Variational Bayesian parameter estimation techniques for the general linear model

Ludger Starke[1] and Dirk Ostwald[1,2]

[1]Computational Cognitive Neuroscience Laboratory
Department of Education and Psychology, Freie Universität Berlin
[2]Center for Adaptive Rationality
Max-Planck Institute for Human Development, Berlin

Abstract

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 modelling 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 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.

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 (Ashburner, 2012; Penny et al., 2011). 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; Stephan

et al., 2008; Marreiros 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., Kiebel et al., 2008; Friston et al., 2008b; 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 modelling 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 interests, model parameters are either treated as random variables, in which case they are endowed with prior and posterior uncertainty modelled 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., Mumford and Nichols, 2008; Zarahn et al., 1997; Purdon and Weisskoff, 1998; Woolrich et al., 2001; Friston et al., 2002b).

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*.

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. We do, however, distinguish between the conditioning of a probability distribution of a random variable y on a (commonly unobserved) random variable θ, which we denote by $p(y|\theta)$, and the parameterization of a probability distribution of a random variable y by a (non-random) parameter θ, which we denote by $p_\theta(y)$. Importantly, in the former case, θ is conceived of as random variable, while in the latter case, it is not. Equivalently, if θ^* denotes a value that the random variable θ 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, $\mathrm{tr}(\cdot)$ 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 $x = a$. 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 in Supplementary Material S1.

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_λ is necessarily of the form $\sigma^2 I_n$ with $\sigma^2 > 0$, i.e. we allow for non-sphericity of the error terms. In (2), $\lambda_1, \ldots, \lambda_k$, is a set of *covariance component parameters* and $Q_1, \ldots, Q_k \in \mathbb{R}^{n\times n}$ is a set of *covariance basis matrices*. We assume throughout, that the true, but unknown, values of $\lambda_1, \ldots, \lambda_k$ are such that V_λ is positive-definite. In line with the common denotation in the neuroimaging literature, we refer to (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 (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 (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 haemodynamic response function and a constant offset. This renders the parameter entries β_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_λ employs $k = 2$ covariance component parameters λ_1 and λ_2 and corresponding covariance basis matrices

$$Q_1 := I_n \text{ and } Q_2 := (Q_2)_{ij} := \begin{cases} \exp\left(-\frac{1}{\tau}|i-j|\right), & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases}. \tag{3}$$

This specific form of the error covariance matrix encodes exponentially decaying correlations between neighbouring data points, and, with $\tau := 0.2$, corresponds

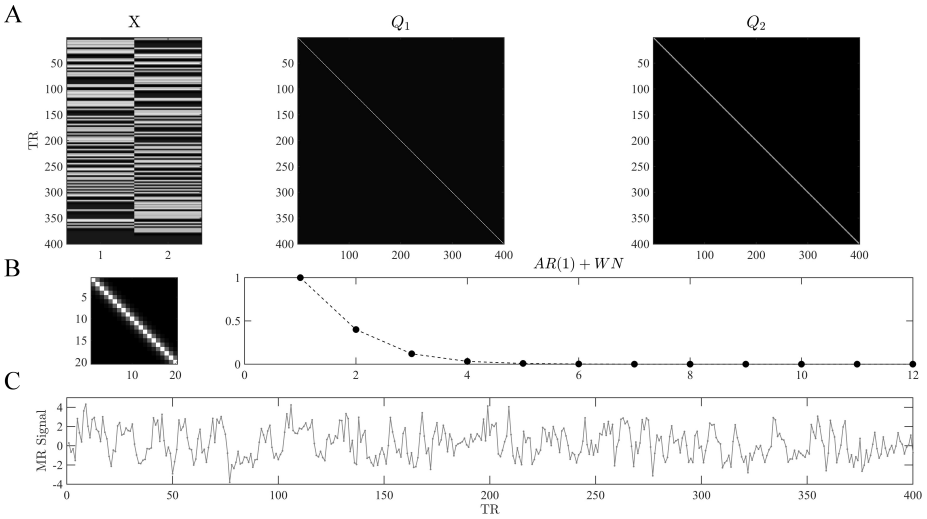

Figure 1: (A) **Example design and covariance basis matrices.** The upper panels depict the design matrix $X \in \mathbb{R}^{400\times 2}$ and the covariance basis matrices $Q_1 \in \mathbb{R}^{400\times 400}$ used in the example applications of the current section. The design matrix encodes the onset functions of two hypothetical experimental conditions which were convolved with the canonical haemodynamic response function. Events of each condition are presented approximately every 6 seconds, and $n = 400$ data points with a TR of 2 seconds are modelled. The covariance basis matrices are specified in eq.(3) and shown here for $n = 400$. (B) The left panel depicts a magnification of the first 20 entries of Q_2. The right panel depicts the the entries of the first row of Q_2 for 12 columns. For $\tau = 0.2$ the entries model exponentially decaying error correlations. (C) A data realization of the ensuing GLM model with true, but unknown, values of $\beta = (2, -1)^T$ and $\lambda = (-0.5, -2)^T$. Note that we do not model a signal offset, or equivalently, set the beta parameter for the signal offset to zero. For implementational details, please see *vbg_ 1.m*.

to the widely used approximation to the *AR(1) + white noise* model in the analysis of fMRI data (Purdon and Weisskoff, 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 seconds (standard deviation 1 second) during an fMRI recording session comprising $n = 400$ volumes and with a TR of 2 seconds. 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 (Phillips et al., 2002; Friston et al., 2002a) 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 dis-

tributions in fMRI time-series analysis (Purdon and Weisskoff, 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 eqs. (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 (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 β and σ^2:

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

Based on the probabilistic model (6), 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 β and σ^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 β and σ^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 (6) assumes proper prior distributions for β and σ^2. In this spirit, the closest conjugate Bayesian equivalent is hence the assumption of proper prior distributions. For the case of the model (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 β is parameterized conditional on the value of λ in terms of its covariance Σ_β. 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 (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 (7) eshews the need for iterative approaches and results in the fully analytical solutions of eqs. (8) to (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 neuromaging 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 β and σ^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 (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 β and σ^2 are well-known to evaluate to (e.g., Hocking, 2013)

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

and

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

Note that (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 (14) is a biased estimator for σ^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) and (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 (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 (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 (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 (18) with $\vartheta := \{\beta, \lambda\}$ is provided in Appendix B in Ostwald et al. (2014).

The free energy itself is defined by

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

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 (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 (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.,

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

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)

$$q(\beta, \lambda) := q(\beta)q(\lambda). \tag{22}$$

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

$$p(\beta, \lambda|y) \neq p(\beta|y)p(\lambda|y), \tag{23}$$

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 (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.

Application to the GLM

To demonstrate the fixed-form VB approach to the GLM of eqs. (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:

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

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

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

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

Note that we denote parameters of the prior distributions with Greek and parameters of the variational distributions with Roman letters. Together with eqs. (1) to (3), eqs. (24) to (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 S2 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_λ, and, using mild approximations in its analytical derivation, evaluates to

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

with

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

In (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 eq. (28) with respect to its parameters. Because of the relative simplicity of eq. (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 S2. This analytical approach yields a set of four update equations and, together with the iterative evaluation of the VB free energy function (28), results in a VB algorithm for the current model as documented in Algorithm 1.

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 $F^{VB^{(i)}}$

Initialization: $i := 1$, $m_\beta^{(i)} := \mu_\beta$, $S_\beta^{(i)} := \Sigma_\beta$, $m_\lambda^{(i)} := \mu_\lambda$, $S_\lambda^{(i)} := \Sigma_\lambda$, $\Delta F^{VB^{(i)}} := \infty$, $F^{VB^{(i)}} := F^{VB}\left(m_\beta^{(i)}, S_\beta^{(i)}, m_\lambda^{(i)}, S_\lambda^{(i)}\right)$
while $\Delta F^{VB^{(i)}} > \delta$ **do**
  $i := i + 1$
  evaluate $B_{m_\beta^{(i-1)}, S_\beta^{(i-1)}, m_\lambda^{(i-1)}}$
  $S_\lambda^{(i)} := \left(\frac{1}{2} B_{m_\beta^{(i-1)}, S_\beta^{(i-1)}, m_\lambda^{(i-1)}} + \Sigma_\lambda^{-1}\right)^{-1}$
  $m_\beta^{(i)} := \left(X^T V_{m_\lambda}^{-1} X + \Sigma_\beta^{-1}\right)^{-1} \left(X^T V_{m_\lambda}^{-1} X y + \Sigma_\beta^{-1} \mu_\beta\right)$
  $S_\beta^{(i)} := \left(X^T V_{m_\lambda}^{-1} X + \Sigma_\beta^{-1}\right)^{-1}$
  solve $\frac{\partial}{\partial m_{\lambda_j}} f^{VB}\left(m_\lambda^{(i)}\right) = 0$ for $m_\lambda^{(i)}$
  evaluate $F^{VB^{(i)}} = F^{VB}\left(m_\beta^{(i)}, S_\beta^{(i)}, m_\lambda^{(i)}, S_\lambda^{(i)}\right)$
  $\Delta F^{VB^{(i)}} := F^{VB^{(i)}} - F^{VB^{(i-1)}}$
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 β and λ by setting

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

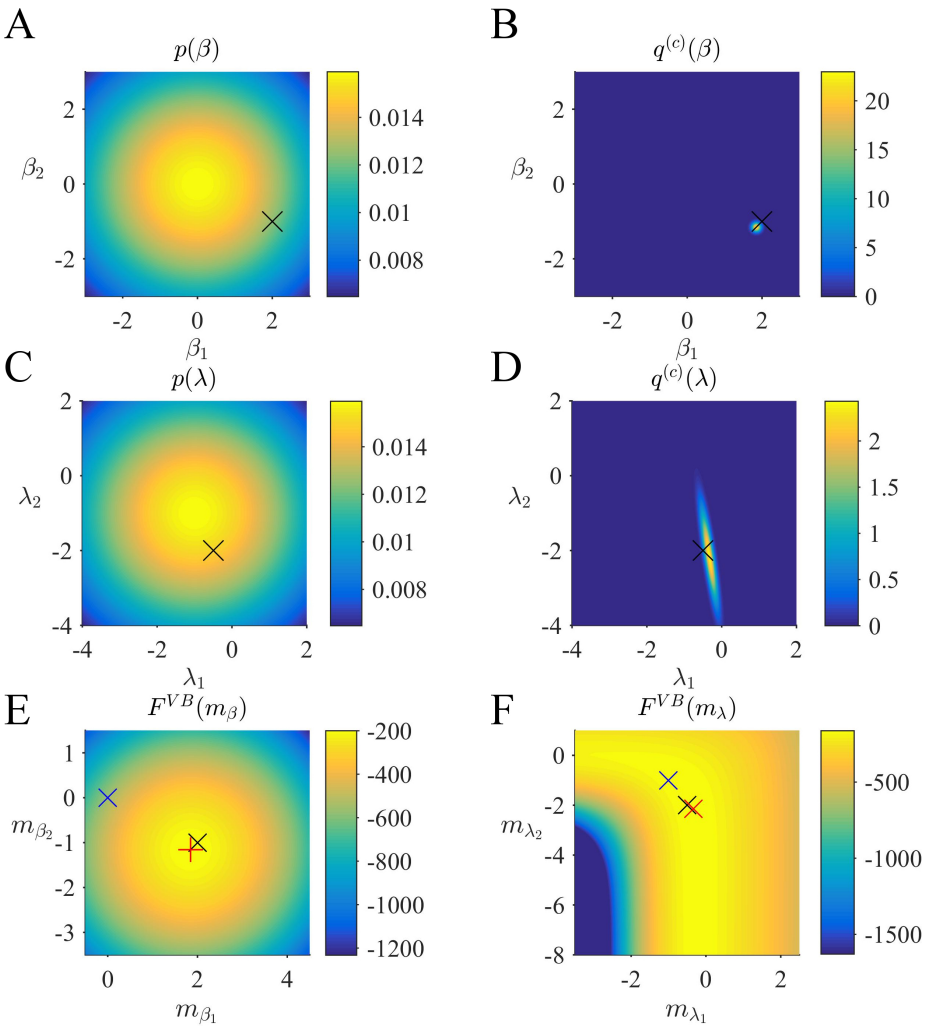

Figure 2: VB estimation. (A) Prior distribution $p(\beta)$ with expectation $\mu_\beta := (0,0)^T$ and covariance $\Sigma_\beta := 10I_2$. Here, and in all subpanels, the black $\times$ marks the true, but unknown, parameter value. (B) Variational approximation $q^{(c)}(\beta)$ to the posterior distribution 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_β. The blue $\times$ indicates the prior expectation parameter and the red $+$ marks the approximated posterior expectation parameter. (F) Variational free energy dependence on m_λ. The blue $\times$ indicates the prior expectation parameter and the red $+$ marks the approximated posterior expectation parameter. For implementational details, please see *vbg_1.m*.

Panel A of Figure 2 depicts the prior distribution over β, and the true, but

unknown, value of β as black $\times$. Panel B depicts the variational distribution over β 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 to 6 iterations. Panels C and D depict the prior distribution over λ and the variational distribution over λ upon convergence, respectively. As for β, the approximation of the posterior distribution is close to the true, but unknown, value of λ. Finally, Panels E and F depict the VB free energy surface as a function of the variational parameters m_β and m_λ, 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_λ is located on an elongated crest.

2.4 Variational Maximum Likelihood (VML)

Variational Maximum Likelihood (Beal, 2003), also referred to as (variational) expectation-maximization (Barber, 2012; McLachlan and Krishnan, 2007), 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 β as the parameter for which a posterior distribution is derived, and λ 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, λ is not treated as a random variable and thus merely parameterizes the joint distribution of β 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 S3, we show how the VML framework can be derived as a special case of VB by assuming a variational distribution that corresponds to the Dirac delta distribution $q(\lambda) := D_{\lambda^*}(\lambda)$, and how under the assumption of a constant, improper prior over λ maximization of the VB free energy is equivalent to maximization of the VML free energy. Importantly, it is the parameter value λ^* of the Dirac delta distribution that corresponds to the parameter λ in the VML framework.

Application to the GLM

In the application of the VML approach to the GLM of eqs. (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.,

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

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

Based on the specifications of eqs. (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_β and S_β and the parameter λ

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

We document the derivation of (39) in Supplementary Material S4. In contrast to the VB free energy (cf. eq. (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 λ. To maximize the VML free energy, we again derived a set of update equations as documented in Supplementary Material S4. These update equations give rise to VML algorithm for the current model, which we document in Algorithm 2.

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 β by setting

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

Algorithm 2 VML Algorithm (for details, see *vbg_est_vml.m*)

Input: data y, prior parameters μ_β, Σ_β, 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 $F^{VML^{(i)}} := F^{VML}\left(m_\beta^{(i)}, S_\beta^{(i)}, \lambda^{(i)}\right)$.
2: **while** $\Delta F^{VML^{(i)}} > \delta$ **do**
3: $\quad i := i + 1$
4: $\quad m_\beta^{(i)} := \left(X^T V_\lambda^{-1} X + \Sigma_\beta^{-1}\right)^{-1}\left(X^T V_\lambda^{-1} X y + \Sigma_\beta^{-1}\mu_\beta\right)$
5: $\quad S_\beta^{(i)} := \left(X^T V_\lambda^{-1} X + \Sigma_\beta^{-1}\right)^{-1}$
6: $\quad$ solve $\frac{\partial}{\partial \lambda_j} f^{VML}\left(\lambda^{(i)}\right) = 0$ for $\lambda^{(i)}$
7: $\quad$ evaluate $F^{VML^{(i)}} := F^{VML}\left(m_\beta^{(i)}, S_\beta^{(i)}, \lambda^{(i)}\right)$
8: $\quad \Delta F^{VML^{(i)}} := F^{VML^{(i)}} - F^{VML^{(i-1)}}$
9: **end while**

and set the initial covariance component estimate to $\lambda^{(1)} = (0, 0)^T$. Panel A of Figure 3 depicts the prior distribution over β and the true, but unknown, value of β. Panel B depicts the variational distribution over β 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 to 6 iterations. Panels C and D depict the VML free energy surface as a function of the variational parameter m_β and the parameter λ, respectively. For the chosen prior distributions, the VML free energy surfaces displays a clear global maximum as a function of m_β, while the maximum location as a function of m_λ 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; Searle et al., 2009; Phillips et al., 2002). In this context and using our denotations, the ReML estimate $\hat{\lambda}_{ReML}$ is defined as the maximizer of the ReML objective function

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

where

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

denotes the ReML objective function and

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

denotes the generalized least-squares estimator for β. Because $\hat{\beta}_{GLS}$ depends on λ in terms of V_λ, maximizing the ReML objective function necessitates iterative

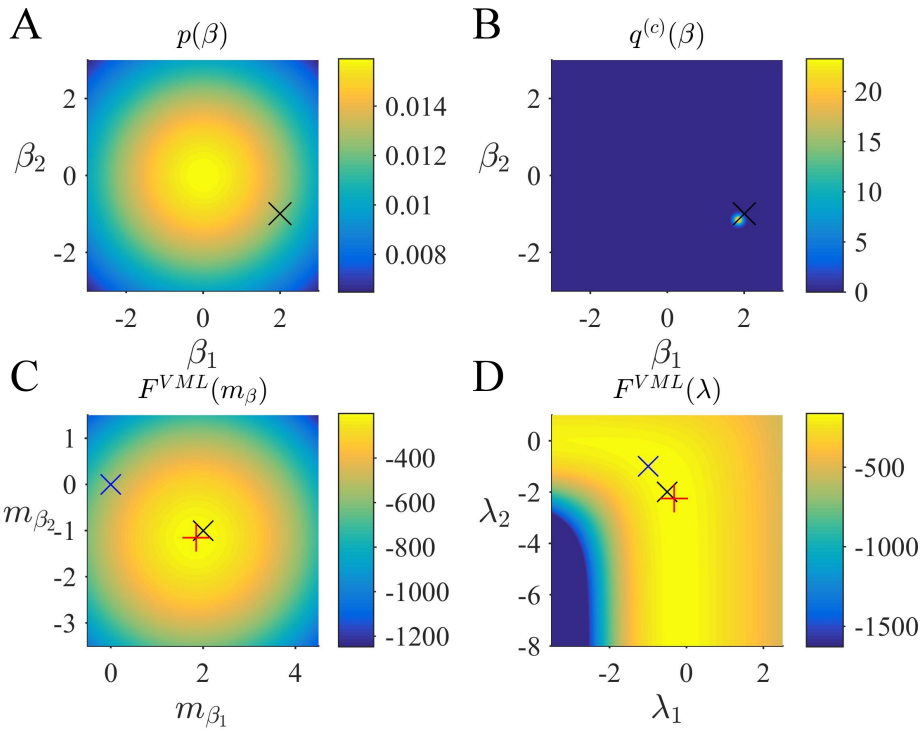

Figure 3: VML estimation.(A) Prior distribution $p(\beta)$ with expectation $\mu_\beta := (0,0)^T$ and covariance $\Sigma_\beta := 10I_2$. Here, and in all subpanels, the black $\times$ marks the true, but unknown, parameter value. (B) Variational approximation $q^{(c)}(\beta)$ to the posterior distribution upon convergence of the algorithm. (C) VML free energy dependence on m_β. The blue $\times$ indicates the prior expectation parameter and the red $+$ marks the approximated posterior expectation parameter. (D) VML free energy dependence on λ. 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*.

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 cancelling out the deterministic part of the model. In Supplementary Material S5 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 β, and an exact update of the VML free energy with respect to the variational distribution of β, i.e., setting $q(\beta) = p_\lambda(\beta|y)$. In other words, for the probabilistic model

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

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 β is endowed with an improper constant prior distribution, the posterior distribution over λ is taken to be the Dirac delta density $D_{\lambda^*}(\lambda)$, and the point estimate of λ^* maximizes the ensuing VML free energy under exact inference of the posterior distribution of β. 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 β 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 (44). In Supplementary Material S5, we show that this ReML free energy can be written as

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

Note that the equivalence of eq. (48) to the constant-augmented ReML objective function of eq. (45) derives from the fact that under the infinitely imprecise prior distribution for β 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_β from its prior expectation, because the infinitely imprecise prior distribution $p(\beta)$ provides no constraints on the estimate of β. To maximize the ReML free energy, we again derived a set of update equations which we document in Algorithm 3.

In Figure 4, we 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 β prior distribution parameters as the initial values for the variational parameters by setting

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

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

Panel A of Figure 4 depicts the converged variational distribution over β and the true, but unknown, value of β for a ReML free energy convergence criterion of $\delta = 10^{-3}$. Panels C and D depict the ReML free energy surface as a function of the variational parameter m_β and λ, respectively. Note that due

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

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.

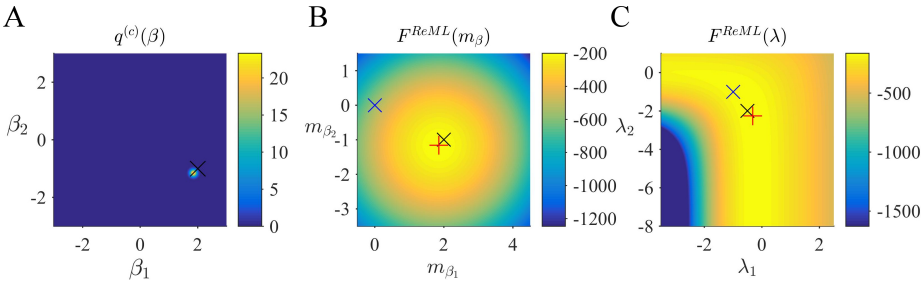

Figure 4: ReML estimation. (A) Variational distribution $q^{(c)}(\beta)$ after convergence based on the initial values $m_\beta := (0,0)^T$ and $S_\beta := 10I_2$ (convergence criterion $\delta = 10^{-3}$). Here, and in all subpanels, the black $\times$ marks the true, but unknown, parameter value. (C) ReML free energy dependence on m_β. Here, and in Panel (C) the blue $\times$ indicates the parameter value at algorithm initialization and the red $+$ marks the parameter value upon algorithm convergence. (C) ReML free energy dependence on λ. For implementational details, please see *vbg_1.m*.

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 delta densities. Specifically, as shown in Supplement S6 the ML estimate

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

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

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

Formally, we thus have

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

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:

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

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 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 $\delta = 10^{-3}$. Panel A depicts the ML free energy maximization with respect to $\beta^{(i)}$ and Panel B 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 $F^{ML^{(i)}}$

Initialization: $i := 1$, $\Delta F^{ML^{(i)}} := \infty$, $F^{ML^{(i)}} := F^{ML}(\beta^{(i)}, \lambda^{(i)})$.
while $\Delta F^{ML^{(i)}} > \delta$ **do**
  $i := i + 1$
  $\beta^{(i)} := \left(X^T V_\lambda^{-1} X)^{-1}\right) X^T V_\lambda^{-1} y$
  solve $\frac{\partial}{\partial \lambda_j} f^{ML}\left(\lambda^{(i)}\right) = 0$ for $\lambda^{(i)}$
  $F^{ML^{(i)}} := F^{ML}\left(\beta^{(i)}, \lambda^{(i)}\right)$
  $\Delta F^{ML^{(i)}} := F^{ML^{(i)}} - F^{ML^{(i-1)}}$
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. eqs. (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 convergence to the true, but unknown, parameter values after a few algorithm iterations.

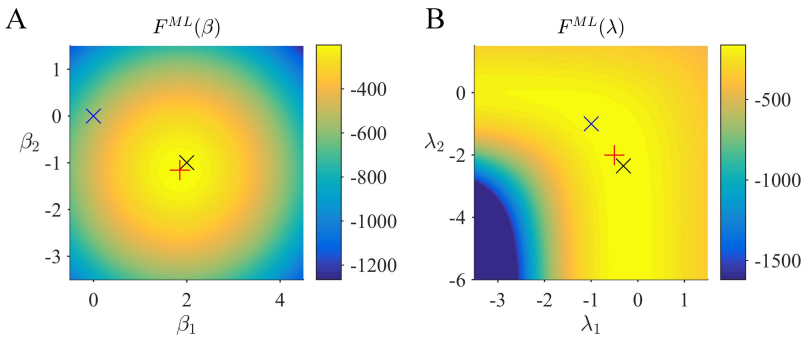

Figure 5: ML estimation. (A) ML free energy dependence on β. Here, and in Panel (B), the black $\times$ marks the true, but unknown parameter value, the blue $\times$ indicates the parameter value at algorithm initialization and the red $+$ marks the parameter value upon algorithm convergence. (B) ML free energy dependence on λ. For implementational details, please see *vbg_ 1.m*.

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 estimators quality (e.g., Lehmann and Casella, 2006). Commonly, these criteria amount to the analytical evaluation of an estimators large sample behaviour. In the current section we adopt the spirit of this approach in simulations. Firstly, we investigate the cumulative average and variance of the β and λ VB, VML, ReML, and ML parameter estimates and secondly, we investigate the ability of each technique's (marginal) likelihood approximation to distinguish between different data generating models. To this end, we adopt an objective Bayesian standpoint (Bernardo, 2003). This means that as in the previous section, we use 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.

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 β are depicted in blue, averages relating to estimates of λ 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 grey 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 λ 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. To assess whether this behaviour 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 S7. In brief, we found a similar covariance component (mis)estimation behaviour 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 behaviour 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. panel B of Figure 5) 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.

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

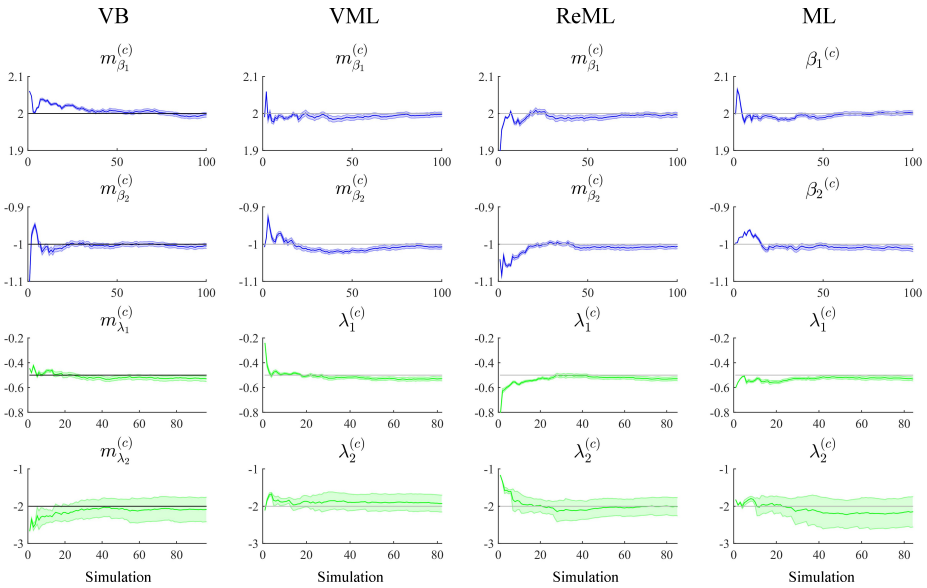

Figure 6: Parameter recovery. The panels along the figure's columns depict the cumulative averages (blue/green lines), cumulative variances (blue/green shaded areas), and true, but unknown, parameter values (grey lines) for VB, VML, ReML, and ML estimation. Parameter estimates relating to the effect sizes β are visualized in blue, parameter estimates relating to the covariance components λ 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*.

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 analysed 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.

In summary, in the reported simulations we tried to validate our implementation of VB, VML, ReML, and ML estimation techniques for a typical neuroimaging data analysis example. We observed generally satisfactory parameter recovery, with the exception of covariance component parameter recovery on a subset of data realizations, and equivalently satisfactory model recovery. 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 thus not easily generalizable. Furthermore, the assessment of the estimator qualities reported here is limited in scope. In the Discussion section, we elaborate on a number of further qualitative checks that may be of interest in future research.

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 partici-

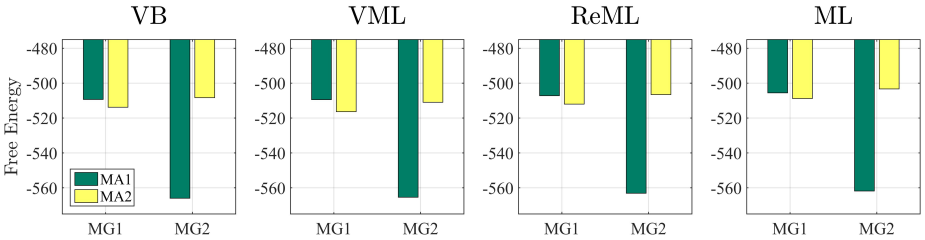

Figure 7: Model recovery. Each panel depicts the average free energies of the indicated estimation technique over 100 data realizations. Two data generating models (MG1 and MG2, panel x-axis) were used and analysed in a cross-over design with two data analysis models (MA1 and MA2, bar color). MG1 and 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*.

pant was presented with a single reversing left hemi-field checkerboard stimulus for 1 second every 16.5 to 21 seconds. 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 x 2.5 x 3 mm resolution and a TR of 1.5 seconds. 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 artefacts, slice scan time correction, re-interpolation to 2 x 2 x 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 panel C of Figure 9. 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 β_1, corresponding to the high stimulus contrast, and the first covariance component parameter λ_1, corresponding to the isotropic error component. The results are visualized in Figures 8 and 9.

Figure 8 visualizes the parameter estimates relating to the effect size parameter β_1. The subpanels of Figure 8 A 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 ventri-

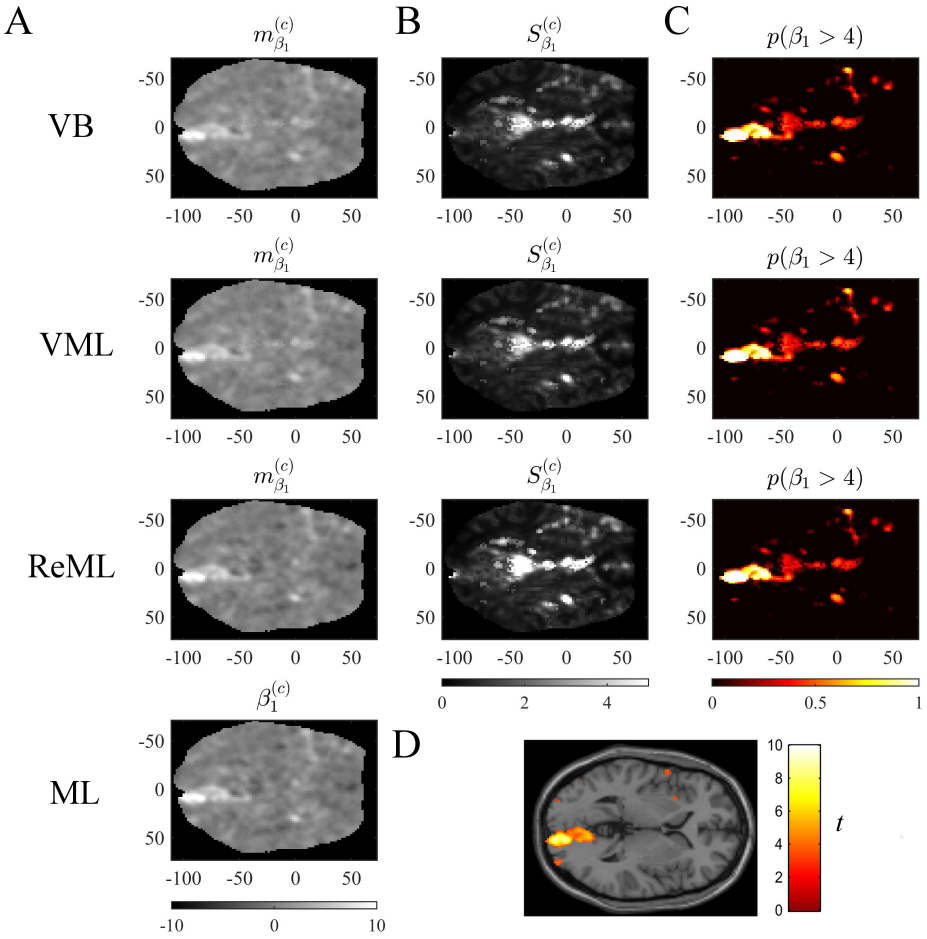

Figure 8: Effect size estimation. The figure panels depict the effect size parameter β_1 estimation results of the VB, VML, ReML, and ML algorithm application to the analysis of a single-participant single-run fMRI data set. This effect size parameter captures the effect of high contrast left visual hemifield checkerboard stimuli as encoded by the first column of the design matrix shown in panel C of Figure 9. The first column (panel A) displays the converged expectation parameter estimates, the second column (panel 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, panel D depicts the result of a standard GLM data analysis of the same data set using SPM12. For implementational details, please see *vbg_4.m*.

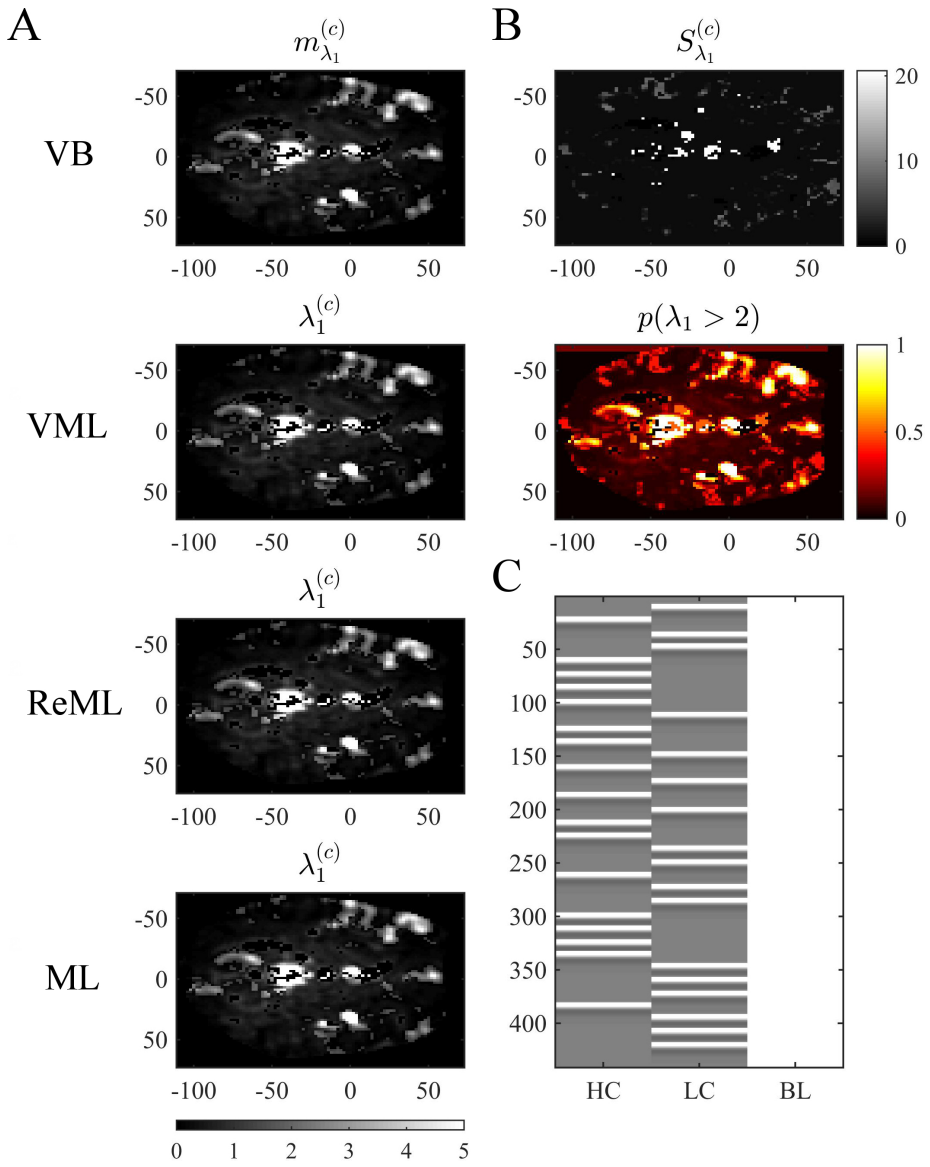

Figure 9: Covariance component parameter estimation. The figure panels depict the covariance component parameter λ_1 estimation results of the VB, VML, ReML, and ML algorithm application to the analysis of a single-participant single-run fMRI data set. This covariance component parameter captures the effect of independently distributed errors. The first column (panel A) displays the converged (expectation) parameter estimates. The second column (panel B) displays the associated variance estimate and posterior probability for $\lambda_1 > 2$, which is only quantifiable under the VB estimation technique. Panel C depicts the GLM design matrix that was used for the fMRI data analysis presented in Figures 8 and 9 (HC: high contrast stimuli regressor, LC: low contrast stimuli regressor, BL: baseline offset regressor). For implementational details, please see *vbg_4.m*.

cles. Panel B depicts the associated variational covariance parameter $S^{(c)}_{\beta_1}$, i.e., the first diagonal entry of the of the variational covariance matrix $S^{(c)}_{\beta} \in \mathbb{R}^{3\times3}$. 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 $m^{(c)}_{\beta_1}$ and $S^{(c)}_{\beta_1}$, Panel C depicts the probability that the true, but unknown, effect size parameter is larger than $\eta = 4$, i.e.

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

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 8D. 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 9 visualizes the variational expectation parameters relating to the effect size parameter λ_1. Here, the subpanels of Figure 9A 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 9B. The first subpanel visualizes the variational covariance parameter $S^{(c)}_{\lambda_1}$, i.e., the first diagonal entry of the variational covariance matrix $S^{(c)}_{\lambda} \in \mathbb{R}^{2\times2}$. The second subpanel visualizes the probability that the true, but unknown, covariance component parameter λ is larger than $\eta = 2$, i.e.

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

which, due to the relatively low uncertainty estimates S_{λ_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 relationship in Figure 10. 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 modelling 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.

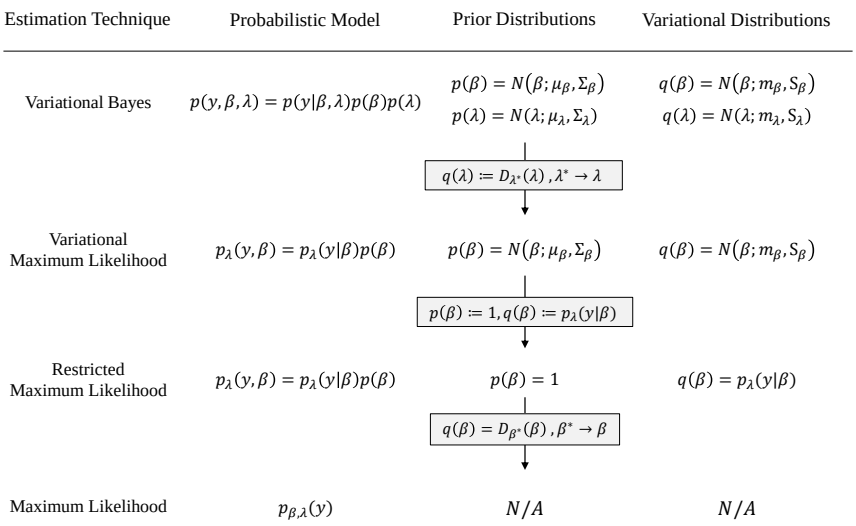

Figure 10: Summary of the relationships between VB, VML, ReML, and ML. Note that the prior and variational distributions shown are formulated with respect to the GLM application. N/A denotes non-applicable.

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 Ghamarani, 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 delta function 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; Murphy, 2012; Tzikas et al., 2008). 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.

Estimator quality

Generally speaking, 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., 2016a), often accompanied by some flurry in the field (e.g., Nichols and Poline, 2009; Vul et al., 2009b; Abbott, 2009; Eklund et al., 2016b; 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., Groeneveld and Kovac, 1990; Boichard et al., 1992; Groeneveld, 1994; Harville, 1977). 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; Witkovskỳ, 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).

Conclusion

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 behavioural data analyses as recently proposed in the context of "computational psychiatry" (e.g., Montague et al., 2012; Stephan et al., 2016a,b,c; Schwartenbeck and Friston, 2016) and recent developments in the analysis of "big data" (e.g., Allenby et al., 2014; Ghahramani, 2015). To conclude, we believe that the mathematization and validation of model estimation techniques employed in the neuroimaging field is an important endeavour as the field matures and we hope to have provided a small step in this direction with the current work.

References

Abbott, A. (2009). Brain imaging studies under fire. *Nature*, 457(7227):245.

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 and Solutions*, 1(3):169–175.

Ashburner, J. (2009). Computational anatomy with the spm software. *Magn Reson Imaging*, 27(8):1163–1174.

Ashburner, J. (2012). Spm: a history. *Neuroimage*, 62(2):791–800.

Ashburner, J. and Friston, K. (2000). Voxel-based morphometry–the methods. *Neuroimage*, 11(6 Pt 1):805–821.

Attias, H. (2000). A variational bayesian framework for graphical models. *Advances in neural information processing systems*, 12(1-2):209–215.

Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press.

Beal, M. and Ghamarani, Z. (2003). *Bayesian Statistics 7*, chapter The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures, pages 1 – 10. Oxford University Press.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. University of London London.

Bernardo, J. M. (2003). *Probability and Statistics*, chapter Bayesian Statistics, pages 1 – 46. Encyclopedia of Life Support Systems (EOLSS), Oxford UK.

Bernardo, J. M. (2009). *Modern Bayesian inference: Foundations and objective methods*, volume 200. Elsevier.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Boichard, D., Schaeffer, L., and Lee, A. (1992). Approximate restricted maximum likelihood and approximate prediction error variance of the mendelian sampling effect. *Genetics Selection Evolution*, 24(4):1.

Broemeling, L. D. (1984). *Bayesian Analysis of Linear Models*. Statistics: A Series of Textbooks and Monographs. Taylor & Francis.

Chappell, M. A., Groves, A. R., Whitcher, B., and Woolrich, M. W. (2009). Variational bayesian inference for a nonlinear forward model. *IEEE Transactions on Signal Processing*, 57(1):223–236.

Chen, C., Kiebel, S., and Friston, K. (2008). Dynamic causal modelling of induced responses. *Neuroimage*, 41(4):1293–1312.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. 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(4):1255–1272.

Draper, N. R. and Smith, H. (2014). *Applied regression analysis*. John Wiley & Sons.

Eklund, A., Nichols, T. E., and Knutsson, H. (2016a). Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates. *Proc Natl Acad Sci U S A*, 113(28):7900–7905.

Eklund, A., Nichols, T. E., and Knutsson, H. (2016b). Correction for eklund et al., cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates. *Proc Natl Acad Sci U S A*.

Foulley, J. (1993). A simple argument showing how to derive restricted maximum likelihood. *Journal of dairy science*, 76(8):2320–2324.

Foulley, J. and van Dyk, D. (2000). The px-em algorithm for fast stable fitting of henderson's mixed model. *Genet Sel Evol*, 32(2):143–163.

Frank, L., Buxton, R., and Wong, E. (1998). Probabilistic analysis of functional magnetic resonance imaging data. *Magn Reson Med*, 39(1):132–148.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Comput Biol*, 4(11):e1000211.

Friston, K., Chu, C., MourÃ£o-Miranda, J., Hulme, O., Rees, G., Penny, W., and Ashburner, J. (2008a). Bayesian decoding of brain images. *Neuroimage*, 39(1):181–205.

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(2):484–512.

Friston, K., Harrison, L., Daunizeau, J., Kiebel, S., Phillips, C., Trujillo-Barreto, N., Henson, R., Flandin, G., and Mattout, J. (2008b). Multiple sparse priors for the m/eeg inverse problem. *Neuroimage*, 39(3):1104–1120.

Friston, K., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., and Penny, W. (2007). Variational free energy and the laplace approximation. *Neuroimage*, 34(1):220–234.

Friston, K., Penny, W., Phillips, C., Kiebel, S., Hinton, G., and Ashburner, J. (2002b). Classical and bayesian inference in neuroimaging: theory. *Neuroimage*, 16(2):465–483.

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. *Human brain mapping*, 2(4):189–210.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459.

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*, pages 1440–1450.

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(1):162–167.

Groeneveld, E. (1994). A reparameterization to improve numerical optimization in multivariate reml (co)variance component estimation. *Genetics Selection Evolution*, 26(6):1–9.

Groeneveld, E. and Kovac, M. (1990). A note on multiple solutions in multivariate restricted maximum likelihood covariance component estimation. *Journal of dairy science*, 73(8):2221–2229.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338.

Henson, R. and Friston, K. (2007). Convolution models for fmri. *Statistical parametric mapping: The analysis of functional brain images*, pages 178–192.

Hocking, R. R. (2013). *Methods and applications of linear models: regression and the analysis of variance*. John Wiley & Sons.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.

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(2):728–741.

Kiebel, S. J. and Friston, K. J. (2004a). Statistical parametric mapping for event-related potentials: I. generic considerations. *Neuroimage*, 22(2):492–502.

Kiebel, S. J. and Friston, K. J. (2004b). Statistical parametric mapping for event-related potentials (ii): a hierarchical temporal model. *Neuroimage*, 22(2):503–520.

LaMotte, L. R. (2007). A direct derivation of the reml likelihood function. *Statistical Papers*, 48(2):321–327.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Lindley, D. V. and Smith, A. F. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–41.

Litvak, V. and Friston, K. (2008). Electromagnetic source reconstruction for group studies. *Neuroimage*, 42(4):1490–1498.

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(1):54–66.

Marreiros, A., Kiebel, S., and Friston, K. (2008). Dynamic causal modelling for fmri: a two-state model. *Neuroimage*, 39(1):269–278.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

Miller, G. (2016). Neuroscience. brain scans are prone to false positives, study says. *Science*, 353(6296):208–209.

Misztal, I. (2008). Reliable computing in estimation of variance components. *J Anim Breed Genet*, 125(6):363–370.

Montague, P. R., Dolan, R. J., Friston, K. J., and Dayan, P. (2012). Computational psychiatry. *Trends Cogn Sci*, 16(1):72–80.

Monti, M. M. (2011). Statistical analysis of fmri time-series: A critical review of the glm approach. *Front Hum Neurosci*, 5:28.

Moran, R., Stephan, K., Seidenbecher, T., Pape, H.-C., Dolan, R., and Friston, K. (2009). Dynamic causal models of steady-state responses. *Neuroimage*, 44(3):796–811.

Mumford, J. A. and Nichols, T. (2006). Modeling and inference of multisubject fmri data. *IEEE Engineering in Medicine and Biology Magazine*, 25(2):42–51.

Mumford, J. A. and Nichols, T. (2009). Simple group fmri modeling and inference. *Neuroimage*, 47(4):1469–1475.

Mumford, J. A. and Nichols, T. E. (2008). Power calculation for group fmri studies accounting for arbitrary design and temporal autocorrelation. *Neuroimage*, 39(1):261–268.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Neal, R. M. and Hinton, G. E. (1998). *A View of the Em Algorithm that Justifies Incremental, Sparse, and other Variants*, pages 355–368. Springer Netherlands, Dordrecht.

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(3):291–293.

Ostwald, D., Kirilina, E., Starke, L., and Blankenburg, F. (2014). A tutorial on variational bayes for latent linear stochastic time-series models. *Journal of Mathematical Psychology*, 60:1–19.

Ostwald, D., Porcaro, C., and Bagshaw, A. P. (2010). An information theoretic approach to eeg-fmri integration of visually evoked responses. *Neuroimage*, 49(1):498–516.

Ostwald, D. and Starke, L. (2016). Probabilistic delay differential equation modeling of event-related potentials. *Neuroimage*, 136:227–257.

Penny, W., Kiebel, S., and Friston, K. (2003). Variational bayesian inference for fmri time series. *Neuroimage*, 19(3):727–741.

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*. Academic press.

Phillips, C., Rugg, M. D., and Fristont, K. J. (2002). Systematic regularization of linear inverse solutions of the eeg source localization problem. *Neuroimage*, 17(1):287–301.

Pinotsis, D., Moran, R., and Friston, K. (2012). Dynamic causal modeling with neural fields. *Neuroimage*, 59(2):1261–1274.

Poline, J.-B. and Brett, M. (2012). The general linear model and fmri: does love last forever? *Neuroimage*, 62(2):871–880.

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(4):239–249.

Rutherford, A. (2001). *Introducing ANOVA and ANCOVA: a GLM approach*. Sage.

Schwartenbeck, P. and Friston, K. (2016). Computational phenotyping in psychiatry: a worked example. *eneuro*, 3(4):ENEURO–0049.

Searle, S. R., Casella, G., and McCulloch, C. E. (2009). *Variance components*, volume 391. John Wiley & Sons.

Shao, J. (2003). *Mathematical Statistics*. Springer Texts in Statistics. Springer.

Stephan, K., Schlagenhauf, F., Huys, Q. J. M., Raman, S., Aponte, E., Brodersen, K., Rigoux, L., Moran, R., Daunizeau, J., Dolan, R., Friston, K., and Heinz, A. (2016a). Computational neuroimaging strategies for single patient predictions. *Neuroimage*.

Stephan, K. E., Bach, D. R., Fletcher, P. C., Flint, J., Frank, M. J., Friston, K. J., Heinz, A., Huys, Q. J. M., Owen, M. J., Binder, E. B., Dayan, P., Johnstone, E. C., Meyer-Lindenberg, A., Montague, P. R., Schnyder, U., Wang, X.-J., and Breakspear, M. (2016b). Charting the landscape of priority problems in psychiatry, part 1: classification and diagnosis. *Lancet Psychiatry*, 3(1):77–83.

Stephan, K. E., Binder, E. B., Breakspear, M., Dayan, P., Johnstone, E. C., Meyer-Lindenberg, A., Schnyder, U., Wang, X.-J., Bach, D. R., Fletcher, P. C., Flint, J., Frank, M. J., Heinz, A., Huys, Q. J. M., Montague, P. R., Owen, M. J., and Friston, K. J. (2016c). Charting the landscape of priority problems in psychiatry, part 2: pathogenesis and aetiology. *Lancet Psychiatry*, 3(1):84–90.

Stephan, K. E., Kasper, L., Harrison, L. M., Daunizeau, J., den Ouden, H. E. M., Breakspear, M., and Friston, K. J. (2008). Nonlinear dynamic causal models for fmri. *Neuroimage*, 42(2):649–662.

Thompson, R. and Mäntysaari, E. A. (1999). Prospects for statistical methods in dairy cattle breeding. *Interbull Bulletin*, (20):71.

Tzikas, D. G., Likas, A. C., and Galatsanos, N. P. (2008). The variational approximation for bayesian inference. *IEEE Signal Processing Magazine*, 25(6):131–146.

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(3):274–290.

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(3):319–324.

Wasserman, L. (2010). *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated.

Witkovskỳ, V. (1996). On variance–covariance components estimation in linear models with ar (1) disturbances. *Acta Math. Univ. Comenianae*, 65(1):129–139.

Woolrich, M., Ripley, B., Brady, M., and Smith, S. (2001). Temporal autocorrelation in univariate linear modeling of fmri data. *Neuroimage*, 14(6):1370–1386.

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(4):1732–1747.

Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., Beckmann, C., Jenkinson, M., and Smith, S. M. (2009). Bayesian analysis of neuroimaging data in fsl. *Neuroimage*, 45(1 Suppl):S173–S186.

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(3):179–197.

Variational Bayesian parameter estimation techniques for the general linear model - Supplementary Material -

Ludger Starke[1] and Dirk Ostwald[1,2]

[1]Computational Cognitive Neuroscience Laboratory
Department of Education and Psychology, Freie Universität Berlin
[2]Center for Adaptive Rationality
Max-Planck Institute for Human Development, Berlin

Abstract

In this Supplementary Material we collect the detailed derivations of the results presented in Starke and Ostwald (2016) "Variational Bayesian parameter estimation techniques for the general linear model".

1 Preliminaries and notational conventions

1.1 Expectations

To ease the notation, we will often write the expectation of a function f of random variable x under the probability distribution $p(x)$ using the expectation operator

$$\langle f(x) \rangle_{p(x)} = \int f(x) p(x) \, dx \tag{1.1}$$

Furthermore, on numerous occasions, we require the following property of expectations of multivariate random variables $x \in \mathbb{R}^d$ under normal distributions: for $x, m, \mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d}$ p.d. and $A \in \mathbb{R}^{d \times d}$ it holds that

$$\langle (x-m)^T A (x-m) \rangle_{N(x;\mu,\Sigma)} = (\mu - m)^T A (\mu - m) + \mathrm{tr}(A\Sigma). \tag{1.2}$$

(see e.g. Petersen and Pedersen (2012), eq. (380))

1.2 Gradient and Hessian

The gradient and Hessian of a real-valued function

$$f : \mathbb{R}^n \to \mathbb{R}, x \mapsto f(x) \tag{1.3}$$

evaluated at a point $a \in \mathbb{R}^n$ will be denoted by

$$\nabla f(a) := \left(\frac{\partial}{\partial x_1} f(a), ..., \frac{\partial}{\partial x_n} f(a) \right)^T \in \mathbb{R}^n \tag{1.4}$$

and

$$H_f(a) := \begin{pmatrix} \frac{\partial^2}{\partial x_1^2} f(a) & \cdots & \frac{\partial^2}{\partial x_1 \partial x_n} f(a) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1} f(a) & \cdots & \frac{\partial^2}{\partial x_n^2} f(a) \end{pmatrix} \in \mathbb{R}^{n \times n}. \tag{1.5}$$

When it eases the notation, we also occasionally denote the partial derivative of f with respect to x_i evaluated at $a \in \mathbb{R}^n$ by $\frac{\partial}{\partial x_i} f|_{x=a}$.

1.3 The Dirac delta distribution

The Dirac delta distribution $D_{x^*}(x)$ is defined by the integral

$$\int_{x^*-\epsilon}^{x^*+\epsilon} f(x)\, D_{x^*}(x)\, \mathrm{d}x = f(x^*)\ , \tag{1.6}$$

given that $f(x)$ is a smooth function with compact support on $[x^* - \epsilon, x^* + \epsilon]$ and $\epsilon > 0$. Importantly, for the constant function $f(x) := 1$, we have

$$\int_{-\infty}^{\infty} D_{x^*}(x)\, dx = 1. \tag{1.7}$$

1.4 Matrix differentiation

The following matrix differentiation rules are used in the subsequent derivations (Petersen and Pedersen, 2012). For a matrix A depending on a scalar parameter x, we have

$$\frac{\partial |A|}{\partial x} = |A| \operatorname{tr}\left(A^{-1} \frac{\partial A}{\partial x}\right) \tag{1.8}$$

$$\frac{\partial \ln |A|}{\partial x} = \operatorname{tr}\left(A^{-1} \frac{\partial A}{\partial x}\right) \tag{1.9}$$

$$\frac{\partial A^{-1}}{\partial x} = -A^{-1} \frac{\partial A}{\partial x} A^{-1} \tag{1.10}$$

$$\frac{\partial \operatorname{tr}(A)}{\partial x} = \operatorname{tr}\left(\frac{\partial A}{\partial x}\right). \tag{1.11}$$

For a matrix A depending on a two-dimensional vector $x = (x_1, x_2)$, the second-order partial derivatives of its inverse are

$$\frac{\partial^2 A^{-1}}{\partial x_1^2} = 2A^{-1} \frac{\partial A}{\partial x_1} A^{-1} \frac{\partial A}{\partial x_1} A^{-1} - A^{-1} \frac{\partial^2 A}{\partial x_1^2} A^{-1} \tag{1.12}$$

$$\frac{\partial^2 A^{-1}}{\partial x_2^2} = 2A^{-1} \frac{\partial A}{\partial x_2} A^{-1} \frac{\partial A}{\partial x_2} A^{-1} - A^{-1} \frac{\partial^2 A}{\partial x_2^2} A^{-1} \tag{1.13}$$

and

$$\frac{\partial^2 A^{-1}}{\partial x_1 \partial x_2} = \frac{\partial^2 A^{-1}}{\partial x_2 \partial x_1} = A^{-1} \frac{\partial A}{\partial x_1} A^{-1} \frac{\partial A}{\partial x_2} A^{-1} + A^{-1} \frac{\partial A}{\partial x_2} A^{-2} \frac{\partial A}{\partial x_1} A^{-1}$$

$$- A^{-1} \frac{\partial^2 A}{\partial x_1 \partial x_2} A^{-1} \quad (1.14)$$

assuming that A has continues second derivatives, such that the symmetry of of second-order derivatives (Schwarz's theorem) holds. For the update equations of the matrix parameters S_β and S_λ, we also need to compute derivatives regarding matrices. We have

$$\frac{\partial \ln(|A|)}{\partial A} = A^{-1} \quad (1.15)$$

and for matrices A and B of matching dimensions

$$\frac{\partial \operatorname{tr}(AB)}{\partial A} = B^T. \quad (1.16)$$

2 The VB free energy and its update equations

To evaluate the VB free energy, we first rewrite it from its definition in eq. (20) in the main text as follows

$$\begin{aligned} F^{VB}(q(\beta)q(\lambda)) &= \langle \ln \left(\frac{p(y,\beta,\lambda)}{q(\beta)q(\lambda)} \right) \rangle_{q(\beta)q(\lambda)} \\ &= \langle \ln p(y|\beta,\lambda) \rangle_{q(\beta)q(\lambda)} + \langle \ln p(\beta) \rangle_{q(\beta)} + \langle \ln p(\lambda) \rangle_{q(\lambda)} \\ &\quad - \langle q(\beta) \rangle_{q(\beta)} - \langle q(\lambda) \rangle_{q(\lambda)}. \end{aligned} \quad (2.1)$$

Using (1.2), the second and third term on the right-hand side of (2.1) can be evaluated exactly, yielding

$$\langle \ln p(\beta) \rangle_{q(\beta)} = -\frac{p}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_\beta| - \frac{1}{2} (m_\beta - \mu_\beta)^T \Sigma_\beta^{-1} (m_\beta - \mu_\beta) - \frac{1}{2} \operatorname{tr}(\Sigma_\beta^{-1} S_\beta) \quad (2.2)$$

and

$$\langle \ln p(\lambda) \rangle_{q(\lambda)} = -\frac{k}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_\lambda| - \frac{1}{2} (m_\lambda - \mu_\lambda)^T \Sigma_\lambda^{-1} (m_\lambda - \mu_\lambda) - \frac{1}{2} \operatorname{tr}(\Sigma_\lambda^{-1} S_\lambda). \quad (2.3)$$

corresponding to terms 6 - 13 of eq. (28) in the main text. The fourth and the fifth term on the right-hand side of (2.1) correspond to the entropies of the variational distributions, which given their Gaussian form are given as function of their respective covariance matrices (e.g., Bishop, 2006)

$$H(q(\beta)) = -\langle \ln q(\beta) \rangle_{q(\beta)} = \frac{p}{2} \ln(2\pi e) + \frac{1}{2} \ln |S_\beta|, \quad (2.4)$$

$$H(q(\lambda)) = -\langle \ln q(\lambda) \rangle_{q(\lambda)} = \frac{k}{2} \ln(2\pi e) + \frac{1}{2} \ln |S_\lambda|. \quad (2.5)$$

Eqs. (2.4) and (2.5) correspond to terms 14 to 16 of eq. (28) in the main text.

Finally, we consider the first term of (2.1). Based on the definition of $p(y|\beta,\lambda)$, the expectation with respect to $q(\beta)$ can be evaluated exactly, yielding

$$\begin{aligned} \langle \ln p(y|\beta,\lambda) \rangle_{q(\beta)q(\lambda)} = &- \frac{n}{2} \ln 2\pi - \frac{1}{2} \langle \ln |V_\lambda| \rangle_{q(\lambda)} \\ &- \frac{1}{2} \langle (y - X m_\beta)^T V_\lambda^{-1} (y - X m_\beta) \rangle_{q(\lambda)} \\ &- \frac{1}{2} \langle \operatorname{tr}(V_\lambda^{-1} X S_\beta X^T) \rangle_{q(\lambda)} \end{aligned} \quad (2.6)$$

To make it possible to evaluate the remaining expectations, we use a second order Taylor approximation. Let

$$f : \mathbb{R}^k \to \mathbb{R}, \lambda \mapsto f(\lambda) \tag{2.7}$$

denote a real-valued function of λ. Then

$$f(\lambda) \approx f(m_\lambda) + (\lambda - m_\lambda)^T \nabla f(m_\lambda) + \frac{1}{2}(\lambda - m_\lambda)^T H_f(m_\lambda)(\lambda - m_\lambda) \tag{2.8}$$

in the vicinity of m_λ. If $q(\lambda)$ is sufficiently narrow, that is, if most of its mass is concentrated close to m_λ, we can thus approximate

$$\begin{aligned}\langle f(\lambda)\rangle_{q(\lambda)} &\approx f(m_\lambda) + \langle(\lambda - m_\lambda)^T \nabla f(m_\lambda)\rangle_{q(\lambda)} + \frac{1}{2}\langle(\lambda - m_\lambda)^T H_f(m_\lambda)(\lambda - m_\lambda)\rangle_{q(\lambda)} \\ &= f(m_\lambda) + \frac{1}{2}\operatorname{tr}\left(H_f(m_\lambda) S_\lambda\right).\end{aligned} \tag{2.9}$$

This approximation needs to be applied to all expectations in equation (2.6). Thus, using the linearity of the trace to subsume all Hessian matrices into

$$\begin{aligned}B_{m_\beta, S_\beta, m_\lambda} = {} & H_{\ln|V_\lambda|}(m_\lambda) + H_{(y - X m_\beta)^T V_\lambda^{-1}(y - X m_\beta)}(m_\lambda) \\ & + H_{\operatorname{tr}(V_\lambda^{-1} X S_\beta X^T)}(m_\lambda),\end{aligned} \tag{2.10}$$

thereby pooling the second-order terms, we arrive at terms 1 - 5 of equation (28) in the main text, and the derivation is complete.

Evaluation of $B_{m_\beta, S_\beta, m_\lambda}$

To estimate the VB free energy in practice, the Hessian matrices on the right-hand side of (2.10) have to be evaluated. For the linear form of the error covariance matrix

$$V_\lambda := \exp(\lambda_1) I_n + \exp(\lambda_2) Q_2 \tag{2.11}$$

the three Hessian matrices of (2.10) can be evaluated analytically:

$H_{\ln|V_\lambda|}$

Using (1.9), the first order partial derivatives are given by

$$\frac{\partial \ln|V_\lambda|}{\partial \lambda_1} = \exp(\lambda_1) \operatorname{tr}\left(V_\lambda^{-1}\right) \tag{2.12}$$

and

$$\frac{\partial \ln|V_\lambda|}{\partial \lambda_2} = \exp(\lambda_2) \operatorname{tr}\left(V_\lambda^{-1} Q_2\right). \tag{2.13}$$

Exploiting the linearity of the trace operator (1.11) and using (1.10) for the derivative of the inverse yields the second order partial derivatives:

$$\begin{aligned}\frac{\partial^2 \ln|V_\lambda|}{\partial \lambda_1^2} &= \exp(\lambda_1) \operatorname{tr}\left(V_\lambda^{-1}\right) - \exp(2\lambda_1) \operatorname{tr}\left(V_\lambda^{-2}\right) \\ &= \exp(\lambda_1) \operatorname{tr}\left(V_\lambda^{-1}\right) - \big(\exp(\lambda_1) \operatorname{tr}\left(V_\lambda^{-2} V_\lambda\right) \\ &\qquad - \exp(\lambda_1 + \lambda_2) \operatorname{tr}\left(V_\lambda^{-2} Q_2\right)\big) \\ &= \exp(\lambda_1 + \lambda_2) \operatorname{tr}\left(V_\lambda^{-2} Q_2\right),\end{aligned} \tag{2.14}$$

$$\begin{aligned}\frac{\partial^2 \ln|V_\lambda|}{\partial\lambda_2^2} &= \exp(\lambda_2)\operatorname{tr}\left(V_\lambda^{-1}Q_2\right) - \exp(2\lambda_2)\operatorname{tr}\left(V_\lambda^{-1}Q_2V_\lambda^{-1}Q_2\right)\\ &= \exp(\lambda_2)\operatorname{tr}\left(V_\lambda^{-1}Q_2\right) - \big(\exp(\lambda_2)\operatorname{tr}\left(V_\lambda^{-1}V_\lambda V_\lambda^{-1}Q_2\right)\\ &\qquad - \exp(\lambda_1+\lambda_2)\operatorname{tr}\left(V_\lambda^{-2}Q_2\right)\big)\\ &= \exp(\lambda_1+\lambda_2)\operatorname{tr}\left(V_\lambda^{-2}Q_2\right),\end{aligned} \tag{2.15}$$

and

$$\frac{\partial^2 \ln|V_\lambda|}{\partial\lambda_1\partial\lambda_2} = \frac{\partial^2 \ln|V_\lambda|}{\partial\lambda_2\partial\lambda_1} = -\exp(\lambda_1+\lambda_2)\operatorname{tr}\left(V_\lambda^{-2}Q_2\right), \tag{2.16}$$

where in the last equation we used that the trace is invariant under cyclic permutations, e.g. $\operatorname{tr}(ABC) = \operatorname{tr}(CAB) = \operatorname{tr}(BCA)$.

$H_{(y-Xm_\beta)^TV_\lambda^{-1}(y-Xm_\beta)}$

The Hessian matrix of $(y-Xm_\beta)^TV_\lambda^{-1}(y-Xm_\beta)$ only depends on the second order partial derivatives of the inverse of V_λ

$$\frac{\partial^2}{\partial\lambda_i\partial\lambda_j}\left((y-Xm_\beta)^TV_\lambda^{-1}(y-Xm_\beta)\right) = (y-Xm_\beta)^T\frac{\partial^2 V_\lambda^{-1}}{\partial\lambda_i\partial\lambda_j}(y-Xm_\beta) \tag{2.17}$$

for $i,j\in\{1,2\}$. Applying (1.12) to (1.14) yields

$$\frac{\partial^2 V_\lambda^{-1}}{\partial\lambda_1^2} = \exp(\lambda_1)V_\lambda^{-2} - 2\exp(2\lambda_1)V_\lambda^{-3}, \tag{2.18}$$

$$\frac{\partial^2 V_\lambda^{-1}}{\partial\lambda_2^2} = \exp(\lambda_2)V_\lambda^{-1}Q_2V_\lambda^{-1} - 2\exp(2\lambda_2)V_\lambda^{-1}Q_2V_\lambda^{-1}Q_2V_\lambda^{-1}, \tag{2.19}$$

and

$$\frac{\partial^2 V_\lambda^{-1}}{\partial x_1\partial x_2} = \frac{\partial^2 A^{-1}}{\partial x_2\partial x_1} = -\exp(\lambda_1+\lambda_2)\left(V_\lambda^{-2}Q_2V_\lambda^{-1} + V_\lambda^{-1}Q_2V_\lambda^{-2}\right). \tag{2.20}$$

$H_{\operatorname{tr}\left(V_\lambda^{-1}XS_\beta X^T\right)}$

Due to the linearity of the trace operator, we have

$$\frac{\partial^2\operatorname{tr}\left(V_\lambda^{-1}XS_\beta X^T\right)}{\partial\lambda_i\partial\lambda_j} = \operatorname{tr}\left(\frac{\partial^2 V_\lambda^{-1}}{\partial\lambda_i\partial\lambda_j}XS_\beta X^T\right) \tag{2.21}$$

for $i,j\in\{1,2\}$. Thus we only have to use (2.18) to (2.20).

Notably, the evaluation of these Hessian matrices will necessitate the inversion of V_λ on every iteration of the optimization algorithm. This inversion can be performed efficiently using the diagonalized form of Q_2. As Q_2 is a real, symmetric matrix by design, there exists a diagonalized form given by $Q_2^D = P^TQ_2P$, where P is a unitary transformation matrix ($P^T = P^{-1}$). The

entries $l_i, i \in \{1, \ldots, n\}$ of Q_2^D are the eigenvalues of Q_2. We thus have

$$\begin{aligned} V_\lambda^{-1} &= \left(\exp(\lambda_1) I_n + \exp(\lambda_2) Q_2\right)^{-1} \\ &= \left(\exp(\lambda_1) P I_n P^T + \exp(\lambda_2) P Q_2^D P^T\right)^{-1} \\ &= \left(P\left(\exp(\lambda_1) I_n + \exp(\lambda_2) Q_2^D\right) P^T\right)^{-1} \\ &= P\left(\exp(\lambda_1) I_n + \exp(\lambda_2) Q_2^D\right)^{-1} P^T. \end{aligned} \tag{2.22}$$

As $\exp(\lambda_1) I_n + \exp(\lambda_2) Q_2^D$ is a diagonal matrix, its inverse is easily evaluated, and the diagonalizing matrix P only needs to be computed once for any given Q_2.

The VB free energy update equations

In this section, we consider the iterative maximization of the VB free energy function with respect to its vector and matrix parameters $m_\beta, S_\beta, m_\lambda$ and S_λ. In each case, we identify the relevant subpart of the VB free energy function depending on the respective parameter, evaluate its gradient with respect to the parameter in question, set the gradient to zero, and, if possible, solve the ensuing equation for a parameter update equation. To emphasize the iterative character of this endeavour, we use the superscript (i) to denote the values of parameters at a given algorithm iteration.

We consider the update with respect to S_λ first. The relevant subpart of $F^{VB}\left(m_\beta^{(i)}, S_\beta^{(i)}, m_\lambda^{(i)}, S_\lambda^{(i)}\right)$ depending on S_λ is given by

$$f^{VB}(S_\lambda) = -\frac{1}{4}\operatorname{tr}\left(B_{m_\beta^{(i)}, S_\beta^{(i)}, m_\lambda^{(i)}} S_\lambda\right) - \frac{1}{2}\operatorname{tr}(\Sigma_\lambda^{-1} S_\lambda) + \frac{1}{2}\ln|S_\lambda|. \tag{2.23}$$

Using the identities (1.15) and (1.16) considering that $B_{m_\beta^{(i)}, S_\beta^{(i)}, m_\lambda^{(i)}}$ and Σ_λ^{-1} are symmetric, evaluation of the gradient of f^{VB} results in

$$\nabla f^{VB}(S_\lambda) = -\frac{1}{4} B_{m_\beta^{(i)}, S_\beta^{(i)}, m_\lambda^{(i)}} - \frac{1}{2}\Sigma_\lambda^{-1} + \frac{1}{2} S_\lambda^{-1}. \tag{2.24}$$

Setting the gradient to zero and solving for the parameter update $S_\lambda^{(i+1)}$ then yields

$$S_\lambda^{(i+1)} := \left(\frac{1}{2} B_{m_\beta^{(i)}, S_\beta^{(i)}, m_\lambda^{(i)}} + \Sigma_\lambda^{-1}\right)^{-1}. \tag{2.25}$$

Note that with the linearity properties of the trace operator, this update equation implies as a result, that the sum of the two trace terms involving S_λ in the VB free energy (equation (28) of the main text) evaluates to $-\frac{k}{2}$ and the term $B_{m_\beta^{(i)}, S_\beta^{(i)}, m_\lambda^{(i)}}$ does not need to be considered when deriving the update equations for m_β, S_β, and m_λ.

Next, the relevant subpart of $F^{VB}\left(m_\beta^{(i)}, S_\beta^{(i)}, m_\lambda^{(i)}, S_\lambda^{(i+1)}\right)$ depending on m_β is given by

$$f^{VB}(m_\beta) = -\frac{1}{2}(y - X m_\beta)^T V_{m_\lambda}^{-1}(y - X m_\beta) - \frac{1}{2}(m_\beta - \mu_\beta)^T S_\beta^{-1}(m_\beta - \mu_\beta), \tag{2.26}$$

where we omitted iteration superscripts for visual clarity. With (1.2), the gradient of $f^{VB}(m_\beta)$ is given by

$$\begin{aligned}\nabla f^{VB}(m_\beta) &= (y - Xm_\beta)^T V_{m_\lambda}^{-1} X - (m_\beta - \mu_\beta)^T \Sigma_\beta^{-1} \\ &= y^T V_{m_\lambda}^{-1} X - m_\beta^T X^T V_{m_\lambda}^{-1} X - m_\beta^T \Sigma_\beta^{-1} + \mu_\beta^T \Sigma_\beta^{-1}\end{aligned} \tag{2.27}$$

Setting the gradient to zero then yields the update equation

$$m_\beta^{(i+1)} := \left(X^T V_{m_\lambda}^{-1} X + \Sigma_\beta^{-1}\right)^{-1} \left(X^T V_{m_\lambda}^{-1} y + \Sigma_\beta^{-1} \mu_\beta\right) \tag{2.28}$$

Analogously, the relevant subpart of $F^{VB}\left(m_\beta^{(i+1)}, S_\beta^{(i)}, m_\lambda^{(i)}, S_\lambda^{(i+1)}\right)$ depending on S_β is given by

$$f^{VB}(S_\beta) = -\frac{1}{2} \operatorname{tr}\left(X^T V_{m_\lambda}^{-1} X S_\beta\right) - \frac{1}{2} \operatorname{tr}(\Sigma_\beta^{-1} S_\beta) + \frac{1}{2} \ln |S_\beta| \tag{2.29}$$

with gradient

$$\nabla f^{VB}(S_\beta) = -\frac{1}{2} X^T V_{m_\lambda}^{-1} X - \frac{1}{2} \Sigma_\beta^{-1} + \frac{1}{2} S_\beta^{-1} \tag{2.30}$$

and the resulting update equation

$$S_\beta^{(i+1)} := \left(X^T V_{m_\lambda}^{-1} X + \Sigma_\beta^{-1}\right)^{-1}. \tag{2.31}$$

Note that the update equations (2.28) and (2.31) conform to the well-known closed-form expressions for Bayesian inference in the conjugate Gaussian model (cf. eq. (9) of the main text), with the difference of the parametric dependence of the error covariance matrix on $m_\lambda^{(i)}$.

Finally, the relevant subpart of $F^{VB}\left(m_\beta^{(i+1)}, S_\beta^{(i+1)}, m_\lambda^{(i)}, S_\lambda^{(i+1)}\right)$ depending on m_λ is given by, again omitting iteration superscripts for visual clarity,

$$\begin{aligned}f^{VB}(m_\lambda) = &-\frac{1}{2} \ln |V_{m_\lambda}| - \frac{1}{2} (y - Xm_\beta)^T V_{m_\lambda}^{-1} (y - Xm_\beta) \\ &- \frac{1}{2} \operatorname{tr}(X^T V_{m_\lambda}^{-1} X S_\beta) - \frac{1}{2} (m_\lambda - \mu_\lambda)^T \Sigma_\lambda^{-1} (m_\lambda - \mu_\lambda).\end{aligned} \tag{2.32}$$

Evaluation of entries $\frac{\partial}{\partial m_{\lambda_j}} f^{VB}(m_\lambda)$ of the gradient $\nabla f^{VB}(m_\lambda)$ yields

$$\begin{aligned}\frac{\partial}{\partial m_{\lambda_j}} f^{VB}(m_\lambda) = &-\frac{1}{2} \operatorname{tr}\left(V_{m_\lambda}^{-1} \left(\frac{\partial V_{m_\lambda}}{\partial m_{\lambda_j}}\right)\right) \\ &- \frac{1}{2} (y - Xm_\beta)^T \left(\frac{\partial V_{m_\lambda}^{-1}}{\partial m_{\lambda_j}}\right) (y - Xm_\beta) \\ &- \frac{1}{2} \operatorname{tr}\left(\left(\frac{\partial V_{m_\lambda}^{-1}}{\partial m_{\lambda_j}}\right) X S_\beta X^T\right) - \left((m_\lambda - \mu_\lambda)^T \Sigma_\lambda^{-1}\right)_j.\end{aligned} \tag{2.33}$$

The evaluation of these entries for the two-component linear error covariance (2.11) then yields

$$\begin{aligned}\frac{\partial}{\partial m_{\lambda_1}} f^{VB}(m_\lambda) = &-\frac{1}{2} \exp(m_{\lambda_1}) \left(\operatorname{tr}(V_{m_\lambda}^{-1}) - (y - Xm_\beta)^T V_{m_\lambda}^{-2} (y - Xm_\beta)\right. \\ &\left. - \operatorname{tr}\left(V_{m_\lambda}^{-1} X S_\beta X^T V_{m_\lambda}^{-1}\right)\right) - \frac{1}{2} \left((m_\lambda - \mu_\lambda) \Sigma_\lambda^{-1}\right)_1,\end{aligned} \tag{2.34}$$

and

$$\frac{\partial}{\partial m_{\lambda_2}} f^{VB}(m_\lambda) = -\frac{1}{2}\exp(m_{\lambda_2})\left(\mathrm{tr}\left(V_{m_\lambda}^{-1}Q_2\right)\right. \\ - (y - Xm_\beta)^T V_{m_\lambda}^{-1} Q_2 V_{m_\lambda}^{-1}(y - Xm_\beta) \\ \left. - \mathrm{tr}\left(Q_2 V_{m_\lambda}^{-1} X S_\beta X^T V_{m_\lambda}^{-1}\right)\right) - \frac{1}{2}\left((m_\lambda - \mu_\lambda)\Sigma_\lambda^{-1}\right)_2 . \tag{2.35}$$

Lastly, to determine the value $m_\lambda^{(i+1)}$ for which

$$\frac{\partial}{\partial m_{\lambda_j}} f^{VB}\left(m_\lambda^{(i+1)}\right) = 0 \tag{2.36}$$

for $j = 1, 2$, we employ the routine *fsolve.m* provided by Matlab (MATLAB and Optimization Toolbox Release 2014b, The MathWorks, Inc., Natick, Massachusetts, United States). This function implements a trust-region dogleg algorithm for the minimization of nonlinear real-valued functions of multiple variables (Coleman and Li, 1996; Nocedal and Wright, 2006).

3 VML as special case of VB

Consider the VB decomposition of the log marginal likelihood

$$\ln p(y) = F^{VB}(q(\beta, \lambda)) + KL(q(\beta, \lambda)||p(\beta, \lambda|y)) \tag{3.1}$$

and the factorized variational distribution

$$q(\beta, \lambda) = q(\beta)q(\lambda) \text{ with } q(\lambda) := D_{\lambda^*}(\lambda). \tag{3.2}$$

In the current section we show that the substitution of (3.2) in (3.1), i.e.,

$$\ln p(y) = F^{VB}(q(\beta)D_{\lambda^*}(\lambda)) + KL(q(\beta)D_{\lambda^*}(\lambda)||p(\beta, \lambda|y)) \tag{3.3}$$

is equivalent to the VML decomposition of the log marginal likelihood as defined in equations (35) and (36) of the main text. Furthermore, we show that, if additionally a constant improper prior over λ is assumed, maximization of the VB free energy yields VML estimates.

Some notational care is necessary with respect to the exchange of the *random variable* λ in the VB context for a *parameter value* λ in the VML context. To this end, we start out by denoting the random variable by λ and a value that it can assume by λ^*. Towards the end of the derivation, we are only concerned with the case that the random variable λ takes on the value λ^*. We then identify the symbol λ with λ^*, which results in the notation of the VML framework in the main text. Because the alternative would have been to denote λ by different symbols across the estimation frameworks, we reasoned that this approach yields the most parsimonious notation.

To achieve our aim, we first reformulate the VB free energy term on the right-hand side of eq. (3.3) as follows:

$$F^{VB}(q(\beta)D_{\lambda^*}(\lambda)) = \iint q(\beta)D_{\lambda^*}(\lambda)\ln\left(\frac{p(y,\beta,\lambda)}{q(\beta)D_{\lambda^*}(\lambda)}\right)d\beta d\lambda \\ = \iint q(\beta)D_{\lambda^*}(\lambda)\ln p(y,\beta,\lambda)d\beta d\lambda \\ - \iint q(\beta)D_{\lambda^*}(\lambda)\ln\left(q(\beta)D_{\lambda^*}(\lambda)\right)d\beta d\lambda. \tag{3.4}$$

Rearrangement of the double integral terms then yields

$$\begin{aligned} F^{VB}(q(\beta)D_{\lambda^*}(\lambda)) &= \int q(\beta)\left(\int D_{\lambda^*}(\lambda)\ln p(y,\beta,\lambda)\,d\lambda\right)d\beta \\ &\quad - \int q(\beta)\left(\int D_{\lambda^*}(\lambda)\ln\left(q(\beta)D_{\lambda^*}(\lambda)\right)\,d\lambda\right)d\beta \\ &= \int q(\beta)\left(\int D_{\lambda^*}(\lambda)\ln p(y,\beta,\lambda)\,d\lambda\right)d\beta \\ &\quad - \int q(\beta)\left(\int D_{\lambda^*}(\lambda)\ln q(\beta)\,d\lambda\right)d\beta \\ &\quad - \int q(\beta)\left(\int D_{\lambda^*}(\lambda)\ln D_{\lambda^*}(\lambda)\,d\lambda\right)d\beta \\ &= \int q(\beta)\left(\int D_{\lambda^*}(\lambda)\ln p(y,\beta,\lambda)\,d\lambda\right)d\beta \\ &\quad - \int q(\beta)\ln q(\beta)\,d\beta\int D_{\lambda^*}(\lambda)\,d\lambda \\ &\quad - \int D_{\lambda^*}(\lambda)\ln D_{\lambda^*}(\lambda)\,d\lambda. \end{aligned} \tag{3.5}$$

With the properties of the Dirac delta function, we then have

$$\begin{aligned} F^{VB}(q(\beta)D_{\lambda^*}(\lambda)) &= \int q(\beta)\ln p(y,\beta,\lambda=\lambda^*)d\beta \\ &\quad - \int q(\beta)\ln q(\beta)\,d\beta \\ &\quad - \int D_{\lambda^*}(\lambda)\ln D_{\lambda^*}(\lambda)\,d\lambda \\ &= \int q(\beta)\ln\left(\frac{p(y,\beta,\lambda=\lambda^*)}{q(\beta)}\right)d\beta \\ &\quad - \int D_{\lambda^*}(\lambda)\ln D_{\lambda^*}(\lambda)\,d\lambda \\ &= \int q(\beta)\ln\left(\frac{p(y,\beta|\lambda=\lambda^*)p(\lambda=\lambda^*)}{q(\beta)}\right)d\beta \\ &\quad - \int D_{\lambda^*}(\lambda)\ln D_{\lambda^*}(\lambda)\,d\lambda \\ &= \int q(\beta)\ln\left(\frac{p_{\lambda^*}(y,\beta)}{q(\beta)}\right)d\beta + \ln p(\lambda=\lambda^*) \\ &\quad - \int D_{\lambda^*}(\lambda)\ln D_{\lambda^*}(\lambda)\,d\lambda. \end{aligned} \tag{3.6}$$

With the definition of the VML free energy, we thus obtain

$$\begin{aligned} F^{VB}(q(\beta)D_{\lambda^*}(\lambda)) &= F^{VML}(q(\beta),\lambda^*) + \ln p(\lambda=\lambda^*) \\ &\quad - \int D_{\lambda^*}(\lambda)\ln D_{\lambda^*}(\lambda)\,d\lambda. \end{aligned} \tag{3.7}$$

We next reformulate the KL divergence term on the right-hand side of

eq. (3.3) as follows:

$$\begin{aligned} KL(q(\beta)D_{\lambda^*}(\lambda)||p(\beta,\lambda|y)) &= \iint q(\beta)D_{\lambda^*}(\lambda)\ln\left(\frac{q(\beta)D_{\lambda^*}(\lambda)}{p(\beta,\lambda|y)}\right)\,d\beta\,d\lambda \\ &= \iint q(\beta)D_{\lambda^*}(\lambda)\ln\left(q(\beta)D_{\lambda^*}(\lambda)\right)\,d\beta\,d\lambda \\ &\quad - \iint q(\beta)D_{\lambda^*}(\lambda)\ln p(\beta,\lambda|y)\,d\beta\,d\lambda. \end{aligned} \tag{3.8}$$

Rearrangement of the double integral terms then yields

$$\begin{aligned} KL(q(\beta)D_{\lambda^*}(\lambda)||p(\beta,\lambda|y)) &= \int q(\beta)\left(\int D_{\lambda^*}(\lambda)\ln\left(q(\beta)D_{\lambda^*}(\lambda)\right)\,d\lambda\right)\,d\beta \\ &\quad - \int q(\beta)\left(\int D_{\lambda^*}(\lambda)\ln p(\beta,\lambda|y)\,d\lambda\right)\,d\beta \\ &= \int q(\beta)\left(\int D_{\lambda^*}(\lambda)\ln q(\beta)\,d\lambda\right)\,d\beta \\ &\quad + \int q(\beta)\left(\int D_{\lambda^*}(\lambda)\ln D_{\lambda^*}(\lambda)\,d\lambda\right)\,d\beta \\ &\quad - \int q(\beta)\left(\int D_{\lambda^*}(\lambda)\ln p(\beta,\lambda|y)\,d\lambda\right)\,d\beta. \end{aligned} \tag{3.9}$$

With the properties of the Dirac delta function and re-arranging, we then have

$$\begin{aligned} KL(q(\beta)D_{\lambda^*}(\lambda)||p(\beta,\lambda|y)) &= \int q(\beta)\ln q(\beta)\,d\beta \\ &\quad - \int q(\beta)\ln p(\beta,\lambda=\lambda^*|y)\,d\beta \\ &\quad + \int D_{\lambda^*}(\lambda)\ln D_{\lambda^*}(\lambda)\,d\lambda \\ &= \int q(\beta)\ln\left(\frac{q(\beta)}{p(\beta,\lambda=\lambda^*|y)}\right)\,d\beta \\ &\quad + \int D_{\lambda^*}(\lambda)\ln D_{\lambda^*}(\lambda)\,d\lambda \\ &= \int q(\beta)\ln\left(\frac{q(\beta)}{p(\beta|\lambda=\lambda^*,y)}\right)\,d\beta - \ln p(\lambda=\lambda^*|y) \\ &\quad + \int D_{\lambda^*}(\lambda)\ln D_{\lambda^*}(\lambda)\,d\lambda \\ &= \int q(\beta)\ln\left(\frac{q(\beta)}{p_{\lambda^*}(\beta|y)}\right)\,d\beta - \ln p(\lambda=\lambda^*|y) \\ &\quad + \int D_{\lambda^*}(\lambda)\ln D_{\lambda^*}(\lambda)\,d\lambda. \end{aligned} \tag{3.10}$$

With the definition of the KL divergence, we thus obtain

$$\begin{aligned} KL(q(\beta)D_{\lambda^*}(\lambda)||p(\beta,\lambda|y)) &= KL(q(\beta)||p_{\lambda^*}(\beta|y) - \ln p(\lambda=\lambda^*|y) \\ &\quad + \int D_{\lambda^*}(\lambda)\ln D_{\lambda^*}(\lambda)\,d\lambda. \end{aligned} \tag{3.11}$$

Substitution of eq. (3.7) and eq. (3.11) on the right-hand side of eq. (3.3) then yields

$$\begin{aligned}
\ln p(y) &= F^{VML}(q(\beta), \lambda^*) + KL(q(\beta)||p_{\lambda^*}(\beta|y)) + \ln\left(\frac{p(\lambda=\lambda^*)}{p(\lambda=\lambda^*|y)}\right) \\
\Leftrightarrow \ln\left(\frac{p(y)p(\lambda=\lambda^*|y)}{p(\lambda=\lambda^*)}\right) &= F^{VML}(q(\beta), \lambda^*) + KL(q(\beta)||p_{\lambda^*}(\beta|y)) \\
\Leftrightarrow \ln p(y|\lambda=\lambda^*) &= F^{VML}(q(\beta), \lambda^*) + KL(q(\beta)||p_{\lambda^*}(\beta|y)) \\
\Leftrightarrow \ln p_{\lambda^*}(y) &= F^{VML}(q(\beta), \lambda^*) + KL(q(\beta)||p_{\lambda^*}(\beta|y))
\end{aligned} \tag{3.12}$$

Finally, setting $\lambda := \lambda^*$ as discussed above then yields the VML log marginal likelihood decomposition of eqs. (35) and (36) in the main text.

Going back to equation (3.7), one sees that, under the additional assumption of a constant improper prior over λ, the VB free energy is equal to the VML free energy and an additive term C that is independent of either parameter:

$$F^{VB}(q(\beta)D_{\lambda^*}(\lambda)) = F^{VML}(q(\beta), \lambda^*) + C. \tag{3.13}$$

Thus, in this case maximization of the VB free energy yields VML estimates.

4 The VML free energy and its update equations

The VML free energy is defined as

$$\begin{aligned}
F^{VML}(q(\beta), \lambda) =& \langle \ln\left(\frac{p_\lambda(y,\beta)}{q(\beta)}\right)\rangle_{q(\beta)} \\
=& \langle \ln p_\lambda(y|\beta)\rangle_{q(\beta)} + \langle \ln p(\beta)\rangle_{q(\beta)} - \langle \ln q(\beta)\rangle_{q(\beta)}.
\end{aligned} \tag{4.1}$$

The latter two terms on the right-hand side of (4.1) have been evaluated in Section 2. The first term can be evaluated using (1.2), yielding

$$\begin{aligned}
\langle \ln p_\lambda(y|\beta)\rangle_{q(\beta)} =& -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|V_\lambda| - \frac{1}{2}(y - Xm_\beta)^T V_\lambda^{-1}(y - Xm_\beta) \\
& - \frac{1}{2}\operatorname{tr}(X^T V_\lambda^{-1} X S_\beta),
\end{aligned} \tag{4.2}$$

which completes the derivation of the VML free energy as eq. (39) of the main text. To identify the update equations for the maximization of the VML free energy, we proceed as in Section 2. Because the main difference between the VB and VML framework is the parameterization of the error covariance matrix V_λ in terms of λ rather than m_λ and the vanishing of terms relating to the prior and variational distributions of λ, we can keep the discussion very concise.

The relevant subpart of $F^{VML}(m_\beta^{(i)}, S_\beta^{(i)}, \lambda^{(i)})$ depending on m_β is given by

$$f^{VML}(m_\beta) = -\frac{1}{2}(y - Xm_\beta)^T V_\lambda^{-1}(y - Xm_\beta) - \frac{1}{2}(m_\beta - \mu_\beta)^T S_\beta^{-1}(m_\beta - \mu_\beta), \tag{4.3}$$

with gradient

$$\nabla f^{VML}(m_\beta) = y^T V_\lambda^{-1} X - m_\beta^T X^T V_\lambda^{-1} X - m_\beta^T \Sigma_\beta^{-1} + \mu_\beta^T \Sigma_\beta^{-1} \tag{4.4}$$

and ensuing update equation

$$m_\beta^{(i+1)} := \left(X^T V_\lambda^{-1} X + \Sigma_\beta^{-1}\right)^{-1} \left(X^T V_\lambda^{-1} X y + \Sigma_\beta^{-1} \mu_\beta\right). \tag{4.5}$$

Likewise, the relevant subpart of $F^{VML}(m_\beta^{(i+1)}, S_\beta^{(i)}, \lambda^{(i)})$ depending on S_β is given by

$$f^{VML}(S_\beta) = -\frac{1}{2}\operatorname{tr}\left(V_\lambda^{-1} X S_\beta X^T\right) - \frac{1}{2}\operatorname{tr}(\Sigma_\beta^{-1} S_\beta) + \frac{1}{2}\ln|S_\beta| \tag{4.6}$$

with gradient

$$\nabla f^{VML}(S_\beta) = -\frac{1}{2} X^T V_\lambda^{-1} X - \frac{1}{2}\Sigma_\beta^{-1} + \frac{1}{2} S_\beta^{-1} \tag{4.7}$$

and the resulting update equation

$$S_\beta^{(i+1)} := \left(X^T V_\lambda^{-1} X + \Sigma_\beta^{-1}\right)^{-1}. \tag{4.8}$$

Finally, the relevant subpart of $F^{VML}\left(m_\beta^{(i+1)}, S_\beta^{(i+1)}, \lambda^{(i)}\right)$ depending on λ is given by

$$f^{VML}(\lambda) = -\frac{1}{2}\ln|V_\lambda| - \frac{1}{2}(y - X m_\beta)^T V_\lambda^{-1}(y - X m_\beta) - \frac{1}{2}\operatorname{tr}\left(V_\lambda^{-1} X S_\beta X^T\right). \tag{4.9}$$

Here, in analogy to eqs. (2.34) and (2.35), the entries of $\nabla f^{VML}(\lambda)$ for the case of the two-component error covariance matrix of interest (eq. (2.11)) evaluate to

$$\begin{aligned}\frac{\partial}{\partial\lambda_1} f^{VML}(\lambda) = &-\frac{1}{2}\exp(\lambda_1)\left(\operatorname{tr}(V_\lambda^{-1}) - (y - X m_\beta)^T V_\lambda^{-2}(y - X m_\beta)\right) \\ &+ \frac{1}{2}\exp(\lambda_1)\operatorname{tr}\left(V_\lambda^{-2} X S_\beta X^T\right).\end{aligned} \tag{4.10}$$

and

$$\begin{aligned}\frac{\partial}{\partial\lambda_2} f^{VML}(\lambda) = &-\frac{1}{2}\exp(\lambda_2)\left(\operatorname{tr}(V_\lambda^{-1} Q_2) - (y - X m_\beta)^T V_\lambda^{-1} Q_2 V_\lambda^{-1}(y - X m_\beta)\right) \\ &+ \frac{1}{2}\exp(\lambda_2)\operatorname{tr}\left(V_\lambda^{-1} Q_2 V_\lambda^{-1} X S_\beta X^T\right)\end{aligned} \tag{4.11}$$

5 The ReML free energy and its update equations

The ReML objective function as VML free energy

We first show that for the probabilistic model

$$p_\lambda(y,\beta) = p_\lambda(y|\beta)p(\beta) \text{ with } p_\lambda(y|\beta) = N(y; X\beta, V_\lambda) \text{ and } p(\beta) := 1 \tag{5.1}$$

it holds that the VML free energy with variational distribution

$$q(\beta) := p_\lambda(\beta|y) \tag{5.2}$$

evaluates to the ReML objective function

$$\ell_{ReML}(\lambda) := -\frac{1}{2}\ln|V_\lambda| - \frac{1}{2}\ln|X^T V_\lambda^{-1} X| - \frac{1}{2}(y - X\hat{\beta}_{GLS})^T V_\lambda^{-1}(y - X\hat{\beta}_{GLS}) \tag{5.3}$$

up to an additive constant, i.e.

$$F^{VML}(p_\lambda(\beta|y), \lambda) = \ell_{ReML}(\lambda) + c \tag{5.4}$$

with

$$c := -\frac{n}{2}\ln(2\pi) + \frac{p}{2}\ln(2\pi) \tag{5.5}$$

To this end, we first note that for the probabilistic model (5.1) and with the definition of the GLS estimator

$$\hat{\beta}_{GLS} := \left(X^T V_\lambda^{-1} X\right)^{-1} X^T V_\lambda^{-1} y \tag{5.6}$$

it holds that

$$p_\lambda(\beta|y) = N(\beta; m_\beta, S_\beta) = N\left(\beta; \hat{\beta}_{GLS}, \left(X^T V_\lambda^{-1} X\right)^{-1}\right). \tag{5.7}$$

In brief, (5.7) follows as a limiting case of the conditional properties of Gaussian distributions for the case of zero prior precision, i.e. the case of an improper prior $p(\beta) = 1$ (see e.g. (Murphy, 2012) for a more detailed discussion).

Evaluation of the VML free energy in the current scenario then yields

$$\begin{aligned} F^{VML}(p_\lambda(\beta|y), \lambda) &= \left\langle \ln\left(\frac{p_\lambda(y,\beta)}{p_\lambda(\beta|y)}\right)\right\rangle_{p_\lambda(\beta|y)} \\ &= \langle \ln\left(p_\lambda(y|\beta)p(\beta)\right)\rangle_{p_\lambda(\beta|y)} - \langle \ln p_\lambda(\beta|y)\rangle_{p_\lambda(\beta|y)} \\ &= \langle \ln p_\lambda(y|\beta)\rangle_{p_\lambda(\beta|y)} - \langle \ln p_\lambda(\beta|y)\rangle_{p_\lambda(\beta|y)}. \end{aligned} \tag{5.8}$$

Evaluation of the first term on the right-hand side (5.8) yields

$$\begin{aligned}\langle \ln p_\lambda(y|\beta)\rangle_{p_\lambda(\beta|y)} =& -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|V_\lambda| - \frac{1}{2}\langle (y-X\beta)^T V_\lambda^{-1}(y-X\beta)\rangle_{p(\lambda)(\beta|y)} \\ =& -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|V_\lambda| - \frac{1}{2}(y-X\hat{\beta}_{GLS})^T V_\lambda^{-1}(y-X\hat{\beta}_{GLS}) \\ & - \frac{1}{2}\operatorname{tr}\left(V_\lambda^{-1}X(X^T V_\lambda^{-1}X)^{-1}X^T\right) \\ =& -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|V_\lambda| - \frac{1}{2}(y-X\hat{\beta}_{GLS})^T V_\lambda^{-1}(y-X\hat{\beta}_{GLS}) \\ & - \frac{1}{2}\operatorname{tr}\left(X^T V_\lambda^{-1}X(X^T V_\lambda^{-1}X)^{-1}\right) \\ =& -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|V_\lambda| - \frac{1}{2}(y-X\hat{\beta}_{GLS})^T V_\lambda^{-1}(y-X\hat{\beta}_{GLS}) \\ & - \frac{p}{2},\end{aligned} \tag{5.9}$$

where the second equality follows with (1.2). The third equality uses the invariance of the trace under cyclic permutations. The second term on the right hand of (5.8) corresponds to the entropy of the distribution $p_\lambda(\beta|y)$ and thus evaluates to

$$H(p_\lambda(\beta|y)) = -\langle p_\lambda(\beta|y)\rangle_{p_\lambda(\beta|y)} = \frac{p}{2}\ln(2\pi e) + \ln|S_\beta| = \frac{p}{2}\ln(2\pi e) - \frac{1}{2}\ln|X^T V_\lambda^{-1}X| \tag{5.10}$$

We thus have shown that

$$F^{VML}(p_\lambda(\beta|y),\lambda) = \ell_{ReML}(\lambda) - \frac{n}{2}\ln 2\pi + \frac{p}{2}\ln(2\pi e) - \frac{p}{2}, \tag{5.11}$$

which concludes the derivation.

The ReML free energy and its update equations

To align the discussion of ReML with the previous discussions of VB and VML, we next define the ReML free energy function as the VML free energy evaluated for the probabilistic model (5.1) at the exact posterior distribution $p_\lambda(\beta|y)$, i.e.,

$$F^{ReML}(m_\beta, S_\beta, \lambda) := F^{VML}(p_\lambda(\beta|y),\lambda) = \ell_{ReML}(\lambda) + c. \tag{5.12}$$

By noting that with (5.7) the variational parameters are given by

$$m_\beta = \hat{\beta}_{GLS} \text{ and } S_\beta = (X^T V_\lambda^{-1}X)^{-1}, \tag{5.13}$$

we can then rewrite the ReML free energy as in the main text:

$$\begin{aligned}
F^{ReML}(m_\beta, S_\beta, \lambda) &= -\frac{1}{2}\ln|V_\lambda| - \frac{1}{2}\ln|X^T V_\lambda^{-1} X| \\
&\quad - \frac{1}{2}(y - X\hat{\beta}_{GLS})^T V_\lambda^{-1}(y - X\hat{\beta}_{GLS}) \\
&\quad - \frac{n}{2}\ln 2\pi + \frac{p}{2}\ln(2\pi e) - \frac{p}{2} \\
&= -\frac{1}{2}\ln|V_\lambda| + \frac{1}{2}\ln|(X^T V_\lambda^{-1} X)^{-1}| \\
&\quad - \frac{1}{2}(y - Xm_\beta)^T V_\lambda^{-1}(y - Xm_\beta) \\
&\quad - \frac{n}{2}\ln 2\pi + \frac{p}{2}\ln(2\pi e) - \frac{1}{2}\operatorname{tr}\left((X^T V_\lambda^{-1} X)(X^T V_\lambda^{-1} X)^{-1}\right) \\
&= -\frac{1}{2}\ln|V_\lambda| + \frac{1}{2}\ln|S_\beta| \\
&\quad - \frac{1}{2}(y - Xm_\beta)^T V_\lambda^{-1}(y - Xm_\beta) \\
&\quad - \frac{n}{2}\ln 2\pi + \frac{p}{2}\ln(2\pi e) - \frac{1}{2}\operatorname{tr}(S_\beta X^T V_\lambda^{-1} X) \\
&= -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|V_\lambda| - \frac{1}{2}(y - Xm_\beta)^T V_\lambda^{-1}(y - Xm_\beta) \\
&\quad - \frac{1}{2}\operatorname{tr}(S_\beta X^T V_\lambda^{-1} X) \\
&\quad + \frac{p}{2}\ln(2\pi e) + \frac{1}{2}\ln|S_\beta|.
\end{aligned} \tag{5.14}$$

Finally, we derive the update equations for the parameters m_β, S_β, and λ of the ReML free energy. Note that because the ReML objective function is identical to the ReML free energy up to an additive constant which is independent of these parameters, the resulting iterative algorithm also maximizes the ReML objective function.

The relevant subpart of $F^{ReML}(m_\beta^{(i)}, S_\beta^{(i)}, \lambda^{(i)})$ that depends on m_β is given by, omitting iteration superscripts for ease of notation,

$$f^{ReML}(m_\beta) = -\frac{1}{2}(y - Xm_\beta)^T V_\lambda^{-1}(y - Xm_\beta) \tag{5.15}$$

with gradient

$$\nabla f^{ReML}(m_\beta) = y^T V_\lambda^{-1} X - m_\beta^T X^T V_\lambda^{-1} X \tag{5.16}$$

and ensuing update equation

$$m_\beta^{(i+1)} := (X^T V_\lambda^{-1} X)^{-1} X^T V_\lambda^{-1} y. \tag{5.17}$$

Unsurprisingly, this is the GLS estimator. Further, the relevant subpart of $F^{ReML}(m_\beta^{(i+1)}, S_\beta^{(i)}, \lambda^{(i)})$ depending on S_β is given by, again omitting iteration superscripts for ease of notation,

$$f^{ReML}(S_\beta) = -\frac{1}{2}\operatorname{tr}(S_\beta X^T V_\lambda^{-1} X) + \frac{1}{2}\ln|S_\beta| \tag{5.18}$$

with gradient

$$\nabla f^{ReML}(S_\beta) = -\frac{1}{2}X^T V_\lambda^{-1} X + \frac{1}{2}S_\beta^{-1} \tag{5.19}$$

and ensuing update equation

$$S_\beta^{(i+1)} := (X^T V_\lambda^{-1} X)^{-1}. \tag{5.20}$$

Finally, because the subpart of F^{ReML} depending on λ is identical to the subpart of F^{VML} depending on λ, the update procedure for F^{ReML} with respect to λ is identical to that of F^{VML}.

6 The ML free energy and its update equations

In this section, we show how the ML objective function can be conceived as a special case of the VML free energy by evaluating F^{VML} for the probabilistic model

$$p_\lambda(y, \beta) = p_\lambda(y|\beta)p(\beta) \tag{6.1}$$

and variational distribution

$$q_\beta := D_{\beta^*}(\beta). \tag{6.2}$$

As in the discussion of VML as a special case of VB, notational care is necessary with respect to the exchange of the *random variable* β in the VML context with the *parameter value* β in the ML context. Again, we start out by denoting the random variable by β and a value it can assume by β^*. When led to only consider the case that β takes on the value β^* and its associated probability density, we identify the symbol β with β^*, resulting in the notation of the ML framework in the main text.

Substitution of (6.1) and (6.2) in the VML free energy definition yields

$$\begin{aligned}
F^{VML}(D_{\beta^*}(\beta), \lambda) &= \int D_{\beta^*}(\beta) \ln\left(\frac{p_\lambda(y,\beta)}{D_{\beta^*}(\beta)}\right) d\beta \\
&= \int D_{\beta^*}(\beta) \ln p_\lambda(y,\beta) d\beta - \int D_{\beta^*}(\beta) \ln D_{\beta^*}(\beta) d\beta \\
&= \ln p_\lambda\left(y, \beta = \beta^*\right) - \int D_{\beta^*}(\beta) \ln D_{\beta^*}(\beta) d\beta \\
&= \ln p_\lambda(y|\beta = \beta^*) + \ln p(\beta = \beta^*) - \int D_{\beta^*}(\beta) \ln D_{\beta^*}(\beta) d\beta \\
&= \ln p_{\beta^*,\lambda}(y) + \ln p(\beta = \beta^*) - \int D_{\beta^*}(\beta) \ln D_{\beta^*}(\beta) d\beta.
\end{aligned} \tag{6.3}$$

In analogy to (3.12), this leads us to define the ML free energy

$$F^{ML}\left(\beta^*, \lambda\right) = \ln p_{\beta^*,\lambda}\left(y\right) \tag{6.4}$$

which is just the log likelihood, the objective function of ML estimation. Using this definition, we obtain

$$F^{VML}(D_{\beta^*}(\beta), \lambda) = F^{ML}\left(\beta^*, \lambda\right) + \ln p(\beta = \beta^*) - \int D_{\beta^*}(\beta) \ln D_{\beta^*}(\beta) d\beta. \tag{6.5}$$

If additionally a constant improper prior $p(\beta) = 1$ is used and the constant terms are discarded, maximization of the VML free energy is thus equivalent to ML estimation. Finally, we evaluate a consistency check on this result.Substitution of (6.1) and (6.2) in the Kl divergence definition yields

$$\begin{aligned} KL(D_{\beta^*}(\beta)||p_\lambda(\beta|y)) &= \int D_{\beta^*}(\beta) \ln D_{\beta^*}(\beta) d\beta - \int D_{\beta^*}(\beta) \ln p_\lambda(\beta|y)\, d\beta \\ &= \int D_{\beta^*}(\beta) \ln D_{\beta^*}(\beta) d\beta - \ln p_\lambda(\beta = \beta^*|y) \end{aligned} \tag{6.6}$$

Substitution of (6.4) and (6.6) into the VML log marginal likelihood decomposition (eq. (34) in the main text) and receive

$$\begin{aligned} \ln p_\lambda(y) &= F^{VML}(D_{\beta^*}(\beta), \lambda) + KL(D_{\beta^*}(\beta)||p_\lambda(\beta|y)) \\ &= F^{ML}(\beta^*, \lambda) + \ln p(\beta = \beta^*) - \int D_{\beta^*}(\beta) \ln D_{\beta^*}(\beta) d\beta \\ &\quad + \int D_{\beta^*}(\beta) \ln D_{\beta^*}(\beta) d\beta - \ln p_\lambda(\beta = \beta^*|y) \\ &= \ln p_{\beta^*,\lambda}(y) + \ln p(\beta = \beta^*) - \ln p_\lambda(\beta^*|y) \\ &= \ln p_\lambda(y|\beta = \beta^*) + \ln p(\beta = \beta^*) - \ln p_\lambda(\beta = \beta^*|y) \end{aligned} \tag{6.7}$$

which is equivalent with

$$p_\lambda(\beta = \beta^*|y) = \frac{p(y|\beta = \beta^*)p(\beta = \beta^*)}{p_\lambda(y)} \tag{6.8}$$

This shows, together with section 3, that fixing the variational distributions of the VB log evidence decomposition to be Dirac delta distributions leads to a result consistent with the definition of conditional probabilities, and Bayes' theorem in particular.

For the GLM, we have

$$F^{ML}(\beta, \lambda) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |V_\lambda| - \frac{1}{2}(y - X\beta)^T V_\lambda^{-1}(y - X\beta) \tag{6.9}$$

To derive parameter update equations, we consider the dependency of F^{ML} on $\beta^{(i)}$ and $\lambda^{(i)}$ in turn. The relevant subpart of $F^{ML}(\beta^{(i)}, \lambda^{(i)})$ that depends on β is then given by, omitting iteration superscripts for ease of notation,

$$f^{ML}(\beta) = -\frac{1}{2}(y - X\beta)^T V_\lambda^{-1}(y - X\beta) \tag{6.10}$$

with gradient

$$\nabla f^{ML}(\beta) = y^T V_\lambda^{-1} X - \beta^T X^T V_\lambda^{-1} X \tag{6.11}$$

and ensuing update equation

$$\beta^{(i+1)} := (X^T V_\lambda^{-1} X)^{-1} X^T V_\lambda^{-1} y, \tag{6.12}$$

corresponding to the GLS estimator as in the case of ReML. The relevant subpart of $F^{ML}(\beta^{(i+1)}, \lambda^{(i)})$ that depends on λ differs from the VML and

ReML scenarios and is given by, again omitting iteration superscripts for ease of notation,

$$f^{ML}(\lambda) = -\frac{1}{2}\ln|V_\lambda| - \frac{1}{2}(y - X\beta)^T V_\lambda^{-1}(y - X\beta) \tag{6.13}$$

Here, in analogy to eqs. (2.33), (2.34), and (2.35), the entries of $\nabla f^{ML}(\lambda)$ for the case of the two-component error covariance matrix of interest evaluate to

$$\frac{\partial}{\partial\lambda_1} f^{ML}(\lambda) = -\frac{1}{2}\exp(\lambda_1)\left(\mathrm{tr}(V_\lambda^{-1}) - (y - X\beta)^T V_\lambda^{-2}(y - X\beta)\right) \tag{6.14}$$

and

$$\frac{\partial}{\partial\lambda_2} f^{VB}(\lambda) = -\frac{1}{2}\exp(\lambda_2)\left(\mathrm{tr}(V_\lambda^{-1}Q_2) - (y - X\beta)^T V_\lambda^{-1}Q_2 V_\lambda^{-1}(y - X\beta)\right) \tag{6.15}$$

As they correspond to a disregard of prior information and posterior uncertainty about β, equations (6.12), (6.14) and (6.15) can also be attained from the VML update equations (4.5), (4.10) and (4.11) by setting $\Sigma_\beta^{-1} = S_\beta = 0$.

7 SPM12 ReML Covariance Component Estimation

In the parameter recovery assessment of our VB, VML, ReML, and ML implementation, we found that the covariance component parameter estimation fails in a significant number of cases. To investigate whether this behaviour is specific to our implementation, we performed the same analyses using the covariance component parameter estimation functions *spm_reml_sc.m* (Version 4805) and *spm_reml.m* (Version 5223) of the SPM12 distribution. These functions perform a Fisher scoring ascent on the ReML objective function to identify maximum-a-posteriori covariance component parameter estimates, probably documented best in (Friston et al., 2002). The function *spm_reml_sc.m* uses weakly informative log normal priors to ensure the positivity of the covariance component parameter estimates, while the *spm_reml.m* function, which is called by SPM12 central *spm_spm.m* function, does not.

We visualize the results in Figure S1. The panel columns of this figure refer to the two covariance component parameter estimates and the panel rows refer to the different SPM12 functions. In the first row, we visualize the cumulative average and variances of the respective parameter estimates based on the *spm_reml_sc.m* function without the removal of outliers. The performance for λ_1 is acceptable, but for the estimation of λ_2 outliers from approximately the 10th simulation on bias the cumulative average significantly away from the true, but unknown, parameter value and strongly amplify the cumulative variance. This is similar to the behaviour we detected in our implemenation which led us to remove these outliers automatically (Grubbs, 1969). The second row of Figure S1 depicts the parameter recovery performance for *spm_reml_sc.m* after removal of appoximately 15% of outliers. This results in similar performance as in our implementation. Finally, the last row of Figure S1 depicts the parameter recovery performance for the *spm_reml.m* function. Because *spm_reml.m* can return negative covariance components and because the SPM12 procedures assume a covariance structure of the form

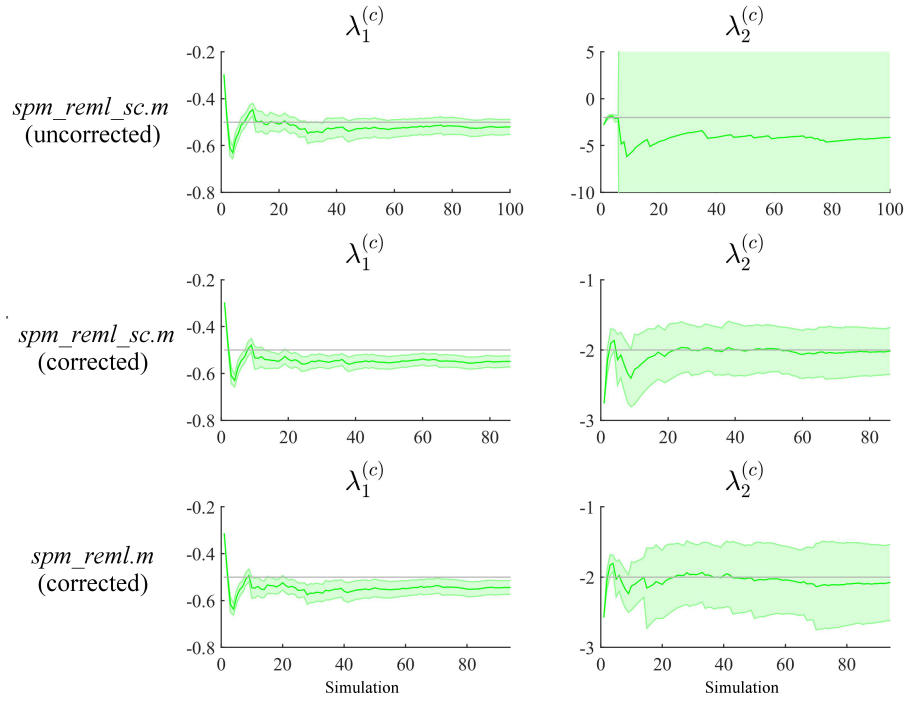

Figure 1: Parameter recovery for SPM12-based covariance component parameter estimation. The panels along the figure's columns depict the cumulative averages (gree line), cumulative variances (green shaded area), and true, but unknown, parameter values (grey) for the first and second covariance component parameters λ_1 and λ_2, respectively. The panels along the figure's rows depict these quantities for the two implementations of covariance component parameter estimation in SPM12 as indicated on the right, and without and with a correction for outliers as indicated. For implementational details, please see *vbg_2.m*.

$V_\lambda = \sum_{i=1}^{k} \lambda_i Q_i$ and not of the form $V_\lambda = \sum_{i=1}^{k} \exp(\lambda_i) Q_i$ as in our implementation, the necessary log transformation of the returned parameter estimates here can result in undefined results. In the data shown, these undefined results have been removed, again rendering the resulting cumulative averages and variances within reasonable bounds of the true, but unknown, parameter values.

In summary, we conclude that the numerical optimization problems that we encountered for the estimation of covariance components based on our implementation of the VB, VML, ReML, and ML estimation techniques are not an uncommon phenomenon in the analysis of neuroimaging data.

References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Coleman, T. F. and Li, Y. (1996). An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on optimization*, 6(2):418–445.

Friston, K., Glaser, D., Henson, R. N. A., Kiebel, S., Phillips, C., and Ashburner, J. (2002). Classical and bayesian inference in neuroimaging: applications. *Neuroimage*, 16(2):484–512.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Petersen, K. B. and Pedersen, M. S. (2012). The matrix cookbook. Version 20121115.