Variational Bayesian parameter estimation techniques for the general linear model

Ludger Starke[1] and Dirk Ostwald[1,2,3]

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

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 general linear model (GLM) with non-spherical error covariance as commonly encountered in the first-level analysis of fMRI data. To this end, we explicitly derive the corresponding free energy objective functions and ensuing iterative algorithms. Finally, in the applied part of our study, we evaluate the parameter and model recovery properties of VB, VML, ReML, and ML, first in an exemplary setting and then in the analysis of experimental fMRI data acquired from a single participant under visual stimulation. (9372 words)

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 interest, 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. Because we are only concerned with parametric probabilistic models of the Gaussian type, we assume throughout the main text that all probability distributions of real random vectors have densities. We do, however, distinguish between the conditioning of a probability distribution of a random variable y on a (commonly unobserved) random variable 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, $\operatorname{tr}(\cdot)$ denotes the trace operator, and p.d. denotes a positive-definite matrix. $H_f(a)$ denotes the Hessian matrix of some real-valued function $f(x)$ evaluated at $x = a$. We denote the probability density function of a Gaussian distributed random vector y with expectation parameter μ and covariance parameter Σ by $N(y; \mu, \Sigma)$. Finally, because of the rather applied character of this note, we formulate functions primarily by means of the definition of the values they take on and eschew formal definitions of their domains and ranges.

Further notational conventions that apply in the context of the mathematical derivations provided in the Supplementary Material are provided therein.

2 Theory

2.1 Model of interest

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

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

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

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

In words, we assume that the error term is distributed according to a Gaussian distribution with expectation parameter $0 \in \mathbb{R}^n$ and positive-definite covariance matrix $V_\lambda \in \mathbb{R}^{n\times n}$. Importantly, we do not assume that V_λ is necessarily of the form $\sigma^2 I_n$, 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*, which are assumed to be fixed and known. 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

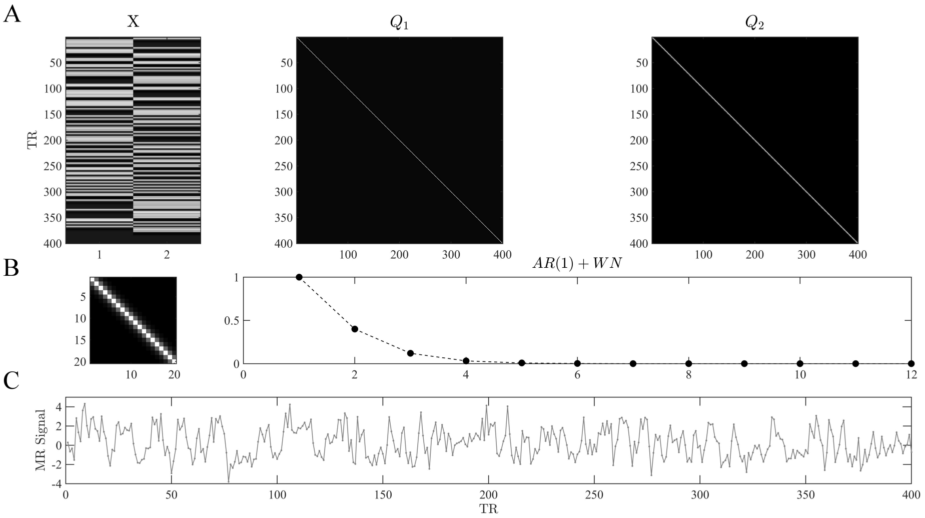


Figure 1: (A) **Example design and covariance basis matrices.** The upper panels depict the design matrix $X \in \mathbb{R}^{400\times2}$ and the covariance basis matrices $Q_1 \in \mathbb{R}^{400\times400}$ used in the example applications of the current section. The design matrix encodes the onset 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 $n = 400$ based on their evaluation using *spm_ Ce.m*. (B) The left panel depicts a magnification of the first 20 entries of Q_2. The right panel depicts 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*.

λ_2 and corresponding covariance basis matrices

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

This specific form of the error covariance matrix encodes exponentially decaying correlations between neighbouring data points, and, with $\tau := 0.2$, corresponds to the widely used approximation to the *AR(1) + white noise* model in the analysis of fMRI data (Purdon and 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 distributions 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

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

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)

$$\hat{\beta} = (X^TX)^{-1}X^Ty \tag{13}$$

and

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

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 S1.2 and here limit ourselves to the presentation of the result: under the given assumptions about the prior, likelihood, and variational distributions, the variational free energy is a function of the variational parameters $m_\beta, S_\beta, m_\lambda$, and S_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

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

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 S1.2. 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. Here, and in all remaining algorithms, convergence is assessed in terms of a vanishing of the free energy increase between successive iterations. This difference is evaluated against a convergence criterion δ, which we set to $\delta = 10^{-3}$ for all reported simulations.

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

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

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
```

Panel A of Figure 2 depicts the prior distribution over β, and the true, but unknown, value of β as black ×. 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

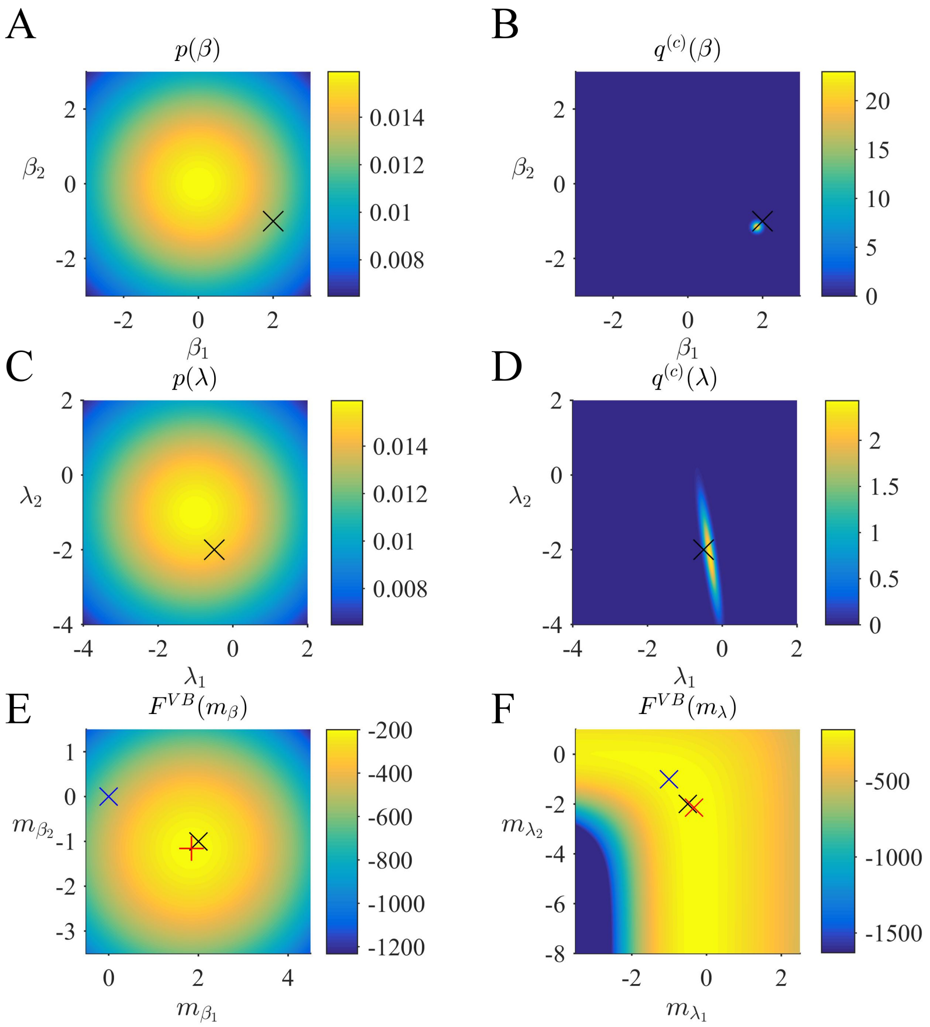


Figure 2: VB estimation. (A) Prior distribution $p(\beta)$ with expectation $\mu_\beta := (0,0)^T$ and covariance $\Sigma_\beta := 10I_2$. Here, and in all subpanels, the black × marks the true, but unknown, parameter value. (B) Variational approximation $q^{(c)}(\beta)$ to the posterior distribution 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 × indicates the prior expectation parameter and the red + marks the approximated posterior expectation parameter. (F) Variational free energy dependence on m_λ. The blue × indicates the prior expectation parameter and the red × marks the approximated posterior expectation parameter. For implementational details, please see *vbg_ 1.m*.

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 S2, we show how the VML framework can be derived as a special case of VB by assuming an improper prior for λ and a Dirac measure δ_{λ^*} for the variational distribution of λ. Importantly, it is the parameter value λ^* of the Dirac measure 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 S1.3. 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 S1.3. These update equations give rise to a VML algorithm for the current model, which we document in Algorithm 2.

Algorithm 2 VML Algorithm (for details, see *vbg_ est_ vml.m*)

```
**Input:** data $y$, prior parameters $\mu_\beta, \Sigma_\beta$, initial value $\lambda^{(1)}$, model $X, Q_1, Q_2$
**Output:** variational parameters $m_\beta^{(i)}, S_\beta^{(i)}, \lambda^{(i)}$, free energy $F^{VML^{(i)}}$
**Initialization**: $i := 1$ and $m_\beta^{(i)} := \mu_\beta$, $S_\beta^{(i)} := \Sigma_\beta$, $\Delta F^{VML^{(i)}} := \infty$, and $F^{VML^{(i)}} := F^{VML}\left(m_\beta^{(i)}, S_\beta^{(i)}, \lambda^{(i)}\right)$.
**while** $\Delta F^{VML^{(i)}} > \delta$ **do**
    $i := i + 1$
    $m_\beta^{(i)} := \left(X^T V_\lambda^{-1} X + \Sigma_\beta^{-1}\right)^{-1} \left(X^T V_\lambda^{-1} X y + \Sigma_\beta^{-1} \mu_\beta\right)$
    $S_\beta^{(i)} := \left(X^T V_\lambda^{-1} X + \Sigma_\beta^{-1}\right)^{-1}$
    solve $\frac{\partial}{\partial \lambda_j} f^{VML}\left(\lambda^{(i)}\right) = 0$ for $\lambda^{(i)}$
    evaluate $F^{VML^{(i)}} := F^{VML}\left(m_\beta^{(i)}, S_\beta^{(i)}, \lambda^{(i)}\right)$
    $\Delta F^{VML^{(i)}} := F^{VML^{(i)}} - F^{VML^{(i-1)}}$
**end while**
```

In Figure 3, we visualize the application of the VML algorithm to an example fMRI time-series realization of the model described in Section 2.1 with true, but unknown, parameter values $\beta = (2, -1)^T$ and $\lambda = (-0.5, -2)^T$. As above, we used an imprecise prior for β by setting

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

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,

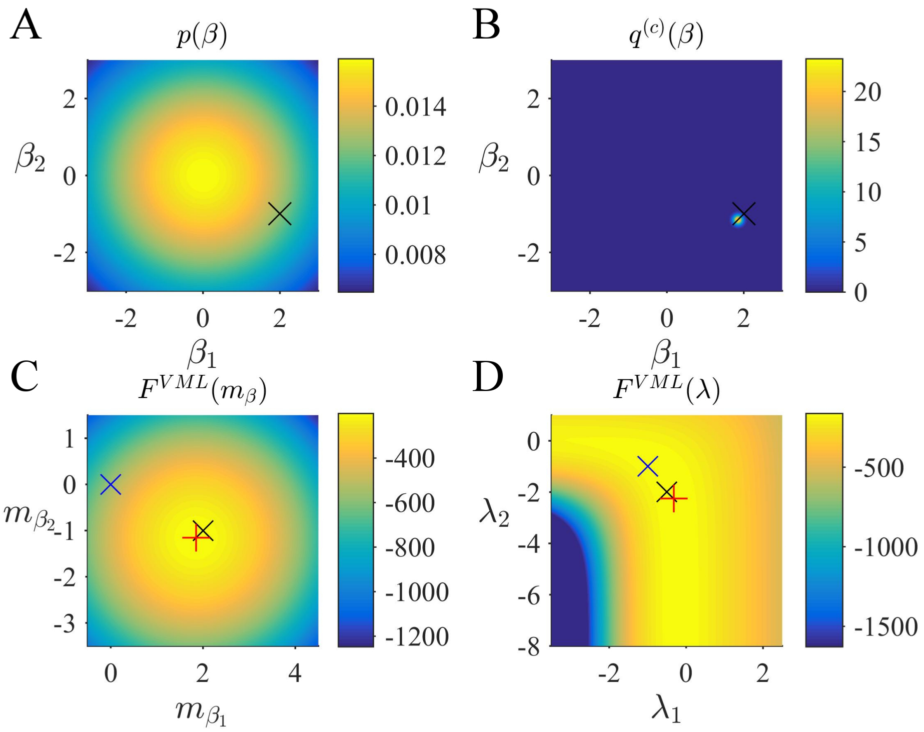


Figure 3: VML estimation.(A) Prior distribution $p(\beta)$ with expectation $\mu_\beta := (0,0)^T$ and covariance $\Sigma_\beta := 10I_2$. Here, and in all subpanels, the black × marks the true, but unknown, parameter value. (B) Variational approximation $q^{(c)}(\beta)$ to the posterior distribution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

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 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 S1.4 we show that, up to an additive constant, the ReML objective function also corresponds to the VML free energy under the assumption of an improper constant prior distribution for 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 measure δ_{λ^*}, 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 S1.4, we show that this ReML free energy can be written as

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

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.

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 $F^{ReML^{(i)}}$
Initialization: $i := 1$, $\Delta F^{ReML^{(i)}} := \infty$, and $F^{ReML^{(i)}} := F^{ReML}\left(m_\beta^{(i)}, S_\beta^{(i)}, \lambda^{(i)}\right)$.
while $\Delta F^{ReML^{(i)}} > \delta$ do
    $m_\beta^{(i)} := \left(X^T V_\lambda^{-1} X\right)^{-1} X^T V_\lambda^{-1} y$
    $S_\beta^{(i)} := \left(X^T V_\lambda^{-1} X\right)^{-1}$
    solve $\frac{\partial}{\partial \lambda_j} f^{ReML}\left(\lambda^{(i)}\right) = 0$ for $\lambda^{(i)}$
    evaluate $F^{ReML^{(i)}} := F^{ReML}\left(m_\beta^{(i)}, S_\beta^{(i)}, \lambda^{(i)}\right)$
    $\Delta F^{ReML^{(i)}} := F^{ReML^{(i)}} - F^{ReML^{(i-1)}}$
end while
```

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 B and C depict the ReML free energy surface as a function of the variational parameter m_β and λ, respectively. Note that due to the imprecise prior distributions in the VB and VML scenarios, the resulting free energy surfaces are almost identical to the ReML free energy surfaces.

2.6 Maximum Likelihood (ML)

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

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

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

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

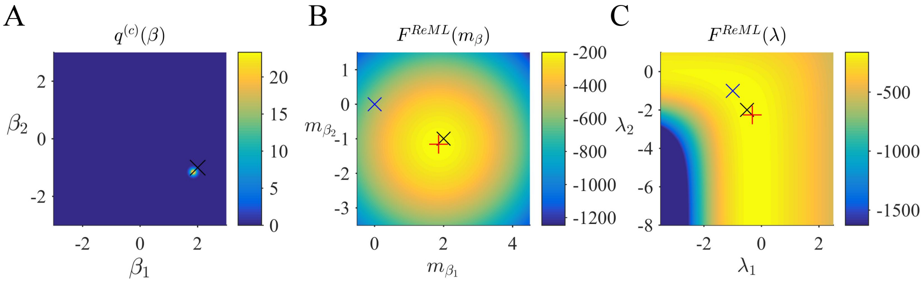


Figure 4: ReML estimation. (A) Variational distribution $q^{(c)}(\beta)$ after convergence based on the initial values $m_\beta := (0,0)^T$ and $S_\beta := 10I_2$ (convergence criterion $\delta = 10^{-3}$). Here, and in all subpanels, the black × marks the true, but unknown, parameter value. (B) ReML free energy dependence on m_β. Here, and in Panel (C) the blue × 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*.

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

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

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 (Supplementary Material S1.5) which we document in Algorithm 4. In Figure 5, we visualize the application of this ML algorithm to an example fMRI time-series realization of the model described in Section 2.1 with true, but unknown, parameter values $\beta = (2,-1)^T$ and $\lambda = (-0.5,-2)^T$, initial parameter settings of $\beta^{(1)} = (0,0)^T$ and $\lambda^{(1)} = (0,0)^T$, and ML free energy convergence criterion $\delta = 10^{-3}$. 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.

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

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
    $i := i + 1$
    $\beta^{(i)} := \left(X^T V_\lambda^{-1} X\right)^{-1} X^T V_\lambda^{-1} y$
    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
```

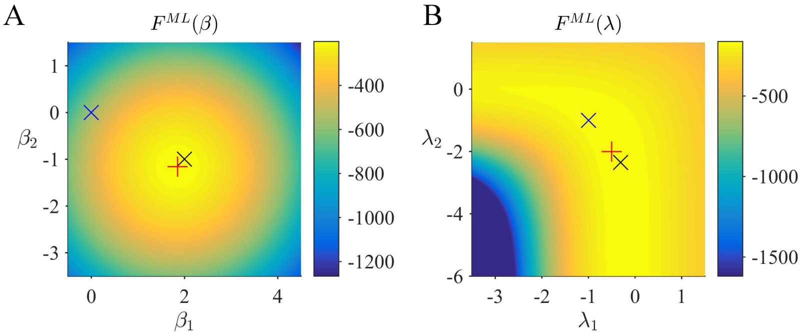


Figure 5: ML estimation. (A) ML free energy dependence on β. Here, and in Panel (B), the black × marks the true, but unknown parameter value, the blue × indicates the parameter value at algorithm initialization and the red + marks the parameter value upon algorithm convergence. (B) ML free energy dependence on λ. For implementational details, please see *vbg_1.m*.

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

3 Applications

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

3.1 Simulations

Classical statistical theory has established a variety of criteria for the assessment of an estimator's quality (e.g., Lehmann and Casella, 2006). Commonly, these criteria amount to the analytical evaluation of an estimators large sample behaviour. In the current section we adopt the spirit of this approach in simulations. To this end, we first capitalize on an objective Bayesian standpoint (Bernardo, 2003) by employing imprecise prior distributions to focus on the estimation techniques' ability to recover the true, but unknown, parameters of the data generating model and the model structure itself. Specifically, we investigate the cumulative average and variance of the β and λ parameter estimates under VB, VML, ReML, and ML and the ability of each technique's (marginal) likelihood approximation to distinguish between different data generating models. In a second step, we then demonstrate exemplarily how parameter prior specifications can induce divergences in the relative estimation qualities of the techniques.

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 (for a summary of the results without outlier removal, please refer to Supplementary Material S3). 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 S4. 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,

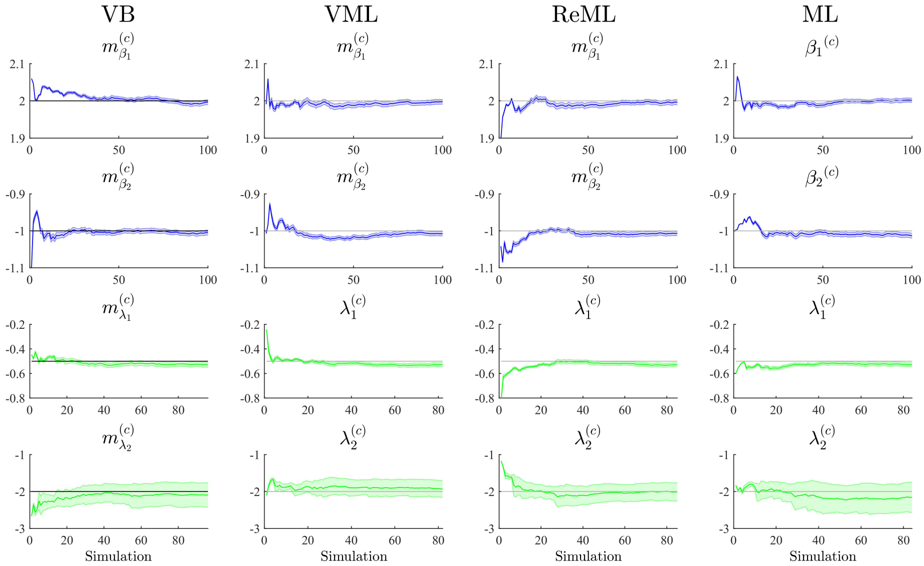


Figure 6: Parameter recovery. The panels along the figure's columns depict the cumulative averages (blue/green lines), cumulative variances (blue/green shaded areas), and true, but unknown, parameter values (grey lines) for VB, VML, ReML, and ML estimation. Parameter estimates relating to the effect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

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 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

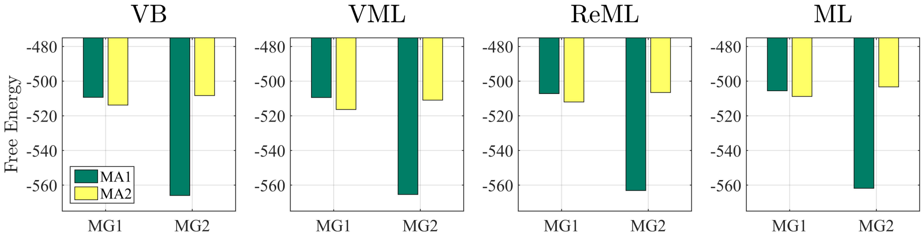


Figure 7: Model recovery. Each panel depicts the average free energies of the indicated estimation technique over 100 data realizations. Two data generating models (MG1 and MG2, panel x-axis) were used and analysed in a cross-over design with two data analysis models (MA1 and MA2, bar color). MG1 and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

data analysis model MA1 assumes the higher log model evidence approximation also in this case. We can thus conclude that model recovery is achieved satisfactorily by all estimation techniques. A more detailed decomposition of the average free energies into the respective free energy's sum terms is provided in Supplementary Material S5.

Estimation quality divergences

Thus far, we have concentrated on the nested character of VML, ReML, and ML in VB and demonstrated that for the current model application the maximum-a-posteriori (MAP) estimates of VB and VML and the point estimates of ReML and ML are able to recover true, but unknown, parameter values. Naturally, the four estimation techniques differ in the information they provide upon estimation: VB estimates quantify posterior uncertainty about both effect size and covariance component parameters, VML estimates quantify posterior uncertainty about effect size parameters only, and ReML and ML do not quantify posterior uncertainty about either parameter class. Beyond these conceptual divergences, an interesting question concerns the qualitative and quantitative differences in estimation that result from the estimation techniques' specific characteristics. In general, while the properties of ML estimates are fairly well understood from a classical frequentist perspective, the same cannot be said for the other techniques (e.g. Blei et al., 2016). We return to this point in the Discussion section. In the current section, we demonstrate divergences in the quality of parameter estimation that emerge in high noise scenarios, which are able to uncover prior distribution induced regularization effects. We demonstrate this for both effect size (Figure 8A) and covariance component parameters (Figure 8B) in the example model described in Section 2.1.

The panels in Figure 8A depict simulation estimates of the the root-mean-square-error (RMSE) $E(||\beta^{\text{MAX}} - \beta||_2)$ (uppermost panel) and biases of the effect size parameter entries $E(\beta_1^{\text{MAX}} - \beta_1)$ and $E(\beta_2^{\text{MAX}} - \beta_2)$ (middle and lowermost panel, respectively) over a range of values of the first covariance component pa-

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

Table 1: Parameter initialization for the simulations reported in Figure 8A and 8B design.

rameter λ_1. Here, $\beta^{\text{MAX}} = (\beta_1^{\text{MAX}}, \beta_2^{\text{MAX}})^T$ denotes the MAP estimates resulting under the VB and VML techniques, and the maximum (restricted) likelihood estimates resulting under ReML and ML, β denotes the true, but unknown, effect size parameter, $E(\cdot)$ denotes the expectation parameter, $\hat{E}(\cdot)$ the estimation of an expectation by means of an average, and $||\cdot||_2$ denotes the Euclidean norm of a vector. The results for the different estimation techniques are color- and linewidth-coded and were obtained under the following simulation: the true, but unknown, effect size parameter values were set to $\beta = (1,1)^T$ and the true, but unknown, parameter value of the second covariance component parameter was constant at $\lambda_2 = -2$. Varying the true, but unknown, value λ_1 on the interval $[6, 12]$ thus increased the contribution of independent and identically distributed noise to the data. For each estimation technique, the respective effect size estimates were initialized as specified in Table 1. In brief, the estimates for β_1 were initialized to the true, but unknown, value and β_2 to zero. Crucially, VB and VML allow for the specification of prior distributions over β. Here, we used a precise prior covariance of $\Sigma_{\beta_1} = 10^{-2}$ and an imprecise variance of $\Sigma_{\beta_2} = 10^1$. Note that these algorithm parameters do not exist in ReML and ML. For each setting of λ_1, 100 realizations of the model were obtained, subjected to all four estimation techniques, and the RMSE and biases estimated by averaging over realizations. The following pattern of results emerges: in terms of the RMSE (upper panel), VB and VML exhibit a more stable estimation of β, with a lower deviation from zero compared to the trend of ReML and ML estimates, at higher noise levels. In more detail, this pattern results from the following effects on the individual β_1 and β_2 estimates: first, for VB and VML, the estimates β_1 exhibit virtually no biases, because their precise prior distribution fixes them at the true, but unknown value, (middle panel). For β_2 this regularization of β_1 results in more stable estimates as compared to ReML and ML, but for higher levels of noise also results in a downward bias (lowermost panel). Taken together, this simulation demonstrates, how, in the case of prior knowledge about the effect size parameters, the endowment of their estimates with precise priors in VB and VML can stabilize the overall effect size estimation and yield better estimates compared to ReML and ML.

In a second simulation, visualized in Figure 8B, we investigated the interaction between prior regularization and estimation quality for the covariance component parameters. As in Figure 8A, the uppermost panel depicts the estimated RMSE for the λ parameters, and the middle and lowermost panels the biases of each component parameter. As in the previous simulation, the true, but unknown, effect size parameter values were set to $\beta = (1, 1)$ and $\lambda_2 = -2$ and λ_1 was varied on the interval $[-1, 1]$. The initial parameters for each estimation technique are documented in Table 1. In brief, all effect size parameter estimates (expectations) were initialized to zero, and isotropic, imprecise prior covariance matrices were employed for VB and VML. The only estimation technique that endows λ estimates with a prior distribution is VB. Here, we employ the imprecise prior covariance $\Sigma_\lambda := 10^1 I_2$, which is, however, "precise enough" to exert some stabilization effects: as shown in the uppermost panel of Figure 8B, only the RMSE of the VB technique remains largely constant over the investigated space of λ_1 values, while for all other estimation techniques the RMSE increases linearly. Two things are noteworthy here. First, at the level of the β estimates all techniques perform equally well in a bias-free manner (data not shown). Second, the λ_1 parameter space investigated includes a region (around 0.5) for which also the VB estimation quality declines, but recovers thereafter, suggesting an interaction between the structural model properties and the parameter regime. For the individual entries of λ, the decline in estimation quality in VML, ReML, and ML is not uniform: for λ_1, the estimation quality remains largely constant up to the critical region around 0.5, whereas the estimation quality of λ_2 deteriorates with increasing values of λ_1 and recovers briefly in the critical region around 0.5. Note that for both simulations of Figure 8 we did not attempt to remove potential estimation outliers, because their definition in high noise scenarios is virtually impossible. It is thus likely, that the convergence failures observed in the first set of simulations contribute to the observed estimation errors. However, because these failures also afflict the VB and VML techniques which displayed improved estimation behaviour in the simulations reported in Figure 8, it is likely that the observed pattern of results is indicative of qualitative estimation divergences.

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

3.2 Application to real data

Having validated the VB, VML, ReML, and ML implementation in simulations, we were interested in their application to real experimental data with the main aim of demonstrating the possible parameter inferences that can (and cannot) be made with each technique. To this end, we applied VB, VML, ReML, and ML to a single participant fMRI data set acquired under visual checkerboard stimulation as originally reported in (Ostwald et al., 2010). In brief, the participant was presented with a single reversing left hemi-field checkerboard stimulus for 1 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

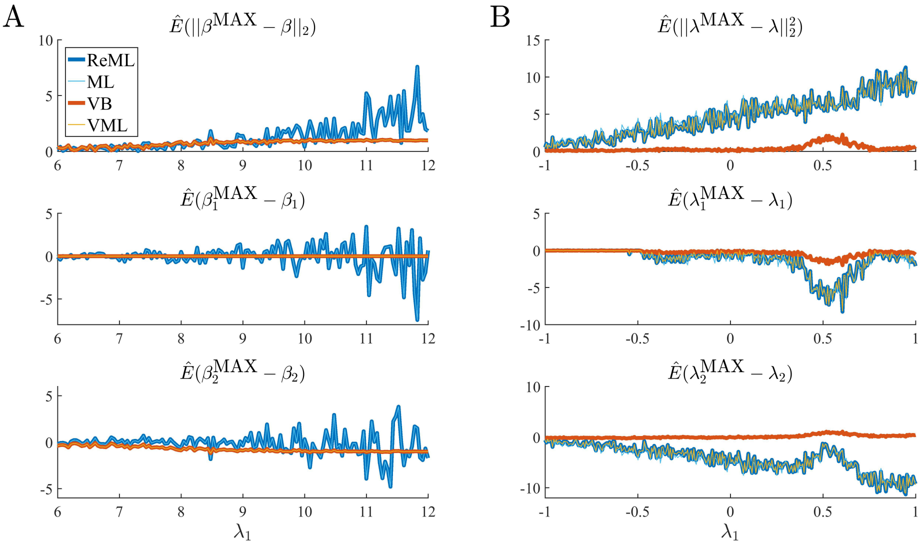


Figure 8: Estimation quality divergences. Each panel depicts the estimated RMSE and estimation bias for all four estimation techniques over a range of noise levels parameterized by λ_1. The estimation techniques are color and linewidth coded. Panel A visualizes a simulation with focus on the effect size parameter estimates β, Panel B visualizes a simulation with focus on the covariance component parameters λ. For a detailed description of the simulation, please refer to the main text and for implementational details, please see *vbg_4.m*. Note that for Panel A, the results of VB and VML and the results of ReML and ML coincide, and for Panel B the results of ReML and VML coincide.

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 10. 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. In line with the common practice in neuroimaging data analysis, no outlier removal was performed for the latter parameter. The results are visualized in Figures 9 and 10.

Figure 9 visualizes the parameter estimates relating to the effect size parameter β_1. The subpanels of Figure 10A depict the resulting two-dimensional map of converged variational parameter estimates, which differs only minimally between the four estimation techniques as indicated on the left of each panel. The variational parameter estimates are highest in the area of the right primary visual cortex, and lowest in the area of the cisterna ambiens/lower lateral ventricles. Panel B depicts the associated variational covariance parameter $S_{\beta_1}^{(c)}$, i.e., the first diagonal entry of the of the variational covariance matrix $S_\beta^{(c)} \in \mathbb{R}^{3\times3}$. 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 $m_{\beta_1}^{(c)}$ and $S_{\beta_1}^{(c)}$, 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 9D. This analysis also identifies the right lateral primary visual cortex as area of strongest activation - but in contrast to the VB, VML,

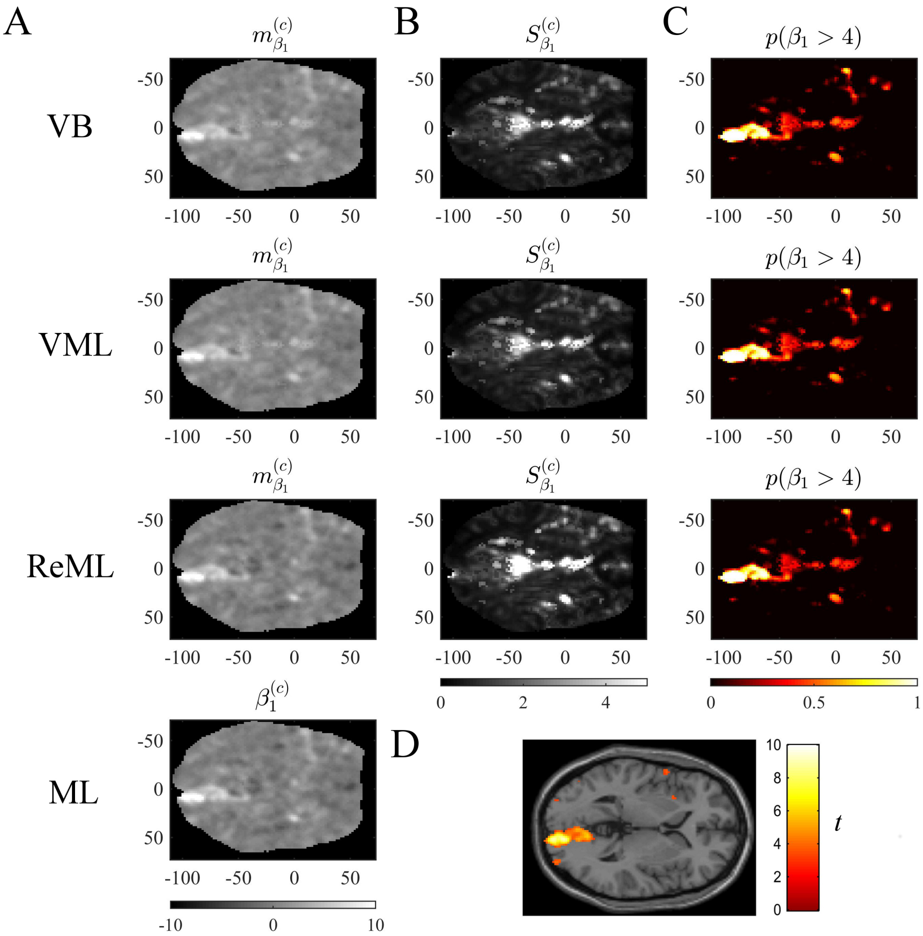


Figure 9: Effect size estimation. The figure panels depict the effect size parameter β_1 estimation results of the VB, VML, ReML, and ML algorithm application to the analysis of a single-participant single-run fMRI data set. This effect size parameter captures the effect of high contrast left visual 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_ 5.m.*

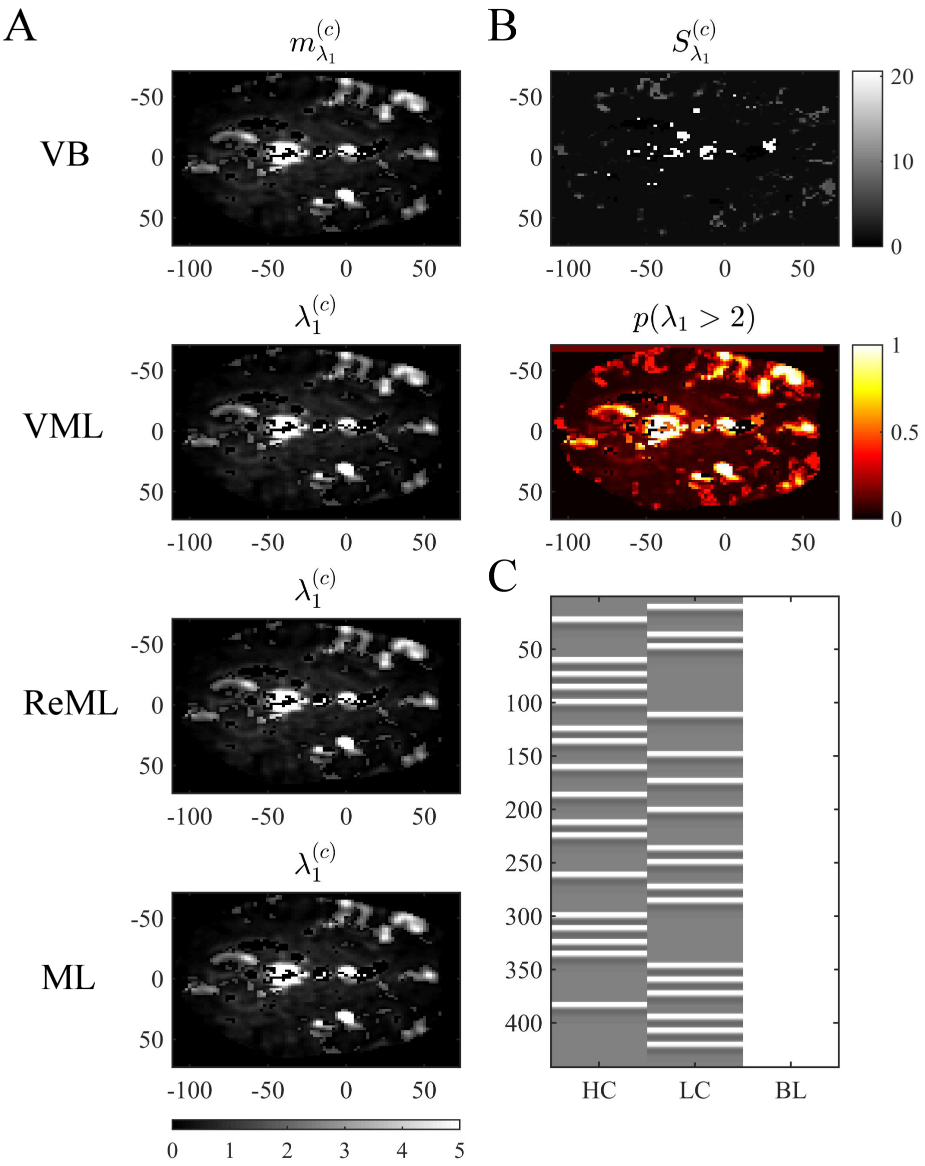


Figure 10: Covariance component parameter estimation. The figure panels depict the covariance component parameter λ_1 estimation results of the VB, VML, ReML, and ML algorithm application to the analysis of a single-participant single-run fMRI data set. This covariance component parameter captures the 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_ 5.m*.

and ReML results does not provide a visual account of the uncertainty associated with the parameter estimates and ensuing T-statistics. In summary, the VB, VML, and ReML-based quantification of effect sizes and their associated uncertainty revealed biologically meaningful results.

Figure 10 visualizes the variational expectation parameters relating to the effect size parameter λ_1. Here, the subpanels of Figure 10A visualize the variational (expectation) parameters across the four estimation techniques. High values for this covariance component are observed in the areas covering cerebrospinal fluid (cisterna ambiens, lateral and third ventricles), lateral frontal areas, and the big arteries and veins. Notably, also in right primary visual cortex, the covariance component estimate is relatively large, indicating that the design matrix does not capture all stimulus-induced variability. The only estimation technique that also quantifies the uncertainty about the covariance component parameters is VB. The results of this quantification are visualized in 10B. The first subpanel visualizes the variational covariance parameter $S_{\lambda_1}^{(c)}$, i.e., the first diagonal entry of the variational covariance matrix $S_{\lambda}^{(c)} \in \mathbb{R}^{2\times 2}$. The second subpanel visualizes the probability that the true, but unknown, covariance component parameter λ is larger than $\eta = 2$, i.e.

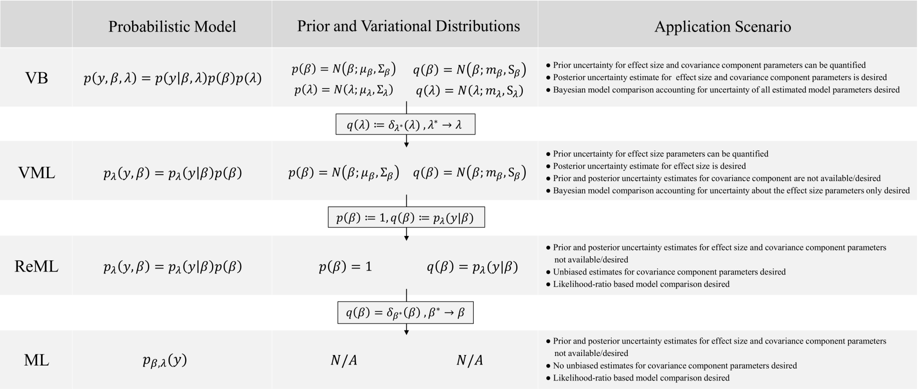
$$p(\lambda_1 > \eta) = 1 - N_{cdf}(\eta; m_{\lambda_1}, S_{\lambda_1}), \tag{56}$$

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 relationships and the non-technical application scenarios in which each technique corresponds to the method of choice in Figure 11. Further, we have provided a detailed documentation of the application of these four estimation techniques to the GLM with non-spherical, linearly decomposable error covariance, a fundamental 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.

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 measure and do not make the connection to ReML. Furthermore, these studies do not focus on the GLM and its application in the analysis of fMRI data. Finally, a number of treatises have considered the application of VB to linear regression models (e.g., Bishop, 2006; 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.



	Probabilistic Model	Prior and Variational Distributions		Application Scenario
VB	$p(y,\beta,\lambda) = p(y\|\beta,\lambda)p(\beta)p(\lambda)$	$p(\beta) = N(\beta;\mu_\beta,\Sigma_\beta)$ $p(\lambda) = N(\lambda;\mu_\lambda,\Sigma_\lambda)$	$q(\beta) = N(\beta;m_\beta,S_\beta)$ $q(\lambda) = N(\lambda;m_\lambda,S_\lambda)$	• Prior uncertainty for effect size and covariance component parameters can be quantified • Posterior uncertainty estimate for effect size and covariance component parameters is desired • Bayesian model comparison accounting for uncertainty of all estimated model parameters desired
		$q(\lambda) \coloneqq \delta_{\lambda^*}(\lambda)\,,\lambda^* \to \lambda$		
VML	$p_\lambda(y,\beta) = p_\lambda(y\|\beta)p(\beta)$	$p(\beta) = N(\beta;\mu_\beta,\Sigma_\beta)$	$q(\beta) = N(\beta;m_\beta,S_\beta)$	• Prior uncertainty for effect size parameters can be quantified • Posterior uncertainty estimate for effect size is desired • Prior and posterior uncertainty estimates for covariance component are not available/desired • Bayesian model comparison accounting for uncertainty about the effect size parameters only desired
		$p(\beta) \coloneqq 1, q(\beta) \coloneqq p_\lambda(y\|\beta)$		
ReML	$p_\lambda(y,\beta) = p_\lambda(y\|\beta)p(\beta)$	$p(\beta) = 1$	$q(\beta) = p_\lambda(y\|\beta)$	• Prior and posterior uncertainty estimates for effect size and covariance component parameters not available/desired • Unbiased estimates for covariance component parameters desired • Likelihood-ratio based model comparison desired
		$q(\beta) = \delta_{\beta^*}(\beta)\,,\beta^* \to \beta$		
ML	$p_{\beta,\lambda}(y)$	N/A	N/A	• Prior and posterior uncertainty estimates for effect size and covariance component parameters not available/desired • No unbiased estimates for covariance component parameters desired • Likelihood-ratio based model comparison desired

Figure 11: VB, VML, ReML, and ML relationships and application scenarios. N/A denotes non-applicable.

Estimator quality

Model estimation techniques yield estimators. Estimators are functions of observed data that return estimates of true, but unknown, model parameters, be it the point-estimates of classical frequentist statistics or the posterior distributions of the Bayesian paradigm (e.g., Wasserman, 2010). An important issue in the development of estimation techniques is hence the quality of estimators to recover true, but unknown, model parameters and model structure. While this issue re-appears in the functional neuroimaging literature in various guises every couple of years (e.g., Vul et al., 2009a; Eklund et al., 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). 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).

On a more general level, the relative merits of the parameter estimation techniques discussed herein form an important field for future research. Ideally, the statistical properties of estimators resulting from variational approaches were understood for the model of interest, and known properties of their specialized cases, such as the bias-free covariance component parameter estimation under ReML with respect to ML, would be deducible from these. However, as pointed out by Blei et al. (2016), the statistical properties of variational approaches are not yet well understood. Nevertheless, there exists a few results on the statistical properties of variational approaches, typically in terms of the variational expectations upon convergence and for fairly specific model classes. Of relevance for the model class considered herein is the recent work by You et al. (2014), who could show the consistency of the variational expectation in the frequentist sense, albeit for spherical covariance matrices and a gamma distribution for the covariance component parameter. For a broader model class with posterior support in real space (including the current model class of interest), Westling (2017) have worked towards establishing the consistency and asymptotic normality of variational expectation estimates. Finally, a number of authors have addressed consistency and asymptotic properties in selected model classes, such as Poisson-mixed effect models, stochastic block models, and Gaussian mixture models (Hall et al., 2011; Celisse et al., 2012; Bickel et al., 2013; Wang et al., 2006).

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

Conclusion

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

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.

Bickel, P., Choi, D., Chang, X., and Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, pages 1922–1943.

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

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2016). Variational inference: A review for statisticians. *arXiv preprint arXiv:1601.00670*.

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

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

Celisse, A., Daudin, J.-J., Pierre, L., et al. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6:1847–1899.

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.

Gal, Y. and Ghahramani, Z. (2017). On modern deep learning and variational inference. *Advances in Approximate Bayesian Inference: NIPS 2016 Workshop*.

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.

Hall, P., Pham, T., Wand, M. P., Wang, S. S., et al. (2011). Asymptotic normality and valid inference for gaussian variational approximation. *The Annals of Statistics*, 39(5):2502–2532.

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.

Wang, B., Titterington, D., et al. (2006). Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model. *Bayesian Analysis*, 1(3):625–650.

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

Westling, T., M. T. (2017). Consistency, calibration, and efficiency of variational inference. *arXiv:1510.08151v3*.

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.

You, C., Ormerod, J. T., and Müller, S. (2014). On variational bayes estimation and variational information criteria for linear regression models. *Australian & New Zealand Journal of Statistics*, 56(1):73–87.

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.

Supplementary Material

Contents

S1 Free energy algorithm derivations

In this section, we evaluate the VB, VML, ReML, and ML free energies for the GLM and derive update equations for their maximization. The notation follows the applied approach used in the main text. We commence with some remarks on additional notation and matrix differentation.

S1.1 Preliminaries

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

$$\langle f(x)\rangle_{p(x)} = \int f(x)p(x)\,dx \tag{S1.1}$$

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

$$\langle (x-m)^T A (x-m) \rangle_{N(x;\mu,\Sigma)} = (\mu - m)^T A(\mu - m) + \mathrm{tr}(A\Sigma) \tag{S1.2}$$

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

Gradient and Hessian

The gradient and Hessian of a real-valued function

$$f : \mathbb{R}^n \to \mathbb{R}, x \mapsto f(x) \tag{S1.3}$$

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

$$\nabla f(a) := \left(\frac{\partial}{\partial x_1} f(a), ..., \frac{\partial}{\partial x_n} f(a) \right)^T \in \mathbb{R}^n \tag{S1.4}$$

and

$$H_f(a) := \begin{pmatrix} \frac{\partial^2}{\partial x_1^2} f(a) & \cdots & \frac{\partial^2}{\partial x_1 \partial x_n} f(a) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1} f(a) & \cdots & \frac{\partial^2}{\partial x_n^2} f(a) \end{pmatrix} \in \mathbb{R}^{n\times n}. \tag{S1.5}$$

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 $\frac{\partial}{\partial x_i} f|_{x=a}$.

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

$$\frac{\partial |A|}{\partial x} = |A| \,\mathrm{tr}\left(A^{-1} \frac{\partial A}{\partial x} \right) \tag{S1.6}$$

$$\frac{\partial \ln |A|}{\partial x} = \mathrm{tr}\left(A^{-1} \frac{\partial A}{\partial x} \right) \tag{S1.7}$$

$$\frac{\partial A^{-1}}{\partial x} = -A^{-1} \frac{\partial A}{\partial x} A^{-1} \tag{S1.8}$$

$$\frac{\partial \,\mathrm{tr}\,(A)}{\partial x} = \mathrm{tr}\left(\frac{\partial A}{\partial x} \right). \tag{S1.9}$$

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

$$\frac{\partial^2 A^{-1}}{\partial x_1^2} = 2A^{-1} \frac{\partial A}{\partial x_1} A^{-1} \frac{\partial A}{\partial x_1} A^{-1} - A^{-1} \frac{\partial^2 A}{\partial x_1^2} A^{-1} \tag{S1.10}$$

$$\frac{\partial^2 A^{-1}}{\partial x_2^2} = 2A^{-1} \frac{\partial A}{\partial x_2} A^{-1} \frac{\partial A}{\partial x_2} A^{-1} - A^{-1} \frac{\partial^2 A}{\partial x_2^2} A^{-1} \tag{S1.11}$$

and

$$\frac{\partial^2 A^{-1}}{\partial x_1 \partial x_2} = \frac{\partial^2 A^{-1}}{\partial x_2 \partial x_1} = A^{-1}\frac{\partial A}{\partial x_1}A^{-1}\frac{\partial A}{\partial x_2}A^{-1} + A^{-1}\frac{\partial A}{\partial x_2}A^{-2}\frac{\partial A}{\partial x_1}A^{-1} - A^{-1}\frac{\partial^2 A}{\partial x_1 \partial x_2}A^{-1} \tag{S1.12}$$

assuming that A has continuous second derivatives, such that the symmetry 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

$$\frac{\partial \ln(|A|)}{\partial A} = A^{-1} \tag{S1.13}$$

and for matrices A and B of matching dimensions

$$\frac{\partial \operatorname{tr}(AB)}{\partial A} = B^T. \tag{S1.14}$$

S1.2 Variational Bayes

Evaluation of the VB free energy

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

$$\begin{aligned} F^{VB}(q(\beta)q(\lambda)) &= \langle \ln\left(\frac{p(y,\beta,\lambda)}{q(\beta)q(\lambda)}\right)\rangle_{q(\beta)q(\lambda)} \\ &= \langle \ln p(y|\beta,\lambda)\rangle_{q(\beta)q(\lambda)} + \langle \ln p(\beta)\rangle_{q(\beta)} + \langle \ln p(\lambda)\rangle_{q(\lambda)} \\ &\quad - \langle q(\beta)\rangle_{q(\beta)} - \langle q(\lambda)\rangle_{q(\lambda)}. \end{aligned} \tag{S1.15}$$

Using (S1.2), the second and third term on the right-hand side of (S1.15) can be evaluated exactly, yielding

$$\langle \ln p(\beta)\rangle_{q(\beta)} = -\frac{p}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_\beta| - \frac{1}{2}(m_\beta - \mu_\beta)^T\Sigma_\beta^{-1}(m_\beta - \mu_\beta) - \frac{1}{2}\operatorname{tr}(\Sigma_\beta^{-1}S_\beta) \tag{S1.16}$$

and

$$\langle \ln p(\lambda)\rangle_{q(\lambda)} = -\frac{k}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_\lambda| - \frac{1}{2}(m_\lambda - \mu_\lambda)^T\Sigma_\lambda^{-1}(m_\lambda - \mu_\lambda) - \frac{1}{2}\operatorname{tr}(\Sigma_\lambda^{-1}S_\lambda). \tag{S1.17}$$

corresponding to terms 6 - 13 of eq. (28) in the main text. The fourth and the fifth term on the right-hand side of (S1.15) 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)

$$H(q(\beta)) = -\langle \ln q(\beta)\rangle_{q(\beta)} = \frac{p}{2}\ln(2\pi e) + \frac{1}{2}\ln|S_\beta|, \tag{S1.18}$$

$$H(q(\lambda)) = -\langle \ln q(\lambda)\rangle_{q(\lambda)} = \frac{k}{2}\ln(2\pi e) + \frac{1}{2}\ln|S_\lambda|. \tag{S1.19}$$

Eqs. (S1.18) and (S1.19) correspond to terms 14 to 16 of eq. (28) in the main text.

Finally, we consider the first term of (S1.15). Based on the definition of $p(y|\beta,\lambda)$, the expectation with respect to $q(\beta)$ can be evaluated exactly, yielding

$$\begin{aligned}\langle \ln p(y|\beta,\lambda)\rangle_{q(\beta)q(\lambda)} = &-\frac{n}{2}\ln 2\pi - \frac{1}{2}\langle \ln |V_\lambda|\rangle_{q(\lambda)} \\ &-\frac{1}{2}\langle (y-Xm_\beta)^T V_\lambda^{-1}(y-Xm_\beta)\rangle_{q(\lambda)} \\ &-\frac{1}{2}\langle \operatorname{tr}(V_\lambda^{-1}XS_\beta X^T)\rangle_{q(\lambda)}.\end{aligned} \tag{S1.20}$$

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

$$f:\mathbb{R}^k \to \mathbb{R}, \lambda \mapsto f(\lambda) \tag{S1.21}$$

denote a real-valued function of λ. Then

$$f(\lambda) \approx f(m_\lambda) + (\lambda - m_\lambda)^T \nabla f(m_\lambda) + \frac{1}{2}(\lambda - m_\lambda)^T H_f(m_\lambda)(\lambda - m_\lambda) \tag{S1.22}$$

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

$$\begin{aligned}\langle f(\lambda)\rangle_{q(\lambda)} &\approx f(m_\lambda) + \langle(\lambda - m_\lambda)^T \nabla f(m_\lambda)\rangle_{q(\lambda)} + \frac{1}{2}\langle(\lambda - m_\lambda)^T H_f(m_\lambda)(\lambda - m_\lambda)\rangle_{q(\lambda)} \\ &= f(m_\lambda) + \frac{1}{2}\operatorname{tr}\left(H_f(m_\lambda)S_\lambda\right).\end{aligned} \tag{S1.23}$$

This approximation needs to be applied to all expectations in equation (S1.20). Thus, using the linearity of the trace to subsume all Hessian matrices into

$$\begin{aligned}B_{m_\beta,S_\beta,m_\lambda} = &H_{\ln|V_\lambda|}(m_\lambda) + H_{(y-Xm_\beta)^T V_\lambda^{-1}(y-Xm_\beta)}(m_\lambda) \\ &+ H_{\operatorname{tr}(V_\lambda^{-1}XS_\beta X^T)}(m_\lambda),\end{aligned} \tag{S1.24}$$

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 (S1.24) have to be evaluated. For the linear form of the error covariance matrix

$$V_\lambda := \exp(\lambda_1)I_n + \exp(\lambda_2)Q_2 \tag{S1.25}$$

the three Hessian matrices of (S1.24) can be evaluated analytically:

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

Using (S1.7), the first order partial derivatives are given by

$$\frac{\partial \ln|V_\lambda|}{\partial\lambda_1} = \exp(\lambda_1)\operatorname{tr}\left(V_\lambda^{-1}\right) \tag{S1.26}$$

and

$$\frac{\partial \ln|V_\lambda|}{\partial \lambda_2} = \exp(\lambda_2)\,\mathrm{tr}\left(V_\lambda^{-1}Q_2\right). \tag{S1.27}$$

Exploiting the linearity of the trace operator (S1.9) and using (S1.8) for the derivative of the inverse yields the second order partial derivatives:

$$\begin{aligned}\frac{\partial^2 \ln|V_\lambda|}{\partial \lambda_1^2} &= \exp(\lambda_1)\,\mathrm{tr}\left(V_\lambda^{-1}\right) - \exp(2\lambda_1)\,\mathrm{tr}\left(V_\lambda^{-2}\right)\\ &= \exp(\lambda_1)\,\mathrm{tr}\left(V_\lambda^{-1}\right) - \big(\exp(\lambda_1)\,\mathrm{tr}\left(V_\lambda^{-2}V_\lambda\right)\\ &\qquad\qquad - \exp(\lambda_1+\lambda_2)\,\mathrm{tr}\left(V_\lambda^{-2}Q_2\right)\big)\\ &= \exp(\lambda_1+\lambda_2)\,\mathrm{tr}\left(V_\lambda^{-2}Q_2\right)\ ,\end{aligned} \tag{S1.28}$$

$$\begin{aligned}\frac{\partial^2 \ln|V_\lambda|}{\partial \lambda_2^2} &= \exp(\lambda_2)\,\mathrm{tr}\left(V_\lambda^{-1}Q_2\right) - \exp(2\lambda_2)\,\mathrm{tr}\left(V_\lambda^{-1}Q_2V_\lambda^{-1}Q_2\right)\\ &= \exp(\lambda_2)\,\mathrm{tr}\left(V_\lambda^{-1}Q_2\right) - \big(\exp(\lambda_2)\,\mathrm{tr}\left(V_\lambda^{-1}V_\lambda V_\lambda^{-1}Q_2\right)\\ &\qquad\qquad - \exp(\lambda_1+\lambda_2)\,\mathrm{tr}\left(V_\lambda^{-2}Q_2\right)\big)\\ &= \exp(\lambda_1+\lambda_2)\,\mathrm{tr}\left(V_\lambda^{-2}Q_2\right),\end{aligned} \tag{S1.29}$$

and

$$\frac{\partial^2 \ln|V_\lambda|}{\partial \lambda_1 \partial \lambda_2} = \frac{\partial^2 \ln|V_\lambda|}{\partial \lambda_2 \partial \lambda_1} = -\exp(\lambda_1+\lambda_2)\,\mathrm{tr}\left(V_\lambda^{-2}Q_2\right), \tag{S1.30}$$

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

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

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

$$\frac{\partial^2}{\partial\lambda_i\partial\lambda_j}\Big((y-Xm_\beta)^T V_\lambda^{-1}(y-Xm_\beta)\Big) = (y-Xm_\beta)^T\frac{\partial^2 V_\lambda^{-1}}{\partial\lambda_i\partial\lambda_j}(y-Xm_\beta) \tag{S1.31}$$

for $i,j \in \{1,2\}$. Applying (S1.10) to (S1.12) yields

$$\frac{\partial^2 V_\lambda^{-1}}{\partial\lambda_1^2} = \exp(\lambda_1)\,V_\lambda^{-2} - 2\exp(2\lambda_1)\,V_\lambda^{-3}\,, \tag{S1.32}$$

$$\frac{\partial^2 V_\lambda^{-1}}{\partial\lambda_2^2} = \exp(\lambda_2)\,V_\lambda^{-1}Q_2V_\lambda^{-1} - 2\exp(2\lambda_2)\,V_\lambda^{-1}Q_2V_\lambda^{-1}Q_2V_\lambda^{-1}, \tag{S1.33}$$

and

$$\frac{\partial^2 V_\lambda^{-1}}{\partial x_1\partial x_2} = \frac{\partial^2 A^{-1}}{\partial x_2\partial x_1} = -\exp(\lambda_1+\lambda_2)\left(V_\lambda^{-2}Q_2V_\lambda^{-1} + V_\lambda^{-1}Q_2V_\lambda^{-2}\right). \tag{S1.34}$$

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

Due to the linearity of the trace operator, we have

$$\frac{\partial^2\,\mathrm{tr}\left(V_\lambda^{-1}XS_\beta X^T\right)}{\partial\lambda_i\partial\lambda_j} = \mathrm{tr}\left(\frac{\partial^2 V_\lambda^{-1}}{\partial\lambda_i\partial\lambda_j}XS_\beta X^T\right) \tag{S1.35}$$

for $i, j \in \{1, 2\}$. Thus we only have to use (S1.32) to (S1.34).

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 $Q_2^D = P^T Q_2 P$, 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

$$\begin{aligned} V_\lambda^{-1} &= \left(\exp(\lambda_1) I_n + \exp(\lambda_2) Q_2\right)^{-1} \\ &= \left(\exp(\lambda_1) P I_n P^T + \exp(\lambda_2) P Q_2^D P^T\right)^{-1} \\ &= \left(P\left(\exp(\lambda_1) I_n + \exp(\lambda_2) Q_2^D\right) P^T\right)^{-1} \\ &= P\left(\exp(\lambda_1) I_n + \exp(\lambda_2) Q_2^D\right)^{-1} P^T. \end{aligned} \tag{S1.36}$$

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

$$f^{VB}(S_\lambda) = -\frac{1}{4}\operatorname{tr}\left(B_{m_\beta^{(i)}, S_\beta^{(i)}, m_\lambda^{(i)}} S_\lambda\right) - \frac{1}{2}\operatorname{tr}(\Sigma_\lambda^{-1} S_\lambda) + \frac{1}{2}\ln|S_\lambda|. \tag{S1.37}$$

Using the identities (S1.13), (S1.14), and 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

$$\nabla f^{VB}(S_\lambda) = -\frac{1}{4} B_{m_\beta^{(i)}, S_\beta^{(i)}, m_\lambda^{(i)}} - \frac{1}{2}\Sigma_\lambda^{-1} + \frac{1}{2} S_\lambda^{-1}. \tag{S1.38}$$

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

$$S_\lambda^{(i+1)} := \left(\frac{1}{2} B_{m_\beta^{(i)}, S_\beta^{(i)}, m_\lambda^{(i)}} + \Sigma_\lambda^{-1}\right)^{-1}. \tag{S1.39}$$

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

$$f^{VB}(m_\beta) = -\frac{1}{2}(y - Xm_\beta)^T V_{m_\lambda}^{-1}(y - Xm_\beta) - \frac{1}{2}(m_\beta - \mu_\beta)^T S_\beta^{-1}(m_\beta - \mu_\beta), \tag{S1.40}$$

where we omitted iteration superscripts for visual clarity. With (S1.2), the gradient of $f^{VB}(m_\beta)$ is given by

$$\begin{aligned}\nabla f^{VB}(m_\beta) &= (y - Xm_\beta)^T V_{m_\lambda}^{-1} X - (m_\beta - \mu_\beta)^T \Sigma_\beta^{-1} \\ &= y^T V_{m_\lambda}^{-1} X - m_\beta^T X^T V_{m_\lambda}^{-1} X - m_\beta^T \Sigma_\beta^{-1} + \mu_\beta^T \Sigma_\beta^{-1}\end{aligned} \tag{S1.41}$$

Setting the gradient to zero then yields the update equation

$$m_\beta^{(i+1)} := \left(X^T V_{m_\lambda}^{-1} X + \Sigma_\beta^{-1}\right)^{-1} \left(X^T V_{m_\lambda}^{-1} y + \Sigma_\beta^{-1} \mu_\beta\right) \tag{S1.42}$$

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

$$f^{VB}(S_\beta) = -\frac{1}{2}\operatorname{tr}\left(X^T V_{m_\lambda}^{-1} X S_\beta\right) - \frac{1}{2}\operatorname{tr}(\Sigma_\beta^{-1} S_\beta) + \frac{1}{2}\ln|S_\beta| \tag{S1.43}$$

with gradient

$$\nabla f^{VB}(S_\beta) = -\frac{1}{2} X^T V_{m_\lambda}^{-1} X - \frac{1}{2}\Sigma_\beta^{-1} + \frac{1}{2} S_\beta^{-1} \tag{S1.44}$$

and the resulting update equation

$$S_\beta^{(i+1)} := \left(X^T V_{m_\lambda}^{-1} X + \Sigma_\beta^{-1}\right)^{-1}. \tag{S1.45}$$

Note that the update equations (S1.42) and (S1.45) 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,

$$\begin{aligned}f^{VB}(m_\lambda) = &-\frac{1}{2}\ln|V_{m_\lambda}| - \frac{1}{2}(y - Xm_\beta)^T V_{m_\lambda}^{-1}(y - Xm_\beta) \\ &- \frac{1}{2}\operatorname{tr}(X^T V_{m_\lambda}^{-1} X S_\beta) - \frac{1}{2}(m_\lambda - \mu_\lambda)^T \Sigma_\lambda^{-1}(m_\lambda - \mu_\lambda).\end{aligned} \tag{S1.46}$$

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

$$\begin{aligned}\frac{\partial}{\partial m_{\lambda_j}} f^{VB}(m_\lambda) = &-\frac{1}{2}\operatorname{tr}\left(V_{m_\lambda}^{-1}\left(\frac{\partial V_{m_\lambda}}{\partial m_{\lambda_j}}\right)\right) \\ &- \frac{1}{2}(y - Xm_\beta)^T\left(\frac{\partial V_{m_\lambda}^{-1}}{\partial m_{\lambda_j}}\right)(y - Xm_\beta) \\ &- \frac{1}{2}\operatorname{tr}\left(\left(\frac{\partial V_{m_\lambda}^{-1}}{\partial m_{\lambda_j}}\right) X S_\beta X^T\right) - \left((m_\lambda - \mu_\lambda)^T \Sigma_\lambda^{-1}\right)_j.\end{aligned} \tag{S1.47}$$

The evaluation of these entries for the two-component linear error covariance (S1.25) then yields

$$\begin{aligned}\frac{\partial}{\partial m_{\lambda_1}} f^{VB}(m_\lambda) = & -\frac{1}{2}\exp(m_{\lambda_1})\left(\operatorname{tr}(V_{m_\lambda}^{-1}) - (y - Xm_\beta)^T V_{m_\lambda}^{-2}(y - Xm_\beta)\right. \\ & \left. - \operatorname{tr}\left(V_{m_\lambda}^{-1} X S_\beta X^T V_{m_\lambda}^{-1}\right)\right) - \frac{1}{2}\left((m_\lambda - \mu_\lambda)\Sigma_\lambda^{-1}\right)_1,\end{aligned} \tag{S1.48}$$

and

$$\begin{aligned}\frac{\partial}{\partial m_{\lambda_2}} f^{VB}(m_\lambda) = & -\frac{1}{2}\exp(m_{\lambda_2})\left(\operatorname{tr}\left(V_{m_\lambda}^{-1} Q_2\right)\right. \\ & - (y - Xm_\beta)^T V_{m_\lambda}^{-1} Q_2 V_{m_\lambda}^{-1}(y - Xm_\beta) \\ & \left. - \operatorname{tr}\left(Q_2 V_{m_\lambda}^{-1} X S_\beta X^T V_{m_\lambda}^{-1}\right)\right) - \frac{1}{2}\left((m_\lambda - \mu_\lambda)\Sigma_\lambda^{-1}\right)_2.\end{aligned} \tag{S1.49}$$

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

$$\frac{\partial}{\partial m_{\lambda_j}} f^{VB}\left(m_\lambda^{(i+1)}\right) = 0 \tag{S1.50}$$

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).

S1.3 Variational maximum likelihood

Evaluation of the VML free energy

The VML free energy is defined as

$$\begin{aligned}F^{VML}(q(\beta), \lambda) = & \langle \ln\left(\frac{p_\lambda(y, \beta)}{q(\beta)}\right)\rangle_{q(\beta)} \\ & = \langle \ln p_\lambda(y|\beta)\rangle_{q(\beta)} + \langle \ln p(\beta)\rangle_{q(\beta)} - \langle \ln q(\beta)\rangle_{q(\beta)}.\end{aligned} \tag{S1.51}$$

The latter two terms on the right-hand side of (S1.51) have been evaluated in Section S1.2. The first term can be evaluated using (S1.2), yielding

$$\begin{aligned}\langle \ln p_\lambda(y|\beta)\rangle_{q(\beta)} = & -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|V_\lambda| - \frac{1}{2}(y - Xm_\beta)^T V_\lambda^{-1}(y - Xm_\beta) \\ & - \frac{1}{2}\operatorname{tr}(X^T V_\lambda^{-1} X S_\beta),\end{aligned} \tag{S1.52}$$

which completes the derivation of the VML free energy as eq. (39) of the main text.

The VML free energy update equations

To identify the update equations for the maximization of the VML free energy, we proceed as in Section S1.2. 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

$$f^{VML}(m_\beta) = -\frac{1}{2}(y - Xm_\beta)^T V_\lambda^{-1}(y - Xm_\beta) - \frac{1}{2}(m_\beta - \mu_\beta)^T S_\beta^{-1}(m_\beta - \mu_\beta), \quad \text{(S1.53)}$$

with gradient

$$\nabla f^{VML}(m_\beta) = y^T V_\lambda^{-1} X - m_\beta^T X^T V_\lambda^{-1} X - m_\beta^T \Sigma_\beta^{-1} + \mu_\beta^T \Sigma_\beta^{-1} \quad \text{(S1.54)}$$

and ensuing update equation

$$m_\beta^{(i+1)} := \left(X^T V_\lambda^{-1} X + \Sigma_\beta^{-1}\right)^{-1} \left(X^T V_\lambda^{-1} X y + \Sigma_\beta^{-1} \mu_\beta\right). \quad \text{(S1.55)}$$

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

$$f^{VML}(S_\beta) = -\frac{1}{2}\operatorname{tr}\left(V_\lambda^{-1} X S_\beta X^T\right) - \frac{1}{2}\operatorname{tr}(\Sigma_\beta^{-1} S_\beta) + \frac{1}{2}\ln|S_\beta| \quad \text{(S1.56)}$$

with gradient

$$\nabla f^{VML}(S_\beta) = -\frac{1}{2} X^T V_\lambda^{-1} X - \frac{1}{2}\Sigma_\beta^{-1} + \frac{1}{2} S_\beta^{-1} \quad \text{(S1.57)}$$

and the resulting update equation

$$S_\beta^{(i+1)} := \left(X^T V_\lambda^{-1} X + \Sigma_\beta^{-1}\right)^{-1}. \quad \text{(S1.58)}$$

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

$$f^{VML}(\lambda) = -\frac{1}{2}\ln|V_\lambda| - \frac{1}{2}(y - Xm_\beta)^T V_\lambda^{-1}(y - Xm_\beta) - \frac{1}{2}\operatorname{tr}\left(V_\lambda^{-1} X S_\beta X^T\right). \quad \text{(S1.59)}$$

Here, in analogy to eqs. (S1.48) and (S1.49), the entries of $\nabla f^{VML}(\lambda)$ for the case of the two-component error covariance matrix of interest (eq. (S1.25)) evaluate to

$$\begin{aligned}\frac{\partial}{\partial\lambda_1} f^{VML}(\lambda) = &-\frac{1}{2}\exp(\lambda_1)\left(\operatorname{tr}(V_\lambda^{-1}) - (y - Xm_\beta)^T V_\lambda^{-2}(y - Xm_\beta)\right)\\ &+ \frac{1}{2}\exp(\lambda_1)\operatorname{tr}\left(V_\lambda^{-2} X S_\beta X^T\right).\end{aligned} \quad \text{(S1.60)}$$

and

$$\begin{aligned}\frac{\partial}{\partial\lambda_2} f^{VML}(\lambda) = &-\frac{1}{2}\exp(\lambda_2)\left(\operatorname{tr}(V_\lambda^{-1} Q_2) - (y - Xm_\beta)^T V_\lambda^{-1} Q_2 V_\lambda^{-1}(y - Xm_\beta)\right)\\ &+ \frac{1}{2}\exp(\lambda_2)\operatorname{tr}\left(V_\lambda^{-1} Q_2 V_\lambda^{-1} X S_\beta X^T\right)\end{aligned} \quad \text{(S1.61)}$$

S1.4 Restricted maximum likelihood

The ReML objective function as VML free energy

We first show that for the probabilistic model

$$p_\lambda(y,\beta) = p_\lambda(y|\beta)p(\beta) \text{ with } p_\lambda(y|\beta) = N(y; X\beta, V_\lambda) \text{ and } p(\beta) := 1 \quad \text{(S1.62)}$$

it holds that the VML free energy with variational distribution

$$q(\beta) := p_\lambda(\beta|y) \quad \text{(S1.63)}$$

evaluates to the ReML objective function

$$\ell_{ReML}(\lambda) := -\frac{1}{2}\ln|V_\lambda| - \frac{1}{2}\ln|X^T V_\lambda^{-1} X| - \frac{1}{2}(y - X\hat{\beta}_{GLS})^T V_\lambda^{-1}(y - X\hat{\beta}_{GLS}) \quad \text{(S1.64)}$$

up to an additive constant, i.e.

$$F^{VML}(p_\lambda(\beta|y), \lambda) = \ell_{ReML}(\lambda) + c \quad \text{(S1.65)}$$

with

$$c := -\frac{n}{2}\ln(2\pi) + \frac{p}{2}\ln(2\pi) \quad \text{(S1.66)}$$

To this end, we first note that for the probabilistic model (S1.62) and with the definition of the GLS estimator

$$\hat{\beta}_{GLS} := \left(X^T V_\lambda^{-1} X\right)^{-1} X^T V_\lambda^{-1} y \quad \text{(S1.67)}$$

it holds that

$$p_\lambda(\beta|y) = N(\beta; m_\beta, S_\beta) = N\left(\beta; \hat{\beta}_{GLS}, \left(X^T V_\lambda^{-1} X\right)^{-1}\right). \quad \text{(S1.68)}$$

In brief, (S1.68) 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

$$\begin{aligned} F^{VML}(p_\lambda(\beta|y), \lambda) &= \left\langle \ln\left(\frac{p_\lambda(y,\beta)}{p_\lambda(\beta|y)}\right)\right\rangle_{p_\lambda(\beta|y)} \\ &= \langle \ln\left(p_\lambda(y|\beta)p(\beta)\right)\rangle_{p_\lambda(\beta|y)} - \langle \ln p_\lambda(\beta|y)\rangle_{p_\lambda(\beta|y)} \\ &= \langle \ln p_\lambda(y|\beta)\rangle_{p_\lambda(\beta|y)} - \langle \ln p_\lambda(\beta|y)\rangle_{p_\lambda(\beta|y)}. \end{aligned} \quad \text{(S1.69)}$$

Evaluation of the first term on the right-hand side (S1.69) yields

$$\begin{aligned}\langle \ln p_\lambda(y|\beta)\rangle_{p_\lambda(\beta|y)} =& -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|V_\lambda| - \frac{1}{2}\langle (y-X\beta)^T V_\lambda^{-1}(y-X\beta)\rangle_{p(\lambda)(\beta|y)}\\ =& -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|V_\lambda| - \frac{1}{2}(y-X\hat{\beta}_{GLS})^T V_\lambda^{-1}(y-X\hat{\beta}_{GLS})\\ & -\frac{1}{2}\operatorname{tr}\left(V_\lambda^{-1}X(X^TV_\lambda^{-1}X)^{-1}X^T\right)\\ =& -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|V_\lambda| - \frac{1}{2}(y-X\hat{\beta}_{GLS})^T V_\lambda^{-1}(y-X\hat{\beta}_{GLS})\\ & -\frac{1}{2}\operatorname{tr}\left(X^TV_\lambda^{-1}X(X^TV_\lambda^{-1}X)^{-1}\right)\\ =& -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|V_\lambda| - \frac{1}{2}(y-X\hat{\beta}_{GLS})^T V_\lambda^{-1}(y-X\hat{\beta}_{GLS})\\ & -\frac{p}{2},\end{aligned} \tag{S1.70}$$

where the second equality follows with (S1.2). The third equality uses the invariance of the trace under cyclic permutations. The second term on the right hand of (S1.69) corresponds to the entropy of the distribution $p_\lambda(\beta|y)$ and thus evaluates to

$$H(p_\lambda(\beta|y)) = -\langle p_\lambda(\beta|y)\rangle_{p_\lambda(\beta|y)} = \frac{p}{2}\ln(2\pi e) + \ln|S_\beta| = \frac{p}{2}\ln(2\pi e) - \frac{1}{2}\ln|X^TV_\lambda^{-1}X| \tag{S1.71}$$

We thus have shown that

$$F^{VML}(p_\lambda(\beta|y),\lambda) = \ell_{ReML}(\lambda) - \frac{n}{2}\ln 2\pi + \frac{p}{2}\ln(2\pi e) - \frac{p}{2}, \tag{S1.72}$$

which concludes the derivation.

Evaluation of the ReML free energy function

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 (S1.62) at the exact posterior distribution $p_\lambda(\beta|y)$, i.e.,

$$F^{ReML}(m_\beta, S_\beta, \lambda) := F^{VML}(p_\lambda(\beta|y),\lambda) = \ell_{ReML}(\lambda) + c. \tag{S1.73}$$

By noting that with (S1.68) the variational parameters are given by

$$m_\beta = \hat{\beta}_{GLS} \text{ and } S_\beta = (X^TV_\lambda^{-1}X)^{-1}, \tag{S1.74}$$

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

$$\begin{aligned}
F^{ReML}(m_\beta, S_\beta, \lambda) &= -\frac{1}{2}\ln|V_\lambda| - \frac{1}{2}\ln|X^T V_\lambda^{-1} X| \\
&\quad - \frac{1}{2}(y - X\hat{\beta}_{GLS})^T V_\lambda^{-1}(y - X\hat{\beta}_{GLS}) \\
&\quad - \frac{n}{2}\ln 2\pi + \frac{p}{2}\ln(2\pi e) - \frac{p}{2} \\
&= -\frac{1}{2}\ln|V_\lambda| + \frac{1}{2}\ln|(X^T V_\lambda^{-1} X)^{-1}| \\
&\quad - \frac{1}{2}(y - Xm_\beta)^T V_\lambda^{-1}(y - Xm_\beta) \\
&\quad - \frac{n}{2}\ln 2\pi + \frac{p}{2}\ln(2\pi e) - \frac{1}{2}\operatorname{tr}\left((X^T V_\lambda^{-1} X)(X^T V_\lambda^{-1} X)^{-1}\right) \\
&= -\frac{1}{2}\ln|V_\lambda| + \frac{1}{2}\ln|S_\beta| \\
&\quad - \frac{1}{2}(y - Xm_\beta)^T V_\lambda^{-1}(y - Xm_\beta) \\
&\quad - \frac{n}{2}\ln 2\pi + \frac{p}{2}\ln(2\pi e) - \frac{1}{2}\operatorname{tr}(S_\beta X^T V_\lambda^{-1} X) \\
&= -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|V_\lambda| - \frac{1}{2}(y - Xm_\beta)^T V_\lambda^{-1}(y - Xm_\beta) \\
&\quad - \frac{1}{2}\operatorname{tr}(S_\beta X^T V_\lambda^{-1} X) \\
&\quad + \frac{p}{2}\ln(2\pi e) + \frac{1}{2}\ln|S_\beta|.
\end{aligned} \tag{S1.75}$$

The ReML free energy update equations

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,

$$f^{ReML}(m_\beta) = -\frac{1}{2}(y - Xm_\beta)^T V_\lambda^{-1}(y - Xm_\beta). \tag{S1.76}$$

with gradient

$$\nabla f^{ReML}(m_\beta) = y^T V_\lambda^{-1} X - m_\beta^T X^T V_\lambda^{-1} X \tag{S1.77}$$

and ensuing update equation

$$m_\beta^{(i+1)} := (X^T V_\lambda^{-1} X)^{-1} X^T V_\lambda^{-1} y. \tag{S1.78}$$

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,

$$f^{ReML}(S_\beta) = -\frac{1}{2}\operatorname{tr}(S_\beta X^T V_\lambda^{-1} X) + \frac{1}{2}\ln|S_\beta| \tag{S1.79}$$

with gradient

$$\nabla f^{ReML}(S_\beta) = -\frac{1}{2}X^T V_\lambda^{-1} X + \frac{1}{2}S_\beta^{-1} \tag{S1.80}$$

and ensuing update equation

$$S_\beta^{(i+1)} := (X^T V_\lambda^{-1} X)^{-1}. \tag{S1.81}$$

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}.

S1.5 Maximum likelihood

The ML free energy update equations

For the GLM, we have by definition

$$F^{ML}(\beta, \lambda) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|V_\lambda| - \frac{1}{2}(y - X\beta)^T V_\lambda^{-1}(y - X\beta) \tag{S1.82}$$

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,

$$f^{ML}(\beta) = -\frac{1}{2}(y - X\beta)^T V_\lambda^{-1}(y - X\beta) \tag{S1.83}$$

with gradient

$$\nabla f^{ML}(\beta) = y^T V_\lambda^{-1} X - \beta^T X^T V_\lambda^{-1} X \tag{S1.84}$$

and ensuing update equation

$$\beta^{(i+1)} := (X^T V_\lambda^{-1} X)^{-1} X^T V_\lambda^{-1} y, \tag{S1.85}$$

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,

$$f^{ML}(\lambda) = -\frac{1}{2}\ln|V_\lambda| - \frac{1}{2}(y - X\beta)^T V_\lambda^{-1}(y - X\beta) \tag{S1.86}$$

Here, in analogy to eqs. (S1.47), (S1.48), and (S1.49), the entries of $\nabla f^{ML}(\lambda)$ for the case of the two-component error covariance matrix of interest evaluate to

$$\frac{\partial}{\partial \lambda_1} f^{ML}(\lambda) = -\frac{1}{2}\exp(\lambda_1)\left(\operatorname{tr}(V_\lambda^{-1}) - (y - X\beta)^T V_\lambda^{-2}(y - X\beta)\right) \tag{S1.87}$$

and

$$\frac{\partial}{\partial \lambda_2} f^{VB}(\lambda) = -\frac{1}{2}\exp(\lambda_2)\left(\operatorname{tr}(V_\lambda^{-1} Q_2) - (y - X\beta)^T V_\lambda^{-1} Q_2 V_\lambda^{-1}(y - X\beta)\right) \tag{S1.88}$$

As they correspond to a disregard of prior information and posterior uncertainty about β, equations (S1.85), (S1.87) and (S1.88) can also be attained from the VML update equations (S1.55), (S1.60) and (S1.61) by setting $\Sigma_\beta^{-1} = S_\beta = 0$.

S2 Foundations of variational Bayes

In this section we formulate a probability-theoretic model of the probabilistic model considered in the main text in order to derive the VML and ML scenarios as special cases of VB. By "probability-theoretic" we mean a measure theory-based approach to probabilistic concepts, as prevalent in contemporary mathematics (e.g. Billingsley, 2012; Shao, 2003; Fristedt and Gray, 1997). This approach is rather uncommon in the neuroimaging and machine learning literature, where many application-oriented developments on VB have taken place (Blei et al., 2016). In the current context, it is necessitated by the fact that "point probability masses" cannot be represented by probability density functions. This implies that to derive VML and ML under VB requires a careful differentiation between those random variables whose distribution can and cannot be represented by probability density functions. This is afforded by the measure theory-based approach. We assume that the reader is familiar with the measure-theoretic viewpoint of probability theory, including Lebesgue integration. To establish notation and prepare some aspects of the discussion to follow, we provide a brief summary of key elements in Section S2.1. In Section S2.2 we then review a selection of entropy formulations which will be required for the formulation of VB and VML in probability-theoretic terms. Finally, in Section S2.3 we formulate the VB, VML, and ML scenarios is probability-theoretic terms and discuss their mutual relationships.

S2.1 Preliminaries

Measurable, measure, and probability spaces

Our formulation rests on the concepts of *measurable*, *measure*, and *probability spaces*. A *measurable space* is a pair $(\Omega, \mathcal{F})$, where Ω denotes a set and $\mathcal{F}$ denotes a σ-field on Ω. An important measurable space in the following will be $(\mathbb{R}^d, \mathcal{B}^d)$, where $\mathcal{B}^d$ denotes the d-dimensional Borel σ-field on $\mathbb{R}^d$. A *measure space* is a triple $(\Omega, \mathcal{F}, \mu)$, where μ denotes a measure, i.e. a mapping $\mu : \mathcal{F} \to [0, \infty]$ with properties

(M1) $\mu(\emptyset) = 0$, and

(M2) for every pairwise disjoint sequence $\{A_i\}_{i\in\mathbb{N}}$ with $A_i \in \mathcal{F}, i \in \mathbb{N}$ it holds that $\mu\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$.

An important measure space in the following will be $(\mathbb{R}^d, \mathcal{B}^d, \mu_l^d)$, where $\mathcal{B}^d$ denotes the d-dimensional *Borel σ-field* and μ_l^d denotes the d-dimensional *Lebesgue measure*

$$\mu_l^d : \mathcal{B}^d \to [0, \infty], \times_{i=1}^d [a_i, b_i[\mapsto \mu_l^d \left(\times_{i=1}^d [a_i, b_i[\right) := \prod_{i=1}^d (b_i - a_i). \qquad \text{(S2.1)}$$

Please note that we do not use the more conventional notation λ^d for the Lebesgue measure to avoid confusion with the covariance component parameter vector λ. Similarly, a *probability space* is a triple $(\Omega, \mathcal{F}, P)$, where P denotes a probability measure, i.e. a mapping $P : \mathcal{F} \to [0, 1]$, with properties

(P1) $P(\emptyset) = 0, P(\Omega) = 1$, and

(P2) for every pairwise disjoint sequence $\{A_i\}_{i\in\mathbb{N}}$ with $A_i \in \mathcal{F}, i \in \mathbb{N}$ it holds that $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

An important probability space in the following will be $(\mathbb{R}^d, \mathcal{B}^d, \delta_x)$, where δ_x denotes the *Dirac measure*, defined as

$$\delta_x : \mathcal{B}^d \to [0,1], B \mapsto \delta_x(B) := \begin{cases} 1 \text{ if } x \in B \\ 0 \text{ if } x \notin B \end{cases} . \tag{S2.2}$$

Of key importance in the derivations to follow is the fact that the Lebesgue integral of a measurable function $f : \mathbb{R}^d \to \mathbb{R}^d$ with respect to the Dirac measure δ_x is readily evaluated as (e.g. Lieb and Loss, 2001)

$$\int f \, d\delta_x = f(x). \tag{S2.3}$$

Random variables and distributions

Let $(\Omega, \mathcal{F}, P)$ and $(\Gamma, \mathcal{S})$ denote a probability space and a measurable space, respectively. A *random variable* is a function

$$X : (\Omega, \mathcal{F}, P) \to (\Gamma, \mathcal{S}), \tag{S2.4}$$

which is measurable, i.e. for which

$$X^{-1}(S) := \{\omega \in \Omega | X(\omega) \in S\} \in \mathcal{F} \text{ for all } S \in \mathcal{S}. \tag{S2.5}$$

A random variable induces a probability measure

$$P_X : \mathcal{S} \to [0,1], S \mapsto P_X(S) := P(\{\omega \in \Omega | X(\omega) \in S\}), \tag{S2.6}$$

on $\mathcal{S}$. The probability measure P_X is referred to as the *distribution* of the random variable X and renders $(\Gamma, \mathcal{S}, P_X)$ a probability space (e.g. Fristedt and Gray, 1997, Chapter 2).

Probability density functions

For a measure space $(\Omega, \mathcal{F}, \mu)$ any quasiintegrable function $f : \Omega \to \mathbb{R}_{\geq 0}$ is a *density function* and defines a measure ν on $\mathcal{F}$ by means of its Lebesgue integral for $A \in \mathcal{F}$, i.e.

$$\nu : \mathcal{F} \to [0, \infty], A \mapsto \nu(A) := \int_A f \, d\mu. \tag{S2.7}$$

We say that "ν is a measure with density function f with respect to the measure μ" and write $\nu = f\mu$ for short. Recall that a $\mathcal{F}$-measurable function $g : \Omega \to \bar{\mathbb{R}}$ is integrable with respect to ν, if the function product $g \cdot f$ is integrable with respect to μ, and that in this case

$$\int_\Omega g \, d\nu = \int_\Omega g \cdot f \, d\mu \tag{S2.8}$$

(e.g. Billingsley, 2012, Theorem 16.11). As noted above, we will be primarily concerned with the measure space $(\mathbb{R}^d, \mathcal{B}^d, \mu_l^d)$, where μ_l^d denotes the Lebesgue measure on $\mathbb{R}^d$. In this case, if the Lebesgue integral of f with respect to μ_l^d equals 1, f is referred to as *probability density function*. Furthermore, in this case we have for $\nu = f\mu_l^d$

$$\int g \, d\nu = \int g \cdot f \, d\mu_l^d = \int g(x) f(x) \, dx. \tag{S2.9}$$

Crucially, this implies that one can evaluate Lebesgue integrals using Riemann integration as done throughout Section S1 (right-hand side of (S2.9), for details see e.g. Schmidt (2011), Chapter 9).

We further require the notion of conditional probability density functions. For a random variable (X, Y) on a product measure space $(\Omega, \mathcal{F}, \mu) := (\Omega_x \times \Omega_y, \mathcal{F}_x \times \mathcal{F}_y, \mu_x \times \mu_y)$ with joint probability density function $f_{X,Y} : \Omega \to \mathbb{R}_{>0}$ with respect to μ, the conditional probability density function of X given $Y = y$ is defined as

$$f_{X|Y} : \Omega \to \mathbb{R}_{>0}, (x, y) \mapsto f_{X|Y}(x, y) := f_{X,Y}(x, y)/f_Y(y) \tag{S2.10}$$

where

$$f_Y : \Omega_2 \to \mathbb{R}_{>0}, y \mapsto f_Y(y) := \int f_{X,Y}(x, y) \, d\mu_x \tag{S2.11}$$

is the marginal probability density function of Y with respect to μ_y (e.g. Shao, 2003, Chapter 1.4).

Discrete, continuous, and mixed random vectors

Because we are considering multivariate random entities in the application of VB, VML, and ML to the GLM, we also require the notion of *random vectors* as the multivariate extension of random variables. More specifically, we require the concepts of *discrete, continuous* and *mixed random vectors*, which we introduce in the following.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. A d-dimensional *discrete random vector* is a function

$$X : \Omega \to \mathbb{R}^d, \omega \mapsto X(\omega) := (X_1(\omega), ..., X_d(\omega))^T \tag{S2.12}$$

whose range space or *alphabet* (Gray, 2011; Cover and Thomas, 2012)

$$\mathcal{X} := \{x_i\}_{i=1}^n := X(\Omega) \subset \mathbb{R}^d \tag{S2.13}$$

is finite. A discrete random vector has an associated *probability mass function* p_X given by

$$p_X : \mathbb{R}^d \to [0, 1], p_X(x) := \begin{cases} P(X = x), & \text{if } x \in \mathcal{X} \\ 0, & \text{if } x \notin \mathcal{X} \end{cases}, \tag{S2.14}$$

with the notational convention

$$P(X = x) := P(\{X = x\}) = P(\{\omega \in \Omega | X(\omega) = x\}). \tag{S2.15}$$

A special discrete random vector required in the following is a *constant random vector*, which for a fixed $x^* \in \mathbb{R}^d$ we write as

$$X^* : \Omega \to \mathbb{R}^d, \omega \mapsto X^*(\omega) := x^*. \tag{S2.16}$$

In this case the alphabet $\mathcal{X}^* = \{x^*\}$ of X^* comprises a single element, the associated probability mass of which is given by

$$p_{X^*}(x^*) = P(\{\omega \in \Omega | X^*(\omega) = x^*\}) = P(\Omega) = 1. \tag{S2.17}$$

Analogously to a random variable, a random vector $X : \Omega \to \mathbb{R}^d$ induces a probability measure P_X on the measurable space $(\mathbb{R}^d, \mathcal{B}^d)$. Notably, the induced probability measures of a constant random vector is the Dirac measure (e.g. Bauer, 1991, p. 25), i.e. with (S2.2) and (S2.16)

$$P_{X^*} = \delta_{x^*}. \tag{S2.18}$$

A d-dimensional *continuous random vector* is a function Y of the form (S2.12) whose induced probability measure P_Y on $(\mathbb{R}^d, \mathcal{B}^d)$ is absolutely continuous with respect to Lebesgue measure and can thus be represented by a probability density function $f_Y : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$.

Finally, we construct the concept of a *mixed random vector* as follows. Set $d := d_1 + d_2$ $(d_1, d_2 \in \mathbb{N})$, let $(\Omega, \mathcal{F})$ be a measurable space, and let $X : \Omega \to \mathbb{R}^{d_1}$ be a discrete random vector, which induces a distribution P_X on $\mathcal{P}(\mathcal{X})$. Let $P_{Y|X}$ be a Markov kernel from $(\mathcal{X}, \mathcal{P}(\mathcal{X}))$ to $(\mathbb{R}^{d_2}, \mathcal{B}^{d_2})$, i.e. $P_{Y|X}$ is a mapping

$$P_{Y|X} : \mathcal{X} \times \mathcal{B}^{d_2} \to \mathbb{R} \tag{S2.19}$$

with the properties

- $P_{Y|X}(x, \cdot) : \mathcal{B}^{d_2} \to [0,1]$ is a probability measure on $\mathcal{B}^{d_2}$ for every $x \in \mathcal{X}$,
- $P_{Y|X}(\cdot, B) : \mathcal{X} \to \mathbb{R}$ is $\mathcal{P}(\mathcal{X})$-measurable for every $B \in \mathcal{B}^{d_2}$.

Then $P_{X,Y} = P_X P_{Y|X}$ is a probability measure on $(\Omega \times \mathbb{R}^{d_2}, \mathcal{F} \otimes \mathcal{B}^{d_2})$ (e.g. Shao, 2003, 1.4.3). Assume in addition that

$$P_{Y|X=x} := P_{Y|X}(x, \cdot) = f_{Y|X=x} \mu_l^{d_2}, \tag{S2.20}$$

i.e.

$$P_{Y|X=x}(B) = \int_B f_{Y|X=x}(y)\, dy \tag{S2.21}$$

for probability density functions $f_{Y|X=x}$ with $x \in \mathcal{X}$. Let Y denote the identity mapping on $\mathbb{R}^{d_2}$ and define $Z := (X, Y)$. Then Z is a d-dimensional random vector on $(\Omega \times \mathbb{R}^{d_2}, \mathcal{F} \otimes \mathcal{B}^{d_2})$ with marginal distributions having the properties

$$P_X(A) = P_{X,Y}(A \times \mathbb{R}^{d_2}) = \int_A P_{Y|X}(x, \mathbb{R}^{d_2})\, dP_X = \sum_{x \in A} p_X(x) \tag{S2.22}$$

for all $A \in \mathcal{P}(\mathcal{X})$ and

$$P_Y(B) = \sum_{x \in \mathcal{X}} p_X(x) \int_B f_{Y|X=x}(y)\, dy = \int_B \sum_{x \in \mathcal{X}} p_X(x) f_{Y|X=x}(y)\, dy =: \int_B f_Y(y)\, dy \tag{S2.23}$$

for all $B \in \mathcal{B}^{d_2}$. In other words, X is (by definition) a discrete d_1-dimensional random vector (which we call the *discrete component* of Z) and Y is (by construction) a continuous d_2-dimensional random vector (which we call the *continuous component* of Z). We call Z a d-dimensional *mixed random vector*.

Note that if we set $P_{Y|X}(x, \cdot) := P_Y$ for every $x \in \mathcal{X}$ with a probability measure P_Y on $\mathcal{B}^{d_2}$, then $P_{X,Y} = P_X P_Y$ is a product measure and the random vectors X and Y are independent. Vice versa, assuming independent d_1- and d_2-dimensional random vectors X and Y, respectively, we can use the construction above to construct a d-dimensional mixed random vector with discrete and continuous components whose marginal distributions are independent.

S2.2 Entropies of distributions of random vectors

Entropy of the distributions of a discrete random vector

Following (Gray, 2011, Chapter 3), we define the entropy of the distribution P_X of a discrete random vector X with alphabet $\mathcal{X}$ as

$$H(P_X) := -\sum_{x \in \mathcal{X}} p_X(x) \ln p_X(x) \tag{S2.24}$$

with the convention $0 \ln 0 := 0$. For later reference we note that the the entropy of the distribution of a constant random vector of the form (S2.16) is zero, because in this case the defining sum (S2.24) comprises a single term which evaluates to zero:

$$H(P_{X^*}) = -p_{X^*}(x^*) \ln p_{X^*}(x^*) = -1 \ln 1 = 0. \tag{S2.25}$$

Entropy of the distribution of a continuous random vector

We define the entropy of the distribution P_Y of a continuous random vector Y as its *differential entropy* (e.g. Cover and Thomas, 2012, Chapter 8), i.e. we set

$$h(P_Y) := -\int_{\mathbb{R}^d} f_Y(y) \ln f_Y(y)\, dy. \tag{S2.26}$$

Entropy of the distribution of a mixed random vector

Finally, following (Nair et al., 2006), we define the entropy of the distribution P_Z of a mixed random vector $Z = (X, Y)$ with the property

$$\int_{\mathbb{R}^{d_2}} |f_{Y|X=x}(y) \ln f_{Y|X=x}(y)| dy < \infty \tag{S2.27}$$

for all $x \in \mathcal{X}$ by

$$\mathcal{H}(P_Z) := -\sum_{x \in \mathcal{X}} \int_{\mathbb{R}^{d_2}} p_X(x) f_{Y|X=x}(y) \ln \left(p_X(x) f_{Y|X=x}(y)\right) dy \tag{S2.28}$$

Note that we can rewrite this definition as

$$\mathcal{H}(P_Z) = -\sum_{x\in\mathcal{X}} p_X(x)\ln p_X(x) - \sum_{x\in\mathcal{X}} p_X(x)\int_{\mathbb{R}^{d_2}} f_{Y|X=x}(y)\ln(f_{Y|X=x}(y))\,dy. \tag{S2.29}$$

$\mathcal{H}(P_Z)$ thus comprises the sum of the entropy of the marginal distribution of the discrete component of Z and a convex combination of the differential entropies of the conditional distributions of the continuous components.

In particular, for the case of independent discrete and continuous components, i.e. $f_{Y|X=x} := f_Y$, we obtain

$$\mathcal{H}(P_Z) = H(P_X) + h(P_Y). \tag{S2.30}$$

More generally, if the components X and Y are independent, and Y may be either continuous or discrete (rendering Z a discrete random vector), we write

$$\mathcal{H}(P_Z) = H(P_X) + \mathbb{H}(P_Y), \tag{S2.31}$$

where $\mathbb{H}$ denotes the entropy of the continuous or discrete component Y.

S2.3 Probability-theoretic variational Bayes

Based on the concepts reviewed in Sections S2.1 and S2.2, we are now in the position to formulate the VB, VML, and ML scenarios discussed in the main text in probability-theoretic terms and to delineate their relationship. We proceed as follows. First, we reformulate the free energy functions F^{VB}, F^{VML} and F^{ML} introduced in the main text in probability-theoretic terms. To distinguish these functions from their counterparts in the main text, they will be denoted by $\mathbb{F}^{VB}, \mathbb{F}^{VML}$, and $\mathbb{F}^{ML}$, respectively. Note that in general, the free energy functions depend on both the realizations of the observed data random variables as well as the distributions (or values in the VML and ML case) of the unobserved parameter random variables (or non-random variables in the VML and ML case). However, in analogy to classical likelihood functions (e.g. Shao, 2003, Chapter 4.4), we conceive of the free energy functions as functions of entities related to the parameter (random) variables only. Intuitively, this corresponds to assumption of a given and fixed data observation - the common scenario in experimental applications of the approaches. Second, upon reformulating the VB, VML, and ML scenarios in probability-theoretic terms, we relate these new formulations to the definitions of the variational free energies in the main text and show their consistency. Finally, we conclude with a theorem on the relationshiop between VB, VML, and ML.

Variational Bayes

To express the VB scenario of the main text in probability-theoretic terms, we set $d_1 := n+p+k$ $(n, p, k \in \mathbb{N})$, consider a probability space $(\Omega, \mathcal{F}, P)$ and the measurable space $(\mathbb{R}^{d_1}, \mathbb{B}^{d_1})$ and define the continuous random vector

$$(Y, B, L) : \Omega \to \mathbb{R}^{d_1}, \omega \mapsto (Y, B, L)(\omega), \tag{S2.32}$$

which induces a probability measure $P_{Y,B,L}$ on $\mathcal{B}^{d_1}$. We thus obtain the probability space $(\mathbb{R}^{d_1}, \mathcal{B}^{d_1}, P_{Y,B,L})$. Because (Y, B, L) is a continuous random vector, $P_{Y,B,L}$ can be represented by a probability density function

$$f_{Y,B,L} : \mathbb{R}^{d_1} \to \mathbb{R}_{>0}, (y, \beta, \lambda) \mapsto f_{Y,B,L}(y, \beta, \lambda) \tag{S2.33}$$

with respect to Lebesgue measure $\mu_l^{d_1}$ on $\mathcal{B}^{d_1}$. Note that $f_{Y,B,L}$ is denoted by $p(y, \beta, \lambda)$ in the main text.

To define the variational Bayes free energy function $\mathbb{F}^{VB}$, we first consider a random vector

$$(\tilde{B}, \tilde{L}) : \Omega \to \mathbb{R}^{p+k}, \, \omega \mapsto (\tilde{B}, \tilde{L})(\omega) \tag{S2.34}$$

whose components are the independent random vectors $\tilde{B} : \Omega \to \mathbb{R}^p$ and $\tilde{L} : \Omega \to \mathbb{R}^k$. This implies that the induced distribution $Q_{(\tilde{B},\tilde{L})}$ on the measurable space $(\mathbb{R}^{p+k}, \mathcal{B}^{p+k})$ of $(\tilde{B}, \tilde{L})$ factorizes, i.e.

$$Q_{(\tilde{B},\tilde{L})} = Q_{\tilde{B}} \otimes Q_{\tilde{L}} \tag{S2.35}$$

where $Q_{\tilde{B}}$ and $Q_{\tilde{L}}$ denote the marginal distribution on $\mathcal{B}^p$ and $\mathcal{B}^k$ induced by $\tilde{B}$ and $\tilde{L}$, respectively (e.g. Fristedt and Gray, 1997, Chapter 9). We hence write $Q_{\tilde{B}\otimes\tilde{L}}$ for $Q_{(\tilde{B},\tilde{L})}$. Let $\mathcal{Q}_{\tilde{B}\otimes\tilde{L}}$ denote the set of all such distributions. For a fixed $y \in \mathbb{R}^n$ we then define

$$\mathbb{F}^{VB} : \mathcal{Q}_{\tilde{B}\otimes\tilde{L}} \to \mathbb{R}, Q_{(\tilde{B}\otimes\tilde{L})} \mapsto \mathbb{F}^{VB}(Q_{\tilde{B}\otimes\tilde{L}}) := \int \ln f_{Y,B,L}(y, \cdot, \cdot) \, dQ_{\tilde{B}\otimes\tilde{L}} + \mathbb{H}(Q_{\tilde{B}\otimes\tilde{L}}). \tag{S2.36}$$

Here the symbol $\mathbb{H}$ denotes the entropy of the distribution $Q_{\tilde{B}\otimes\tilde{L}}$, the evaluation of which depends on the type of random vectors $\tilde{B}$ and $\tilde{L}$, as discussed in Section S2.2.

Variational maximum likelihood

In analogy to the above, we set $d_2 := n + p \; (n, p \in \mathbb{N})$, consider a probability space $(\Omega, \mathcal{F}, P)$ and the measurable space $(\mathbb{R}^{d_2}, \mathcal{B}^{d_2})$, and define the continuous random vector

$$(Y, B) : \Omega \to \mathbb{R}^{d_2}, \omega \mapsto (Y, B)(\omega), \tag{S2.37}$$

which induces the probability measure $P_{Y,B}$ on $\mathcal{B}^{d_2}$. We thus obtain the probability space $(\mathbb{R}^{d_2}, \mathcal{B}^{d_2}, P_{Y,B})$. As in the main text, we assume that $P_{Y,B}$ is represented by a parameter-dependent probability density function

$$f_{Y,B}^{\lambda} : \mathbb{R}^{d_2} \to \mathbb{R}_{>0}, (y, \beta) \mapsto f_{Y,B}^{\lambda}(y, \beta) \tag{S2.38}$$

with respect to Lebesgue measure on $\mathcal{B}^{d_2}$. Note that $f_{Y,B}^{\lambda}$ is denoted by $p_\lambda(y, \beta)$ in the main text.

To define the variational maximum likelihood free energy function $\mathbb{F}^{VML}$, we first consider a random vector $\tilde{B}$

$$\tilde{B} : \Omega \to \mathbb{R}^p, \omega \mapsto \tilde{B}(\omega), \tag{S2.39}$$

which induces a probability measure $Q_{\tilde{B}}$ on the measurable space $(\mathbb{R}^p, \mathcal{B}^p)$. Let $\mathcal{Q}_{\tilde{B}}$ denote the set of all such induced probability measures. For a fixed $y \in \mathbb{R}^n$ we then define

$$\mathbb{F}^{VML} : \mathcal{Q}_{\tilde{B}} \times \mathbb{R}^k \to \mathbb{R}, \, (Q_{\tilde{B}}, \lambda) \mapsto \mathbb{F}^{VML}(Q_{\tilde{B}}, \lambda) := \int \ln f^{\lambda}_{Y,B}(y, \cdot) dQ_{\tilde{B}} + \mathbb{H}(Q_{\tilde{B}}), \quad \text{(S2.40)}$$

where as above the symbol $\mathbb{H}$ denotes the entropy of the distribution $Q_{\tilde{B}}$, the evaluation of which depends on the type of random vector $\tilde{B}$ as discussed in Section S2.2.

Maximum likelihood

Finally, to express the maximum likelihood scenario in probability-theoretic terms, we consider a probability space $(\Omega, \mathcal{F}, P)$, the measurable space $(\mathbb{R}^n, \mathcal{B}^n)$ and define the continuous random vector

$$Y : \Omega \to \mathbb{R}^n, \omega \mapsto Y(\omega), \quad \text{(S2.41)}$$

which induces the probability measure P_Y on $\mathcal{B}^n$ and hence the probability space $(\mathbb{R}^n, \mathcal{B}^n, P_Y)$. As in the main text, we assume that P_Y is represented by a parameter-dependent probability density function

$$f_Y^{\beta,\lambda} : \mathbb{R}^n \to \mathbb{R}_{>0}, y \mapsto f_Y^{\beta,\lambda}(y) \quad \text{(S2.42)}$$

with respect to Lebesgue measure on $\mathcal{B}^n$. Note that $f_Y^{\beta,\lambda}$ is denoted by $p_{\beta,\lambda}(y)$ in the main text. We define the maximum likelihood free energy function $\mathbb{F}^{ML}$ as the standard log-likelihood function of the maximum likelihood scenario, i.e. for fixed $y \in \mathbb{R}^n$, we set

$$\mathbb{F}^{ML} : \mathbb{R}^p \times \mathbb{R}^k \to \mathbb{R}, (\beta, \lambda) \mapsto \mathbb{F}^{ML}(\beta, \lambda) := \ln f_Y^{\beta,\lambda}(y). \quad \text{(S2.43)}$$

Note that we have $\mathbb{F}^{ML} = F^{ML}$ by definition.

This concludes the probability-theoretic formulations of the VB, VML, and ML, scenarios. We next show the consistency of the free energy functions defined in this section with the definitions of the free energy functions considered in the main text in form of the following lemma:

Lemma (Consistency of free energy function definitions).

The definitions of the variational free energy functions $\mathbb{F}^{VB}$ and F^{VB}, as well as $\mathbb{F}^{VML}$ and F^{VML} are consistent. More specifically,

(L1) *if in the definition of the variational Bayes free energy $\mathbb{F}^{VB}$(S2.36) $\tilde{B}$ and $\tilde{L}$ are continuous random vectors represented by probability density functions $q_{\tilde{B}}$ and $q_{\tilde{L}}$ with respect to Lebesgue measures μ_l^p and μ_l^k, respectively, then the definitions of $\mathbb{F}^{VB}$ and F^{VB} are equivalent, and*

(L2) *if in the definition of variational maximum likelihood free energy $\mathbb{F}^{VML}$ (S2.40) $\tilde{B}$ is a continuous random vector represented by a probability density function $q_{\tilde{B}}$ with respect to Lebesgue measure, then the definitions of $\mathbb{F}^{VML}$ and F^{VML} are equivalent.*

Proof of (L1).

Consider the definitions

$$\mathbb{F}^{VB}(Q_{\tilde{B}\otimes\tilde{L}}) := \int \ln f_{Y,B,L}(y, \cdot, \cdot) \, dQ_{\tilde{B}\otimes\tilde{L}} + \mathbb{H}(Q_{\tilde{B}\otimes\tilde{L}}) \quad \text{(S2.44)}$$

and, under the mean-field approximation,

$$F^{VB}(q(\beta)q(\lambda)) = \iint q(\beta)q(\lambda)\ln\left(\frac{p(y,\beta,\lambda)}{q(\beta)q(\lambda)}\right) d\beta d\lambda \tag{S2.45}$$

Note that the integral in (S2.44) denotes a Lebesgue integral with respect to the probability measure $Q_{\tilde{B}\otimes\tilde{L}}$, while the integral in (S2.45) denotes a (double) Riemann integral. If $\tilde{B}$ and $\tilde{L}$ are continuous random vectors with associated probability density functions $q_{\tilde{B}}$ and $q_{\tilde{L}}$, then with (S2.9) and the notational conventions of the main text, we have

$$\begin{aligned}\int \ln f_{Y,B,L}(y,\cdot,\cdot)\, dQ_{\tilde{B}\otimes\tilde{L}} &= \iint q_{\tilde{B}}(\beta)q_{\tilde{L}}(\lambda)\ln f_{Y,B,L}(y,\beta,\lambda)\, d\mu_l^p(\beta)d\mu_l^k(\lambda)\\ &= \iint q_{\tilde{B}}(\beta)q_{\tilde{L}}(\lambda)\ln f_{Y,B,L}(y,\beta,\lambda)\, d\beta d\lambda.\end{aligned} \tag{S2.46}$$

Further, because $(\tilde{B},\tilde{L})$ is a continuous random vector, (S2.26) applies for the entropy of $Q_{\tilde{B}\otimes\tilde{L}}$, and thus

$$\mathbb{H}(Q_{\tilde{B}\otimes\tilde{L}}) = h(Q_{\tilde{B}\otimes\tilde{L}}) = -\iint q_{\tilde{B}}(\beta)q_{\tilde{L}}(\lambda)\ln(q_{\tilde{B}}(\beta)q_{\tilde{L}}(\lambda))\, d\beta d\lambda. \tag{S2.47}$$

Hence, (L1) follows with the linearity of the (Riemann) integral and omission of the subscripts $\tilde{B}$ and $\tilde{L}$ in the denotation of the probability density functions $q_{\tilde{B}}$ and $q_{\tilde{L}}$.

□

Proof of (L2).

Consider the definitions

$$\mathbb{F}^{VML}(Q_{\tilde{B}},\lambda) = \int \ln f^{\lambda}_{Y,B}(y,\cdot)dQ_{\tilde{B}} + \mathbb{H}(Q_{\tilde{B}}) \tag{S2.48}$$

and

$$F^{VML}(q(\beta),\lambda) = \int q(\beta)\ln\left(\frac{p_\lambda(y,\beta)}{q(\beta)}\right) d\beta. \tag{S2.49}$$

Note that the integral in (S2.48) denotes a Lebesgue integral with respect to the probability measure $Q_{\tilde{B}}$, while the integral in (S2.49) denotes a Riemann integral. If $\tilde{B}$ is a continuous random vector with associated density function $q_{\tilde{B}}$ with respect to Lebesgue measure, then with (S2.9) and the notational conventions of the main text, we have

$$\begin{aligned}\int \ln f^{\lambda}_{Y,B}(y,\cdot)dQ_{\tilde{B}} &= \int q_{\tilde{B}}(\beta)\ln f^{\lambda}_{Y,B}(y,\beta)\, d\mu_l^p(\beta)\\ &= \int q_{\tilde{B}}(\beta)\ln p_\lambda(y,\beta)\, d\beta.\end{aligned} \tag{S2.50}$$

Further, because $\tilde{B}$ is a continuous random vector, (S2.26) applies for the entropy of $Q_{\tilde{B}}$, and thus

$$\mathbb{H}(Q_{\tilde{B}}) = h(Q_{\tilde{B}}) = -\int q_{\tilde{B}}(\beta)\ln q_{\tilde{B}}(\beta)\, d\beta. \tag{S2.51}$$

Hence, (L2) follows with the linearity of the (Riemann) integral and omission of the subscript $\tilde{B}$ in the denotation of the probability density function $q_{\tilde{B}}$.

□

Finally, we provide the key result of this section:

Theorem (Relationship of VB, VML, and ML). *Variational maximum likelihood and maximum likelihood are special cases of variational Bayes. More specifically:*

(T1) *For a constant marginal density $f_L(\lambda) := 1$ and the constant random vector*

$$\tilde{L}^* : \Omega \to \mathbb{R}^k, \omega \mapsto \tilde{L}^*(\omega) := \lambda^* \tag{S2.52}$$

it holds that

$$\mathbb{F}^{VB}(Q_{\tilde{B}\otimes\tilde{L}^*}) = \mathbb{F}^{VML}(Q_{\tilde{B}}, \lambda). \tag{S2.53}$$

(T2) *For a constant marginal density $f_B(\beta) := 1$ and the constant random vector*

$$\tilde{B}^* : \Omega \to \mathbb{R}^p, \omega \mapsto \tilde{B}^*(\omega) := \beta^* \tag{S2.54}$$

it holds that

$$\mathbb{F}^{VML}(Q_{\tilde{B}^*}, \lambda) = \mathbb{F}^{ML}(\beta^*, \lambda). \tag{S2.55}$$

Proof of (T1)

We first note that because $\tilde{B}$ and $\tilde{L}^*$ are independent random vectors and because $\tilde{L}^*$ is a constant random vector, we have with (S2.30), (S2.31), and (S2.25)

$$\mathbb{H}(Q_{\tilde{B}\otimes\tilde{L}^*}) = \mathbb{H}(Q_{\tilde{B}}) + H(Q_{\tilde{L}^*}) = \mathbb{H}(Q_{\tilde{B}}). \tag{S2.56}$$

Second, with (S2.18) we have $Q_{\tilde{L}^*} = \delta_{\lambda^*}$. Hence, with Fubini's theorem,

$$\begin{aligned}\int \ln f_{Y,B,L}\, dQ_{\tilde{B}\otimes\tilde{L}^*} &= \int \left(\int \ln f_{Y,B,L}(y,\beta,\lambda)\, d\delta_{\lambda^*} \right) dQ_{\tilde{B}} \\ &= \int \ln f_{Y,B,L}(y,\beta,\lambda^*) dQ_{\tilde{B}} \\ &= \int \ln f_{Y,B|L}(y,\beta|\lambda^*) f_L(\lambda^*) dQ_{\tilde{B}} \\ &= \int \ln f_{Y,B|L}(y,\beta|\lambda^*) dQ_{\tilde{B}},\end{aligned} \tag{S2.57}$$

where the last equality follows with $f_L(\lambda^*) = 1$. Notationally identifying the probability density function

$$f_{Y,B|L}(\cdot,\cdot|\lambda^*) : \mathbb{R}^{n+p} \to \mathbb{R}_{>0}, (y,\beta) \mapsto f_{Y,B|L}(y,\beta|\lambda^*) \tag{S2.58}$$

with the probability density function $f_{Y,B}^{\lambda^*}$ and omission of the asterisk superscript then completes the proof.

□

Proof of (T2)

We first note that because $\tilde{B}^*$ is a constant random vector, we have with (S2.25) $\mathbb{H}(Q_{\tilde{B}^*}) = 0$. Second, with (S2.18) we have $Q_{\tilde{B}^*} = \delta_{\beta^*}$. Hence,

$$\begin{aligned}\int \ln f_{Y,B}^{\lambda}(y,\cdot)\, dQ_{\tilde{B}^*} &= \int \ln f_{Y,B}^{\lambda}(y,\cdot)\, d\delta_{\beta^*} \\ &= \ln f_{Y,B}^{\lambda}(y,\beta^*) \\ &= \ln f_{Y|B}^{\lambda}(y|\beta^*)\end{aligned} \tag{S2.59}$$

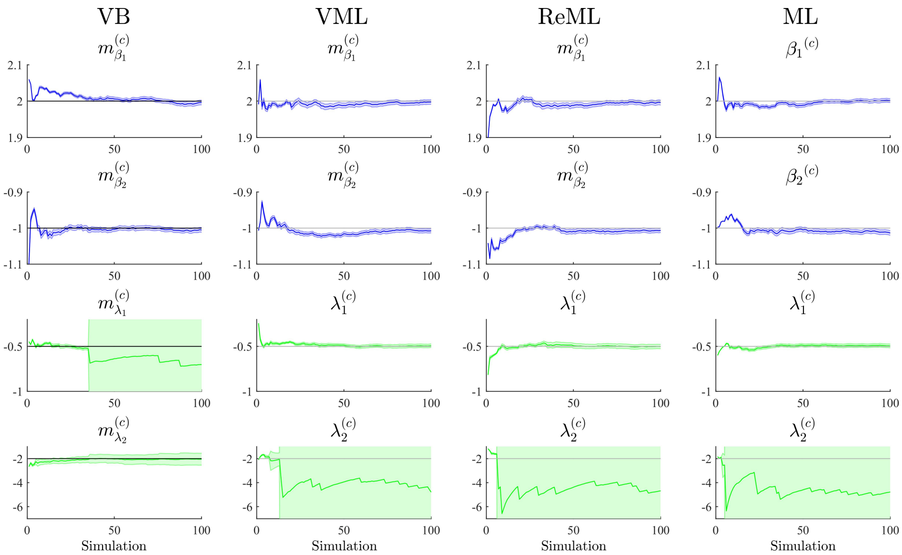


Figure S1: The panels along the figure's columns depict the cumulative averages (blue/green lines), cumulative variances (blue/green shaded areas), and true, but unknown, parameter values (grey lines) for VB, VML, ReML, and ML estimation. Parameter estimates relating to the effect sizes β are visualized 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. As opposed to the data shown in the main text, the covariance component parameter estimates are not corrected for outliers. For implementational details, please see *vbg_2.m*.

where the latter equality follows with $f_B(\beta^*) = 1$. Notationally identifying the probability density function

$$f_{Y|B}^{\lambda}(\cdot|\beta^*) : \mathbb{R}^n \to \mathbb{R}_{\geq 0}, y \mapsto f_{Y|B}^{\lambda}(y|\beta^*) \tag{S2.60}$$

with the probability density function $f_Y^{\beta^*,\lambda}$ and omission of the asterisk superscript then completes the proof.

□

S3 Cumulative averages without outlier removal

In Section 3.1 of the main text, we present a parameter recovery simulation for which we removed 15-20 % of outliers in the estimation of the covariance component parameters. The same results without outlier removal are depicted in Figure S1. As evident from the Figure, outliers affect all four estimation techniques, but primarily one of the covariance components. Retaining the outliers results in negative estimation bias estimates and an increase of the estimation variance estimate.

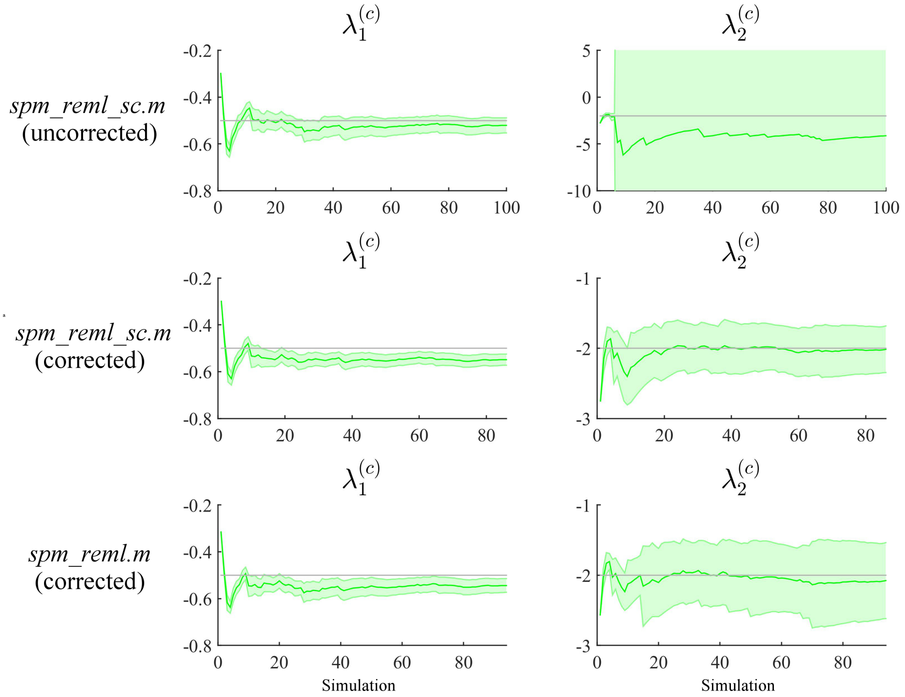


Figure S2: Parameter recovery for SPM12-based covariance component parameter estimation. The panels along the figure's columns depict the cumulative averages (gree line), cumulative variances (green shaded area), and true, but unknown, parameter values (grey) for the first and second covariance component 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*.

S4 SPM12 ReML 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 S2. 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 implementation which led us to remove these outliers automatically (Grubbs, 1969). The second row of Figure S2 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 SS2 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 $V_\lambda = \sum_{i=1}^{k} \lambda_i Q_i$ 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.

S5 Model recovery free energy contributions

To understand the observed pattern of average free energies in Figure 7 in further detail, we tabulated the sum terms of each free energy function (Table S1) and visualize the average term contributions to the overall average free energy in Figure S3. We omit from visualization the first term, which is identical for all free energy functions and evaluates to T1 $= -367.58$. Of the remaining terms, the largest contributions are provided by T3 and T2, reflecting the residual sum of squares and the log determinant of the estimated data covariance matrix, respectively (Figure S3A). The remaining terms T4 - T17, as far as they exist for each free energy function, make smaller contributions (Figure S3B). Notably, the residual sum of squares is virtually identical over all pairings of data generating and data analysis model. This reflects the fact that the two-regressor model MA2 can readily capture the data pattern of the single-regressor model MG1 by estimating β_2 to be approximately zero. The average free energy differences for the two data analysis models in case of MG2 thus appear to be primarily accounted for by the different contributions of T2. It is likely that these differences result from the erroneous allocation of data variance under model MG2 to the covariance components of model MA1. The more subtle differences between the average free energies for MA1 and MA2 in the case of MG1 on the other hand, seem to arise from two factors: firstly, a slight overestimation of the covariance component parameters of MA2 in case of MG1, leading to a persistence of the lower average free energy values in the case of ML estimation, and secondly, from additional contributions of

T	F^{VB}	F^{VML}	F^{ReML}	F^{ML}
1	$-\frac{n}{2}\ln 2\pi$	$-\frac{n}{2}\ln 2\pi$	$-\frac{n}{2}\ln 2\pi$	$-\frac{n}{2}\ln 2\pi$
2	$-\frac{1}{2}\ln \lvert V_{m_\lambda}\rvert$	$-\frac{1}{2}\ln \lvert V_\lambda\rvert$	$-\frac{1}{2}\ln \lvert V_\lambda\rvert$	$-\frac{1}{2}\ln \lvert V_\lambda\rvert$
3	$-\frac{1}{2}(y - Xm_\beta)^T V_{m_\lambda}^{-1}(y - Xm_\beta)$	$-\frac{1}{2}(y - Xm_\beta)^T V_\lambda^{-1}(y - Xm_\beta)$	$-\frac{1}{2}(y - Xm_\beta)^T V_\lambda^{-1}(y - Xm_\beta)$	$-\frac{1}{2}(y - X\beta)^T V_\lambda^{-1}(y - X\beta)$
4	$-\frac{1}{2}\operatorname{tr}(S_\beta X^T V_{m_\lambda}^{-1} X)$	$-\frac{1}{2}\operatorname{tr}(S_\beta X^T V_\lambda^{-1} X)$	$-\frac{1}{2}\operatorname{tr}(S_\beta X^T V_\lambda^{-1} X)$	
5	$-\frac{1}{4}\operatorname{tr}(B_{m_\lambda, S_\beta, m_\lambda} S_\lambda)$			
6	$-\frac{p}{2}\ln 2\pi$	$-\frac{p}{2}\ln 2\pi$		
7	$-\frac{1}{2}\ln \lvert\Sigma_\beta\rvert$	$-\frac{1}{2}\ln \lvert\Sigma_\beta\rvert$		
8	$-\frac{1}{2}(m_\beta - \mu_\beta)^T \Sigma_\beta^{-1}(m_\beta - \mu_\beta)$	$-\frac{1}{2}(m_\beta - \mu_\beta)^T \Sigma_\beta^{-1}(m_\beta - \mu_\beta)$		
9	$-\frac{1}{2}\operatorname{tr}(\Sigma_\beta^{-1} S_\beta)$	$-\operatorname{tr}(\Sigma_\beta^{-1} S_\beta)$		
10	$-\frac{k}{2}\ln 2\pi$			
11	$-\frac{1}{2}\ln \lvert\Sigma_\lambda\rvert$			
12	$-\frac{1}{2}(m_\lambda - \mu_\lambda)^T \Sigma_\lambda^{-1}(m_\lambda - \mu_\lambda)$			
13	$-\frac{1}{2}\operatorname{tr}(\Sigma_\lambda^{-1} S_\lambda)$			
14	$+\frac{p}{2}\ln(2\pi e)$	$+\frac{p}{2}\ln(2\pi e)$	$+\frac{p}{2}\ln(2\pi e)$	
15	$+\frac{1}{2}\ln \lvert S_\beta\rvert$	$+\frac{1}{2}\ln \lvert S_\beta\rvert$	$+\frac{1}{2}\ln \lvert S_\beta\rvert$	
16	$+\frac{k}{2}\ln(2\pi e)$			
17	$+\frac{1}{2}\ln \lvert S_\lambda\rvert$			

Table S1: Free energy sum terms. Note that T1 is identical over all free energy functions, T2 is the negative log determinant of the estimated data covariance matrix, and T3 corresponds to the residual sum of squares. The remaining terms, if they exist relate to the prior and posterior uncertainties over model parameters and are commonly referred to as "model complexity" terms.

.

T4 - T16 for VB, VML, and ReML, as evident from the more negative sums (Σ) of these terms. In summary, for the current data generating and data analysis model comparison, both covariance component overestimation and the free energy model complexity terms T4 - 16 appear to contribute to the identifiability of the true, but unknown, model structure.

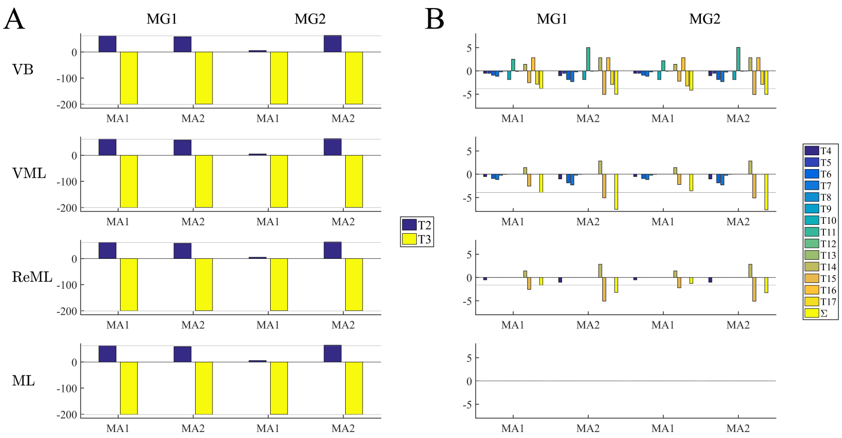


Figure S3: Average free energy term contributions. The figure depicts a decomposition of the average free energy values depicted in the main text Figure 7 according to the terms tabulated in Table S1. Panel A displays the largest contributions, afforded by terms T2 and T3, Panel B displays the remaining term contributions, and the sum (Σ) of these remaining distributions. Note the difference in scale between Panels A and B. Panel rows refer to the four estimation techniques. In each subpanel, the left two bar groups refer to data generated by MG1 analyzed with data analysis models MA1 and MA2, and the right two bar groups refer to data generated by MG2 analyzed with MA1 and MA2. For visual comparison, thin grey lines corresponding to the values obtained under the MG1/MA1 combination are included. For implementational details, please see *vbg_ 3.m*.

References

Bauer, H. (1991). *Wahrscheinlichkeitstheorie.(4. Aufl.) de Gruyter.*

Billingsley, P. (2012). *Probability and Measure, Anniversary Edition.* John Wiley & Sons, Inc.

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

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2016). Variational inference: A review for statisticians. *arXiv preprint arXiv:1601.00670.*

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.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory.* John Wiley & Sons.

Fristedt, B. E. and Gray, L. F. (1997). *A modern approach to probability theory.* Birkhauser.

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.

Gray, R. M. (2011). *Entropy and information theory*. Springer Science & Business Media.

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

Lieb, E. H. and Loss, M. (2001). Analysis, volume 14 of graduate studies in mathematics. *American Mathematical Society, Providence, RI,*, 4.

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

Nair, C., Prabhakar, B., and Shah, D. (2006). On entropy for mixtures of discrete and continuous variables. *arXiv preprint cs/0607075*.

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.

Schmidt, K. D. (2011). *Mass und Wahrscheinlichkeit*. Springer.

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