and alternative model estimation approaches.

In this technical study, we thus revisit the conceptual underpinnings of the aforementioned techniques and provide a detailed account of their mathematical relations and implementational details. Our exposition is guided by the fundamental insight that VML, ReML, and ML can be understood as special cases of VB (Friston et al., 2002a, 2007; Friston, 2008). In the current note, we reiterate and consolidate this conceptualization by paying particular attention to the respective technique's formal treatment of a model's parameter set. Specifically, across the estimation techniques of interest, model parameters are either treated as random variables, in which case they are endowed with prior and posterior uncertainty modeled by parametric probability density functions, or as non-random quantities. In the latter case, prior and posterior uncertainties about the respective parameters' values are left unspecified. Because the focus of the current account is on statistical estimation techniques, we restrict the model of application to a very basic scenario that every neuroimaging practitioner is familiar with: the analysis of a single-participant, single-session EPI time-series in the framework of the general linear model (GLM) (Monti, 2011; Poline and Brett, 2012). Importantly, in line with the standard practice in fMRI data analysis, we do not assume spherical covariance matrices (e.g., Zarahn et al., 1997; Purdon and Weisskoff, 1998; Woolrich et al., 2001; Friston et al., 2002b; Mumford and Nichols, 2008).

We proceed as follows. After some preliminary notational remarks, we begin the theoretical exposition by first introducing the model of application in Section 2.1. We next briefly discuss two standard estimation techniques (conjugate Bayes and ML for spherical covariance matrices) that effectively span the space of VB, VML, ReML, and ML and serve as useful reference points in Section 2.2. After this prelude, we are then concerned with the central estimation techniques of interest herein. In a hierarchical fashion, we subsequently discuss the theoretical background and the practical algorithmic application of VB, VML, ReML, and ML to the GLM in Sections 2.3–2.6. We focus on the central aspects and conceptual relationships of the techniques and present all

mathematical derivations as Supplementary Material. In the applied part of our study (Section 3), we then firstly evaluate VB, VML, ReML, and ML from an objective Bayesian viewpoint (Bernardo, 2009) in simulations; and secondly, apply them to real fMRI data acquired from a single participant under visual stimulation (Ostwald et al., 2010). We close by discussing the relevance and relation of our exposition with respect to previous treatments of the topic matter in Section 4.

In summary, we make the following novel contributions in the current technical study. Firstly, we provide a comprehensive mathematical documentation and derivation of the conceptual relationships between VB, VML, ReML, and ML. Secondly, we derive a collection of explicit algorithms for the application of these estimation techniques to the GLM with non-spherical linearized covariance matrix. Finally, we explore the validity of the ensuing algorithms in simulations and in the application to real experimental fMRI data. We complement our theoretical documentation by the practical implementation of the algorithms and simulations in a collection of Matlab .m files (MATLAB and Optimization Toolbox Release 2014b, The MathWorks, Inc., Natick, MA, United States), which is available from the Open Science Framework (<https://osf.io/c4ux7/>). On occasion, we make explicit reference to these functions, which share the stub *vbg\_\*.m*.

## 1.1. Notation and Preliminary Remarks

A few remarks on our mathematical notation are in order. We formulate VB, VML, ReML, and ML against the background of probabilistic models (e.g., Bishop, 2006; Barber, 2012; Murphy, 2012). By probabilistic models we understand (joint) probability distributions over sets of observed and unobserved random variables. Notationally, we do not distinguish between probability distributions and their associated probability density functions and write, for example,  $p(y, \theta)$  for both. Because we are only concerned with parametric probabilistic models of the Gaussian type, we assume throughout the main text that all probability distributions of real random vectors have densities. We do, however, distinguish between the conditioning of a probability distribution of a random variable  $y$  on a (commonly unobserved) random variable  $\theta$ , which we denote by  $p(y|\theta)$ , and the parameterization of a probability distribution of a random variable  $y$  by a (non-random) parameter  $\theta$ , which we denote by  $p_\theta(y)$ . Importantly, in the former case,  $\theta$  is conceived of as random variable, while in the latter case, it is not. Equivalently, if  $\theta^*$  denotes a value that the random variable  $\theta$  may take on, we set  $p(y|\theta = \theta^*) \Leftrightarrow p_{\theta^*}(y)$ .

Otherwise, we use standard applied mathematical notation. For example, real vectors and matrices are denoted as elements of  $\mathbb{R}^n$  and  $\mathbb{R}^{m \times n}$  for  $n, m \in \mathbb{N}$ ,  $I_n \in \mathbb{R}^{n \times n}$  denotes the  $n$ -dimensional identity matrix,  $|\cdot|$  denotes a matrix determinant,  $\text{tr}(\cdot)$  denotes the trace operator, and p.d. denotes a positive-definite matrix.  $H_f(a)$  denotes the Hessian matrix of some real-valued function  $f(x)$  evaluated at  $x = a$ . We denote the probability density function of a Gaussian distributed random vector  $y$  with expectation parameter  $\mu$  and covariance parameter  $\Sigma$  by  $N(y; \mu, \Sigma)$ . Finally, because of the rather applied character of this note, we formulate functions primarily by means of the definition of the values they

take on and eschew formal definitions of their domains and ranges. Further notational conventions that apply in the context of the mathematical derivations provided in the Supplementary Material are provided therein.

## 2. THEORY

### 2.1. Model of Interest

Throughout this study, we are interested in estimating the parameters of the model

$$y = X\beta + \varepsilon, \tag{1}$$

where  $y \in \mathbb{R}^n$  denotes the data,  $X \in \mathbb{R}^{n \times p}$  denotes a design matrix of full column rank  $p$ , and  $\beta \in \mathbb{R}^p$  denotes a parameter vector. We make the following fundamental assumption about the error term  $\varepsilon \in \mathbb{R}^n$

$$\varepsilon \sim N(\varepsilon; 0, V_\lambda) \text{ with } V_\lambda := \sum_{i=1}^k \exp(\lambda_i) Q_i \in \mathbb{R}^{n \times n} \text{ p.d.} \tag{2}$$

In words, we assume that the error term is distributed according to a Gaussian distribution with expectation parameter  $0 \in \mathbb{R}^n$  and positive-definite covariance matrix  $V_\lambda \in \mathbb{R}^{n \times n}$ . Importantly, we do not assume that  $V_\lambda$  is necessarily of the form  $\sigma^2 I_n$ , i.e., we allow for non-sphericity of the error terms. In Equation (2),  $\lambda_1, \dots, \lambda_k$ , is a set of *covariance component parameters* and  $Q_1, \dots, Q_k \in \mathbb{R}^{n \times n}$  is a set of *covariance basis matrices*, which are assumed to be fixed and known. We assume throughout, that the true, but unknown, values of  $\lambda_1, \dots, \lambda_k$  are such that  $V_\lambda$  is positive-definite. In line with the common denotation in the neuroimaging literature, we refer to Equations (1) and (2) as the *general linear model* (GLM) and its formulation by means of Equations (1) and (2) as its *structural form*.

Models of the form (1) and (2) are widely used in the analysis of neuroimaging data, and, in fact, throughout the empirical sciences (e.g., Rutherford, 2001; Draper and Smith, 2014; Gelman et al., 2014). In the neuroimaging community, models of the form Equations (1) and (2) are used, for example, in the analysis of fMRI voxel time-series at the session and participant-level (Monti, 2011; Poline and Brett, 2012), for the analysis of group effects (Mumford and Nichols, 2006, 2009), or in the context of voxel-based morphometry (Ashburner and Friston, 2000; Ashburner, 2009).

In the following, we discuss the application of VB, VML, ReML, and ML to the general forms of Equations (1) and (2). In our examples, however, we limit ourselves to the application of the GLM in the analysis of a single voxel's time-series in a single fMRI recording (run). In this case,  $y \in \mathbb{R}^n$  corresponds to the voxel's MR values over EPI volume acquisitions and  $n \in \mathbb{N}$  represents the total number of volumes acquired during the session. The design matrix  $X \in \mathbb{R}^{n \times p}$  commonly constitutes a constant regressor and the onset stick functions of different experimental conditions convolved with a hemodynamic response function and a constant offset. This renders the parameter entries  $\beta_j$  ( $j \in \mathbb{N}_p$ ) to correspond to the average session MR signal and condition-specific effects.

Importantly, in the context of fMRI time-series analyses, the most commonly used form of the covariance matrix  $V_\lambda$  employs  $k = 2$  covariance component parameters  $\lambda_1$  and  $\lambda_2$  and corresponding covariance basis matrices

$$Q_1 := I_n \text{ and } Q_2 := (Q_2)_{ij} := \exp\left(-\frac{1}{\tau}|i-j|\right). \tag{3}$$

This specific form of the error covariance matrix encodes exponentially decaying correlations between neighboring data points, and, with  $\tau := 0.2$ , corresponds to the widely used approximation to the *AR(1) + white noise* model in the analysis of fMRI data (Purdon and Weiskoff, 1998; Friston et al., 2002b).

In **Figure 1**, we visualize the exemplary design matrix and covariance basis matrix set that will be employed in the example applications throughout the current section. In the example, we assume two experimental conditions, which have been presented with an expected inter-trial interval of 6 s (standard deviation 1 s) during an fMRI recording session comprising  $n = 400$  volumes and with a TR of 2 s. The design matrix was created using the micro-time resolution convolution and downsampling approach discussed in Henson and Friston (2007).

### 2.2. Conjugate Bayes and ML under Error Sphericity

We start by briefly recalling the fundamental results of conjugate Bayesian and classical point-estimation for the GLM with spherical error covariance matrix. In fact, the introduction of ReML (Friston et al., 2002a; Phillips et al., 2002) and later VB (Friston et al., 2007) to the neuroimaging literature were motivated amongst other things by the need to account for non-sphericity of the error distributions in fMRI time-series analysis (Purdon and Weiskoff, 1998; Woolrich et al., 2001). Further, while not a common approach in fMRI, recalling the conjugate Bayes scenario helps to contrast the probabilistic model of interest in VB from its mathematically more tractable, but perhaps less intuitively plausible, analytical counterpart. Together, the two estimation techniques discussed in the current section may thus be conceived as forming the respective endpoints of the continuum of estimation techniques discussed in the remainder.

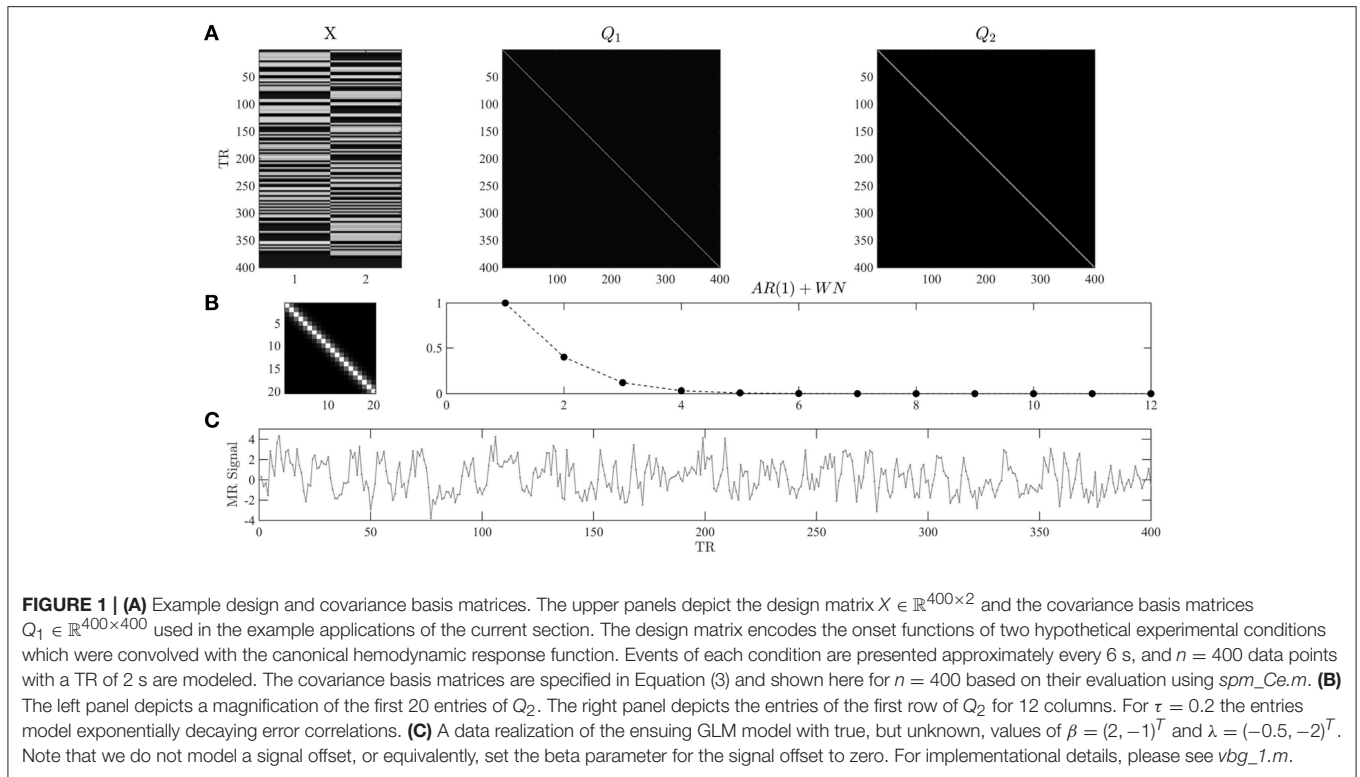
With spherical covariance matrix, the GLM of Equations (1) and (2) simplifies to

$$y = X\beta + \varepsilon, \text{ where } \varepsilon \sim N(\varepsilon; 0, \sigma^2 I_n). \tag{4}$$

A conjugate Bayesian treatment of the GLM considers the structural form Equation (4) as a conditional probabilistic statement about the distribution of the observed random variable  $y$

$$p(y|\beta, \sigma^2) = N(y; X\beta, \sigma^2 I_n), \tag{5}$$

which is referred to as the *likelihood* and requires the specification of the marginal distribution  $p(\beta, \sigma^2)$ , referred to as the *prior*. Together, the likelihood and the prior define the probabilistic model of interest, which takes the form of a joint distribution



over the observed random variable  $y$  and the unobserved random variables  $\beta$  and  $\sigma^2$ :

$$p(y, \beta, \sigma^2) = p(y|\beta, \sigma^2)p(\beta, \sigma^2). \tag{6}$$

Based on the probabilistic model (Equation 5), the two fundamental aims of Bayesian inference are, firstly, to determine the conditional parameter distribution given a value of the observed random variable  $p(\beta, \sigma^2|y)$ , often referred to as the *posterior*, and secondly, to evaluate the marginal probability  $p(y)$  of a value of the observed random variable, often referred to as *marginal likelihood* or *model evidence*. The latter quantity forms an essential precursor for Bayesian model comparison, as discussed for example in further detail in Stephan et al. (2016a). Note that in our treatment of the Bayesian scenario the marginal and conditional probability distributions of  $\beta$  and  $\sigma^2$  are meant to capture our uncertainty about the values of these parameters and not distributions of true, but unknown, parameter values. For the true, but unknown, values of  $\beta$  and  $\sigma^2$  we postulate, as in the classical point-estimation scenario, that they assume fixed values, which are never revealed (but can of course be chosen *ad libitum* in simulations).

The VB treatment of Equation (6) assumes proper prior distributions for  $\beta$  and  $\sigma^2$ . In this spirit, the closest conjugate Bayesian equivalent is hence the assumption of proper prior distributions. For the case of the model (Equation 6), upon reparameterization in terms of a precision parameter  $\lambda := 1/\sigma^2$ , a natural conjugate approach assumes a non-independent prior

distribution of Gaussian-Gamma form,

$$p(\beta, \lambda) = p(\beta|\lambda)p(\lambda) = N(\beta; \mu_\beta, \Sigma_\beta)G(\lambda; a_\lambda, b_\lambda), \tag{7}$$

where  $\mu_\beta \in \mathbb{R}^p$ ,  $\Sigma_\beta := \lambda^{-1}V_\beta$ ,  $a_\lambda, b_\lambda \in \mathbb{R}$  are the prior distribution parameters and  $V_\beta \in \mathbb{R}^{p \times p}$  p.d. is the prior beta parameter covariance structure. For the gamma distribution we use the shape and rate parameterization. Notably, the Gaussian distribution of  $\beta$  is parameterized conditional on the value of  $\lambda$  in terms of its covariance  $\Sigma_\beta$ . Under this prior assumption, it can be shown that the posterior distribution is also of Gaussian-Gamma form,

$$p(\beta, \lambda|y) = N(\beta; \mu_{\beta|y}, \Sigma_{\beta|y})G(\lambda; a_{\lambda|y}, b_{\lambda|y}), \tag{8}$$

with posterior parameters

$$\begin{aligned} \mu_{\beta|y} &= (X^T X + V_\beta^{-1})^{-1}(X^T y + V_\beta^{-1} \mu_\beta) \\ \Sigma_{\beta|y} &= \lambda^{-1} V_{\beta|y} = \lambda^{-1} (X^T X + V_\beta^{-1})^{-1} \\ a_{\lambda|y} &= (2a_\lambda + n)/2 \\ b_{\lambda|y} &= b_\lambda + \frac{1}{2} y^T y + \frac{1}{2} \mu_\beta^T V_\beta^{-1} \mu_\beta - \frac{1}{2} \mu_{\beta|y}^T V_{\beta|y}^{-1} \mu_{\beta|y}. \end{aligned} \tag{9}$$

Furthermore, in this scenario the marginal likelihood evaluates to a multivariate non-central T-distribution

$$p(y) = T(y; \mu_y, \Sigma_y, \nu_y) \tag{10}$$

with expectation, covariance, and degrees of freedom parameters

$$\mu_y = X\mu_\beta, \Sigma_y = \frac{2b}{2a+n-1}(XV_\beta X^T + I_n), \text{ and } \nu_y = 2a+n-1, \tag{11}$$

respectively. For derivations of Equations (8–11) see, for example, Lindley and Smith (1972), Broemeling (1984), and Gelman et al. (2014).

Importantly, in contrast to the VB, VML, ReML, and ML estimation techniques developed in the remainder, the assumption of the prior probabilistic dependency of the effect size parameter on the covariance component parameter in Equation (7) eschews the need for iterative approaches and results in the fully analytical solutions of Equations (8–11). However, as there is no principled reason beyond mathematical convenience that motivates this prior dependency, the fully conjugate framework seems to be rarely used in the analysis of neuroimaging data. Moreover, the assumption of an uninformative improper prior distribution (Frank et al., 1998) is likely more prevalent in the neuroimaging community than the natural conjugate form discussed above. This is due to the implementation of a closely related procedure in FSL’s FLAME software (Woolrich et al., 2004, 2009). However, because VB assumes proper prior distributions, we eschew the details of this approach herein.

In contrast to the probabilistic model of the Bayesian scenario, the classical ML approach for the GLM does not conceive of  $\beta$  and  $\sigma^2$  as unobserved random variables, but as parameters, for which point-estimates are desired. The probabilistic model of the classical ML approach for the structural model (Equation 4) thus takes the form

$$p_{\beta, \sigma^2}(y) = N(y; X\beta, \sigma^2 I_n). \tag{12}$$

The ML point-estimators for  $\beta$  and  $\sigma^2$  are well-known to evaluate to (e.g., Hocking, 2013)

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{13}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^T (y - X\hat{\beta}). \tag{14}$$

Note that Equation (13) also corresponds to the ordinary least-squares estimator. It can be readily generalized for non-spherical error covariance matrices by a “sandwiched” inclusion of the appropriate error covariance matrix, if this is (assumed) to be known, resulting in the generalized least-squares estimator (e.g., Draper and Smith, 2014). Further note that Equation (14) is a biased estimator for  $\sigma^2$  and hence commonly replaced by its restricted maximum likelihood counterpart, which replaces the factor  $n^{-1}$  by the factor  $(n - p)^{-1}$  (e.g., Foulley, 1993).

Having briefly reviewed the conjugate Bayesian and classical point estimation techniques for the GLM parameters under the assumption of a spherical error covariance matrix, we next discuss VB, VML, ReML, and ML for the scenario laid out in Section 2.1.

### 2.3. Variational Bayes (VB)

VB is a computational technique that allows for the evaluation of the primary quantities of interest in the Bayesian paradigm as introduced above: the posterior parameter distribution and the marginal likelihood. For the GLM, VB thus rests on the same probabilistic model as standard conjugate Bayesian inference: the structural form of the GLM (cf. Equations 1, 2) is understood as the parameter conditional likelihood distribution and both parameters are endowed with marginal distributions. The probabilistic model of interest in VB thus takes the form

$$p(y, \beta, \lambda) = p(y|\beta, \lambda)p(\beta, \lambda) \tag{15}$$

with likelihood distribution

$$p(y|\beta, \lambda) = N(y; X\beta, V_\lambda). \tag{16}$$

Above, we have seen that a conjugate prior distribution can be constructed which allows for exact inference in models of the form Equations (1) and (2) based on a conditionally-dependent prior distribution and simple covariance form. In order to motivate the application of the VB technique to the GLM, we here thus assume that the marginal distribution  $p(\beta, \lambda)$  factorizes, i.e., that

$$p(\beta, \lambda) = p(\beta|\lambda)p(\lambda) := p(\beta)p(\lambda). \tag{17}$$

Under this assumption, exact Bayesian inference for the GLM is no longer possible and approximate Bayesian inference is clearly motivated (Murphy, 2012).

To compute the marginal likelihood and obtain an approximation to the posterior distribution over parameters  $p(\beta, \lambda|y)$ , VB uses the following decomposition of the log marginal likelihood into two information theoretic quantities (Cover and Thomas, 2012), the *free energy* and a *Kullback-Leibler (KL) divergence*

$$\ln p(y) = F^{VB}(q(\beta, \lambda)) + KL(q(\beta, \lambda)||p(\beta, \lambda|y)). \tag{18}$$

We discuss the constituents of the right-hand side of Equation (18) in turn. Firstly,  $q(\beta, \lambda)$  denotes the so-called *variational distribution*, which will constitute the approximation to the posterior distribution and is of parameterized form, i.e., governed by a probability density. We refer to the parameters of the variational distribution as *variational parameters*. Secondly, the non-negative KL-divergence is defined as the integral

$$KL(q(\beta, \lambda)||p(\beta, \lambda|y)) = \iint q(\beta, \lambda) \ln \left( \frac{q(\beta, \lambda)}{p(\beta, \lambda|y)} \right) d\beta d\lambda. \tag{19}$$

Note that, formally, the KL-divergence is a functional, i.e., a function of functions, in this case the probability density functions  $q(\beta, \lambda)$  and  $p(\beta, \lambda|y)$ , and returns a scalar number. Intuitively, it measures the dissimilarity between its two input distributions: the more similar the variational distribution  $q(\beta, \lambda)$  is to the posterior distribution  $p(\beta, \lambda|y)$ , the smaller the divergence becomes. It is of fundamental importance for the VB technique that the KL-divergence is always positive and zero if,

and only if,  $q(\beta, \lambda)$  and  $p(\beta, \lambda|y)$  are equal. For a proof of these properties, see Appendix A in Ostwald et al. (2014). Together with the log marginal likelihood decomposition Equation (18) the properties of the KL-divergence equip the free energy with its central properties for the VB technique, as discussed below. A proof of Equation (18) with  $\vartheta := \{\beta, \lambda\}$  is provided in Appendix B in Ostwald et al. (2014).

The free energy itself is defined by

$$F^{VB}(q(\beta, \lambda)) = \iint q(\beta, \lambda) \ln \left( \frac{p(y, \beta, \lambda)}{q(\beta, \lambda)} \right) d\beta d\lambda. \quad (20)$$

Due to the non-negativity of the KL-divergence, the free energy is always smaller than or equal to the log marginal likelihood—the free energy thus forms a lower bound to the log marginal likelihood. Note that in Equation (20), the data  $y$  is assumed to be fixed, such that the free energy is a function of the variational distribution only. Because, for a given data observation, the log marginal likelihood  $\ln p(y)$  is a fixed quantity, and because increasing the free energy contribution to the right-hand side of Equation (18) necessarily decreases the KL-divergence between the variational and the true posterior distribution, maximization of the free energy with respect to the variational distribution has two consequences: firstly, it renders the free energy an increasingly better approximation to the log marginal likelihood; secondly, it renders the variational approximation an increasingly better approximation to the posterior distribution.

In summary, VB rests on finding a variational distribution that is as similar as possible to the posterior distribution, which is equivalent to maximizing the free energy with regard to the variational distribution. The maximized free energy then substitutes for the log marginal likelihood and the corresponding variational distribution yields an approximation to the posterior parameter distribution, i.e.,

$$\max_{q(\beta, \lambda)} F^{VB}(q(\beta, \lambda)) \approx \ln p(y) \text{ and } \arg \max_{q(\beta, \lambda)} F^{VB}(q(\beta, \lambda)) \approx p(\beta, \lambda|y). \quad (21)$$

To facilitate the maximization process, the variational distribution is often assumed to factorize over parameter sets, an assumption commonly referred to as *mean-field approximation* (Friston et al., 2007)

$$q(\beta, \lambda) := q(\beta)q(\lambda). \quad (22)$$

Of course, if the posterior does not factorize accordingly, i.e., if

$$p(\beta, \lambda|y) \neq p(\beta|y)p(\lambda|y), \quad (23)$$

the mean-field approximation limits the exactness of the method.

In applications, maximization of the free energy is commonly achieved by either *free-form* or *fixed-form* schemes. In brief, free-form maximization schemes do not assume a specific form of the variational distribution, but employ a fundamental theorem of variational calculus to maximize the free energy and to analytically derive the functional form and parameters of the variational distribution. For more general features of

the free-form approach, please see, for example, Bishop (2006), Chappell et al. (2009), and Ostwald et al. (2014). Fixed-form maximization schemes, on the other hand, assume a specific parametric form for the variational distribution's probability density function from the outset. Under this assumption, the free energy integral (Equation 20) can be evaluated (or at least approximated) analytically and rendered a function of the variational parameters. This function can in turn be optimized using standard nonlinear optimization algorithms. In the following section, we apply a fixed-form VB approach to the current model of interest.

### 2.3.1. Application to the GLM

To demonstrate the fixed-form VB approach to the GLM of Equations (1) and (2), we need to specify the parametric forms of the prior distributions  $p(\beta)$  and  $p(\lambda)$ , as well as the parametric forms of the variational distribution factors  $q(\beta)$  and  $q(\lambda)$ . Here, we assume that all these marginal distributions are Gaussian, and hence specified in terms of their expectation and covariance parameters:

$$p(\beta) = N(\beta; \mu_\beta, \Sigma_\beta), \text{ where } \mu_\beta \in \mathbb{R}^p \text{ and } \Sigma_\beta \in \mathbb{R}^{p \times p} \text{ p.d.} \quad (24)$$

$$p(\lambda) = N(\lambda; \mu_\lambda, \Sigma_\lambda), \text{ where } \mu_\lambda \in \mathbb{R}^k \text{ and } \Sigma_\lambda \in \mathbb{R}^{k \times k} \text{ p.d.} \quad (25)$$

$$q(\beta) = N(\beta; m_\beta, S_\beta), \text{ where } m_\beta \in \mathbb{R}^p \text{ and } S_\beta \in \mathbb{R}^{p \times p} \text{ p.d.} \quad (26)$$

$$q(\lambda) = N(\lambda; m_\lambda, S_\lambda), \text{ where } m_\lambda \in \mathbb{R}^k \text{ and } S_\lambda \in \mathbb{R}^{k \times k} \text{ p.d.} \quad (27)$$

Note that we denote parameters of the prior distributions with Greek and parameters of the variational distributions with Roman letters. Together with Equations (1–3), Equations (24–27) specify all distributions necessary to evaluate the free energy integral and render the free energy a function of the variational parameters. We document this derivation in Supplementary Material S1.2 and here limit ourselves to the presentation of the result: under the given assumptions about the prior, likelihood, and variational distributions, the variational free energy is a function of the variational parameters  $m_\beta, S_\beta, m_\lambda$ , and  $S_\lambda$ , and, using mild approximations in its analytical derivation, evaluates to

$$\begin{aligned} F^{VB}(m_\beta, S_\beta, m_\lambda, S_\lambda) = & -\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |V_{m_\lambda}| \\ & - \frac{1}{2} (y - X m_\beta)^T V_{m_\lambda}^{-1} (y - X m_\beta) \\ & - \frac{1}{2} \text{tr}(S_\beta X^T V_{m_\lambda}^{-1} X) - \frac{1}{4} \text{tr}(B_{m_\lambda, S_\beta, m_\lambda} S_\lambda) \\ & - \frac{p}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_\beta| \\ & - \frac{1}{2} (m_\beta - \mu_\beta)^T \Sigma_\beta^{-1} (m_\beta - \mu_\beta) - \frac{1}{2} \text{tr}(\Sigma_\beta^{-1} S_\beta) \\ & - \frac{k}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_\lambda| \\ & - \frac{1}{2} (m_\lambda - \mu_\lambda)^T \Sigma_\lambda^{-1} (m_\lambda - \mu_\lambda) - \frac{1}{2} \text{tr}(\Sigma_\lambda^{-1} S_\lambda) \\ & + \frac{k}{2} \ln(2\pi e) + \frac{1}{2} \ln |S_\beta| \\ & + \frac{p}{2} \ln(2\pi e) + \frac{1}{2} \ln |S_\lambda| \end{aligned} \quad (28)$$

with

$$\begin{aligned}
 B_{m_\beta, S_\beta, m_\lambda} &:= H_{\ln |V_\lambda|} (m_\lambda) \\
 &+ H_{\text{tr}(V_\lambda^{-1} X S_\beta X^T)} (m_\lambda) \\
 &+ H_{(y - X m_\beta)^T V_\lambda^{-1} (y - X m_\beta)} (m_\lambda).
 \end{aligned}
 \tag{29}$$

In Equation (28), the third term may be viewed as an *accuracy term* which measures the deviation of the estimated model prediction from the data, the eighth and twelfth terms may be viewed as *complexity terms*, that measure how far the model can and has to deviate from its prior expectations to account for the data, and the last four terms can be conceived as *maximum entropy* terms that ensure that the posterior parameter uncertainty is as large as possible given the available data (Jaynes, 2003).

In principle, any numerical routine for the maximization of nonlinear functions could be applied to maximize the free energy function of Equation (28) with respect to its parameters. Because of the relative simplicity of Equation (28), we derived explicit update equations by evaluating the VB free energy gradient with respect to each of the parameters and setting to zero as documented in Supplementary Material S1.2. This analytical approach yields a set of four update equations and, together with the iterative evaluation of the VB free energy function (Equation 28), results in a VB algorithm for the current model as documented in Algorithm 1. Here, and in all remaining algorithms, convergence is assessed in terms of a vanishing of the free energy increase between successive iterations. This difference is evaluated against a convergence criterion  $\delta$ , which we set to  $\delta = 10^{-3}$  for all reported simulations.

---

**Algorithm 1** VB Algorithm (for details, see *vbg\_est\_vb.m*)

---

**Input:** data  $y$ , prior parameters  $\mu_\beta, \Sigma_\beta, \mu_\lambda, \Sigma_\lambda$ , model components  $X, Q_1, Q_2$

**Output:** variational parameters  $m_\beta^{(i)}, S_\beta^{(i)}, m_\lambda^{(i)}, S_\lambda^{(i)}$ , free energy  $F^{VB(i)}$

- 1: **Initialization:**  $i := 1, m_\beta^{(i)} := \mu_\beta, S_\beta^{(i)} := \Sigma_\beta, m_\lambda^{(i)} := \mu_\lambda, S_\lambda^{(i)} := \Sigma_\lambda, \Delta F^{VB(i)} := \infty, F^{VB(i)} := F^{VB}(m_\beta^{(i)}, S_\beta^{(i)}, m_\lambda^{(i)}, S_\lambda^{(i)})$
  - 2: **while**  $\Delta F^{VB(i)} > \delta$  **do**
  - 3:  $i := i + 1$
  - 4: **evaluate**  $B_{m_\beta^{(i-1)}, S_\beta^{(i-1)}, m_\lambda^{(i-1)}}$
  - 5:  $S_\lambda^{(i)} := \left( \frac{1}{2} B_{m_\beta^{(i-1)}, S_\beta^{(i-1)}, m_\lambda^{(i-1)}} + \Sigma_\lambda^{-1} \right)^{-1}$
  - 6:  $m_\beta^{(i)} := \left( X^T V_{m_\lambda}^{-1} X + \Sigma_\beta^{-1} \right)^{-1} \left( X^T V_{m_\lambda}^{-1} X y + \Sigma_\beta^{-1} \mu_\beta \right)$
  - 7:  $S_\beta^{(i)} := \left( X^T V_{m_\lambda}^{-1} X + \Sigma_\beta^{-1} \right)^{-1}$
  - 8: **solve**  $\frac{\partial}{\partial m_\lambda} f^{VB}(m_\lambda^{(i)}) = 0$  **for**  $m_\lambda^{(i)}$
  - 9: **evaluate**  $F^{VB(i)} = F^{VB}(m_\beta^{(i)}, S_\beta^{(i)}, m_\lambda^{(i)}, S_\lambda^{(i)})$
  - 10:  $\Delta F^{VB(i)} := F^{VB(i)} - F^{VB(i-1)}$
  - 11: **end while**
- 

In **Figure 2**, we visualize the application of the VB algorithm to an example fMRI time-series realization from the model described in Section 2.1 with true, but unknown, parameter values  $\beta = (2, -1)^T$  and  $\lambda = (-0.5, -2)^T$ . We used imprecise priors for both  $\beta$  and  $\lambda$  by setting

$$\begin{aligned}
 p(\beta) &:= N\left(\beta; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}\right) \text{ and} \\
 p(\lambda) &:= N\left(\lambda; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}\right).
 \end{aligned}
 \tag{30}$$

**Figure 2A** depicts the prior distribution over  $\beta$ , and the true, but unknown, value of  $\beta$  as black  $\times$ . **Figure 2B** depicts the variational distribution over  $\beta$  after convergence for a VB free energy convergence criterion of  $\delta = 10^{-3}$ . Given the imprecise prior distribution, this variational distribution falls close to the true, but unknown, value. In general, convergence of the algorithm is achieved within 4–6 iterations. **Figures 2C,D** depict the prior distribution over  $\lambda$  and the variational distribution over  $\lambda$  upon convergence, respectively. As for  $\beta$ , the approximation of the posterior distribution is close to the true, but unknown, value of  $\lambda$ . Finally, **Figures 2E,F** depict the VB free energy surface as a function of the variational parameters  $m_\beta$  and  $m_\lambda$ , respectively. For the chosen prior distributions, the VB free energy surfaces display clear global maxima, which the VB algorithm can identify. Note, however, that the maximum of the VB free energy as a function of  $m_\lambda$  is located on an elongated crest.

## 2.4. Variational Maximum Likelihood (VML)

Variational Maximum Likelihood (Beal, 2003), also referred to as (variational) expectation-maximization (McLachlan and Krishnan, 2007; Barber, 2012), can be considered a semi-Bayesian estimation approach. For a subset of model parameters, VML determines a Bayesian posterior distribution, while for the remaining parameters maximum-likelihood point estimates are evaluated. As discussed below, VML can be derived as a special case of VB under specific assumptions about the posterior distribution of the parameter set for which only point estimates are desired. If for this parameter set additionally a constant, improper prior is assumed, variational Bayesian inference directly yields VML estimates. In its application to the GLM, we here choose to treat  $\beta$  as the parameter for which a posterior distribution is derived, and  $\lambda$  as the parameter for which a point-estimate is desired.

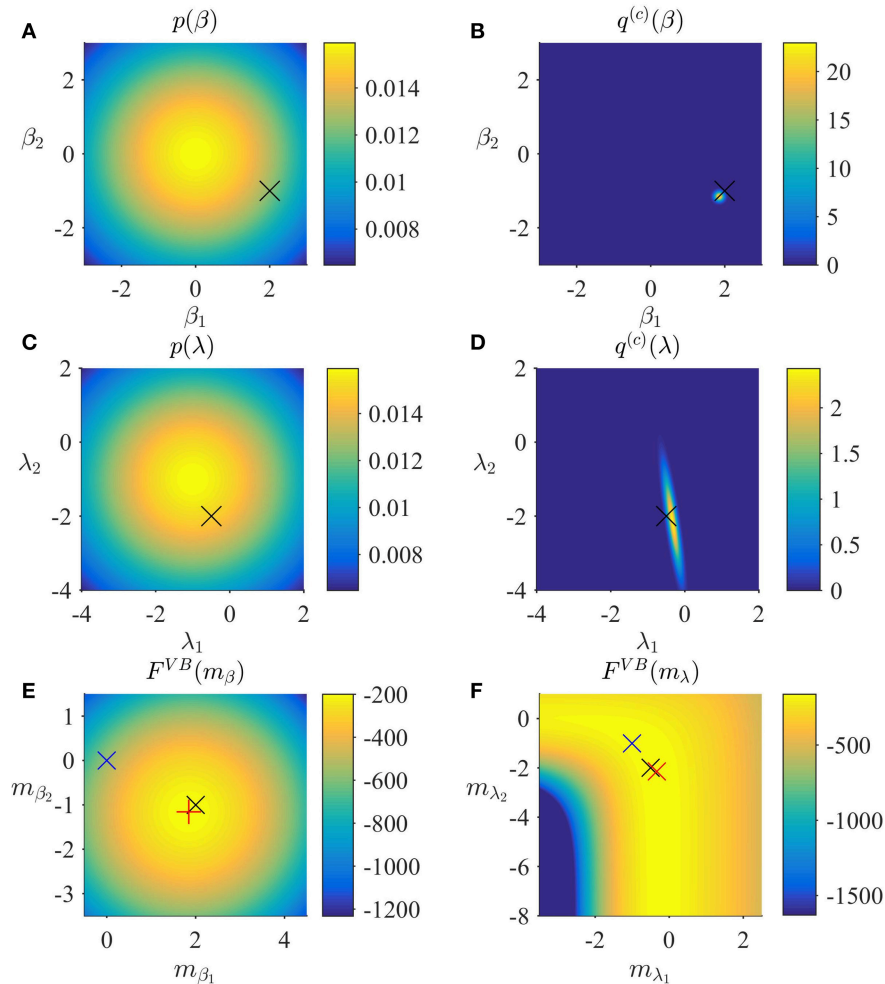
The current probabilistic model of interest thus takes the form

$$p_\lambda(y, \beta) = p_\lambda(y|\beta)p(\beta)
 \tag{31}$$

with likelihood distribution

$$p_\lambda(y|\beta) = N(y; X\beta, V_\lambda).
 \tag{32}$$

Note that in contrast to the probabilistic model underlying VB estimation,  $\lambda$  is not treated as a random variable and thus merely



**FIGURE 2 |** VB estimation. **(A)** Prior distribution  $p(\beta)$  with expectation  $\mu_\beta := (0, 0)^T$  and covariance  $\Sigma_\beta := 10I_2$ . Here, and in all subpanels, the black  $\times$  marks the true, but unknown, parameter value. **(B)** Variational approximation  $q^{(c)}(\beta)$  to the posterior distribution upon convergence ( $\delta = 10^{-3}$ ). **(C)** Prior distribution  $p(\lambda)$  with expectation  $\mu_\lambda := (0, 0)^T$  and covariance  $\Sigma_\lambda = 10I_2$ . **(D)** Variational approximation  $q^{(c)}(\lambda)$  to the posterior distribution upon convergence. **(E)** Variational free energy dependence on  $m_\beta$ . The blue  $\times$  indicates the prior expectation parameter and the red  $\times$  marks the approximated posterior expectation parameter. **(F)** Variational free energy dependence on  $m_\lambda$ . The blue  $\times$  indicates the prior expectation parameter and the red  $\times$  marks the approximated posterior expectation parameter. For implementational details, please see *vbg\_1.m*.

parameterizes the joint distribution of  $\beta$  and  $y$ . Similar to VB, VML rests on a decomposition of the log marginal likelihood

$$\ln p_\lambda(y) = \int p_\lambda(y, \beta) d\beta \tag{33}$$

into a free energy and a KL-divergence term

$$\ln p_\lambda(y) = F^{VML}(q(\beta), \lambda) + KL(q(\beta)||p_\lambda(\beta|y)). \tag{34}$$

In contrast to the VB free energy, the VML free energy is defined by

$$F^{VML}(q(\beta), \lambda) = \int q(\beta) \ln \left( \frac{p_\lambda(y, \beta)}{q(\beta)} \right) d\beta, \tag{35}$$

while the KL divergence term takes the form

$$KL(q(\beta)||p_\lambda(\beta|y)) = \int q(\beta) \ln \left( \frac{q(\beta)}{p_\lambda(\beta|y)} \right) d\beta. \tag{36}$$

In Supplementary Material S2, we show how the VML framework can be derived as a special case of VB by assuming an improper prior for  $\lambda$  and a Dirac measure  $\delta_{\lambda^*}$  for the variational distribution of  $\lambda$ . Importantly, it is the parameter value  $\lambda^*$  of the Dirac measure that corresponds to the parameter  $\lambda$  in the VML framework.

### 2.4.1. Application to the GLM

In the application of the VML approach to the GLM of Equations (1) and (2) we need to specify the parametric forms of the prior distribution  $p(\beta)$  and the parametric form of the variational



distribution  $q(\beta)$ . As above, we assume that these distributions are Gaussian, i.e.,

$$p(\beta) = N(\beta; \mu_\beta, \Sigma_\beta), \text{ where } \mu_\beta \in \mathbb{R}^p \text{ and } \Sigma_\beta \in \mathbb{R}^{p \times p} \text{ p.d.} \quad (37)$$

$$q(\beta) = N(\beta; m_\beta, S_\beta), \text{ where } m_\beta \in \mathbb{R}^p \text{ and } S_\beta \in \mathbb{R}^{p \times p} \text{ p.d.} \quad (38)$$

Based on the specifications of Equations (37) and (38), the integral definition of the VML free energy can be analytically evaluated under mild approximations, which yields the VML free energy function of the variational parameters  $m_\beta$  and  $S_\beta$  and the parameter  $\lambda$

$$\begin{aligned} F^{VML}(m_\beta, S_\beta, \lambda) = & -\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |V_\lambda| - \frac{1}{2} (y - X m_\beta)^T V_\lambda^{-1} (y - X m_\beta) \\ & - \frac{1}{2} \text{tr}(S_\beta X^T V_\lambda^{-1} X) \\ & - \frac{p}{2} \ln 2\pi - \frac{1}{2} \ln |S_\beta| \\ & - \frac{1}{2} (m_\beta - \mu_\beta)^T \Sigma_\beta^{-1} (m_\beta - \mu_\beta) - \text{tr}(\Sigma_\beta^{-1} S_\beta) \\ & + \frac{p}{2} \ln(2\pi e) + \frac{1}{2} \ln |S_\beta|. \end{aligned} \quad (39)$$

We document the derivation of Equation (39) in Supplementary Material S1.3. In contrast to the VB free energy (cf. Equation 28), the VML free energy for the GLM is characterized by the absence of terms relating to the prior and posterior uncertainty about the covariance component parameter  $\lambda$ . To maximize the VML free energy, we again derived a set of update equations as documented in Supplementary Material S1.3. These update equations give rise to a VML algorithm for the current model, which we document in Algorithm 2.

---

**Algorithm 2** VML Algorithm (for details, see *vbg\_est\_vml.m*)

---

**Input:** data  $y$ , prior parameters  $\mu_\beta, \Sigma_\beta$ , initial value  $\lambda^{(1)}$ , model  $X, Q_1, Q_2$

**Output:** variational parameters  $m_\beta^{(i)}, S_\beta^{(i)}, \lambda^{(i)}$ , free energy  $F^{VML(i)}$

- 1: **Initialization:**  $i := 1$  and  $m_\beta^{(i)} := \mu_\beta, S_\beta^{(i)} := \Sigma_\beta, \Delta F^{VML(i)} := \infty$ , and  $F^{VML(i)} := F^{VML}(m_\beta^{(i)}, S_\beta^{(i)}, \lambda^{(i)})$ .
  - 2: **while**  $\Delta F^{VML(i)} > \delta$  **do**
  - 3:  $i := i + 1$
  - 4:  $m_\beta^{(i)} := (X^T V_\lambda^{-1} X + \Sigma_\beta^{-1})^{-1} (X^T V_\lambda^{-1} X y + \Sigma_\beta^{-1} \mu_\beta)$
  - 5:  $S_\beta^{(i)} := (X^T V_\lambda^{-1} X + \Sigma_\beta^{-1})^{-1}$
  - 6: solve  $\frac{\partial}{\partial \lambda} f^{VML}(\lambda^{(i)}) = 0$  for  $\lambda^{(i)}$
  - 7: evaluate  $F^{VML(i)} := F^{VML}(m_\beta^{(i)}, S_\beta^{(i)}, \lambda^{(i)})$
  - 8:  $\Delta F^{VML(i)} := F^{VML(i)} - F^{VML(i-1)}$
  - 9: **end while**
- 

In **Figure 3**, we visualize the application of the VML algorithm to an example fMRI time-series realization of the model described in Section 2.1 with true, but unknown, parameter

values  $\beta = (2, -1)^T$  and  $\lambda = (-0.5, -2)^T$ . As above, we used an imprecise prior for  $\beta$  by setting

$$p(\beta) := N\left(\beta; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}\right). \quad (40)$$

and set the initial covariance component estimate to  $\lambda^{(1)} = (0, 0)^T$ . **Figure 3A** depicts the prior distribution over  $\beta$  and the true, but unknown, value of  $\beta$ . **Figure 3B** depicts the variational distribution over  $\beta$  after convergence with a VML free energy convergence criterion of  $\delta = 10^{-3}$ . As in the VB scenario, given the imprecise prior distribution, this variational distribution falls close to the true, but unknown, value and convergence is usually achieved within 4–6 iterations. **Figures 3C,D** depict the VML free energy surface as a function of the variational parameter  $m_\beta$  and the parameter  $\lambda$ , respectively. For the chosen prior distributions, the VML free energy surfaces displays a clear global maximum as a function of  $m_\beta$ , while the maximum location as a function of  $m_\lambda$  is located on an elongated crest.

## 2.5. Restricted Maximum Likelihood (ReML)

ReML is commonly viewed as a generalization of the maximum likelihood approach, which in the case of the GLM yields unbiased, rather than biased, covariance component parameter estimates (Harville, 1977; Phillips et al., 2002; Searle et al., 2009). In this context and using our denotations, the ReML estimate  $\hat{\lambda}_{ReML}$  is defined as the maximizer of the ReML objective function

$$\hat{\lambda}_{ReML} := \arg \max_{\lambda} \ell_{ReML}(\lambda), \quad (41)$$

where

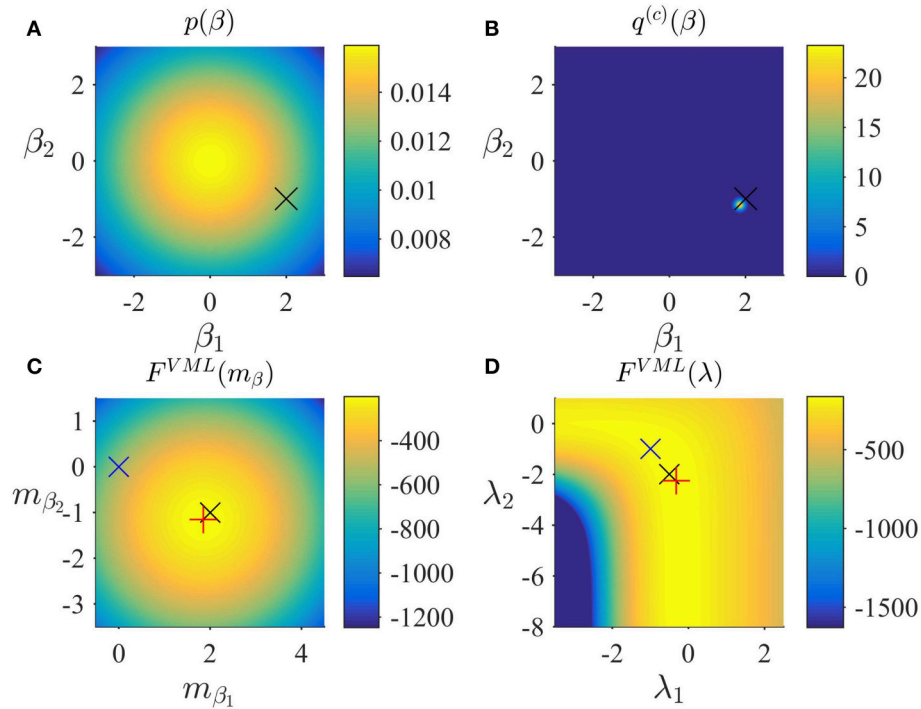
$$\ell_{ReML}(\lambda) := -\frac{1}{2} \ln |V_\lambda| - \frac{1}{2} \ln |X^T V_\lambda^{-1} X| - \frac{1}{2} (y - X \hat{\beta}_{GLS})^T V_\lambda^{-1} (y - X \hat{\beta}_{GLS}) \quad (42)$$

denotes the ReML objective function and

$$\hat{\beta}_{GLS} := (X^T V_\lambda X)^{-1} X^T V_\lambda^{-1} y \quad (43)$$

denotes the generalized least-squares estimator for  $\beta$ . Because  $\hat{\beta}_{GLS}$  depends on  $\lambda$  in terms of  $V_\lambda$ , maximizing the ReML objective function necessitates iterative numerical schemes. Traditional derivations of the ReML objective function, such as provided by LaMotte (2007) and Hocking (2013), are based on mixed-effects linear models and the introduction of a contrast matrix  $A$  with the property that  $A^T X = 0$  and then consider the likelihood of  $A^T y$  after canceling out the deterministic part of the model. In Supplementary Material S1.4 we show that, up to an additive constant, the ReML objective function also corresponds to the VML free energy under the assumption of an improper constant prior distribution for  $\beta$ , and an exact update of the VML free energy with respect to the variational distribution of  $\beta$ , i.e., setting  $q(\beta) = p_\lambda(\beta|y)$ . In other words, for the probabilistic model

$$\begin{aligned} p_\lambda(y, \beta) &= p_\lambda(y|\beta)p(\beta) \text{ with } p_\lambda(y|\beta) = N(y; X\beta, V_\lambda) \text{ and} \\ p(\beta) &:= 1 \end{aligned} \quad (44)$$



**FIGURE 3 |** VML estimation. **(A)** Prior distribution  $p(\beta)$  with expectation  $\mu_\beta := (0, 0)^T$  and covariance  $\Sigma_\beta := 10I/2$ . Here, and in all subpanels, the black  $\times$  marks the true, but unknown, parameter value. **(B)** Variational approximation  $q^{(c)}(\beta)$  to the posterior distribution upon convergence of the algorithm. **(C)** VML free energy dependence on  $m_\beta$ . The blue  $\times$  indicates the prior expectation parameter and the red  $+$  marks the approximated posterior expectation parameter. **(D)** VML free energy dependence on  $\lambda$ . The blue  $\times$  indicates the parameter value at algorithm initialization and the red  $+$  marks the parameter value upon algorithm convergence. For implementational details, please see *vbg\_1.m*.

it holds that

$$F^{VML}(p_\lambda(\beta|y), \lambda) = \ell_{ReML}(\lambda) + c, \tag{45}$$

where

$$c := -\frac{n}{2} \ln 2\pi + \frac{p}{2} \ln(2\pi), \tag{46}$$

and thus

$$\hat{\lambda}_{ReML} = \arg \max_{\lambda} F^{VML}(p_\lambda(\beta|y), \lambda). \tag{47}$$

ReML estimation of covariance components in the context of the general linear model can thus be understood as the special case of VB, in which  $\beta$  is endowed with an improper constant prior distribution, the posterior distribution over  $\lambda$  is taken to be the Dirac measure  $\delta_{\lambda^*}$ , and the point estimate of  $\lambda^*$  maximizes the ensuing VML free energy under exact inference of the posterior distribution of  $\beta$ . In this view, the additional term of the ReML objective function with respect to the ML objective function obtains an intuitive meaning:  $-\frac{1}{2} \ln |X^T V_\lambda^{-1} X|$  corresponds to the entropy of the posterior distribution  $p_\lambda(\beta|y)$  which is maximized by the ReML estimate  $\hat{\lambda}_{ReML}$ . The ReML objective function thus accounts for the uncertainty that stems from estimating of the parameter  $\beta$  by assuming that is as large as possible under the constraints of the data observed.

In line with the discussion of VB and VML, we may define a ReML free energy, by which we understand the VML free energy function evaluated at  $p_\lambda(\beta|y)$  for the probabilistic model (Equation 44). In Supplementary Material S1.4, we show that this ReML free energy can be written as

$$\begin{aligned} F^{ReML}(m_\beta, S_\beta, \lambda) = & -\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |V_\lambda| \\ & - \frac{1}{2} (y - X m_\beta)^T V_\lambda^{-1} (y - X m_\beta) \\ & - \frac{1}{2} \text{tr}(S_\beta X^T V_\lambda^{-1} X) \\ & + \frac{p}{2} \ln(2\pi e) + \frac{1}{2} \ln |S_\beta|. \end{aligned} \tag{48}$$

Note that the equivalence of Equation (48) to the constant-augmented ReML objective function of Equation (45) derives from the fact that under the infinitely imprecise prior distribution for  $\beta$  the variational expectation and covariance parameters evaluate to

$$m_\beta = \hat{\beta}_{GLS} \text{ and } S_\beta = (X^T V_\lambda^{-1} X)^{-1}, \tag{49}$$

respectively. With respect to the general VML free energy, the ReML free energy is characterized by the absence of a term that penalizes the deviation of the variational parameter  $m_\beta$

from its prior expectation, because the infinitely imprecise prior distribution  $p(\beta)$  provides no constraints on the estimate of  $\beta$ . To maximize the ReML free energy, we again derived a set of update equations which we document in Algorithm 3. In **Figure 4**, we

---

**Algorithm 3** ReML Algorithm (for details, see *vbg\_est\_reml.m*)

---

**Input:** data  $y$ , initial values  $m_\beta^{(1)}, S_\beta^{(1)}, \lambda^{(1)}$ , model  $X, Q_1, Q_2$   
**Output:** variational parameters  $m_\beta^{(i)}, S_\beta^{(i)}, \lambda^{(i)}$ , free energy  $F^{ReML^{(i)}}$

- 1: **Initialization:**  $i := 1, \Delta F^{ReML^{(i)}} := \infty$ , and  $F^{ReML^{(i)}} := F^{ReML}(m_\beta^{(i)}, S_\beta^{(i)}, \lambda^{(i)})$ .
- 2: **while**  $\Delta F^{ReML^{(i)}} > \delta$  **do**
- 3:  $m_\beta^{(i)} := (X^T V_\lambda^{-1} X)^{-1} X^T V_\lambda^{-1} y$
- 4:  $S_\beta^{(i)} := (X^T V_\lambda^{-1} X)^{-1}$
- 5: solve  $\frac{\partial}{\partial \lambda_j} f^{ReML}(\lambda^{(i)}) = 0$  for  $\lambda^{(i)}$
- 6: evaluate  $F^{ReML^{(i)}} := F^{ReML}(m_\beta^{(i)}, S_\beta^{(i)}, \lambda^{(i)})$
- 7:  $\Delta F^{ReML^{(i)}} := F^{ReML^{(i)}} - F^{ReML^{(i-1)}}$
- 8: **end while**

---

visualize the application of the ReML algorithm to an example fMRI time-series realization of the model described in Section 2.1 with true, but unknown, parameter values  $\beta = (2, -1)^T$  and  $\lambda = (-0.5, -2)^T$ . Here, we chose the  $\beta$  prior distribution parameters as the initial values for the variational parameters by setting

$$m_\beta^{(1)} := \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ and } S_\beta^{(1)} := \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}, \quad (50)$$

and as above, set the initial covariance component estimate to  $\lambda^{(1)} = (0, 0)^T$ .

**Figure 4A** depicts the converged variational distribution over  $\beta$  and the true, but unknown, value of  $\beta$  for a ReML free energy convergence criterion of  $\delta = 10^{-3}$ . **Figures 4B,C** depict the ReML free energy surface as a function of the variational parameter  $m_\beta$  and  $\lambda$ , respectively. Note that due to the imprecise prior distributions in the VB and VML scenarios, the resulting free energy surfaces are almost identical to the ReML free energy surfaces.

## 2.6. Maximum Likelihood (ML)

Finally, also the ML objective function can be viewed as the special case of the VB log marginal likelihood decomposition for variational distributions  $q(\beta)$  and  $q(\lambda)$  both conforming to Dirac measures. Specifically, as shown in Supplement Material S2 the ML estimate

$$(\hat{\beta}_{ML}, \hat{\lambda}_{ML}) := \arg \max_{\beta, \lambda} \ell^{ML}(\beta, \lambda) := \arg \max_{\beta, \lambda} \ln N(y; X\beta, V_\lambda) \quad (51)$$

corresponds to the maximizer of the VML free energy for the probabilistic model

$$p_\lambda(y, \beta) = p_\lambda(y|\beta)p(\beta) \text{ with } q(\beta) = \delta_{\beta^*}(\beta) \text{ and } p(\beta) = 1, \quad (52)$$

i.e., a Dirac measure  $\delta_{\beta^*}$  for the variational distribution and an improper and constant prior density for the parameter  $\beta$ . Formally, we thus have

$$(\hat{\beta}_{ML}, \hat{\lambda}_{ML}) := \arg \max_{\beta, \lambda} F^{VML}(\delta_{\beta^*}(\beta), \lambda). \quad (53)$$

To align the discussion of ML with the discussion of VB, VML, and ReML, we may define the thus evaluated VML free energy as the *ML free energy*, which is just the standard log likelihood function of the GLM:

$$F^{ML}(\beta, \lambda) = -\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |V_\lambda| - \frac{1}{2} (y - X\beta)^T V_\lambda^{-1} (y - X\beta). \quad (54)$$

Note that the posterior approximation  $q(\beta)$  does not encode any uncertainty in this case, and thus the additional term corresponding to the entropy of this distribution in the ReML free energy vanishes for the case of ML. Finally, to maximize the ML free energy we again derived a set of update equations (Supplementary Material S1.5) which we document in Algorithm 4. In **Figure 5**, we visualize the application of this ML algorithm to an example fMRI time-series realization of the model described in Section 2.1 with true, but unknown, parameter values  $\beta = (2, -1)^T$  and  $\lambda = (-0.5, -2)^T$ , initial parameter settings of  $\beta^{(1)} = (0, 0)^T$  and  $\lambda^{(1)} = (0, 0)^T$ , and ML free energy convergence criterion  $\delta = 10^{-3}$ . **Figure 5A** depicts the ML free energy maximization with respect to  $\beta^{(i)}$  and **Figure 5B** depicts the ML free energy maximization with respect to  $\lambda^{(i)}$ . Note the similarity to the equivalent free energy surfaces in the VB, VML, and ReML scenarios.

---

**Algorithm 4** ML Algorithm (for details, see *vbg\_est\_ml.m*)

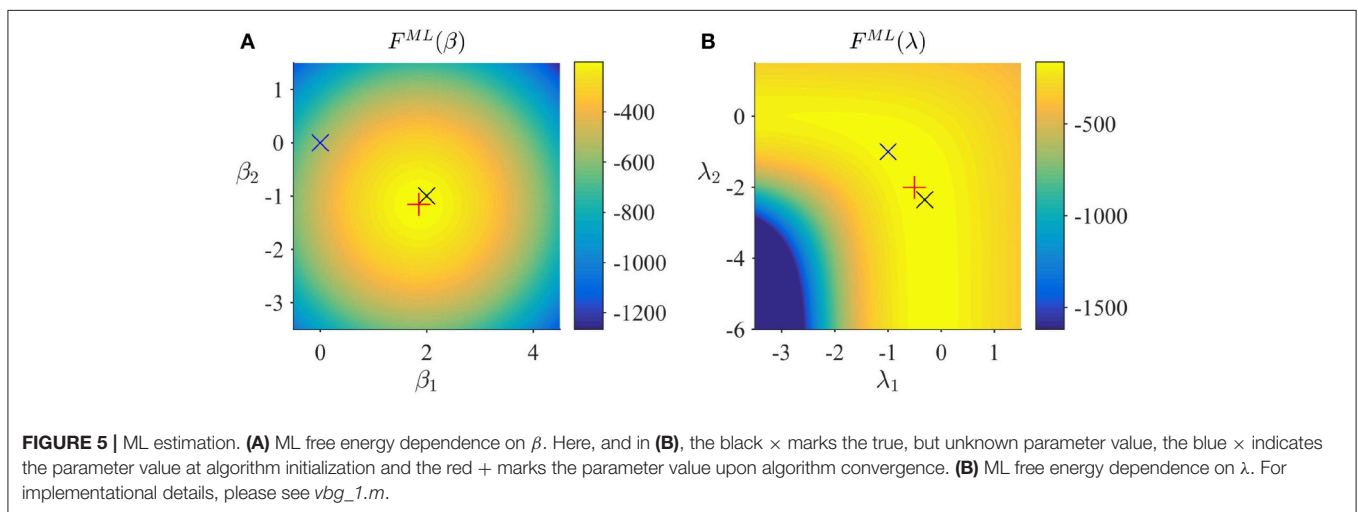
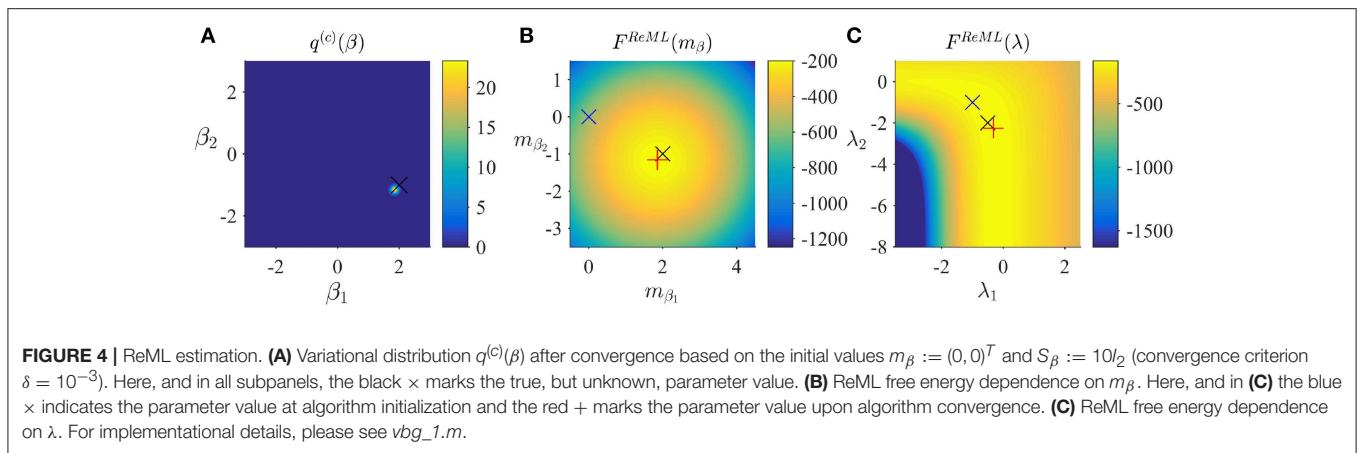
---

**Input:** data  $y$ , initial values  $\beta^{(1)}, \lambda^{(1)}$ , model  $X, Q_1, Q_2$   
**Output:** parameter estimates  $\beta^{(i)}, \lambda^{(i)}$ , free energy  $F^{ML^{(i)}}$

- 1: **Initialization:**  $i := 1, \Delta F^{ML^{(i)}} := \infty, F^{ML^{(i)}} := F^{ML}(\beta^{(i)}, \lambda^{(i)})$ .
- 2: **while**  $\Delta F^{ML^{(i)}} > \delta$  **do**
- 3:  $i := i + 1$
- 4:  $\beta^{(i)} := (X^T V_\lambda^{-1} X)^{-1} X^T V_\lambda^{-1} y$
- 5: solve  $\frac{\partial}{\partial \lambda_j} f^{ML}(\lambda^{(i)}) = 0$  for  $\lambda^{(i)}$
- 6:  $F^{ML^{(i)}} := F^{ML}(\beta^{(i)}, \lambda^{(i)})$
- 7:  $\Delta F^{ML^{(i)}} := F^{ML^{(i)}} - F^{ML^{(i-1)}}$
- 8: **end while**

---

In summary, in this section we have shown how VML, ReML, and ML estimation can be understood as special case of VB estimation. In the application to the GLM, the hierarchical nature of these estimation techniques yields a nested set of free energy objective functions, in which gradually terms that quantify uncertainty about parameter subsets are eliminated (cf. Equations 28, 39, 48, and 54). In turn, the iterative maximization of these objective functions yields a nested set of numerical algorithms, which assume gradually less complex formats (**Algorithms 1–4**). As shown by the numerical examples,



under imprecise prior distributions, the resulting free energy surfaces and variational (expectation) parameter estimates are highly consistent across the estimation techniques. Finally, for all techniques, the relevant parameter estimates converge to the true, but unknown, parameter values after a few algorithm iterations.

### 3. APPLICATIONS

In Section 2 we have discussed the conceptual relationships and the algorithmic implementation of VB, VML, ReML, and ML in the context of the GLM and demonstrated their validity for a single simulated data realization. In the current section, we are concerned with their performance over a large number of simulated data realizations (Section 3.1) and their exemplary application to real experimental data (Section 3.2).

#### 3.1. Simulations

Classical statistical theory has established a variety of criteria for the assessment of an estimator’s quality (e.g., Lehmann and Casella, 2006). Commonly, these criteria amount to the analytical evaluation of an estimator’s large sample behavior. In the current section we adopt the spirit of this approach in simulations. To

this end, we first capitalize on an objective Bayesian standpoint (Bernardo, 2003) by employing imprecise prior distributions to focus on the estimation techniques’ ability to recover the true, but unknown, parameters of the data generating model and the model structure itself. Specifically, we investigate the cumulative average and variance of the  $\beta$  and  $\lambda$  parameter estimates under VB, VML, ReML, and ML and the ability of each technique’s (marginal) likelihood approximation to distinguish between different data generating models. In a second step, we then demonstrate exemplarily how parameter prior specifications can induce divergences in the relative estimation qualities of the techniques.

##### 3.1.1. Parameter Recovery

To study each estimation technique’s ability to recover true, but unknown, model parameters, we drew 100 realizations of the example model discussed in Section 2.1 and focussed our evaluation on the cumulative averages and variances of the converged (variational) parameter estimates  $m_\beta^{(c)} \in \mathbb{R}^2$  (VB, VML, ReML),  $\beta^{(c)} \in \mathbb{R}^2$  (ML),  $m_\lambda^{(c)} \in \mathbb{R}^2$  (VB), and  $\lambda^{(c)} \in \mathbb{R}^2$  (VML, ReML, ML). The simulations are visualized

in **Figure 6**. Each panel column of **Figure 6** depicts the results for one of the estimation techniques, and each panel row depicts the results for one of the four parameter values of interest. Each panel displays the cumulative average of the respective parameter estimate. Averages relating to estimates of  $\beta$  are depicted in blue, averages relating to estimates of  $\lambda$  are depicted in green. In addition to the cumulative average, each panel shows the cumulative variance of the parameter estimates as shaded area around the cumulative average line, and the true, but unknown, values  $\beta = (2, 1)^T$  and  $\lambda = (-0.5, -2)^T$  as gray line. Overall, parameter recovery as depicted here is within acceptable bounds and the estimates variances are tolerable. While there are no systematic differences in parameter recovery across the four estimation techniques, there are qualitative differences in the recovery of effect size and covariance component parameters. For all techniques, the recovery of the effect size parameters is unproblematic and highly reliable. The recovery of covariance component recovery, however, fails in a significant amount of approximately 15–20% of data realizations. In the panels relating to estimates of  $\lambda$  in **Figure 6**, these cases have been removed using an automated outlier detection approach (Grubbs, 1969). In the outlying cases, the algorithms converged to vastly different values, often deviating from the true, but unknown, values by an order of magnitude (for a summary of the results without outlier removal, please refer to Supplementary Material S3). To assess whether this behavior was specific to our implementation of the algorithms, we also evaluated the de-facto neuroimaging community standard for covariance component estimation, the *spm\_reml.m* and *spm\_reml\_sc.m* functions of the SPM12 suite in the same model scenario. We report these simulations as Supplementary Material S4. In brief, we found a similar covariance component (mis)estimation behavior as in our implementation.

Further research revealed that the relative unreliability of algorithmic covariance component estimation is a well-known phenomenon in the statistical literature (e.g., Groeneveld and Kovac, 1990; Boichard et al., 1992; Groeneveld, 1994; Foulley and van Dyk, 2000). We see at least two possible explanations in the current case. Firstly, we did not systematically explore the behavior of the algorithmic implementation for different initial values. It is likely, that the number of estimation outliers can be reduced by optimizing, for each data realization, the algorithm's starting conditions. However, also in this case, an automated outlier detection approach would be necessary to optimize the respective initial values. Secondly, we noticed already in the demonstrative examples in Section 2, that the free energy surface with respect to the covariance components is not as well-behaved as for the effect sizes. Specifically, the maximum is located on an elongated crest of the function, which is relatively flat (see e.g., **Figure 5B**) and hence impedes the straight-forward identification of the maximizing parameter value (see also Figure 4 of Groeneveld and Kovac, 1990 for a very similar covariance component estimation objective function surface). In the Discussion section, we suggest a number of potential remedies for the observed outlier proneness of the covariance

component estimation aspect of the VB, VML, ReML, and ML estimation techniques.

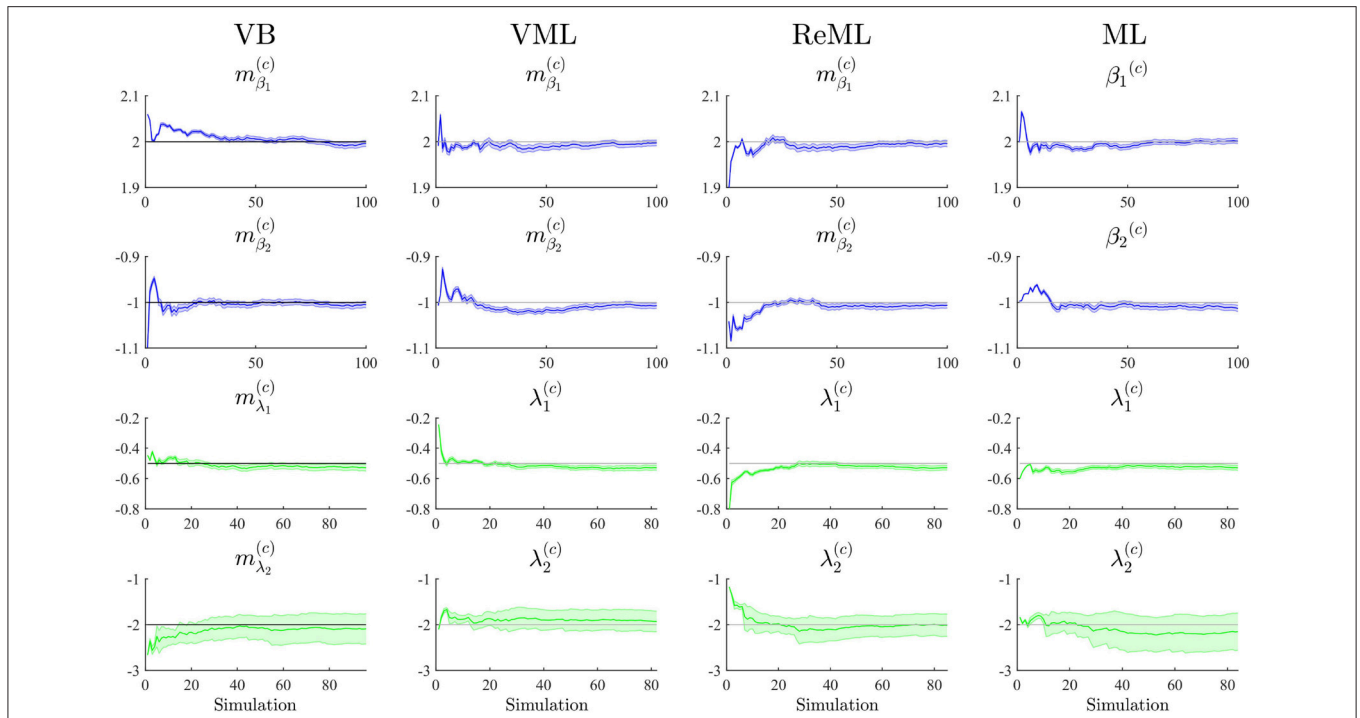
### 3.1.2. Model Recovery

Having established overall reasonable parameter recovery properties for our implementation of the VB, VML, ReML, and ML estimation techniques, we next investigated the ability of the respective techniques' (marginal) log likelihood approximations to recover true, but unknown, model structures. We here focussed on the comparison of two data generating models that differ in the design matrix structure and have identical error covariance structures. Model MG1 corresponds to the first column of the example design matrix of **Figure 1** and thus is parameterized by a single effect size parameter. Model MG2 corresponds to the model used in all previous applications comprising two design matrix columns. To assess the model recovery properties of the different estimation techniques, we generated 100 data realizations based on each of these two models with true, but unknown, effect size parameter values of  $\beta_1 = 2$  (MG1 and MG2) and  $\beta_2 = -1$  (MG2 only), and covariance component parameters  $\lambda = (-0.5, -2)^T$  (MG1 and MG2), as in the previous simulations. We then analyzed each model's data realizations with data analysis models that corresponded to only the single data-generating design matrix regressor (MA1) or both regressors (MA2) for each of the four estimation techniques.

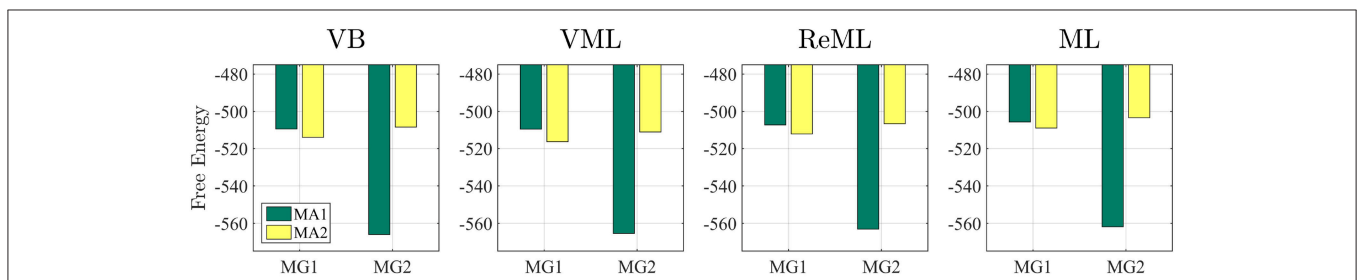
The results of this simulation are visualized in **Figure 7**. For each estimation technique (panels), the average free energies, after exclusion of outlier estimates for the covariance component parameters, are visualized as bars. The data-generating models MG1 and MG2 are grouped on the x-axis and the data-analysis models are grouped by bar color (MA1 green, MA2 yellow). As evident from **Figure 7**, the correct analysis model obtained the higher free energy, i.e., log model evidence approximation, for both data-generating models across all estimation techniques. This difference was more pronounced when analysing data generated by model MG2 than when analysing data generated by model MG1. In this case, the observed data pattern is clearly better described by MA2. In the case of the data-generating model MG1, data analysis model MA2 can naturally account for the observed data by estimating the second effect size parameter to be approximately zero. Nevertheless, this additional model flexibility is penalized correctly by all algorithms, such that the more parsimonious data analysis model MA1 assumes the higher log model evidence approximation also in this case. We can thus conclude that model recovery is achieved satisfactorily by all estimation techniques. A more detailed decomposition of the average free energies into the respective free energy's sum terms is provided in Supplementary Material S5.

### 3.1.3. Estimation Quality Divergences

Thus far, we have concentrated on the nested character of VML, ReML, and ML in VB and demonstrated that for the current model application the maximum-a-posteriori (MAP) estimates of VB and VML and the point estimates of ReML and ML are able to recover true, but unknown, parameter values. Naturally, the four estimation techniques differ in the information they provide upon estimation: VB estimates quantify posterior uncertainty



**FIGURE 6 |** Parameter recovery. The panels along the figure’s columns depict the cumulative averages (blue/green lines), cumulative variances (blue/green shaded areas), and true, but unknown, parameter values (gray lines) for VB, VML, ReML, and ML estimation. Parameter estimates relating to the effect sizes  $\beta$  are visualized in blue, parameter estimates relating to the covariance components  $\lambda$  are visualized in green. The panels along the figure’s rows depict the parameter recovery performance for the subcomponents of the effect size parameters (row 1 and 2) and covariance component parameters (row 3 and 4), respectively. The covariance component parameter estimates are corrected for outliers as discussed in the main text. For implementational details, please see *vbg\_2.m*.

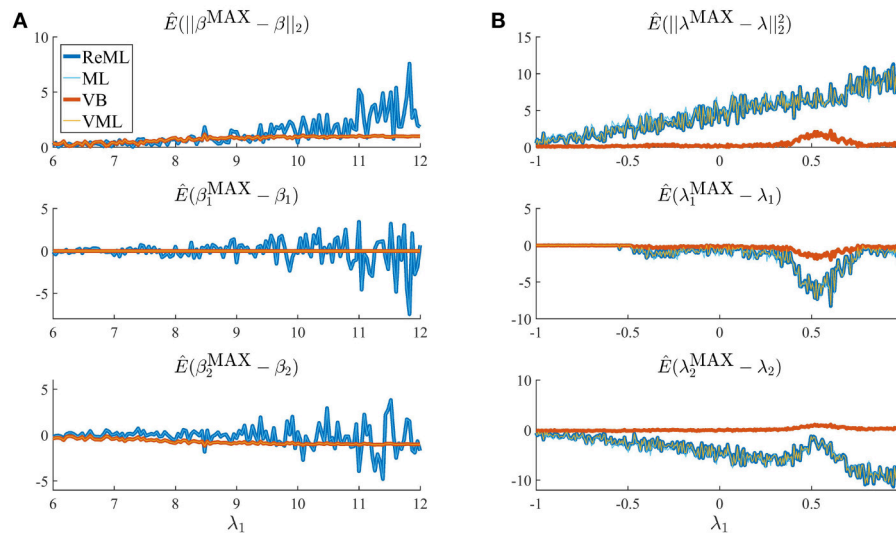


**FIGURE 7 |** Model recovery. Each panel depicts the average free energies of the indicated estimation technique over 100 data realizations. Two data generating models (MG1 and MG2, panel x-axis) were used and analyzed in a cross-over design with two data analysis models (MA1 and MA2, bar color). MG1 and MA1 comprise the same single column design matrix, and MG2 and MA2 comprise the same two column design matrix. Models MG1 and MA1 are nested in MG2 and MA2. Across all estimation techniques, the correct data generating model is identified as indexed by the respective higher free energy log model evidence approximation. For implementational details, please see *vbg\_3.m*.

about both effect size and covariance component parameters, VML estimates quantify posterior uncertainty about effect size parameters only, and ReML and ML do not quantify posterior uncertainty about either parameter class. Beyond these conceptual divergences, an interesting question concerns the qualitative and quantitative differences in estimation that result from the estimation techniques’ specific characteristics. In general, while the properties of ML estimates are fairly well understood from a classical frequentist perspective, the same cannot be said for the other techniques (e.g., Blei et al., 2016).

We return to this point in the Discussion section. In the current section, we demonstrate divergences in the quality of parameter estimation that emerge in high noise scenarios, which are able to uncover prior distribution induced regularization effects. We demonstrate this for both effect size (Figure 8A) and covariance component parameters (Figure 8B) in the example model described in Section 2.1.

The panels in Figure 8A depict simulation estimates of the the root-mean-square-error (RMSE)  $E(\|\beta^{MAX} - \beta\|_2)$  (uppermost panel) and biases of the effect size parameter entries  $E(\beta_1^{MAX} - \beta_1)$



**FIGURE 8 |** Estimation quality divergences. Each panel depicts the estimated RMSE and estimation bias for all four estimation techniques over a range of noise levels parameterized by  $\lambda_1$ . The estimation techniques are color and linewidth coded. **(A)** Visualizes a simulation with focus on the effect size parameter estimates  $\beta$ , **(B)** visualizes a simulation with focus on the covariance component parameters  $\lambda$ . For a detailed description of the simulation, please refer to the main text and for implementational details, please see *vbg\_4.m*. Note that for **(A)**, the results of VB and VML and the results of ReML and ML coincide, and for **(B)** the results of ReML and VML coincide.

and  $E(\beta_2^{\text{MAX}} - \beta_2)$  (middle and lowermost panel, respectively) over a range of values of the first covariance component parameter  $\lambda_1$ . Here,  $\beta^{\text{MAX}} = (\beta_1^{\text{MAX}}, \beta_2^{\text{MAX}})^T$  denotes the MAP estimates resulting under the VB and VML techniques, and the maximum (restricted) likelihood estimates resulting under ReML and ML,  $\beta$  denotes the true, but unknown, effect size parameter,  $E(\cdot)$  denotes the expectation parameter,  $\hat{E}(\cdot)$  the estimation of an expectation by means of an average, and  $\|\cdot\|_2$  denotes the Euclidean norm of a vector. The results for the different estimation techniques are color- and linewidth-coded and were obtained under the following simulation: the true, but unknown, effect size parameter values were set to  $\beta = (1, 1)^T$  and the true, but unknown, parameter value of the second covariance component parameter was constant at  $\lambda_2 = -2$ . Varying the true, but unknown, value  $\lambda_1$  on the interval  $[6, 12]$  thus increased the contribution of independent and identically distributed noise to the data. For each estimation technique, the respective effect size estimates were initialized as specified in **Table 1**. In brief, the estimates for  $\beta_1$  were initialized to the true, but unknown, value and  $\beta_2$  to zero. Crucially, VB and VML allow for the specification of prior distributions over  $\beta$ . Here, we used a precise prior covariance of  $\Sigma_{\beta_1} = 10^{-2}$  and an imprecise variance of  $\Sigma_{\beta_2} = 10^1$ . Note that these algorithm parameters do not exist in ReML and ML. For each setting of  $\lambda_1$ , 100 realizations of the model were obtained, subjected to all four estimation techniques, and the RMSE and biases estimated by averaging over realizations. The following pattern of results emerges: in terms of the RMSE (upper panel), VB and VML exhibit a more stable estimation of  $\beta$ , with a lower deviation from zero compared to the trend of ReML and ML estimates, at higher noise levels. In more detail, this pattern results from the following

effects on the individual  $\beta_1$  and  $\beta_2$  estimates: first, for VB and VML, the estimates  $\beta_1$  exhibit virtually no biases, because their precise prior distribution fixes them at the true, but unknown value, (middle panel). For  $\beta_2$  this regularization of  $\beta_1$  results in more stable estimates as compared to ReML and ML, but for higher levels of noise also results in a downward bias (lowermost panel). Taken together, this simulation demonstrates, how, in the case of prior knowledge about the effect size parameters, the endowment of their estimates with precise priors in VB and VML can stabilize the overall effect size estimation and yield better estimates compared to ReML and ML.

In a second simulation, visualized in **Figure 8B**, we investigated the interaction between prior regularization and estimation quality for the covariance component parameters. As in **Figure 8A**, the uppermost panel depicts the estimated RMSE for the  $\lambda$  parameters, and the middle and lowermost panels the biases of each component parameter. As in the previous simulation, the true, but unknown, effect size parameter values were set to  $\beta = (1, 1)$  and  $\lambda_2 = -2$  and  $\lambda_1$  was varied on the interval  $[-1, 1]$ . The initial parameters for each estimation technique are documented in **Table 1**. In brief, all effect size parameter estimates (expectations) were initialized to zero, and isotropic, imprecise prior covariance matrices were employed for VB and VML. The only estimation technique that endows  $\lambda$  estimates with a prior distribution is VB. Here, we employ the imprecise prior covariance  $\Sigma_{\lambda} := 10^1 I_2$ , which is, however, “precise enough” to exert some stabilization effects: as shown in the uppermost panel of **Figure 8B**, only the RMSE of the VB technique remains largely constant over the investigated space of  $\lambda_1$  values, while for all other estimation techniques the RMSE increases linearly. Two things are noteworthy here. First,

**TABLE 1** | Parameter initialization for the simulations reported in **Figures 8A,B** design.

	VB				VML			ReML		ML	
	$m_{\beta}^{(1)}$	$S_{\beta}^{(1)}$	$m_{\lambda}^{(1)}$	$S_{\lambda}^{(1)}$	$m_{\beta}^{(1)}$	$S_{\beta}^{(1)}$	$\lambda^{(1)}$	$\beta^{(1)}$	$\lambda^{(1)}$	$\beta^{(1)}$	$\lambda^{(1)}$
<b>8A</b>	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0.01 & 0 \\ 0 & 10 \end{pmatrix}$	$\begin{pmatrix} -1 \\ -1 \end{pmatrix}$	$\begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0.01 & 0 \\ 0 & 10 \end{pmatrix}$	$\begin{pmatrix} -1 \\ -1 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -1 \\ -1 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -1 \\ -1 \end{pmatrix}$
<b>8B</b>	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}$	$\begin{pmatrix} -1 \\ -1 \end{pmatrix}$	$\begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}$	$\begin{pmatrix} -1 \\ -1 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -1 \\ -1 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -1 \\ -1 \end{pmatrix}$

at the level of the  $\beta$  estimates all techniques perform equally well in a bias-free manner (data not shown). Second, the  $\lambda_1$  parameter space investigated includes a region (around 0.5) for which also the VB estimation quality declines, but recovers thereafter, suggesting an interaction between the structural model properties and the parameter regime. For the individual entries of  $\lambda$ , the decline in estimation quality in VML, ReML, and ML is not uniform: for  $\lambda_1$ , the estimation quality remains largely constant up to the critical region around 0.5, whereas the estimation quality of  $\lambda_2$  deteriorates with increasing values of  $\lambda_1$  and recovers briefly in the critical region around 0.5. Note that for both simulations of **Figure 8** we did not attempt to remove potential estimation outliers, because their definition in high noise scenarios is virtually impossible. It is thus likely, that the convergence failures observed in the first set of simulations contribute to the observed estimation errors. However, because these failures also afflict the VB and VML techniques which displayed improved estimation behavior in the simulations reported in **Figure 8**, it is likely that the observed pattern of results is indicative of qualitative estimation divergences.

In summary, in the reported simulations we tried to evaluate our implementation of VB, VML, ReML, and ML estimation techniques for a typical neuroimaging data analysis example. In our first simulation set, we observed generally satisfactory parameter recovery for imprecise priors, with the exception of covariance component parameter recovery on a subset of data realizations. In our second simulation, we additionally observed satisfactory model recovery. In our last set of simulations, we probed for estimation quality divergences between the techniques and could show how regularizing prior distributions of the advanced estimation techniques VB and VML can aid to stabilize effect size and covariance component parameter estimation. Naturally, the reported simulations are conditional on our chosen model structure, the true, but unknown, parameter values, and the algorithm initial conditions (prior distributions), and hence not easily generalizable.

### 3.2. Application to Real Data

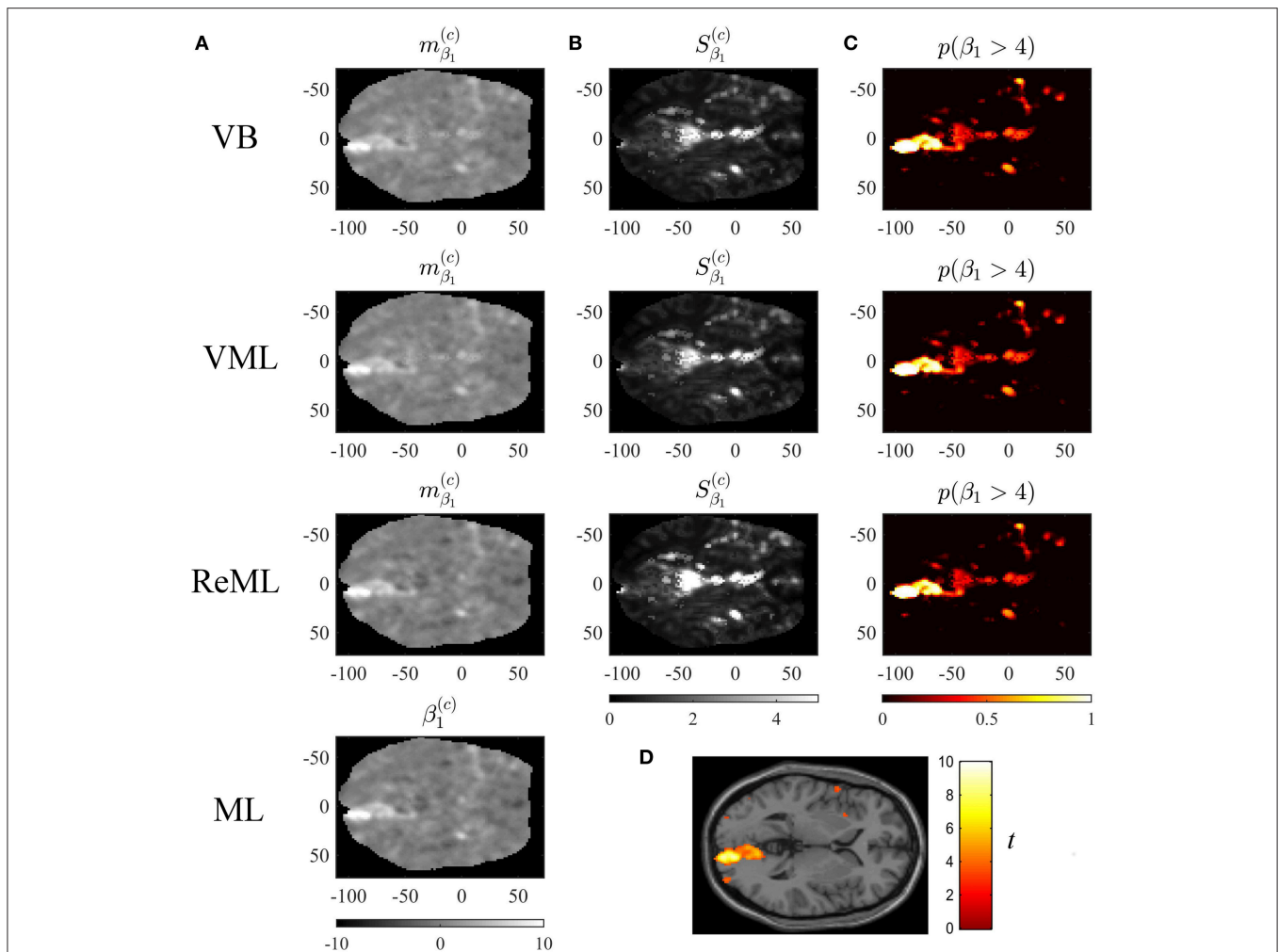
Having validated the VB, VML, ReML, and ML implementation in simulations, we were interested in their application to real experimental data with the main aim of demonstrating the possible parameter inferences that can (and cannot) be made with each technique. To this end, we applied VB, VML, ReML, and ML to a single participant fMRI data set acquired under visual checkerboard stimulation as originally reported in

Ostwald et al. (2010). In brief, the participant was presented with a single reversing left hemi-field checkerboard stimulus for 1 s every 16.5–21 s. These relatively long inter-stimulus intervals were motivated by the fact that the data was acquired as part of an EEG-fMRI study that investigated trial-by-trial correlations between EEG and fMRI evoked responses. Stimuli were presented at two contrast levels and there were 17 stimulus presentations per contrast level. 441 volumes of T2\*-weighted functional data were acquired from 20 slices with  $2.5 \times 2.5 \times 3$  mm resolution and a TR of 1.5 s. The slices were oriented parallel to the AC-PC axis and positioned to cover the entire visual cortex. Data preprocessing using SPM5 included anatomical realignment to correct for motion artifacts, slice scan time correction, re-interpolation to  $2 \times 2 \times 2$  mm voxels, anatomical normalization, and spatial smoothing with a 5 mm FWHM Gaussian kernel. For full methodological details, please see Ostwald et al. (2010).

To demonstrate the application of VB, VML, ReML, and ML to this data set, we used the SPM12 facilities to create a three-column design matrix for the mass-univariate analysis of voxel time-course data. This design matrix included HRF-convolved stimulus onset functions for both stimulus contrast levels and a constant offset. The design matrix is visualized in **Figure 10C**. We then selected one slice of the preprocessed fMRI data (MNI plane  $z = 2$ ) and used our implementation of the four estimation techniques to estimate the corresponding three effect size parameters  $\beta \in \mathbb{R}^3$  and the covariance component parameters  $\lambda \in \mathbb{R}^2$  of the two covariance basis matrices introduced in Section 2.1 for each voxel. We focus our evaluation on the resulting variational parameter estimates of the effect size parameter  $\beta_1$ , corresponding to the high stimulus contrast, and the first covariance component parameter  $\lambda_1$ , corresponding to the isotropic error component. In line with the common practice in neuroimaging data analysis, no outlier removal was performed for the latter parameter. The results are visualized in **Figures 9, 10**.

**Figure 9** visualizes the parameter estimates relating to the effect size parameter  $\beta_1$ . The subpanels of **Figure 10A** depict the resulting two-dimensional map of converged variational parameter estimates, which differs only minimally between the four estimation techniques as indicated on the left of each panel. The variational parameter estimates are highest in the area of the right primary visual cortex, and lowest in the area of the cisterna ambiens/lower lateral ventricles. **Figure 10B** depicts the associated variational covariance parameter  $S_{\beta_1}^{(c)}$ , i.e., the first





**FIGURE 9 |** Effect size estimation. The figure panels depict the effect size parameter  $\beta_1$  estimation results of the VB, VML, ReML, and ML algorithm application to the analysis of a single-participant single-run fMRI data set. This effect size parameter captures the effect of high contrast left visual hemifield checkerboard stimuli as encoded by the first column of the design matrix shown in (C). The first column (A) displays the converged expectation parameter estimates, the second column (B) the associated variance estimates, and the third column (C) the posterior probability for the true, but unknown, effect size parameter to assume values larger than 4. For visual comparison, (D) depicts the result of a standard GLM data analysis of the same data set using SPM12. For implementational details, please see *vbg\_5.m*.

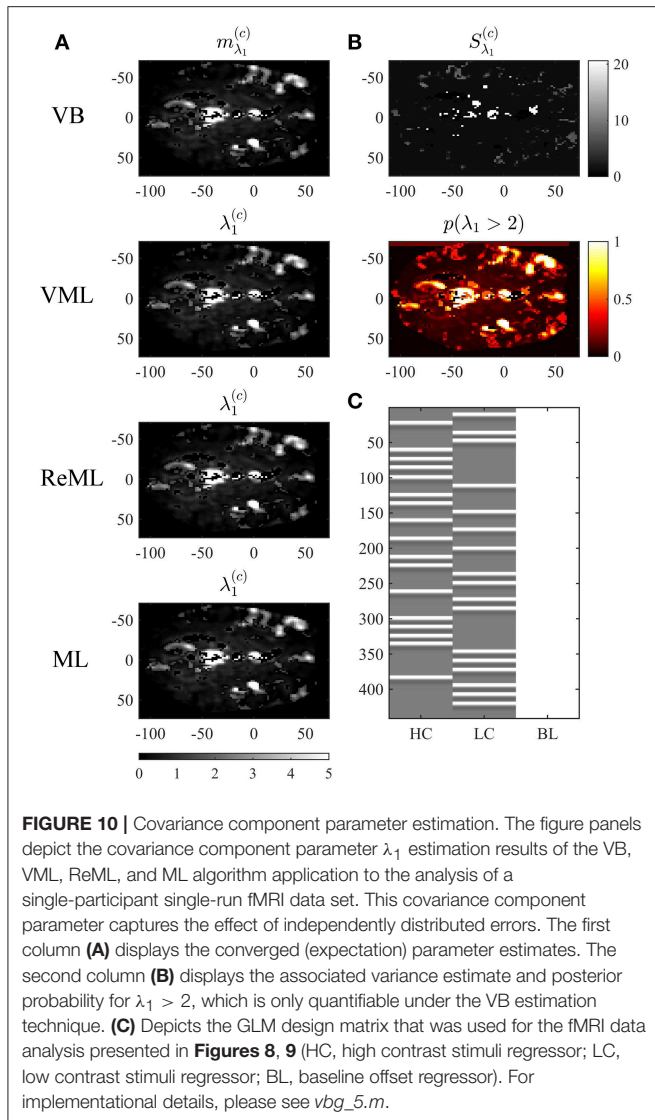
diagonal entry of the of the variational covariance matrix  $S_{\beta}^{(c)} \in \mathbb{R}^{3 \times 3}$ . Here, the highest uncertainty is observed for ventricular locations and the right medial cerebral artery. Overall, the uncertainty estimates are marginally more pronounced for the VB and VML techniques compared to the ReML estimates. Note that the ML technique does not quantify the uncertainty of the GLM effect size parameters. Based on the variational parameters  $m_{\beta_1}^{(c)}$  and  $S_{\beta_1}^{(c)}$ , **Figure 10C** depicts the probability that the true, but unknown, effect size parameter is larger than  $\eta = 4$ , i.e.,

$$p(\beta_1 > \eta) = 1 - N_{cdf}(\eta; m_{\beta_1}, S_{\beta_1}), \quad (55)$$

where  $N_{cdf}$  denotes the univariate Gaussian cumulative density function. Here, the stimulus-contralateral right hemispheric primary visual cortex displays the highest values and the differences between VB, VML, and ReML are marginal. For

comparison, we depict the result of a classical GLM analysis with contrast vector  $c = (1, 0, 0)^T$  at an uncorrected cluster-defining threshold of  $p < 0.001$  and voxel number threshold of  $k = 0$  overlaid on the canonical single participant T1 image in **Figure 9D**. This analysis also identifies the right lateral primary visual cortex as area of strongest activation—but in contrast to the VB, VML, and ReML results does not provide a visual account of the uncertainty associated with the parameter estimates and ensuing T-statistics. In summary, the VB, VML, and ReML-based quantification of effect sizes and their associated uncertainty revealed biologically meaningful results.

**Figure 10** visualizes the variational expectation parameters relating to the effect size parameter  $\lambda_1$ . Here, the subpanels of **Figure 10A** visualize the variational (expectation) parameters across the four estimation techniques. High values for this covariance component are observed in the areas covering



cerebrospinal fluid (cisterna ambiens, lateral and third ventricles), lateral frontal areas, and the big arteries and veins. Notably, also in right primary visual cortex, the covariance component estimate is relatively large, indicating that the design matrix does not capture all stimulus-induced variability. The only estimation technique that also quantifies the uncertainty about the covariance component parameters is VB. The results of this quantification are visualized in **Figure 10B**. The first subpanel visualizes the variational covariance parameter  $S_{\lambda_1}^{(c)}$ , i.e., the first diagonal entry of the variational covariance matrix  $S_{\lambda}^{(c)} \in \mathbb{R}^{2 \times 2}$ . The second subpanel visualizes the probability that the true, but unknown, covariance component parameter  $\lambda$  is larger than  $\eta = 2$ , i.e.,

$$p(\lambda_1 > \eta) = 1 - N_{cdf}(\eta; m_{\lambda_1}, S_{\lambda_1}), \quad (56)$$

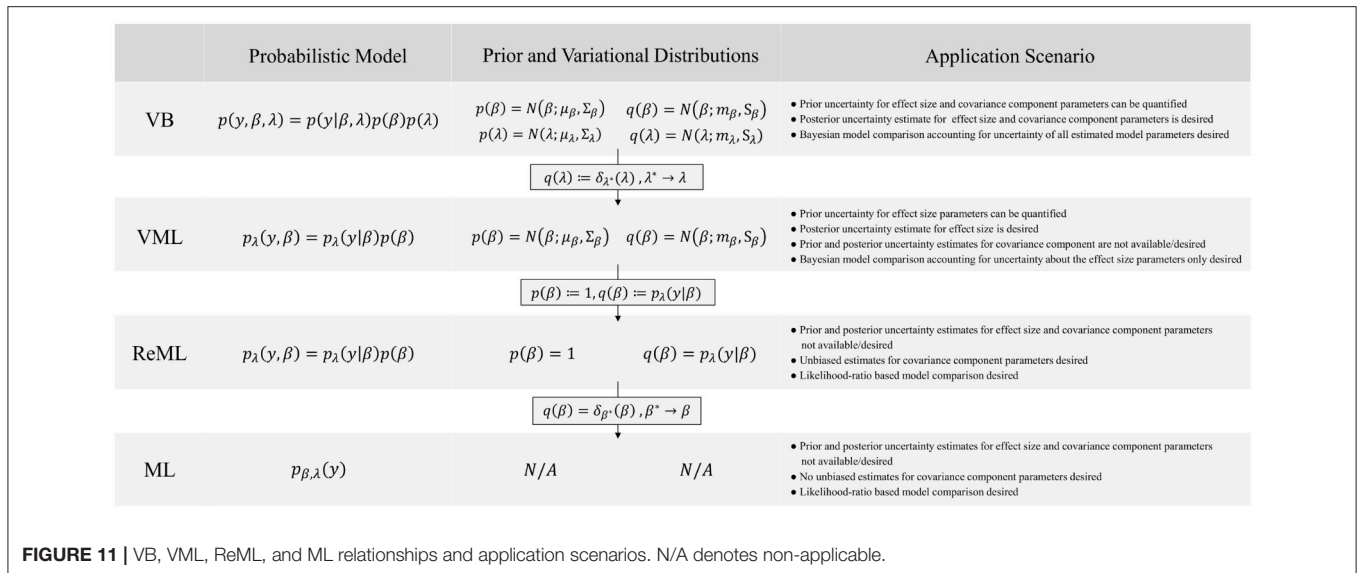
which, due to the relatively low uncertainty estimates  $S_{\lambda_1}$  shows high similarity with the variational expectation parameter map.

In summary, our exemplary application of VB, VML, ReML, and ML to real experimental data revealed biologically sensible results for both effect size and covariance component parameter estimates.

### 4. DISCUSSION

In this technical study, we have reviewed the mathematical foundations of four major parametric statistical parameter estimation techniques that are routinely employed in the analysis of neuroimaging data. We have detailed, how VML (expectation-maximization), ReML, and ML parameter estimation can be viewed as special cases of the VB paradigm. We summarize these relationships and the non-technical application scenarios in which each technique corresponds to the method of choice in **Figure 11**. Further, we have provided a detailed documentation of the application of these four estimation techniques to the GLM with non-spherical, linearly decomposable error covariance, a fundamental modeling scenario in the analysis of fMRI data. Finally, we validated the ensuing iterative algorithms with respect to both simulated and real experimental fMRI data. In the following, we relate our exposition to previous treatments of similar topic matter, discuss potential future work on the qualitative properties of VB parameter estimation techniques, and finally comment on the general relevance of the current study.

The relationships between VB, VML, ReML, and ML have been previously pointed out in Friston et al. (2002a) and Friston et al. (2007). In contrast to the current study, however, Friston et al. (2002a) and Friston et al. (2007) focus on high-level general results and provide virtually no derivations. Moreover, when introducing VB in Friston et al. (2007), the GLM with non-spherical, linearly decomposable error covariance is treated as one of a number of model applications and is not studied in detail across all estimation techniques. From this perspective, the current study can be understood as making many of the implicit results in Friston et al. (2002a) and Friston et al. (2007) explicit and filling in many of the detailed connections and consequences, which are implied by Friston et al. (2002a) and Friston et al. (2007). The relationship between VB and VML has been noted already from outset of the development of the VB paradigm (Beal, 2003; Beal and Ghahramani, 2003). In fact, VB was originally motivated as a generalization of the EM algorithm (Neal and Hinton, 1998; Attias, 2000). However, these treatments do not provide an explicit derivation of VML from VB based on the Dirac measure and do not make the connection to ReML. Furthermore, these studies do not focus on the GLM and its application in the analysis of fMRI data. Finally, a number of treatises have considered the application of VB to linear regression models (e.g., Bishop, 2006; Tzikas et al., 2008; Murphy, 2012). However, these works do not consider non-spherical linearly decomposable error covariance matrices and also do not make the connection to classical statistical estimation using ReML for functional neuroimaging. Taken together, the current study complements the existing literature with its emphasis on the mathematical traceability of the relationship between VB,



VML, ReML, and ML, its focus on the GLM application, and its motivation from a functional neuroimaging background.

### 4.1. Estimator Quality

Model estimation techniques yield estimators. Estimators are functions of observed data that return estimates of true, but unknown, model parameters, be it the point-estimates of classical frequentist statistics or the posterior distributions of the Bayesian paradigm (e.g., Wasserman, 2010). An important issue in the development of estimation techniques is hence the quality of estimators to recover true, but unknown, model parameters and model structure. While this issue re-appears in the functional neuroimaging literature in various guises every couple of years (e.g., Vul et al., 2009a; Eklund et al., 2016), often accompanied by some flurry in the field (e.g., Abbott, 2009; Nichols and Poline, 2009; Vul et al., 2009b; Eklund et al., 2016; Miller, 2016), it is perhaps true to state that the systematic study of estimator properties for functional neuroimaging data models is not the most matured research field. From an analytical perspective, this is likely due to the relative complexity of functional neuroimaging data models as compared to the fundamental scenarios that are studied in mathematical statistics (e.g., Shao, 2003). In the current study, we used simulations to study both parameter and model recovery, and while obtaining overall satisfiable results, we found that the estimation of covariance component parameters can be deficient for a subset of data realizations. As pointed out in Section 3, this finding is not an unfamiliar result in the statistical literature (e.g., Harville, 1977; Groeneveld and Kovac, 1990; Boichard et al., 1992; Groeneveld, 1994). We see two potential avenues for improving on this issue in future research. Firstly, there exist a variety of covariance component estimation algorithm variants in the literature (e.g., Gilmour et al., 1995; Witkovský, 1996; Thompson and Mäntysaari, 1999; Foulley and van Dyk, 2000; Misztal, 2008) and research could be devoted to applying insights from this literature in the neuroimaging context. Secondly, as the deficient estimation primarily concerns

the covariance component parameter that scales the AR(1) + WN model covariance basis matrix, it remains to be seen, whether the inclusion of a variety of physiological regressors in the deterministic aspect of the GLM will eventually supersede the need for covariance component parameter estimation in the analysis of first-level fMRI data altogether (e.g., Glover et al., 2000; Lund et al., 2006). Finally, we presented the application of VB, VML, ReML, and ML in the context of fMRI time-series analysis. As pointed out in Section 1, the very same statistical estimation techniques are of eminent importance for a wide range of other functional neuroimaging data models. Moreover, together with the GLM, they also form a fundamental building block of model-based behavioral data analyses as recently proposed in the context of “computational psychiatry” (e.g., Montague et al., 2012; Schwartenbeck and Friston, 2016; Stephan et al., 2016a,b,c) and recent developments in the analysis of “big data” (e.g., Allenby et al., 2014; Ghahramani, 2015).

On a more general level, the relative merits of the parameter estimation techniques discussed herein form an important field for future research. Ideally, the statistical properties of estimators resulting from variational approaches were understood for the model of interest, and known properties of their specialized cases, such as the bias-free covariance component parameter estimation under ReML with respect to ML, would be deducible from these. However, as pointed out by Blei et al. (2016), the statistical properties of variational approaches are not yet well understood. Nevertheless, there exists a few results on the statistical properties of variational approaches, typically in terms of the variational expectations upon convergence and for fairly specific model classes. Of relevance for the model class considered herein is the recent work by You et al. (2014), who could show the consistency of the variational expectation in the frequentist sense, albeit for spherical covariance matrices and a gamma distribution for the covariance component parameter. For a broader model class with posterior support in real space (including the current model class of interest), Westling (2017)

have worked toward establishing the consistency and asymptotic normality of variational expectation estimates. Finally, a number of authors have addressed consistency and asymptotic properties in selected model classes, such as Poisson-mixed effect models, stochastic block models, and Gaussian mixture models (Wang and Titterton, 2006; Hall et al., 2011; Celisse et al., 2012; Bickel et al., 2013).

In summary, understanding the qualitative statistical properties of variational Bayesian estimators and their relative merits with respect to more specialized approaches forms a burgeoning field of research. New impetus in this direction may also arise from recent attempts to understand the properties of deep learning algorithms from a probabilistic variational perspective (Gal and Ghahramani, 2017).

## 5. CONCLUSION

To conclude, we believe that the mathematization and validation of model estimation techniques employed in the neuroimaging

field is an important endeavor as the field matures. With the current work, we attempted to provide a small step in this direction. We further hope to be able to contribute to a better understanding of the statistical properties of the parameter estimation techniques for neuroimaging-relevant model classes in our future work.

## AUTHOR CONTRIBUTIONS

LS and DO conceptualized and designed the work, performed theoretical derivations, simulations and analyses, and drafted and revised the paper.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fnins.2017.00504/full#supplementary-material>

## REFERENCES

- Abbott, A. (2009). Brain imaging studies under fire. *Nature* 457:245. doi: 10.1038/457245a
- Allenby, G. M., Bradlow, E. T., George, E. I., Liechty, J., and McCulloch, R. E. (2014). Perspectives on bayesian methods and big data. *Customer Needs Solut.* 1, 169–175. doi: 10.1007/s40547-014-0017-9
- Ashburner, J. (2009). Computational anatomy with the spm software. *Magn. Reson. Imaging* 27, 1163–1174. doi: 10.1016/j.mri.2009.01.006
- Ashburner, J. (2012). Spm: a history. *Neuroimage* 62, 791–800. doi: 10.1016/j.neuroimage.2011.10.025
- Ashburner, J., and Friston, K. (2000). Voxel-based morphometry—the methods. *Neuroimage* 11(6 Pt 1), 805–821. doi: 10.1006/nimg.2000.0582
- Attias, H. (2000). A variational bayesian framework for graphical models. *Adv. Neural Inform. Process. Syst.* 12, 209–215.
- Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge: Cambridge University Press.
- Beal, M., and Ghahramani, Z. (2003). “The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures,” in *Bayesian Statistics 7*, eds J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West (Oxford: Oxford University Press), 1–10.
- Beal, M. J. (2003). *Variational Algorithms for Approximate Bayesian Inference*. London: University of London.
- Bernardo, J. M. (2003). “Bayesian Statistics,” in *Probability and Statistics*, ed R. Viertl (Oxford: Encyclopedia of Life Support Systems (EOLSS)), 1–46.
- Bernardo, J. M. (2009). *Modern Bayesian Inference: Foundations and Objective Methods*, Vol. 200. Valencia: Elsevier.
- Bickel, P., Choi, D., Chang, X., and Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Stat.* 41, 1922–1943. doi: 10.1214/13-AOS1124
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ: Springer-Verlag New York, Inc.
- Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2016). Variational inference: a review for statisticians. *arXiv preprint arXiv:1601.00670*.
- Boichard, D., Schaeffer, L., and Lee, A. (1992). Approximate restricted maximum likelihood and approximate prediction error variance of the mendelian sampling effect. *Genet. Select. Evol.* 24:1. doi: 10.1186/1297-9686-24-4-331
- Broemeling, L. D. (1984). *Bayesian Analysis of Linear Models. Statistics: A Series of Textbooks and Monographs*. New York, NY: Taylor & Francis.
- Celisse, A., Daudin, J.-J., and Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electr. J. Stat.* 6, 1847–1899. doi: 10.1214/12-EJS729
- Chappell, M. A., Groves, A. R., Whitcher, B., and Woolrich, M. W. (2009). Variational bayesian inference for a nonlinear forward model. *IEEE Trans. Signal Process.* 57, 223–236. doi: 10.1109/TSP.2008.2005752
- Chen, C., Kiebel, S., and Friston, K. (2008). Dynamic causal modelling of induced responses. *Neuroimage* 41, 1293–1312. doi: 10.1016/j.neuroimage.2008.03.026
- Cover, T. M., and Thomas, J. A. (2012). *Elements of Information Theory*. Hoboken, NJ: John Wiley & Sons.
- David, O., Kiebel, S. J., Harrison, L. M., Mattout, J., Kilner, J. M., and Friston, K. J. (2006). Dynamic causal modeling of evoked responses in EEG and MEG. *Neuroimage* 30, 1255–1272. doi: 10.1016/j.neuroimage.2005.10.045
- Draper, N. R., and Smith, H. (2014). *Applied Regression Analysis*. Hoboken, NJ: John Wiley & Sons.
- Eklund, A., Nichols, T. E., and Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. U.S.A.* 113, 7900–7905. doi: 10.1073/pnas.1602413113
- Foulley, J. (1993). A simple argument showing how to derive restricted maximum likelihood. *J. Dairy Sci.* 76, 2320–2324. doi: 10.3168/jds.S0022-0302(93)77569-4
- Foulley, J., and van Dyk, D. (2000). The px-em algorithm for fast stable fitting of henderson’s mixed model. *Genet. Sel. Evol.* 32, 143–163. doi: 10.1186/1297-9686-32-2-143
- Frank, L., Buxton, R., and Wong, E. (1998). Probabilistic analysis of functional magnetic resonance imaging data. *Magn. Reson. Med.* 39, 132–148. doi: 10.1002/mrm.1910390120
- Friston, K. (2008). Hierarchical models in the brain. *PLoS Comput. Biol.* 4:e1000211. doi: 10.1371/journal.pcbi.1000211
- Friston, K., Chu, C., Mourão-Miranda, J., Hulme, O., Rees, G., Penny, W., et al. (2008a). Bayesian decoding of brain images. *Neuroimage* 39, 181–205. doi: 10.1016/j.neuroimage.2007.08.013
- Friston, K., Glaser, D., Henson, R. N. A., Kiebel, S., Phillips, C., and Ashburner, J. (2002a). Classical and bayesian inference in neuroimaging: applications. *Neuroimage* 16, 484–512. doi: 10.1006/nimg.2002.1091
- Friston, K., Harrison, L., Daunizeau, J., Kiebel, S., Phillips, C., Trujillo-Barreto, N., et al. (2008b). Multiple sparse priors for the M/EEG inverse problem. *Neuroimage* 39, 1104–1120. doi: 10.1016/j.neuroimage.2007.09.048
- Friston, K., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *Neuroimage* 19, 1273–1302. doi: 10.1016/S1053-8119(03)00202-7
- Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., and Penny, W. (2007). Variational free energy and the laplace approximation. *Neuroimage* 34, 220–234. doi: 10.1016/j.neuroimage.2006.08.035

- Friston, K., Penny, W., Phillips, C., Kiebel, S., Hinton, G., and Ashburner, J. (2002b). Classical and bayesian inference in neuroimaging: theory. *Neuroimage* 16, 465–483. doi: 10.1006/nimg.2002.1090
- Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* 2, 189–210. doi: 10.1002/hbm.460020402
- Gal, Y., and Ghahramani, Z. (2017). “On modern deep learning and variational inference,” in *Advances in Approximate Bayesian Inference: NIPS 2016 Workshop* (Cambridge).
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian Data Analysis*, Vol. 2. Boca Raton, FL: Chapman & Hall/CRC.
- Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature* 521, 452–459. doi: 10.1038/nature14541
- Gilmour, A. R., Thompson, R., and Cullis, B. R. (1995). Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* 51, 1440–1450. doi: 10.2307/2533274
- Glover, G., Li, T., and Ress, D. (2000). Image-based method for retrospective correction of physiological motion effects in fMRI: retroicor. *Magn. Reson. Med.* 44, 162–167. doi: 10.1002/1522-2594(200007)44:1<162::AID-MRM23>3.0.CO;2-E
- Groeneveld, E. (1994). A reparameterization to improve numerical optimization in multivariate reml (co)variance component estimation. *Genet. Select. Evol.* 26, 1–9. doi: 10.1186/1297-9686-26-6-537
- Groeneveld, E., and Kovac, M. (1990). A note on multiple solutions in multivariate restricted maximum likelihood covariance component estimation. *J. Dairy Sci.* 73, 2221–2229. doi: 10.3168/jds.S0022-0302(90)78902-3
- Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics* 11, 1–21. doi: 10.1080/00401706.1969.10490657
- Hall, P., Pham, T., Wand, M. P., and Wang, S. S. (2011). Asymptotic normality and valid inference for gaussian variational approximation. *Ann. Stat.* 39, 2502–2532. doi: 10.1214/11-AOS908
- Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.* 72, 320–338. doi: 10.1080/01621459.1977.10480998
- Henson, R., and Friston, K. (2007). Convolution models for fMRI. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, eds W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols (London: Academic Press), 178–192. doi: 10.1016/B978-012372560-8/50014-0
- Hocking, R. R. (2013). *Methods and Applications of Linear Models: Regression and the Analysis of Variance*. Hoboken, NJ: John Wiley & Sons.
- Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press. doi: 10.1017/CBO9780511790423
- Kiebel, S. J., Daunizeau, J., Phillips, C., and Friston, K. J. (2008). Variational bayesian inversion of the equivalent current dipole model in EEG/MEG. *Neuroimage* 39, 728–741. doi: 10.1016/j.neuroimage.2007.09.005
- Kiebel, S. J., and Friston, K. J. (2004a). Statistical parametric mapping for event-related potentials: I. generic considerations. *Neuroimage* 22, 492–502. doi: 10.1016/j.neuroimage.2004.02.012
- Kiebel, S. J., and Friston, K. J. (2004b). Statistical parametric mapping for event-related potentials (II): a hierarchical temporal model. *Neuroimage* 22, 503–520. doi: 10.1016/j.neuroimage.2004.02.013
- LaMotte, L. R. (2007). A direct derivation of the reml likelihood function. *Stat. Papers* 48, 321–327. doi: 10.1007/s00362-006-0335-6
- Lehmann, E. L., and Casella, G. (2006). *Theory of Point Estimation*. New York, NY: Springer Science & Business Media.
- Lindley, D. V., and Smith, A. F. (1972). Bayes estimates for the linear model. *J. R. Stat. Soc. B (Methodol.)* 34, 1–41. doi: 10.2307/2985048
- Litvak, V., and Friston, K. (2008). Electromagnetic source reconstruction for group studies. *Neuroimage* 42, 1490–1498. doi: 10.1016/j.neuroimage.2008.06.022
- Lund, T. E., Madsen, K. H., Sidaros, K., Luo, W.-L., and Nichols, T. E. (2006). Non-white noise in fMRI: does modelling have an impact? *Neuroimage* 29, 54–66. doi: 10.1016/j.neuroimage.2005.07.005
- Marreiros, A., Kiebel, S., and Friston, K. (2008). Dynamic causal modelling for fMRI: a two-state model. *Neuroimage* 39, 269–278. doi: 10.1016/j.neuroimage.2007.08.019
- McLachlan, G., and Krishnan, T. (2007). *The EM Algorithm and Extensions*, Vol. 382. Hoboken, NJ: John Wiley & Sons.
- Miller, G. (2016). Neuroscience. Brain scans are prone to false positives, study says. *Science* 353, 208–209. doi: 10.1126/science.353.6296.208
- Misztal, I. (2008). Reliable computing in estimation of variance components. *J. Anim. Breed. Genet.* 125, 363–370. doi: 10.1111/j.1439-0388.2008.00774.x
- Montague, P. R., Dolan, R. J., Friston, K. J., and Dayan, P. (2012). Computational psychiatry. *Trends Cogn. Sci.* 16, 72–80. doi: 10.1016/j.tics.2011.11.018
- Monti, M. M. (2011). Statistical analysis of fmri time-series: a critical review of the glm approach. *Front. Hum. Neurosci.* 5:28. doi: 10.3389/fnhum.2011.00028
- Moran, R., Stephan, K., Seidenbecher, T., Pape, H.-C., Dolan, R., and Friston, K. (2009). Dynamic causal models of steady-state responses. *Neuroimage* 44, 796–811. doi: 10.1016/j.neuroimage.2008.09.048
- Mumford, J. A., and Nichols, T. (2006). Modeling and inference of multisubject fMRI data. *IEEE Eng. Med. Biol. Mag.* 25, 42–51. doi: 10.1109/EMEMB.2006.1607668
- Mumford, J. A., and Nichols, T. (2009). Simple group fmri modeling and inference. *Neuroimage* 47, 1469–1475. doi: 10.1016/j.neuroimage.2009.05.034
- Mumford, J. A., and Nichols, T. E. (2008). Power calculation for group fmri studies accounting for arbitrary design and temporal autocorrelation. *Neuroimage* 39, 261–268. doi: 10.1016/j.neuroimage.2007.07.061
- Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge: MIT Press.
- Neal, R. M., and Hinton, G. E. (1998). *A View of the Em Algorithm that Justifies Incremental, Sparse, and other Variants*. Dordrecht: Springer.
- Nichols, T. E., and Poline, J.-B. (2009). Commentary on vul et al.’s (2009) “puzzlingly high correlations in fmri studies of emotion, personality, and social cognition.” *Perspect. Psychol. Sci.* 4, 291–293. doi: 10.1111/j.1745-6924.2009.01126.x
- Ostwald, D., Kirilina, E., Starke, L., and Blankenburg, F. (2014). A tutorial on variational bayes for latent linear stochastic time-series models. *J. Math. Psychol.* 60, 1–19. doi: 10.1016/j.jmp.2014.04.003
- Ostwald, D., Porcaro, C., and Bagshaw, A. P. (2010). An information theoretic approach to EEG-fMRI integration of visually evoked responses. *Neuroimage* 49, 498–516. doi: 10.1016/j.neuroimage.2009.07.038
- Ostwald, D., and Starke, L. (2016). Probabilistic delay differential equation modeling of event-related potentials. *Neuroimage* 136, 227–257. doi: 10.1016/j.neuroimage.2016.04.025
- Penny, W., Kiebel, S., and Friston, K. (2003). Variational bayesian inference for fMRI time series. *Neuroimage* 19, 727–741. doi: 10.1016/S1053-8119(03)00071-5
- Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., and Nichols, T. E. (2011). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. London: Academic Press.
- Phillips, C., Rugg, M. D., and Friston, K. J. (2002). Systematic regularization of linear inverse solutions of the EEG source localization problem. *Neuroimage* 17, 287–301. doi: 10.1006/nimg.2002.1175
- Pinotsis, D., Moran, R., and Friston, K. (2012). Dynamic causal modeling with neural fields. *Neuroimage* 59, 1261–1274. doi: 10.1016/j.neuroimage.2011.08.020
- Poline, J.-B., and Brett, M. (2012). The general linear model and fMRI: does love last forever? *Neuroimage* 62, 871–880. doi: 10.1016/j.neuroimage.2012.01.133
- Purdon, P., and Weisskoff, R. (1998). Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI. *Hum. Brain Mapp.* 6, 239–249. doi: 10.1002/(SICI)1097-0193(1998)6:4<239::AID-HBM4>3.0.CO;2-4
- Rutherford, A. (2001). *Introducing ANOVA and ANCOVA: a GLM Approach*. London: SAGE.
- Schwartenbeck, P., and Friston, K. (2016). Computational phenotyping in psychiatry: a worked example. *eneuro* 3:ENEURO-0049. doi: 10.1523/ENEURO.0049-16.2016
- Searle, S. R., Casella, G., and McCulloch, C. E. (2009). *Variance Components*, Vol. 391. Hoboken, NJ: John Wiley & Sons.
- Shao, J. (2003). *Mathematical Statistics*. New York, NY: Springer Texts in Statistics. Springer.
- Stephan, K. E., Schlagenhaut, F., Huys, Q. J. M., Raman, S., Aponte, E., Brodersen, K., et al. (2016a). Computational neuroimaging strategies for single patient predictions. *Neuroimage*. 145(Pt B), 180–199. doi: 10.1016/j.neuroimage.2016.06.038

- Stephan, K. E., Bach, D. R., Fletcher, P. C., Flint, J., Frank, M. J., Friston, K. J., et al. (2016b). Charting the landscape of priority problems in psychiatry, part 1: classification and diagnosis. *Lancet Psychiatry* 3, 77–83. doi: 10.1016/S2215-0366(15)00361-2
- Stephan, K. E., Binder, E. B., Breakspear, M., Dayan, P., Johnstone, E. C., Meyer-Lindenberg, A., et al. (2016c). Charting the landscape of priority problems in psychiatry, part 2: pathogenesis and aetiology. *Lancet Psychiatry* 3, 84–90. doi: 10.1016/S2215-0366(15)00360-0
- Stephan, K. E., Kasper, L., Harrison, L. M., Daunizeau, J., den Ouden, H. E. M., Breakspear, M., et al. (2008). Nonlinear dynamic causal models for fMRI. *Neuroimage* 42, 649–662. doi: 10.1016/j.neuroimage.2008.04.262
- Thompson, R., and Mäntysaari, E. A. (1999). Prospects for statistical methods in dairy cattle breeding. *Interbull Bull.* 71, 1–8.
- Tzikas, D. G., Likas, A. C., and Galatsanos, N. P. (2008). The variational approximation for bayesian inference. *IEEE Signal Process. Mag.* 25, 131–146. doi: 10.1109/MSP.2008.929620
- Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009a). Puzzlingly high correlations in fmri studies of emotion, personality, and social cognition. *Perspect. Psychol. Sci.* 4, 274–290. doi: 10.1111/j.1745-6924.2009.01125.x
- Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009b). Reply to comments on “puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition.” *Perspect. Psychol. Sci.* 4, 319–324. doi: 10.1111/j.1745-6924.2009.01132.x
- Wang, B., and Titterton, D. M. (2006). Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model. *Bayesian Anal.* 1, 625–650. doi: 10.1214/06-BA121
- Wasserman, L. (2010). *All of Statistics: A Concise Course in Statistical Inference*. New York, NY: Springer Publishing Company, Inc.
- Westling, T. M. T. (2017). Consistency, calibration, and efficiency of variational inference. *arXiv:1510.08151v3*.
- Witkovský, V. (1996). On variance–covariance components estimation in linear models with ar (1) disturbances. *Acta Math. Univ. Comenianae* 65, 129–139.
- Woolrich, M., Ripley, B., Brady, M., and Smith, S. (2001). Temporal autocorrelation in univariate linear modeling of fMRI data. *Neuroimage* 14, 1370–1386. doi: 10.1006/nimg.2001.0931
- Woolrich, M. W., Behrens, T. E. J., Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2004). Multilevel linear modelling for fMRI group analysis using bayesian inference. *Neuroimage* 21, 1732–1747. doi: 10.1016/j.neuroimage.2003.12.023
- Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., et al. (2009). Bayesian analysis of neuroimaging data in FSL. *Neuroimage* 45(1 Suppl.), S173–S186. doi: 10.1016/j.neuroimage.2008.10.055
- You, C., Ormerod, J. T., and Müller, S. (2014). On variational bayes estimation and variational information criteria for linear regression models. *Aust. New Zealand J. Stat.* 56, 73–87. doi: 10.1111/anzs.12063
- Zarahn, E., Aguirre, G., and D’Esposito, M. (1997). Empirical analyses of bold fmri statistics. I. Spatially unsmoothed data collected under null-hypothesis conditions. *Neuroimage* 5, 179–197. doi: 10.1006/nimg.1997.0263

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Starke and Ostwald. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.