NeuroImage 179 (2018) 604–619

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Variational Bayesian inversion for hierarchical unsupervised generative embedding (HUGE)

Yu Yao [a,*], Sudhir S. Raman [a], Michael Schiek [b], Alex Leff [c], Stefan Frässle [a], Klaas E. Stephan [a,c]

[a] Translational Neuromodeling Unit, Institute for Biomedical Engineering, University of Zurich and ETH Zurich, 8032, Zurich, Switzerland
[b] Central Institute ZEA-2 Electronic Systems, Research Center Jülich, 52425 Jülich, Germany
[c] Wellcome Trust Centre for Neuroimaging, University College London, London, WC1N 3BG, United Kingdom

ARTICLE INFO

Keywords:
fMRI
DCM
Variational Bayes
Translational neuromodeling
Computational psychiatry
Spectrum diseases
Clustering

ABSTRACT

A recently introduced hierarchical generative model unified the inference of effective connectivity in individual subjects and the unsupervised identification of subgroups defined by connectivity patterns. This hierarchical unsupervised generative embedding (HUGE) approach combined a hierarchical formulation of dynamic causal modelling (DCM) for fMRI with Gaussian mixture models and relied on Markov chain Monte Carlo (MCMC) sampling for inference. While well suited for the inversion of complex hierarchical models, MCMC-based sampling suffers from a computational burden that is prohibitive for many applications.

To address this problem, this paper derives an efficient variational Bayesian (VB) inversion scheme for HUGE that simultaneously provides approximations to the posterior distribution over model parameters and to the log model evidence. The face validity of the VB scheme was tested using two synthetic fMRI datasets with known ground truth. Additionally, an empirical fMRI dataset of stroke patients and healthy controls was used to evaluate the practical utility of the method in application to real-world problems.

Our analyses demonstrate good performance of our VB scheme, with a marked speed-up of model inversion by two orders of magnitude compared to MCMC, while maintaining a similar level of accuracy. Notably, additional acceleration would be possible if parallel computing techniques were applied. Generally, our VB implementation of HUGE is fast enough to support multi-start procedures for whole-group analyses, a useful strategy to ameliorate problems with local extrema. HUGE thus represents a potentially useful practical solution for an important problem in clinical neuromodeling and computational psychiatry, i.e., the unsupervised detection of subgroups in heterogeneous populations that are defined by effective connectivity.

Introduction

Generative models of neuroimaging (Friston et al., 2003; Harrison et al., 2015; Havlicek et al., 2017; Hinne et al., 2014; Langs et al., 2014) or behavioral (Behrens et al., 2007; Friston et al., 2017; Mathys et al., 2014) data have become important pillars of computational and cognitive neuroscience. This type of analysis has the advantage of inferring putative mechanisms underlying neurophysiological and cognitive processes from neuroimaging and behavioral measurements. Such mechanistic accounts are not only highly beneficial for understanding the healthy human brain (for examples, see Piray et al., 2017; Rae et al., 2015; van Leeuwen et al., 2011) but also for identifying possible disease mechanisms in psychiatric disorders and for guiding differential diagnosis in individual patients (for review, see Stephan et al., 2017). Ultimately, this might help to overcome the lack of predictive validity of current symptom-based diagnostic schemes (DSM-5 or ICD-11) of psychiatric disorders (Stephan et al., 2015).

An important intermediate task relates to the stratification of clinical spectrum disorders (e.g., schizophrenia) into physiologically more homogenous subgroups. A neuromodeling strategy to addressing this challenge is offered by generative embedding (Brodersen et al., 2014; Brodersen et al., 2011). Classically, generative embedding is a two-step procedure: first, generative models of measured data are inverted to infer the hidden (latent) parameter values of a system (e.g., neuronal circuit) of interest (Bishop, 2006). For example, for neuroimaging data, the most frequently used generative model is dynamic causal modelling

* Corresponding author. Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich & ETH Zurich, Wilfriedstrasse 6, CH 8032, Zurich, Switzerland.
E-mail address: yao@biomed.ee.ethz.ch (Y. Yao).

https://doi.org/10.1016/j.neuroimage.2018.06.073
Received 3 January 2018; Received in revised form 24 May 2018; Accepted 27 June 2018
Available online 28 June 2018

(DCM) (Friston et al., 2003), which describes neuronal dynamics as a function of the effective (directed) connectivity among neuronal populations. In a second step, summary statistics of posterior parameter distributions can be extracted. These represent a compact summary of the mechanisms that generated the data – effectively, a model-based dimensionality reduction – and serve as features for subsequent supervised (Brodersen et al., 2011) or unsupervised (Brodersen et al., 2014) learning at the second (between-subject or group) level. In the supervised case, the goal is to predict a clinically relevant outcome variable, such as treatment response or future symptom score, from the inferred mechanisms that are embodied by the model parameter estimates (for examples, see Brodersen et al., 2014; Harlé et al., 2015; Lomakina et al., 2015; Wiecki et al., 2015). By contrast, in the unsupervised case, the aim is to detect mechanistically distinct subgroups within heterogeneous diseases (Brodersen et al., 2014). Generative embedding frequently results in more accurate classification/regression/clustering than conventional (un)supervised learning. As importantly, it allows for a meaningful interpretation of the results. This is because the achieved accuracy or purity can now be understood in relation to latent mechanisms that are encoded by the model's parameters.

Instead of inverting the generative model for each subject separately and then applying group-level learning to first-level posterior estimates, an alternative is to construct a fully hierarchical generative model that simultaneously describes individual data generation and assigns subjects to groups or clusters. This has the advantage that model inversion at the single-subject level can be informed by group-level results. More specifically, this corresponds to an empirical Bayesian approach that allows the model to learn the prior distribution from the data (Banerjee et al., 2015). In this paper, we focus on the unsupervised variant of this approach; we refer to this as hierarchical unsupervised generative embedding (HUGE).

Raman et al. (2016) introduced HUGE to neuroimaging, integrating a hierarchical formulation of DCM with a mixture of Gaussians clustering model (Bishop, 2006). Notably, Raman et al. (2016) used Markov chain Monte Carlo (MCMC) sampling for the inversion of this hierarchical model. While MCMC-based sampling is asymptotically exact (i.e., in the limit of infinite samples), it also suffers from a number of practical limitations. Most importantly, MCMC is computationally very costly, requiring run times for complex models that can be prohibitively long for many applications. This problem is particularly acute when, as in DCM, the likelihood of the generative model rests on differential equations that require integration for each sampling step (but see Aponte et al., 2016). Additionally, failure of convergence of the MCMC chains can be difficult to detect. Finally, it is challenging to obtain accurate and robust estimates of the (log) model evidence in MCMC (Calderhead et al., 2009; Gelman et al., 1998; Raftery et al., 2007).

In this article, we present an approximate inference scheme for HUGE that is based on variational Bayes (VB) (Attias, 1999). Casting model inversion in a VB framework promises increased computational efficiency and may render hierarchical generative embedding a practical tool for clinically relevant applications, such as the stratification of spectrum disorders. The structure of this article is as follows: after an introduction to classical DCM and the hierarchical extension by Raman et al. (2016), we proceed to deriving the variational update equations for our hierarchical model. We then test our VB scheme using both synthetic and empirical data. First, we provide an example of how a "standard" empirical Bayesian analysis of DCM (without clustering) can be implemented using HUGE. Second, we demonstrate face validity of the VB scheme using two synthetic fMRI datasets for which ground truth was known. Finally, we proceed to an empirical dataset from an fMRI experiment on speech recognition in aphasic patients and healthy controls (Schofield et al., 2012), which has not been analyzed with HUGE to date but was used in the original work on generative embedding (Brodersen et al., 2011). Here, we return to this dataset and evaluate the performance of HUGE. Based on these results, we discuss the advantages and disadvantages of our variational inversion compared to the MCMC-based approach from Raman et al. (2016).

Methods

Classical single-subject DCM

Dynamic causal modelling (DCM) is a generative modelling framework for estimating effective (directed) connectivity between neuronal populations from neuroimaging data (Friston et al., 2003). Originally developed for functional magnetic resonance imaging (fMRI) data, DCM has since been adapted to other modalities as well, including magneto-/electroencephalography (M/EEG) (David et al., 2006). DCM represents a generative model where neuroimaging data y is generated from hidden (latent) neuronal activity x. Here, the dynamics of x are a function of the effective connectivity between neuronal populations and some experimental manipulations u (e.g., sensory stimulation, task demands) to the network. Typically, a system of bilinear differential equations is used to model the dynamics of neuronal activity (for its derivation, see Stephan et al., 2008):

$$\dot{x} = Ax + \sum_{l=1}^{L} u_l B^{(l)} x + Cu \tag{1}$$

The bilinear formulation of neuronal dynamics in Eq. (1) allows for a straightforward interpretation of its parameters. Specifically, the elements of matrix A can be interpreted as the endogenous (intrinsic) connectivity between regions. Elements of C represent the strengths of the direct influence of experimental manipulations u on neuronal activity (i.e., external driving influences). The elements of the matrix B^l can be interpreted as the strengths of modulatory effects on the endogenous connectivity A by the l-th experimental manipulation u_l. Graphical representations of DCMs are shown in section 2.4. The matrices A, $B = \{B^{(l)}\}_{l=1}^{L}$ and C are typically of primary interest in DCM analyses and will be referred to as (neuronal) connectivity parameters $\theta^{(c)} = \{A, B, C\}$.

The neuronal model is then coupled to a forward observation model that maps hidden neuronal dynamics x to measured data y. For fMRI data, this forward model includes a cascade of differential equations describing how changes in neuronal dynamics induce changes in blood volume and deoxyhemoglobin content (Friston et al., 2000). The latter then enter a nonlinear static observation equation of regional blood oxygen level dependent (BOLD) signals (Buxton et al., 1998; Stephan et al., 2007). This observation model depends on a second set of parameters, commonly called hemodynamic parameters $\theta^{(h)}$. The full set of DCM parameters is defined as the concatenation of connectivity and hemodynamic parameters $\theta = \{\theta^{(c)}, \theta^{(h)}\}$.

Hence, in order to predict the BOLD response y from experimental inputs u and parameters θ using DCM, one has to integrate the dynamics of neuronal activity (Eq. (1)) to obtain x, given the connectivity parameters $\theta^{(c)}$and inputs u. The neuronal activity x and the hemodynamic parameters $\theta^{(h)}$ then enter the forward observation model. Consequently, the data generating process of DCM, which we will call $g(u, \theta)$ in the following, does not have a closed form.

In addition, DCM assumes a Gaussian noise model

$$\eta \sim N\left(0, \Lambda^{-1}\right) \tag{2}$$

where the noise precision Λ is shaped by hyperparameters (Friston et al., 2003): Λ is assumed to have diagonal structure, with region-specific noise precision λ_r on the diagonal (Raman et al., 2016). In combination, the data generating process $g(u, \theta)$ and the noise model yield a probabilistic forward mapping from experimental inputs to measured fMRI data and thus specify a likelihood function:

$$p(y|\theta) = N\left(g(u, \theta), \Lambda^{-1}\right) \tag{3}$$

In order to estimate neuronal connectivity and hemodynamic

parameters, a multivariate Gaussian prior distribution is placed over these parameters. Together, likelihood and prior yield a generative model that can be inverted using standard Bayesian inference techniques, such as variational Bayes (VB, Friston et al., 2007).

In brief, VB for DCM provides an estimate of two quantities simultaneously: (i) an approximation to the true posterior density over model parameters, and (ii) the negative free energy, which serves as a lower-bound approximation to the log model evidence (i.e., the log probability of the data given the model). The model evidence serves as a principled measure of model "goodness", taking into account both the accuracy and complexity of a model. Within a Bayesian setting, the model evidence allows one to test competing hypothesis about network architecture (corresponding to DCMs with different A, B and C matrices) or different hemodynamic models by means of Bayesian model selection (Friston et al., 2007; Penny et al., 2004; Stephan et al., 2009, Stephan et al., 2007).

A more detailed introduction to DCM, as well as model inversion and model comparison techniques, can be found elsewhere (Daunizeau et al., 2011; Friston et al., 2003; Friston et al., 2007; Stephan et al., 2007).

Hierarchical unsupervised generative embedding (HUGE)

This section summarizes the HUGE framework introduced by Raman et al. (2016). HUGE combines a hierarchical formulation of DCM with a Gaussian mixture model in order to unify the two-step procedure of generative embedding (Brodersen et al., 2014; Brodersen et al., 2011) into the inversion of a single (hierarchical) model. Additionally, the hierarchical framework of the model allows for an empirical Bayesian approach, where single-subject analyses are informed by group-level results.

Unlike single-subject DCM, where fixed Gaussian prior distributions are placed over the parameters and hyperparameters of the model (Friston et al., 2007), the hierarchical DCM in HUGE assumes that each individual from a population of N subjects belongs to one of K subgroups or clusters. The DCM connectivity parameters $\theta^{(c)} = \{A, B, C\}$ for all subjects from one cluster are assumed to be normally distributed, where each cluster k has a distinct mean μ_k and covariance matrix Σ_k:

$$p\left(\theta_n^{(c)} \middle| d_n = k, \mu_k, \Sigma_k\right) = N\left(\theta_n^{(c)} \middle| \mu_k, \Sigma_k\right) \tag{4}$$

Here, $\theta_n^{(c)}$ are the connectivity parameters of subject n. This cluster-specific normal distribution effectively means that different priors apply over subjects, depending on which subgroup they belong to. The assignment indicator d_n assigns subject n to one of the K clusters and is modelled using a categorical distribution (i.e., the special case of the multinomial distribution for a single drawing):

$$p(d_n = k|\pi) = \mathrm{Cat}(k|\pi) = \pi_k \tag{5}$$

where probability π_k is the "weight" of cluster k and π is the probability vector consisting of all weights:

$$\pi = (\pi_1, \ldots, \pi_K)^T \text{ with } \sum_{k=1}^{K} \pi_k = 1 \tag{6}$$

For the hemodynamic parameters, the hierarchical model retains the fixed global Gaussian prior from the classical DCM formulation:

$$p\left(\theta_n^{(h)} \middle| \mu_h, \Sigma_h\right) = N\left(\theta_n^{(h)} \middle| \mu_h, \Sigma_h\right) \tag{7}$$

As in single-subject DCM, the measured BOLD data y is described by means of a probabilistic forward model that relates experimental inputs u and model parameters θ, via neuronal and hemodynamic states x, to observed fMRI time series:

$$y_n = g(u, \theta_n) + \eta_n \text{ with } \theta_n = \left\{\theta_n^{(c)}, \theta_n^{(h)}\right\} \tag{8}$$

The measurement noise is assumed to be additive white Gaussian noise with zero mean:

$$\begin{aligned} \eta_n &\sim N\left(0, \Lambda_n^{-1}\right) \\ \Lambda_n &= \sum_{r=1}^{R} \lambda_{n,r} Q_r \end{aligned} \tag{9}$$

where the noise precision is now also subject-dependent. In Eq. (9), Q_r is a diagonal region-indicator matrix, whose diagonal entries are one if they belong to region r and zero otherwise (Raman et al., 2016). In other words, the model allows for both subject and region-specific precisions of observation noise.

To obtain a full generative model, prior distributions over model parameters and hyperparameters have to be introduced. Eqs. (4) and (7) specify the form of these priors for neuronal and hemodynamic parameters, respectively. The values of the prior parameters used throughout this paper are specified in the Supplementary Material (section S5).

Concerning the prior over cluster weights π, we follow the original implementation of HUGE (Raman et al., 2016) in using a Dirichlet distribution:

$$\pi \sim \mathrm{D}(\pi|\alpha_0) \tag{10}$$

Here, the parameter α_0 is a vector of dimension K containing only positive elements $\alpha_{0,k}$. Furthermore, the prior for the cluster parameters is given by a normal-inverse-Wishart distribution:

$$\mu_k, \Sigma_k \sim \mathrm{NW}^{-1}(\mu_k, \Sigma_k | m_0, \tau_0, \nu_0, S_0) \tag{11}$$

where m_0 is the mean, τ_0 the precision, ν_0 the degrees of freedom and S_0 the scale matrix of the normal-inverse-Wishart distribution.

Finally, Raman et al. (2016) chose a log-normal distribution as the prior over noise precisions: $\lambda_{n,r} \sim \log N\left(\lambda_{n,r} \middle| \mu_0, \sigma_0\right)$. Here, we deviate from this choice by introducing a gamma prior distribution over noise precisions:

$$\lambda_{n,r} \sim \mathrm{Gam}(\lambda_{n,r}|a_0, b_0) \tag{12}$$

where a_0 and b_0 are the rate and inverse shape parameters, respectively. This is the only modification to the original specification and was motivated by the fact that gamma priors serve as conjugate priors on precision for a Gaussian likelihood (Bishop, 2006). This simplifies the derivation of the VB update equations for the posterior density below.

Under this choice of likelihood and priors, the joint probability distribution takes the following form:

$$\begin{aligned} &p\left(\left\{y_n, d_n, \theta_n, \Lambda_n\right\}_{n=1}^{N}, \left\{\mu_k, \Sigma_k\right\}_{k=1}^{K}, \pi \middle| m_0, \tau_0, \nu_0, S_0, a_0, b_0, \alpha_0\right) = \\ &\prod_{n=1}^{N} \left(\left(\mathrm{N}\left(y_n \middle| g(\theta_n, u), \Lambda_n^{-1}\right) \mathrm{N}\left(\theta_n^{(c)} \middle| \mu_{d_n}, \Sigma_{d_n}\right) \mathrm{Cat}(d_n|\pi) \right. \right. \\ &\left. \left. \cdot \mathrm{N}\left(\theta_n^{(h)} \middle| \mu_h, \Sigma_h\right) \prod_{r=1}^{R} \mathrm{Gam}(\lambda_{n,r}|a_0, b_0) \right) \right. \\ &\left. \cdot \prod_{k=1}^{K} \left(\mathrm{NW}^{-1}(\mu_k, \Sigma_k | m_0, \tau_0, \nu_0, S_0)\right) \mathrm{D}(\pi|\alpha_0) \right. \end{aligned} \tag{13}$$

which, except for the gamma distribution over $\lambda_{n,r}$, is identical to the joint distribution proposed in Raman et al. (2016). Fig. 1 shows the graphical model of HUGE, where filled circles indicate variables with fixed values (e.g., parameters of prior distributions or data).

For a comprehensive review of the mathematical details of the probability distributions introduced in this section, we refer to Gelman (2014).

In Raman et al. (2016), inversion of this hierarchical generative model used MCMC. In this paper, we propose a variational Bayesian approach to derive a computationally more efficient approximate inversion scheme.

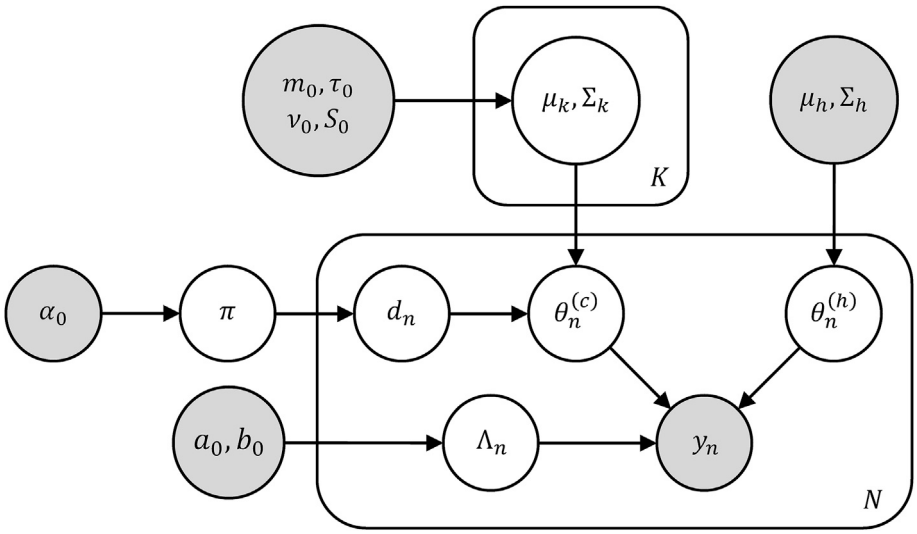

Fig. 1. Graphical model of HUGE. Filled nodes indicate variables with fixed or known values, such as parameters of prior distributions or data.

Variational inversion of HUGE

Having specified the generative model, we now present the variational update equations for HUGE. A general introduction to VB is provided in the Supplementary Material (section S2). In brief, VB attempts to fit an approximate posterior distribution $q(\vartheta)$ over latent variables ϑ by maximizing the negative free energy F (Friston et al., 2007). This implicitly minimizes the Kullback-Leibler divergence between approximate $q(\vartheta)$ and true $p(\vartheta|y)$ posterior distributions. To make the computation of the negative free energy tractable, the complexity of $q(\vartheta)$ can be restricted by means of the Laplace and mean field approximations (see Supplementary Material, section S2). In contrast to classical implementations of DCM, HUGE uses conjugate priors wherever possible, leading to analytical update equations. The derivations of these equations are presented in the Supplementary Material.

Applying the mean field approximation to HUGE, we assume a factorization of the approximate posterior $q(\vartheta)$ over the following disjoint subsets of model parameters: the DCM parameters $\Theta_1 = \{\theta_n\}_{n=1}^N$, the noise precisions $\Theta_2 = \{\Lambda_n\}_{n=1}^N$, the assignment indicators $\Theta_3 = \{d_n\}_{n=1}^N$, the cluster weights $\Theta_4 = \{\pi\}$ and the cluster parameters $\Theta_5 = \{\mu_k, \Sigma_k\}_{k=1}^K$. This leads to the following factorized distribution:

$$\begin{aligned} q\left(\{d_n, \theta_n, \Lambda_n\}_{n=1}^N, \{\mu_k, \Sigma_k\}_{k=1}^K, \pi\right) &= q(\pi) q\left(\{\theta_n\}_{n=1}^N\right) q\left(\{\Lambda_n\}_{n=1}^N\right) q\left(\{d_n\}_{n=1}^N\right) \\ q\left(\{\mu_k, \Sigma_k\}_{k=1}^K\right) &= \prod_{j=1}^5 q(\Theta_j) \end{aligned} \tag{14}$$

Some factors in Eq. (14) will further decompose into products of independent distributions, due to the inherent structure of the model (for details, see Supplementary Material, section S3).

Additionally, we will apply the Laplace approximation (Friston et al., 2007) to the variational distribution over DCM parameters (or, in the case of some hemodynamic parameters, their logs) and thus restrict its parametric form to a normal distribution: $q(\theta_n) := \mathrm{N}(\theta_n|\mu_n, \Sigma_n)$.

As outlined in the Supplementary Material (section S2), the optimal approximate posterior densities $q_j^*(\Theta_j)$ that maximize the negative free energy with respect to the j-th subset of model parameters can be found according to Eq. (8) of the Supplementary Material. This yields one update equation per subset, with update equations for different subsets being mutually dependent on each other. For optimization, we thus iterate these updates until convergence; this tightens the negative free energy bound on the log model evidence and renders the approximate distribution an optimal proxy to the true posterior distribution (Bishop, 2006). In the following, we present the update equations for each subset of parameters. The detailed derivation of these equations is provided in the Supplementary Material (section S3).

Cluster weights (π)

According to the derivation given in the Supplementary Material section S3.1.1, the optimal variational density over the cluster weights $q^*(\pi)$ is a Dirichlet distribution with parameters:

$$\alpha_k = \alpha_{0,k} + \sum_{n=1}^N q_{nk} - 1 \tag{15}$$

The variable q_{nk} in Eq. (15) denotes the probability that subject n belongs to cluster k. The expression for q_{nk} is given below in Eq. (18).

Cluster mean and covariance (μ_k and Σ_k)

In section S3.1.2 of the Supplementary Material, we show that the optimal variational distributions over cluster mean and covariance factorizes over clusters $q^*(\{\mu_k, \Sigma_k\}_{k=1}^K) = \prod_{k=1}^K q^*(\mu_k, \Sigma_k)$. Due to conjugacy of the prior (compare section 2.2), each factor $q^*(\mu_k, \Sigma_k)$ is given by a normal-inverse-Wishart distribution with parameters:

$$\begin{aligned} m_k &= \frac{q_k \mu_k^{(c)} + \tau_0 m_0}{q_k + \tau_0} \\ \tau_k &= q_k + \tau_0 \\ \nu_k &= q_k + \nu_0 \\ S_k &= \Sigma_k^{(c)} + \sum_{n=1}^N q_{nk} \left(\mu_n^{(c)} - \mu_k^{(c)}\right)\left(\mu_n^{(c)} - \mu_k^{(c)}\right)^T \\ &\quad + \frac{q_k \tau_0}{q_k + \tau_0} \left(\mu_k^{(c)} - m_0\right)\left(\mu_k^{(c)} - m_0\right)^T + S_0, \end{aligned} \tag{16}$$

where we have defined the following auxiliary variables:

$$\begin{aligned} q_k &= \sum_{n=1}^N q_{nk} \\ \mu_k^{(c)} &= \frac{1}{q_k} \sum_{n=1}^N q_{nk} \mu_n^{(c)} \\ \Sigma_k^{(c)} &= \sum_{n=1}^N q_{nk} \Sigma_n^{(c)}. \end{aligned} \tag{17}$$

The vector $\mu_n^{(c)}$ denotes the sub-vector of the variational mean μ_n of the model parameters θ_n associated with the DCM connectivity parameters. Equivalently, $\Sigma_n^{(c)}$ is defined as the sub-matrix of the entire covariance matrix Σ_n which is associated with the DCM connectivity parameters (see Eq. (19)).

Cluster assignments (d_n)

The approximate posterior probability of subject n belonging to cluster k is given by:

$$\begin{aligned}\log q^*(d_n = k) &= -\frac{1}{2}\log|S_k| + \frac{1}{2}\Psi_{p_c}(\nu_k) - \frac{p_c}{2\tau_k} - \frac{\nu_k}{2}\mathrm{tr}\left(S_k^{-1}\Sigma_n^{(c)}\right) \\ &\quad -\frac{\nu_k}{2}\left(\mu_n^{(c)} - m_k\right)^T S_k^{-1}\left(\mu_n^{(c)} - m_k\right) + \Psi(\alpha_k) + const \\ &=: \log q_{nk}\end{aligned} \tag{18}$$

(see Supplementary Material section S3.1.3 for detailed derivation). Notably, Eq. (18) determines q_{nk} only up to a constant factor. However, since q_{nk} defines a distribution over the categorical variable d_n, the sum over all possible values of k has to equal one. Hence, the unknown scaling factor can be determined via normalization. Here, $\mathrm{tr}(X)$ denotes the trace operation (i.e. the sum over all diagonal elements of a matrix) and $\Psi(x)$ the digamma function (i.e. the derivative of the logarithm of the gamma function $\Gamma(x)$, see also (Abramowitz et al., 1972) or Eq. (1) of the Supplementary Material). Furthermore, $\Psi_p(x)$ denotes the expression defined in Eq. (2) of the Supplementary Material.

DCM parameters (θ_n)

Similar to the cluster assignments, section S3.1.4 of the Supplementary Material shows that an optimal approximate density over the DCM parameters $q^*(\{\theta_n\}_{n=1}^N)$ factors over subjects. Due to the Laplace approximation, the factors $q^*(\theta_n)$ are normally distributed with mean and covariance given by:

$$\begin{aligned}\Sigma_n &= \left(G_n^T\overline{\Lambda}_n G_n + \Lambda'_n\right)^{-1} \\ \mu_n &= \Sigma_n\left(G_n^T\overline{\Lambda}_n(\varepsilon_n + G_n\theta_0) + \mu'_n\right)\end{aligned} \tag{19}$$

Note that we have defined the following auxiliary variables:

$$\begin{aligned}\Lambda'_n &= \begin{pmatrix}\sum_{k=1}^K q_{nk}\nu_k S_k^{-1} & 0 \\ 0 & \Sigma_h^{-1}\end{pmatrix} \\ \mu'_n &= \begin{pmatrix}\sum_{k=1}^K q_{nk}\nu_k S_k^{-1} m_k \\ \Sigma_h^{-1}\mu_h\end{pmatrix} \\ \varepsilon_n &= y_n - g(\theta_0, u) \\ G_n &= \left.\frac{\partial g(\theta,u)}{\partial\theta}\right|_{\theta=\theta_0}\end{aligned} \tag{20}$$

where θ_0 denotes the current expansion point of a Taylor approximation to the data generating process $g(\theta_n, u)$ (see Eq. (3)), which is typically chosen as the mean μ_n from the last iteration of the variational update scheme. G_n is the Jacobian matrix of $g(\theta_n, u)$ with respect to θ and ε_n can be interpreted as the current prediction error of the model for subject n. Additionally, the matrix $\overline{\Lambda}_n$ denotes the mean noise precision, i.e. the mean of Λ_n under the variational distribution, for which an expression is given below in Eq. (23).

Noise precision (Λ_n)

As defined in Eq. (9), the subject-specific noise precision matrix Λ_n is parameterized in terms of its region-specific diagonal elements $\lambda_{n,r}$ and a set of region indicator matrices Q_r. It is shown in section S3.1.5 of the Supplementary Material that the approximate posterior over noise precisions factorizes into a product over regions and subjects $q^*(\{\Lambda_n\}_{n=1}^N) = \prod_{r=1}^R\prod_{n=1}^N q^*(\lambda_{n,r})$, where $q^*(\lambda_{n,r})$ is given by a Gamma distribution with parameters:

$$\begin{aligned}a_{n,r} &= a_0 + \frac{\mathrm{tr}(Q_r)}{2} \\ b_{n,r} &= b_0 + \frac{b'_{n,r}}{2}\end{aligned} \tag{21}$$

Here, $\mathrm{tr}(Q_r)$ is the number of ones in Q_r (or, in other words, the number of data points per brain region). Furthermore, using ε_n and G_n defined in Eq. (20), we have introduced the auxiliary variable:

$$b'_{n,r} = \varepsilon_n^T Q_r \varepsilon_n + \mathrm{tr}\left(G_n^T Q_r G_n \Sigma_n\right) \tag{22}$$

Additionally, we can now define the mean noise precision matrix mentioned in the last section:

$$\begin{aligned}\overline{\Lambda}_n &= \sum_{r=1}^R \overline{\lambda}_{n,r} Q_r \\ \overline{\lambda}_{n,r} &= \frac{a_{n,r}}{b_{n,r}}\end{aligned} \tag{23}$$

where the second line follows from the mean of the Gamma distribution.

Negative free energy (F)

The negative free energy for HUGE is derived by solving the general expression in Eq. (58) of the Supplementary Material for the joint distribution from Eq. (13) and the variational distribution from Eq. (14) (see section S4 of the Supplementary Material for detailed derivation). The resulting expression after extensive simplification is given by:

$$\begin{aligned}F &= \log\Gamma\left(\sum_{k=1}^K \alpha_{0,k}\right) - \log\Gamma\left(\sum_{k=1}^K \alpha_k\right) - \sum_{k=1}^K \log\Gamma(\alpha_{0,k}) \\ &\quad -K\log\Gamma_{p_c}(\nu_0) + \frac{K\nu_0}{2}\log|S_0| + \frac{Kp_c}{2}\log\tau_0 - \frac{N}{2}\log|\Sigma_h| \\ &\quad +NR(a_0\log b_0 - \log\Gamma(a_0)) + \frac{Np}{2} + \frac{Np_c}{2}\log 2 - \frac{Np_y}{2}\log 2\pi \\ &\quad + \sum_{n=1}^N\sum_{r=1}^R(\log\Gamma(a_{n,r}) - a_{n,r}\log b_{n,r}) \\ &\quad + \sum_{k=1}^K \log\Gamma(\alpha_k) - \sum_{n=1}^N\sum_{k=1}^K q_{nk}\log q_{nk} \\ &\quad + \sum_{k=1}^K\left(\log\Gamma_{p_c}(\nu_0) - \frac{\nu_k}{2}\log|S_k| - \frac{p_c}{2}\log\tau_k\right) \\ &\quad + \frac{1}{2}\sum_{n=1}^N\left(\log|\Sigma_n| - \mathrm{tr}\left(\Sigma_h^{-1}\Sigma_n^{(h)}\right) - \left(\mu_n^{(h)} - \mu_h\right)^T\Sigma_h^{-1}\left(\mu_n^{(h)} - \mu_h\right)\right)\end{aligned} \tag{24}$$

Variational update schedule

As mentioned before, the update equations for the different parameters of the variational distribution $q(\Theta_j)$ are mutually dependent on each other. It is therefore necessary to iterate their updates until convergence to obtain the optimal parameters of $q(\Theta_j)$ that maximize the negative free energy. We now briefly outline the general procedure: first, initial values for all parameters are chosen, for instance, initial parameter values could be set to the prior parameters. Next, the update equation for one set of parameters is evaluated using the current estimates of all other parameters. This procedure is successively repeated for each set of parameters until all parameters have been updated. The negative free energy is then evaluated given the new estimates of all parameters. If the negative free energy has increased by more than a pre-set threshold (10^{-10} in the current implementation) compared to the previous iteration, the update procedure is continued for another iteration; otherwise the algorithm has converged. This process is illustrated as a flowchart in Fig. 2. We implemented the variational update equations for HUGE in Matlab. The

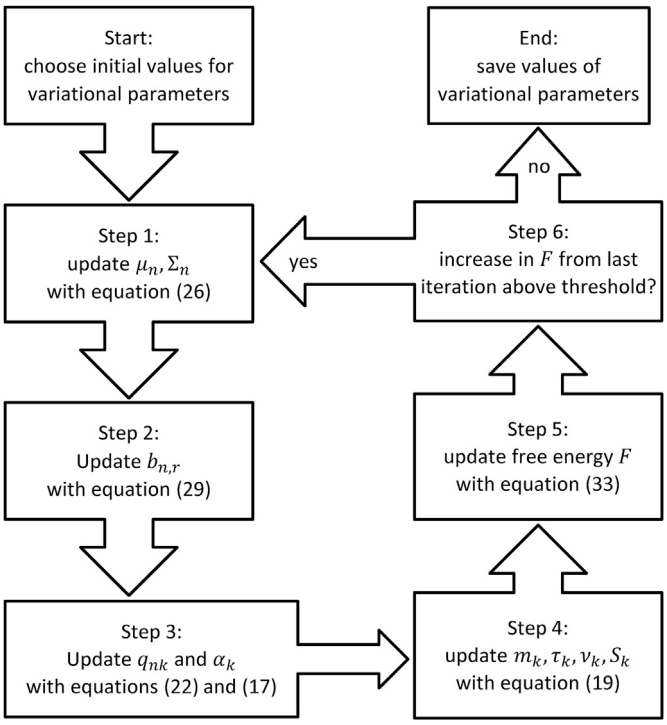

Fig. 2. Flowchart of a possible variational update schedule for the parameters of the variational distribution. Here, "update" refers to an update based on the current value of the other variational parameters.

numerical integration required for evaluating the observation function $g(\theta, u)$ is executed using the same implementation as in Raman et al. (2016), which rests on Euler's method implemented in C for increased computational efficiency. The code for our VB approach to HUGE introduced in this paper will be made available as part of the open source toolbox TAPAS (http://www.translationalneuromodeling.org/tapas).

Datasets

Synthetic datasets

We assessed the face validity of our variational inversion scheme for HUGE using two synthetic and one empirical dataset. The synthetic datasets were based on a two-region linear DCM and a three-region bilinear DCM (Fig. 3), following the same procedures as in Raman et al. (2016). For all simulations, we verified that the chosen parameter values resulted in a stable system by checking that the principal eigenvalue of the coupling matrix was negative.

The two-region DCM used as the basis of our first synthetic dataset was a linear DCM with one driving input per region, one endogenous connection from region 2 to region 1 and inhibitory self-connections on both regions. In a first step, two subgroups were established by defining two different sets of parameters for the DCM; these served as the mean parameter vectors for the two subgroups. For each subgroup, 20 "synthetic subjects" were simulated by sampling the DCM parameter values of each subject independently from an isotropic normal distribution with standard deviation 0.1 centered on the mean DCM parameter vector of the respective subgroup. This process gave rise to a total of 40 synthetic subjects, clustered in two groups with 20 subjects each. For each of these subjects, the set of subject-specific DCM parameters was then used to generate BOLD signal time series with 300 scans per brain region and a repetition time (TR) of two seconds. Finally, white Gaussian measurement noise was added to the BOLD signal. The amplitude of the measurement noise was chosen such that in each region the standard deviation of the noise was equal to the standard deviation of the BOLD signal. This corresponds to a signal-to-noise ratio (SNR) of one and represents a relatively challenging scenario (Welvaert et al., 2013).

The second synthetic dataset is based on the more complex three-region bilinear DCM shown in Fig. 3. Similar to the first dataset, different sets of mean parameter values were used to establish subgroups among the synthetic subjects. However, unlike in Raman et al. (2016), this dataset consists of three subgroups: One subgroup of 40 and two subgroups of 20 subjects each, for a total of 80 subjects. Moreover, the mean parameter vectors of the two smaller subgroups differ only in three of the nine parameters of the DCM. A dataset like this might arise, for example, in a clinical study, where the patient cohort comprises mechanistically distinct subgroups which differ only in a subset of parameters. As with the previous dataset, BOLD signal time series were generated for each subject with 256 scans per brain region and a TR of two seconds. The method used to simulate synthetic subjects and generate measurement noise was the same as for the first dataset. The numerical values of the cluster mean parameters used to generate the two synthetic datasets, as well as the numerical values of the prior parameters used to invert HUGE for all datasets are provided in the Supplementary Material (section S5).

Empirical dataset

After testing the face validity of the variational inversion scheme using synthetic data, we also applied our hierarchical model and the VB inversion to an empirical dataset including stroke patients with aphasia and healthy controls (Schofield et al., 2012). We used this clinical dataset for two reasons: first, the original (supervised) generative embedding analysis of this dataset (Brodersen et al., 2011) sets a challenging benchmark; second, the working memory dataset on patients with schizophrenia used by Raman et al. (2016) is characterized by confounding variables of no interest (such as age and sex) that critically affect cluster solutions (see Brodersen et al., 2014). While accounting for confounding variables is easy to do in the original two-step approach to generative embedding, it is difficult in our hierarchical model since this would require re-deriving the update equations.

Subjects consisted of 26 healthy, right-handed subjects with English as their first language (twelve females; mean age 54.1 years; range 26–72

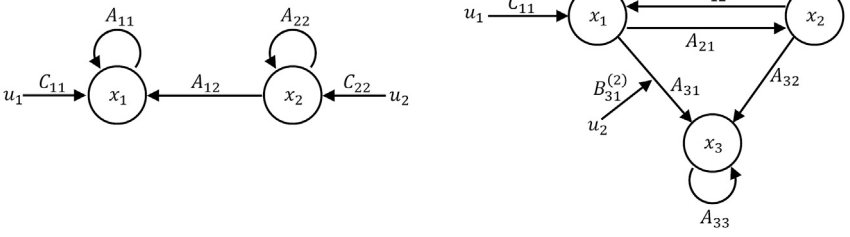

Fig. 3. Graphical representation of the linear two-region DCM used to generate the first synthetic dataset (left) and the bilinear three-region DCM used to generate the second synthetic dataset (right).

years) and eleven stroke patients (one female; mean age 66.1 years; range 45–90 years) with moderate aphasia. Participants were presented with two types of auditory stimuli (normal speech and time-reversed speech) and were asked to report the gender of the speaker for each stimulus. A 1.5 T MR scanner was used to acquire 488 vol (122 vol in four sessions) of functional images using a T2*-weighted echo-planar imaging (EPI) sequence sensitive to the BOLD contrast (in-plane resolution 3 mm × 3 mm; slice thickness 2 mm; inter-slice gap 1 mm; TR = 3.15 s, TE = 50 m s) for each subject. Details are provided by Schofield et al. (2012).

Based on the data by Schofield et al. (2012), a six-region DCM with 22 neuronal parameters was used in the original generative embedding paper to distinguish patients from controls with near-perfect accuracy (Brodersen et al., 2011). Achieving the same with HUGE represents a more challenging scenario. First, the original generative embedding analysis of these data in Brodersen et al. (2011) used a supervised classification method (i.e., support vector machine) with the neuronal connectivity parameter estimates as input features. Hence, the algorithm was aware of the true number of groups (in this case, moderately aphasic patients and healthy controls) and the true group assignment of each subject. In contrast, the HUGE approach discussed in this paper is an unsupervised method – that is, it neither knows the number of groups/clusters in the population nor the assignment of each subject.

Additionally, the DCM used in the original analysis (Brodersen et al., 2011) has 22 neuronal connectivity parameters, which means that the dimensionality of the feature space is relatively high compared to the number of data points (37 subjects). Estimating clusters in high-dimensional space from a limited number of samples is a hard problem (Bishop, 2006), and local extrema of the objective function may pose a serious challenge for local optimization schemes like VB.

Given these considerations, we simplified the DCM used in the analysis of Brodersen et al. (2011) to reduce the dimensionality of the feature space and thus allow for a more graceful performance of the VB scheme. Specifically, we excluded the medial geniculate body (MGB), which, in the original DCM, mainly served as a relay station for auditory input to Heschl's gyrus (HG). The resulting simplified DCM (Fig. 4) contains only four regions and sets the driving input directly to bilateral HG. This simplification reduced the number of neuronal connectivity from 22 to 14. Notably, however, clustering in 14-dimensional space is still a challenging task.

We therefore additionally restarted the VB inversion at random initial positions; this is a common method to reduce the influence of local optima (Bishop, 2006). For the VB scheme for HUGE, we randomize the initial values of the means of the approximate posterior over DCM parameters μ_n and clusters μ_k by setting these parameters to their prior value plus random fluctuations sampled from a Gaussian with a standard deviation of 0.1. We used 100 restarts, each time with a different initial position of the VB scheme. For each of these initial positions, we ran the VB scheme under three different settings for the number of clusters K (i.e., 1, 2, or 3). This resulted in 300 runs of the VB scheme from 100 different initial positions. For details concerning the multi-start approach, see section S8 of the Supplementary Material.

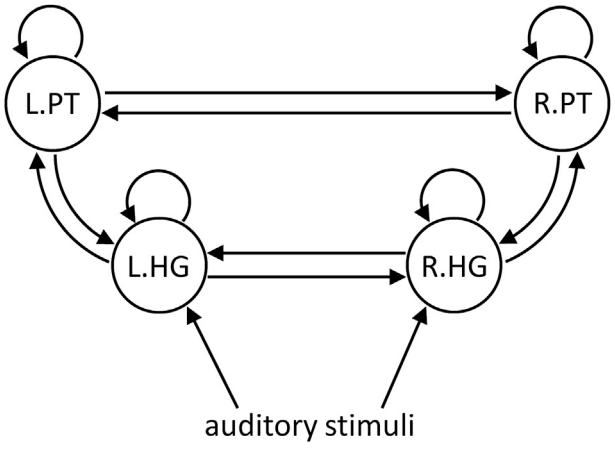

Fig. 4. Graphical representation of the DCM used as the basis for the HUGE analysis of the empirical fMRI dataset (Schofield et al., 2012).

Assessing clustering performance and model fit

Balanced purity

Since the main objective of HUGE is to search for clusters within a heterogeneous subject population, it is necessary to introduce a measure of goodness for clustering results, which allows for a quantitative assessment of the performance of HUGE and the VB inversion. For this purpose, we use the "balanced purity" criterion introduced by Brodersen et al. (2014), which measures the degree of agreement between the inferred cluster labels and the true labels. Balanced purity is a modification of the conventional "purity" criterion (Manning et al., 2009), which corrects for the confounding effects due to imbalanced datasets (i.e., clusters of different sizes). Given a clustering solution $\Omega = (\omega_1, \ldots, \omega_k, \ldots, \omega_K)$, where ω_k contains the indices of all subjects for which cluster k had the highest posterior probability, and the set of true class assignments $\Phi = (c_1, \ldots, c_k, \ldots, c_K)$, the balanced purity is defined as:

$$\begin{aligned} \mathrm{bp}(\Omega, \Phi) &= \left(1 - \frac{1}{K}\right)\left(\frac{\mathrm{purity}(\Omega, \Phi) - \xi}{1 - \xi}\right) + \frac{1}{K} \\ \mathrm{purity}(\Omega, \Phi) &= \frac{1}{N}\sum_{k=1}^{K} \max_j \left|\omega_k \cap c_j\right| \end{aligned} \tag{25}$$

Here, K is the number of clusters, N the number of subjects and $\left|\omega_k \cap c_j\right|$ the number of subjects in cluster k with true label j. The number ξ denotes the degree of imbalance in the data, defined as the fraction of subjects associated with the largest class. The balanced purity is 1 for a perfect clustering result, where the inferred cluster label corresponds to the correct label for each subject. In contrast, if the clustering scheme assigns subjects at random, the balanced purity tends towards 1/K on average.

Bayes factors

In addition to the quality of the clustering result, assessing model fit is also of importance. Specifically, the number of clusters K is a free parameter in the current formulation of HUGE, which necessitates an additional model selection step to determine the value for K that best represents the acquired data.

Fortunately, the negative free energy, which our VB implementation of HUGE provides for free, represents a lower bound approximation to the log-model evidence and thus serves as a principled measure of model fit in form of Bayes factors. These are defined as the ratio between the model evidence of two competing models (e.g., $K = i$ versus $K = j$):

$$B_{ij} = \frac{p(y|m = i)}{p(y|m = j)} \tag{26}$$

Heuristically, one can interpret the Bayes factor as the posterior odds ratio between models i and j for equal prior odds (Penny et al., 2004). Conventionally, a Bayes factor of 20 or higher (equivalent to a free energy difference >3) is considered as strong evidence for the superiority of one model compared to another (Kass et al., 1995).

Results

In this section, we present clustering results obtained with our VB inversion scheme for the synthetic and empirical datasets introduced in section 2.4. In addition, we demonstrate that the HUGE model can also be used to perform a "standard" empirical Bayesian DCM analysis (without subgroup detection) by assuming a single cluster.

Synthetic datasets

Simulations: demonstration of empirical Bayes

First, we demonstrate how to perform a "pure" empirical Bayesian analysis using HUGE and the associated VB scheme. Although designed as a clustering model, HUGE can be adapted to this task by forcing the number of clusters to one. This effectively switches of the clustering and the DCMs for all subjects are inverted while marginalizing out the Gaussian prior distribution.

For this demonstration, we apply HUGE to the first 40 subjects from the second synthetic dataset (see Methods) while fixing the number of clusters to one. The prior distribution was chosen such that the marginal prior distribution over DCM parameters for each subject corresponds approximately to the prior distribution over DCM parameter in SPM (SPM8 r6313, Penny et al., 2007). This choice should maximize the comparability of the HUGE results with those from single-subject model inversions in SPM. The numerical values of the prior parameters are provided in the Supplementary Material (section S5). For comparison, we additionally invert the DCMs for each subject individually using SPM (SPM8 r6313, Penny et al., 2007).

Fig. 5 shows the range of ground truth DCM parameter values, as well as the range of maximum a posteriori (MAP) estimates obtained with SPM for each subject individually and with empirical Bayes (i.e., HUGE with number of clusters fixed to one) in a hierarchical setting. The variability in MAP estimates obtained in a hierarchical setting is consistently smaller than for the individually obtained MAP estimates. The VB scheme, run in empirical Bayes configuration, converged within 33 iterations, corresponding to 1.5 min on a laptop computer (2.8 GHz, 16 GB RAM). Inverting the DCMs for each subject individually with SPM required about 10 min. In principle, the empirical Bayes analysis could also be carried out using the MCMC implementation from Raman et al. (2016) instead of VB, which, however would require significantly more computational resources.

Simulations: two-region linear DCM

For the first synthetic dataset based on the two-region linear DCM shown in Fig. 3, we tested whether the VB inversion method introduced in the previous section could recover both the data-generating parameter values for each subject, as well as the group structure (i.e., two subgroups) in an unsupervised fashion. In addition, we compared all results to those obtained with the MCMC implementation from Raman et al. (2016). For this purpose, we ran five independent chains and pooled the samples from all chains resulting in a total of 500,000 samples. Convergence was monitored with the potential scale reduction factor (PSRF) proposed by Gelman et al. (1992). The results are presented alongside those of VB (for details concerning the MCMC inversion, see Supplementary Material, section S7).

Both the VB and MCMC algorithms correctly identified the existence of two clusters and assigned the synthetic subjects to the correct cluster with high accuracy. Specifically, VB assigned only one subject (i.e., subject 26) to the wrong cluster (cluster 1 instead of the "true" cluster 2), which corresponds to a balanced purity of 97.5% (see Eq. (25) for the definition of balanced purity). In comparison, the MCMC inversion was able to assign all subjects correctly, although the posterior assignment probability for subject 26 was only 65%. Fig. 6 shows the posterior assignment probabilities estimated by VB and MCMC for $K = 2$.

Fig. 7 shows the estimated cluster mean parameter values obtained under VB and MCMC inversion, as well as the true mean parameter values. Both inversion schemes accurately recover most data-generating parameter values, with MCMC delivering slightly more accurate estimates. A notable exception is the A_{12} parameter of the second cluster, which neither VB nor MCMC could estimate reliably (Fig. 7). This is likely due to the structure of the underlying DCM. Specifically, the values of both A_{12} and the input strength C_{22} to region 2 are relatively small,

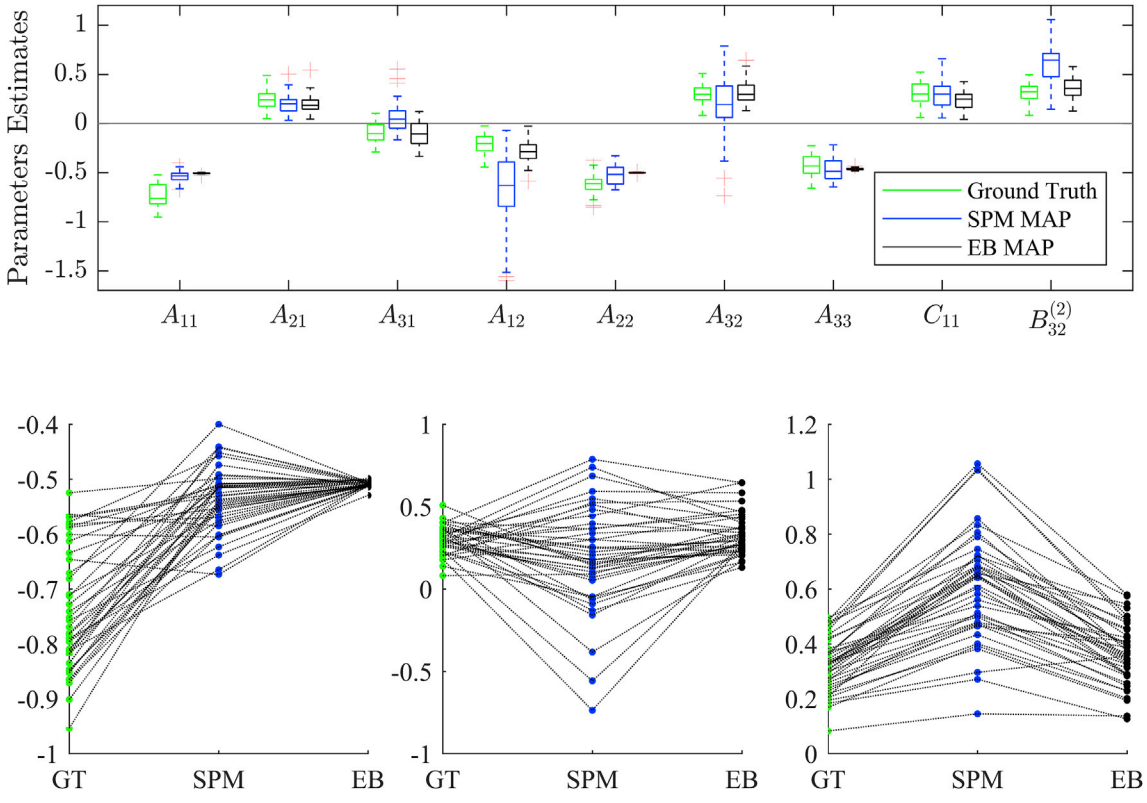

Fig. 5. The regularizing effect of HUGE on parameter estimation demonstrated for the first 40 subjects from the second synthetic dataset. Top panel: Range of ground truth parameter values (green), maximum a posteriori (MAP) estimates obtained for each subject individually with SPM (blue) and MAP estimates obtained with empirical Bayes, i.e. HUGE with K set to one (black). Bottom panels: Actual values of ground truth (green dots), SPM MAP estimates (blue dots) and empirical Bayes MAP estimates (black dots) for DCM parameters A_{11} (Bottom left), A_{32} (Bottom center) and $B_{32}^{(2)}$ (Bottom right).

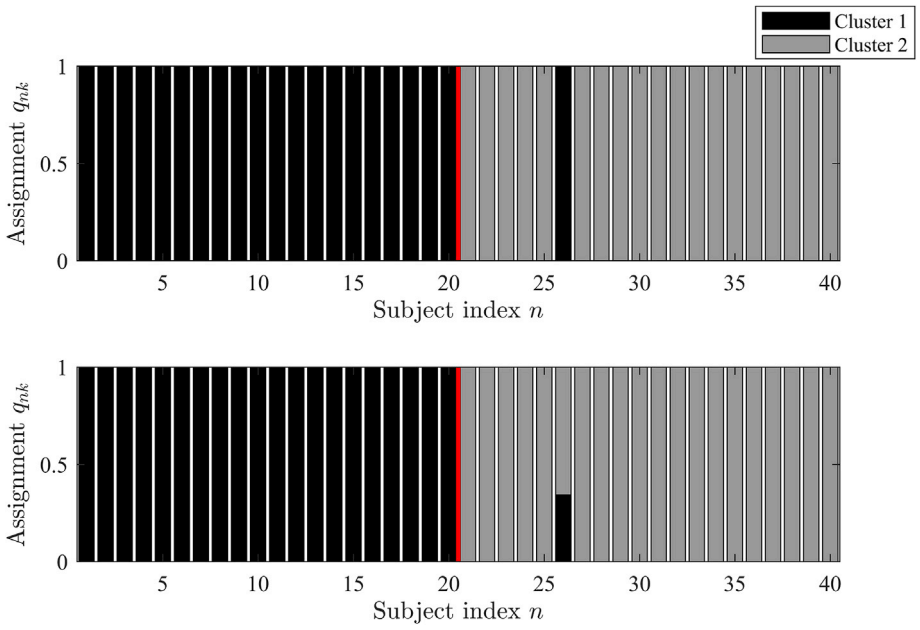

Fig. 6. Synthetic data from the two-region DCM: Estimated assignment probability of subjects to clusters for K = 2 obtained with VB (top panel) and MCMC (bottom panel). Red lines indicate correct assignments: subjects 1–20 – cluster 1 and subjects 21–40 – cluster 2. The balanced purity is 97.5% for VB and 100% for MCMC.

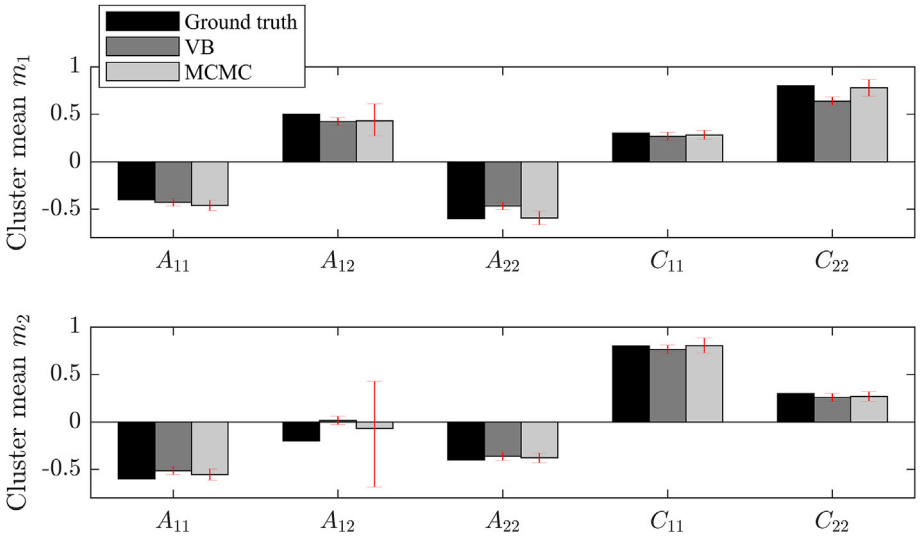

Fig. 7. Synthetic data from the two-region DCM: Cluster mean estimates for K = 2 with top panel showing cluster 1 and bottom panel cluster 2. True (data-generating) cluster means are shown in black, VB estimates in dark grey and MCMC estimates in light grey. Red error bars indicate marginal 95% credible intervals.

making the link from region 2 to region 1 weak and thus challenging to estimate. Interestingly, however, VB and MCMC handle this situation differently. Under the influence of the prior, the VB estimate of A_{12} reverts to the prior mean of 0.0078 (compare Supplementary Material, Table S2), although the size of the error bar indicates that VB is overconfident about this estimate. This is a known issue of the specific form of the KL-divergence used in the negative free energy approximation of VB (for details, see Bishop, 2006). On the other hand, MCMC seems to deliver the expected result, i.e., mean estimate between the prior and the true parameter, with large posterior variance. However, closer inspection reveals that the large size of the error bars is due to the five chains not converging properly for this particular parameter (for details, see Supplementary Material, section S7.1).

Next, we addressed the question regarding the optimal number of clusters, given the observed data. Since the number of clusters has to be pre-specified in the current formulation of HUGE, we repeated the VB inversion under various settings of the number of mixture components K. Specifically, here we tested $K = [1,2,3,4]$ and then compared the negative free energies for the different settings (Fig. 8). We observed that a model with two mixture components outperformed models with less or more components. This can be quantified using Bayes factors (see Eq. (26)), which for the current dataset are $B_{21} = 3.6 \times 10^8$, $B_{23} = 22.3$ and $B_{24} = 324.3$. In summary, using the negative free energy obtained with the VB inversion scheme, HUGE correctly detected that the data were generated from two distinct clusters, with relatively small computational overhead. Notably, the same analysis with MCMC inversion would require the use of thermodynamic integration (Calderhead et al., 2009) as the current gold standard for computing the model evidence, leading to prohibitive demands on computational resources.

A plot of the posterior parameters α_k of the Dirichlet distribution over the cluster weights for the two-, three- and four-component model reveals that for the models with $K > 2$ only two of their components make a

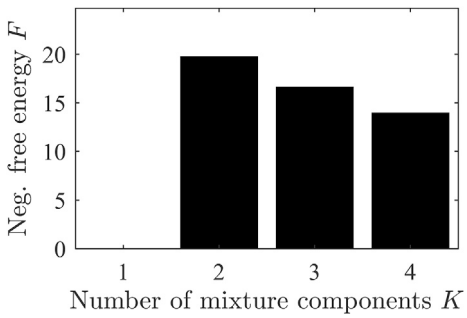

Fig. 8. Model comparison for two-region DCM (simulated data): Negative free energy differences (relative to the worst model) as a function of the number of mixture components in the model.

non-negligible contribution (Fig. 9). Hence, the additional flexibility of these models does not allow for a better explanation of the data. At the same time, the additional complexity of including these superfluous components imposes a penalty, which leads to the observed decrease in negative free energy for $K > 2$.

VB inversion for $K = 2$ required 254 iterations of the update equations, corresponding to 6.5 min on a laptop computer (2.8 GHz, 16 GB RAM). In contrast, the MCMC-based inversion for $K = 2$ on the same dataset using the same computer with the settings reported in Raman et al. (2016) – that is, 200,000 samples including 100,000 samples burn-in – required 5.5 h per chain. This corresponds roughly to a speed-up of two orders of magnitude by the VB scheme proposed in this paper.

Simulations: three-region bilinear DCM

The second synthetic dataset was based on the three-region bilinear DCM shown in Fig. 3. This dataset represents a more challenging scenario than the first dataset for the following reasons: (i) it includes bilinear (i.e., modulatory) effects, (ii) it includes three clusters of subjects, two of which differ only in a subset of parameters, and (iii) the number of DCM connectivity parameters per subject increased from five to nine. Generally, clustering becomes more difficult with increasing dimensionality of the feature space (Bishop, 2006). We applied both the VB and MCMC inversion schemes with $K = 3$ to the 80 sets of synthetic BOLD data and again found that most subjects were assigned to the correct clusters (Fig. 10). The high clustering accuracy is also reflected by the balanced purity of 98.3% (VB and MCMC). Note that as with the previous dataset, we ran five independent MCMC chains and pooled the samples from all chains. Convergence was monitored with the PSRF proposed by Gelman et al. (1992) (for details, see Supplementary Material S7.2).

Fig. 11 shows that for most of the parameters, the cluster means could be accurately recovered. Again, the MCMC estimate (derived from 500,000 samples) delivered slightly more accurate estimates than VB. This is most evident for parameters A_{31} and A_{32}, which is due to the intrinsic difficulty of disentangling the contribution of two potential causes (activity in regions 1 and 2) for a single observation (activity in region 3).

For this dataset, we again varied the number of mixture components and compared the negative free energy across these different settings. The result is consistent with our observations for the first (simpler) dataset and suggests that, even for more challenging scenarios (i.e., larger dimensionality of the feature space), our VB inversion scheme for HUGE is able to accurately detect the correct number of distinct clusters (Fig. 12). The free energy values shown in Fig. 12 correspond to Bayes factors of $B_{31} = 3.4 \times 10^{15}$, $B_{32} = 2.2 \times 10^{4}$ and $B_{34} = 26$. Comparing the computation times between VB-based and MCMC-based inversion schemes, VB inversion required 38 iterations of the update equations, corresponding to 3.5 min (for $K = 3$) on a laptop computer (2.8 GHz, 16 GB RAM), while the MCMC-based inversion required 13.8 h per chain on the same computer.

Interestingly, for $K = 2$, both VB and MCMC converged to a reasonable solution where subjects 1–40, which originated from the most distinct cluster, were assigned to one cluster and subjects 41–80, which originated from the other two more similar clusters, were assigned to the remaining cluster (result not shown).

Empirical fMRI dataset

Next, we applied our VB inversion scheme for HUGE to the empirical dataset described in section 2.4. As noted above, we used a multi-start approach and selected the result with the highest negative free energy for each setting of K. The setting with two clusters outperformed the other settings in terms of the negative free energy (Fig. 13). The Bayes factors between models with different number of clusters are $B_{21} = 2.2 \times 10^{8}$ and $B_{23} = 1.0 \times 10^{4}$. The assignment probabilities for the two-cluster case is shown in Fig. 13. The resulting balanced purity of 95.5% indicates excellent separation of aphasic patients and healthy controls. For details on the multi-start approach, see section S8 in the Supplementary Material.

Finally, we inspected the estimates of the cluster means for the maximum negative free energy solution (i.e., $K = 2$) in order to identify the parameters that were discriminative between healthy controls and aphasic patients (Fig. 13). From visual inspection, it appears that these parameters include particularly the self-connection of left HG, left PT to left HG, the self-connection of left PT, right HG to left HG, right HG to right PT, right PT to left PT, right PT to right HG, the input strength to right HG and to a lesser extend also left HG to left PT and left HG to right HG. Notably, this list includes the interhemispheric connections from right to left hemisphere and the connection from left HG to left PT. These parameters belong to the subset of discriminative features that were found to be sufficient to distinguish between patients and healthy controls in the original supervised generative embedding analysis by Brodersen et al. (2011).

As before, we inverted the dataset with the MCMC implementation

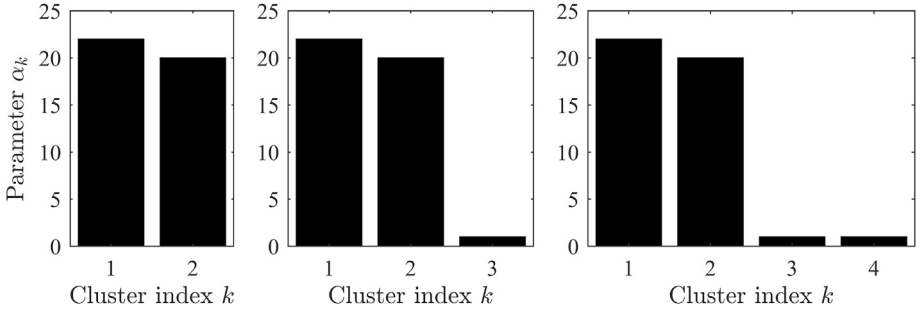

Fig. 9. Synthetic data from the two-region DCM: Parameters of the posterior distribution over cluster weights for models with K = 2 (left panel), K = 3 (center panel) and K = 4 (right panel) components. The parameter α_k corresponds approximately to the effective number of subjects assigned to that cluster. For the models with K > 2, no subjects were assigned to the clusters beyond two; however, α_k is non-zero due to the prior, which assigns pseudo-observations to all clusters.

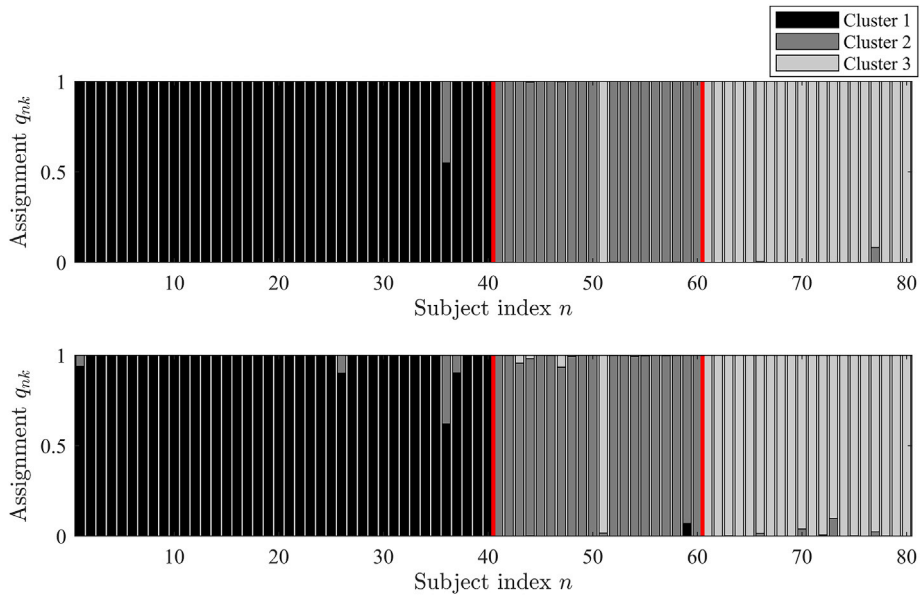

Fig. 10. Synthetic data from the three-region DCM: Estimated assignment probability of subjects to clusters for K = 3 obtained with VB (top panel) and MCMC (bottom panel). Red lines indicate correct assignments: subjects 1–40 – cluster 1, subjects 41–60 – cluster 2 and subjects 61–80 – cluster 3. The balanced purity is 98.3% for both VB and MCMC.

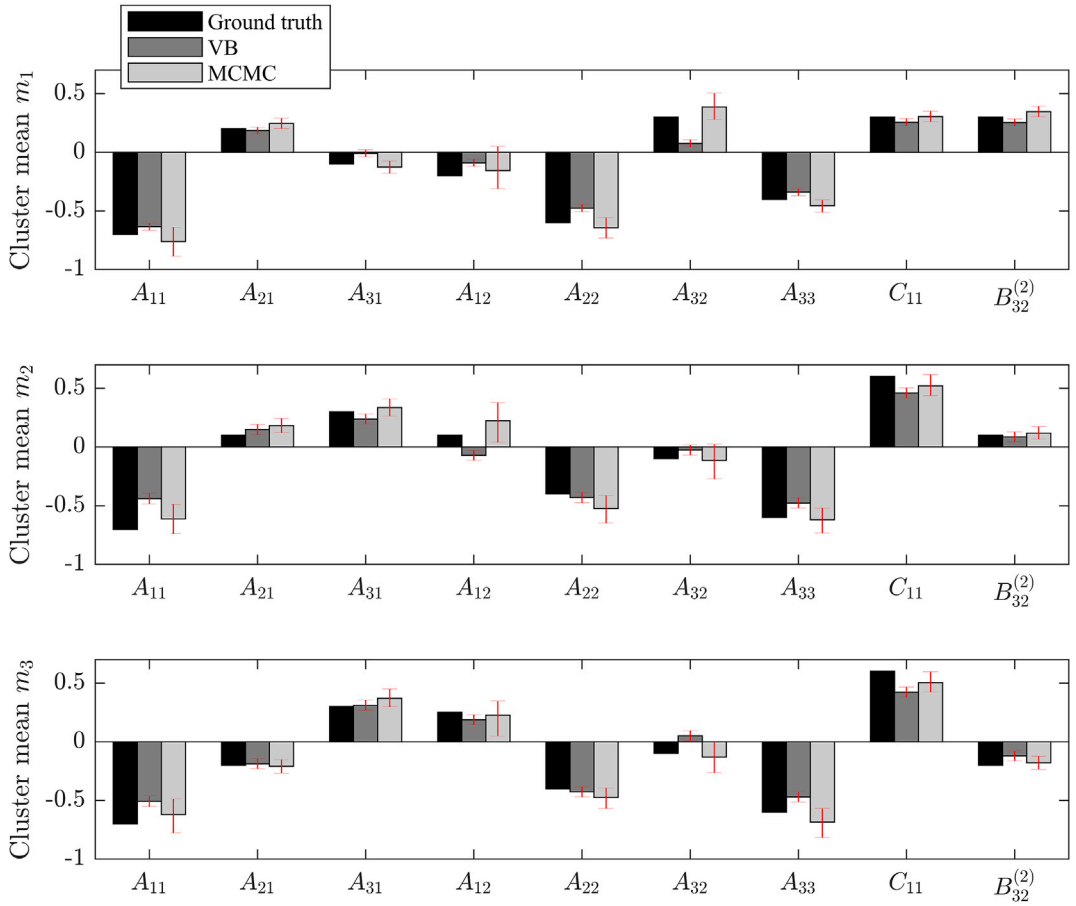

Fig. 11. Synthetic data from the three-region DCM: Cluster mean estimates for K = 3 with top panel showing cluster 1, middle panel cluster 2 and bottom panel cluster 3. True (data generating) cluster means are shown in black, VB estimates in dark grey and MCMC estimates in light grey. Red error bars indicate marginal 95% credible intervals.

from Raman et al. (2016). For this purpose, we ran four independent chains with $K = 2$ and 800,000 samples each (including 100,000 samples burn-in). Visual inspection of Fig. 14, which shows subject assignment and cluster mean estimates obtained by pooling the samples from all

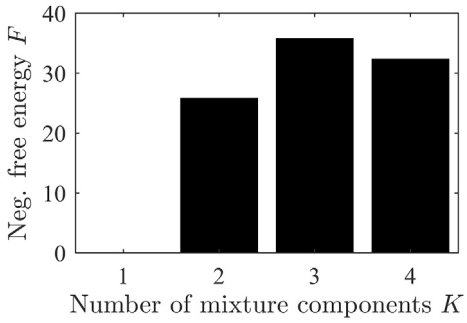

Fig. 12. Model comparison for three-region DCM (simulated data): Negative free energy differences (relative to the worst model) as a function of the number of mixture components in the model.

less than 220 iterations. This translates into an average computation time of about 30 minutes[1]. Repeating the VB inversion on a laptop computer (2.8 GHz, 16 GB RAM) for the starting positions that yielded the highest negative free energy for all three cases (K = [1,2,3]) resulted in computation times that were 9.7 min (73 iterations) for K = 1, 15 min (112 iterations) for K = 2 and 22.5 min (167 iterations) for K = 3. The increase in computation time compared to the synthetic datasets is due to the increased dimensionality of the feature space. Model inversion under the MCMC scheme (for $K = 2$) was performed on the same HCP cluster as the multi-start scheme for VB and required on average 52 h per chain.

Discussion

The approach described in this paper – hierarchical unsupervised generative embedding (HUGE) – unifies two important streams of

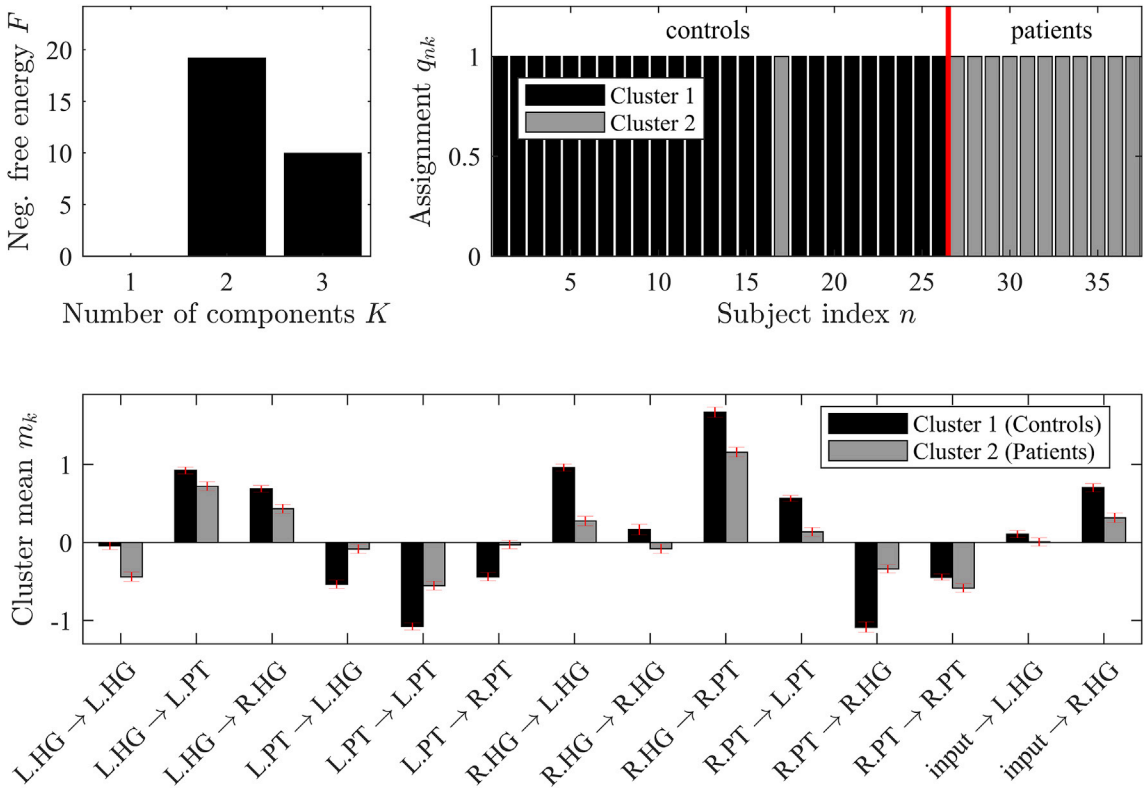

Fig. 13. VB results for the aphasia dataset: Top left: Negative free energy differences relative to the worst model. Values shown here represent the maximum negative free energy obtained for each of the settings K = [1,2,3] from the 100 restarts. Top right: Estimated assignment probability of subjects to clusters for K = 2 (balanced purity: 95.5%). Bottom panel: Estimated cluster means from the maximum negative free energy solution with K = 2. Red error bars indicate marginal 95% credible intervals.

chains, indicates a plausible result with balanced purity of 100%. Furthermore, most cluster mean estimates appear to be consistent with VB (Fig. 13). Notably, these include the self-connection of left HG and the connections from left HG to left PT and left PT to left HG, which seem to be highly discriminative between controls and patients in both MCMC and VB based analyses. On the other hand, the PSRF revealed that, despite being significantly longer, the different chains did not converge as consistently for this dataset as they did for the synthetic datasets, which is also the reason behind the relatively wide error bars in Fig. 14. Hence, the posterior estimates of the MCMC inversion should be interpreted with caution. A detailed discussion of this result is provided in section S7.3 in the Supplementary Material.

The 300 instances of the VB-based inversion of HUGE for the empirical dataset were performed on a computer cluster, which could run all inversions in parallel. One inversion required on average 143 iterations of the VB update equations with 90% of all inversions converging in

development in neuroimaging: (i) hierarchical models for empirical Bayesian analyses of multi-subject fMRI data (Friston et al., 2016; Lindquist et al., 2017; Sanyal et al., 2012), and (ii) combining generative models of single-subject fMRI data with (un)supervised learning for clinical predictions (Brodersen et al., 2014; Brodersen et al., 2011; Stephan et al., 2017). An early version of HUGE was based on computationally demanding MCMC sampling (Raman et al., 2016). In this paper, we derived a novel and efficient VB inversion scheme for hierarchical unsupervised generative embedding (HUGE), evaluated its face validity using simulations, and demonstrated its practical utility for empirical Bayesian analyses of DCM. Specifically, the results on the synthetic datasets indicate that VB is able to achieve an accuracy comparable to

[1] This number represents only a rough ballpark figure, since the processors of the cluster have different performance characteristics.

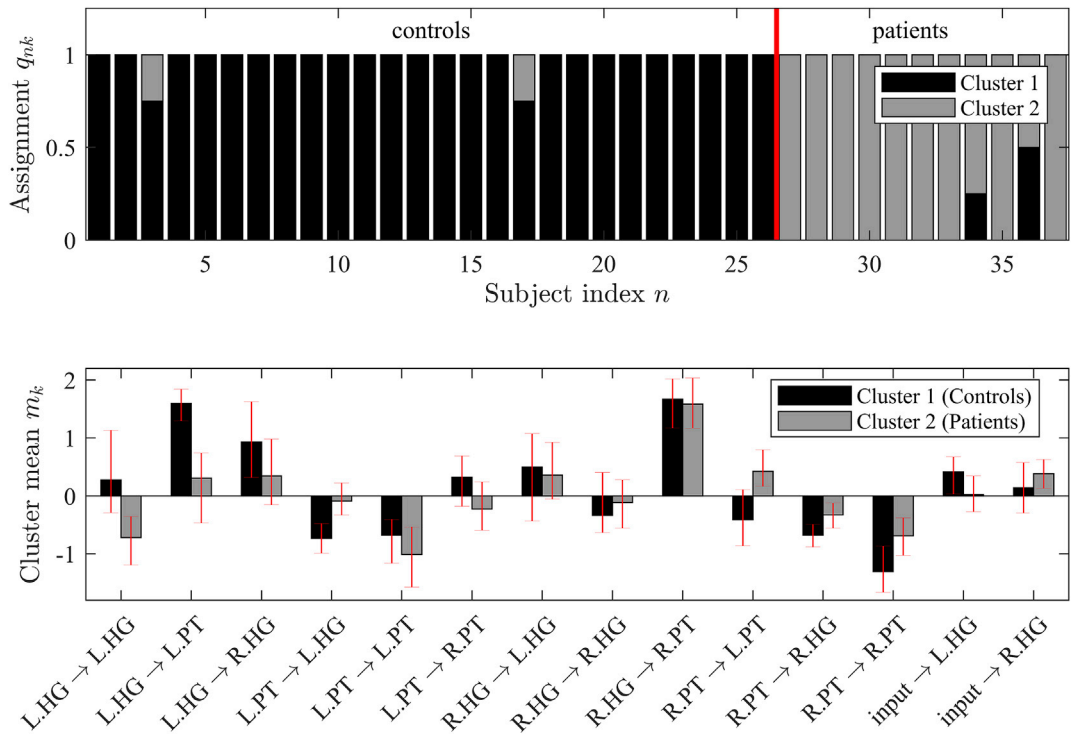

Fig. 14. MCMC results for the aphasia dataset: Top panel: Assignment estimates (balanced purity: 100%; The probability of subject 36 being in cluster 2 is barely above 50%). Bottom panel: cluster mean estimates. Red error bars indicate marginal 95% credible intervals.

MCMC, despite its dependence on approximations which are not present in the MCMC scheme. Based on 500,000 samples for the synthetic datasets and 2,800,000 samples for the empirical dataset (excluding burn-in), MCMC delivers only slightly more accurate results in terms of balanced purity and cluster mean estimates. In addition to the simulations, we also showed that the VB framework can identify the group structure in a real-world dataset. In the following, we discuss novelty, advantages and disadvantages of HUGE and its VB-based inversion, the computational complexity of the VB inversion scheme and potential additional savings that could be obtained using parallel computing techniques.

Generative embedding exploits a key advantage of generative models (i.e., providing a low-dimensional approximation to how high-dimensional data were generated) in order to obtain a compact and interpretable feature space for subsequent (un)supervised learning (Brodersen et al., 2014; Brodersen et al., 2011). Its unsupervised variant was introduced as a strategy to address a central problem in computational psychiatry: the need to stratify heterogeneous spectrum disorders into pathophysiologically more homogenous subgroups and thus enhance the predictive validity of diagnoses (Stephan et al., 2017). HUGE unifies the original two-step procedure of generative embedding into the inversion of a single hierarchical model. This is not only mathematically more elegant (and challenging) but offers several important conceptual advantages: (i) it allows the prior to be learned from the data (empirical Bayes), (ii) it enables subgroup-specific regularization (i.e. subgroup-specific priors), and (iii) the detection of clusters takes uncertainty about connectivity parameter estimates into account. In addition, (iv) HUGE uses a specifically derived and efficient VB implementation that, wherever possible, exploits conjugacy to obtain fast, analytical update equations. By combining these four aspects, the HUGE implementation presented in this paper represents a first method with which it becomes feasible in practice (with acceptably short runtimes even for larger datasets) to detect, in a completely unsupervised manner, subgroups in heterogeneous populations that are defined by effective connectivity.

The novel aspects of HUGE may be best appreciated by juxtaposing it to other hierarchical models of fMRI data. In recent years, the fMRI literature has seen the emergence of several hierarchical models of brain activity and connectivity (e.g., Bielczyk et al., 2018; Ktena et al., 2018; Mandke et al., 2018; Richiardi et al., 2011; Vidaurre et al., 2017). We briefly comment on three schemes in a bit more detail since their comparison with HUGE may usefully illustrate unique contributions by the present work. First, Janssen et al. (2015) also include a mixture model in a hierarchical model (see also Hinne et al., 2015), but with important differences to HUGE. Janssen et al. (2015) used a non-parametric Bayesian approach (an infinite Gaussian mixture model) to cluster resting-state fMRI time series; in contrast, the mixture model component in HUGE acts on latent variables (DCM parameters) which serve as a model-based dimensionality reduction layer between the mixture model and the fMRI observations. Furthermore, to obtain subject-specific results (in their case, parcellations), Janssen et al. (2015) use a two-step approach. By contrast, a single inversion of the HUGE model yields both subject-specific parameter estimates, as well as group-level (i.e., clustering) results. Finally, the model by Janssen et al. (2015) does not directly estimate connectivity (but is primarily interested in assigning voxels to clusters based on their time series, with functional connectivity examined post hoc once clusters are determined), whereas HUGE provides estimates of effective connectivity. Second, Benozzo et al. (2017) present a hierarchical model that also provides estimates of effective connectivity but is based on a different formalism (Granger causality) and uses a different approximate Bayesian inference scheme (Expectation Propagation). In this model, hierarchy has a different purpose than in HUGE and serves to induce sparsity in model coefficients. A final key distinction is that this model, unlike HUGE, does not possess a hemodynamic forward model but directly operates on measured BOLD signals. Third, the model conceptually closest to HUGE is the work by Friston et al. (2016) on parametric empirical Bayes (PEB) for DCM. PEB-DCM represents an empirical Bayesian approach for inverting a hierarchical model of multi-subject DCMs. PEB-DCM inverts a “full” (i.e., maximally parameterized) model, using the Variational Laplace algorithm (Friston et al., 2007), and uses Bayesian model reduction for selecting a (nested) submodel. Unlike PEB-DCM, HUGE does not universally employ the

Laplace approximation, but, wherever possible, uses conjugate priors to derive analytical update equations. Furthermore, PEB-DCM is a supervised method and requires that group labels are known; by contrast, HUGE enables the unsupervised detection of subgroups in the population studied and allows for empirical Bayes to unfold separately for each of the (initially unknown) subgroups.

Returning to the comparison of HUGE to the classical two-step generative embedding procedure using DCM, several advances are notable. Concerning the first point mentioned above, classical DCM is a fully Bayesian approach that requires specifying priors for the various model parameters, raising the question how inference may depend on the particular choice of priors. HUGE allows for an empirical Bayesian approach to this problem by introducing a distribution over priors (i.e., a hyperprior). Although this shifts the choice to the level of hyperpriors, it enables the model to marginalize over a range of prior settings and to adjust for general trends in the population by providing an additional degree of freedom; thus, alleviating the influence of prior assumptions on results for individual subjects.

Regarding the second point, the ability to learn different priors for different subgroups enables HUGE to exert subgroup-specific regularization. This is particularly helpful for dealing with heterogeneous clinical populations that are thought to consist of numerous (but typically poorly known) subgroups (Brodersen et al., 2014; Stephan et al., 2017). By contrast, even in the presence of prior evidence for the existence of multiple subgroups in the population, classical generative embedding would use the same priors for the inversion of the DCM of all subjects. In HUGE, inference on subject-specific parameters in HUGE is guided by inference on subject-wise cluster assignment; critically, this unfolds automatically without the need to specify prior assignment preferences for individual subjects.

With regard to the third point, HUGE neither performs clustering on observed data nor on point estimates, but on the full posterior densities of DCM parameters (which are estimated in parallel). In the classical two-step generative embedding approach, point estimates (e.g., MAPs) of the posterior are obtained from subject-wise model inversions and used as input features for (un)supervised learning. By contrast, in HUGE, the clustering step takes the uncertainty of the DCM parameter estimation into account. This is evident from the VB update equations (specifically, the term $-0.5\nu_k \mathrm{tr}(S_k^{-1}\Sigma_n^{(c)})$ in the expression for the cluster assignments, see Eq. (18)). The VB inference scheme tends to prefer solutions where most subjects are assigned to one large cluster. Generally, this is a desirable property: for noisy input, the more principled approach of HUGE yields a conservative solution with regard to cluster assignments. However, in application contexts where sensitivity of subgroup detection is paramount, priors for the mixture model component may have to be adapted compared to our current configuration.

Fourth, HUGE benefits from a novel VB inversion scheme that was specifically derived for its purpose (details of the derivation are provided by the Supplementary Material). A significant benefit afforded by model inversion under VB is that it automatically delivers an approximation to the log model evidence (the negative free energy). In a Bayesian setting, the model evidence is a principled metric of model goodness and is routinely used to distinguish between competing models (Bishop, 2006; MacKay, 2004). One example of the utility of Bayesian model selection in the context of HUGE concerns the choice of the optimal number of clusters, as demonstrated in section 3. Furthermore, different connectivity structures of the underlying DCM can also be compared. By contrast, computing the model evidence with Monte Carlo-based methods poses additional computational demands. Simple sampling-based estimators of the model evidence (e.g., prior arithmetic mean, posterior harmonic mean) have serious limitations; a superior alternative is thermodynamic integration (TI) (Calderhead et al., 2009; Lartillot et al., 2006; Paquet, 2008). Critically, TI requires numerous parallel MCMC chains at different temperatures, rendering TI computationally very expensive. Only recently, parallel computing techniques have been introduced that exploit the processing power of GPUs and begin to make TI feasible (Aponte et al., 2016). In addition, detecting failure of convergence of the MCMC chain, as seen for the aphasia dataset in our examples above, is a nontrivial task (Gelman et al., 1992).

In the previous paragraphs, we have highlighted some advantages of HUGE over the original (two-step) generative embedding approach. However, more complex models incur an increase in computational demands, which may threaten the practical utility of a model. For HUGE, approximate inversion methods like VB or expectation propagation represent a promising alternative to earlier MCMC-based formulations (Raman et al., 2016). The VB inversion presented here provides a speed-up of two orders of magnitude compared to the MCMC-based inversion presented in Raman et al. (2016). On the other hand, considering that multi-start procedures may be necessary to protect against local optima, one might wonder whether in practice the VB inversion scheme provides any computational advantage over MCMC. However, the different instances (i.e., random restarts) of VB can be parallelized, while a single MCMC chain cannot (for a detailed analysis, see Supplementary Material section S8). Furthermore, the structure of the VB update equations allows for additional savings by applying parallel processing techniques to the numerical calculation of the Jacobians G_n (see Eq. (20)). A detailed analysis of the computation time of the VB inversion scheme for HUGE revealed that the bottleneck in the current implementation is the evaluation of the neuronal and hemodynamic state equations $g(\theta_n, u)$ (Eq. (8)), which accounts for about 90% of the computation time. This is mainly due to the complexity of the numerical integration involved in evaluating the neuronal and hemodynamic equations, but also because the Jacobian G_n is presently evaluated with the finite difference method. Our analysis indicated that exploitation of opportunities for parallelization would allow for a further speed-up by nearly one order of magnitude. This is something we will pursue in future work. A detailed analysis of this topic is provided in the Supplementary Material section S6.

Another advantage of our approach is that VB is not affected by the so-called label-switching problem for mixture models. Label-switching refers to the phenomena that, in a mixture model with symmetric priors, permuting the numbering of the clusters does not change the posterior distribution. When applying Monte Carlo methods to mixture models, the label-switching problem prevents obtaining cluster-related estimates (e.g., the cluster mean) by simply taking the ergodic averages of samples (Celeux, 1998). Label-switching is commonly solved by either introducing constraints on the parameters (e.g., forcing an ordering on the cluster means) (Richardson et al., 1997) or using re-labelling schemes (Celeux, 1998). However, both approaches have their limitations, and using VB avoids the label-switching problem altogether.

Despite the advantages highlighted above, VB also has a number of limitations. The most severe limitation is its susceptibility to local extrema, which we also observed in our analysis of the empirical dataset. Specifically, clustering solutions obtained under VB can depend strongly on the choice of the initial (starting) values of the algorithm. This is a well-known problem of VB, which is aggravated by the complexity of the hierarchical DCM which induces strong posterior correlations among different parameters. This problem puts a limit on the accuracy achievable with our VB scheme in particular and the HUGE model in general. For the empirical dataset, we addressed the problem of local maxima by running the optimization repeatedly from random starting positions. However, this multi-start strategy becomes less effective with increasing dimensionality of the problem (Bishop, 2006) and does not guarantee convergence to the global maximum. For a more detailed analysis of the multi-start approach for the empirical dataset, see section S8 in the Supplementary Material.

Having said this, we would like to emphasize that the empirical dataset utilized in this paper constitutes a challenging scenario for clustering because of the small sample size and the relatively high dimensionality of the parameter space. Unfortunately, in fMRI studies, conditions like these are encountered frequently. Generally, one should

strive for larger numbers of subjects and compact single-subject models (e.g., reducing the size of the networks studied and thus the dimensionality of the parameter space in which clustering takes place) in order to create graceful conditions for generative embedding analyses with HUGE. However, it is worth emphasizing that the requirement of many subjects relative to features is not specific for HUGE but similarly applies to any unsupervised learning approach (Bishop, 2006). Deciding between different potential DCMs (including network structures with different numbers of regions) can be done straightforwardly by comparing which model optimizes the balanced purity with respect to the external criteria of interest (e.g., clinical diagnoses or outcomes). Notably, there is no overfitting issue here: these criteria are completely independent from the model and its estimation.

A final and important point raised by Raman et al. (2016) concerns the representation of potentially confounding effects by covariates like age, gender or handedness in the hierarchical model. These effects can overshadow effects of interest, such as differences between subgroups of patients, and taking these confounds into account can be essential for obtaining meaningful clustering results (Brodersen et al., 2014). Notably, while correcting for confounding covariates is straightforward in the two-step procedure of generative embedding (Brodersen et al., 2014), this is a non-trivial endeavor for the hierarchical model presented here because introducing covariates affects the update equations of the full model. Nevertheless, given the importance of this issue for clinical applications, future extensions of HUGE will incorporate covariates into the hierarchical model.

Acknowledgements

The work presented in this paper was supported by the Central Institute ZEA-2—Electronic Systems at Research Center Jülich, Germany (to YY), the ETH Zurich Postdoctoral Fellowship and the Marie Curie Actions for People COFUND Program (to SF), as well as the René und Susanne Braginsky Foundation and the University of Zurich (to KES). In addition, we would like to thank Prof. Stefan van Waasen of the Central Institute ZEA-2—Electronic Systems at Research Center Jülich, Germany for his generous support, as well as discussion and suggestions.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.neuroimage.2018.06.073.

References

Abramowitz, M., Stegun, I.A. (Eds.), 1972. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Dover Publications, New York.
Aponte, E.A., Raman, S., Sengupta, B., Penny, W.D., Stephan, K.E., Heinzle, J., 2016. mpdcm: a toolbox for massively parallel dynamic causal modeling. J. Neurosci. Meth. 257, 7–16.
Attias, H., 1999. Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc, Stockholm, Sweden, pp. 21–30.
Banerjee, S., Carlin, B.P., Gelfand, A.E., 2015. Hierarchical Modeling and Analysis for Spatial Data, second ed. CRC Press, Boca Raton.
Behrens, T.E.J., Woolrich, M.W., Walton, M.E., Rushworth, M.F.S., 2007. Learning the value of information in an uncertain world. Nat. Neurosci. 10, 1214–1221.
Benozzo, D., Jylänki, P., Olivetti, E., Avesani, P., van Gerven, M.A.J., 2017. Bayesian estimation of directed functional coupling from brain recordings. PLoS One 12, e0177359.
Bielczyk, N.Z., Walocha, F., Ebel, P.W., Haak, K.V., Llera, A., Buitelaar, J.K., Glennon, J.C., Beckmann, C.F., 2018. Thresholding functional connectomes by means of mixture modeling. Neuroimage 171, 402–414.
Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer, Cambridge.
Brodersen, K.H., Deserno, L., Schlagenhauf, F., Lin, Z., Penny, W.D., Buhmann, J.M., Stephan, K.E., 2014. Dissecting psychiatric spectrum disorders by generative embedding. Neuroimage: Clinica 4, 98–111.
Brodersen, K.H., Schofield, T.M., Leff, A.P., Ong, C.S., Lomakina, E.I., Buhmann, J.M., Stephan, K.E., 2011. Generative embedding for model-based classification of fMRI data. PLoS Comput. Biol. 7.
Buxton, R.B., Wong, E.C., Frank, L.R., 1998. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. Magn. Reson. Med. 39, 855–864.
Calderhead, B., Girolami, M., 2009. Estimating Bayes factors via thermodynamic integration and population MCMC. Comput. Stat. Data Anal. 53, 4028–4045.
Celeux, G., 1998. Bayesian inference for mixture: the label switching problem. In: Payne, R., Green, P. (Eds.), COMPSTAT: Proceedings in Computational Statistics 13th Symposium Held in Bristol, Great Britain, 1998. Physica-Verlag HD, Heidelberg, pp. 227–232.
Daunizeau, J., David, O., Stephan, K.E., 2011. Dynamic causal modelling: a critical review of the biophysical and statistical foundations. Neuroimage 58, 312–322.
David, O., Kiebel, S.J., Harrison, L.M., Mattout, J., Kilner, J.M., Friston, K.J., 2006. Dynamic causal modeling of evoked responses in EEG and MEG. Neuroimage 30, 1255–1272.
Friston, K.J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G., 2017. Active inference: a process theory. Neural Comput. 29, 1–49.
Friston, K.J., Harrison, L., Penny, W.D., 2003. Dynamic causal modelling. Neuroimage 19, 1273–1302.
Friston, K.J., Litvak, V., Oswal, A., Razi, A., Stephan, K.E., van Wijk, B.C.M., Ziegler, G., Zeidman, P., 2016. Bayesian model reduction and empirical Bayes for group (DCM) studies. Neuroimage 128, 413–431.
Friston, K.J., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W.D., 2007. Variational free energy and the Laplace approximation. Neuroimage 34, 220–234.
Friston, K.J., Mechelli, A., Turner, R., Price, C.J., 2000. Nonlinear responses in fMRI: the balloon model, Volterra Kernels, and other hemodynamics. Neuroimage 12, 466–477.
Gelman, A., 2014. Bayesian Data Analysis, third ed. CRC Press, Boca Raton, Fla.
Gelman, A., Meng, X.L., 1998. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. Stat. Sci. 13, 163–185.
Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7, 457–472.
Harlé, K.M., Stewart, J.L., Zhang, S., Tapert, S.F., Yu, A.J., Paulus, M.P., 2015. Bayesian neural adjustment of inhibitory control predicts emergence of problem stimulant use. Brain 138, 3413–3426.
Harrison, S.J., Woolrich, M.W., Robinson, E.C., Glasser, M.F., Beckmann, C.F., Jenkinson, M., Smith, S.M., 2015. Large-scale probabilistic functional modes from resting state fMRI. Neuroimage 109, 217–231.
Havlicek, M., Roebroeck, A., Friston, K.J., Gardumi, A., Ivanov, D., Uludag, K., 2017. On the importance of modeling fMRI transients when estimating effective connectivity: a dynamic causal modeling study using ASL data. Neuroimage 155, 217–233.
Hinne, M., Ambrogioni, L., Janssen, R.J., Heskes, T., van Gerven, M.A.J., 2014. Structurally-informed Bayesian functional connectivity analysis. Neuroimage 86, 294–305.
Hinne, M., Ekman, M., Janssen, R.J., Heskes, T., van Gerven, M.A.J., 2015. Probabilistic clustering of the human connectome identifies communities and hubs. PLoS One 10 e0117179.
Janssen, R.J., Jylänki, P., Kessels, R.P.C., van Gerven, M.A.J., 2015. Probabilistic model-based functional parcellation reveals a robust, fine-grained subdivision of the striatum. Neuroimage 119, 398–405.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795.
Ktena, S.I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., Rueckert, D., 2018. Metric learning with spectral graph convolutions on brain connectivity networks. Neuroimage 169, 431–442.
Langs, G., Sweet, A., Lashkari, D., Tie, Y., Rigolo, L., Golby, A.J., Golland, P., 2014. Decoupling function and anatomy in atlases of functional connectivity patterns: language mapping in tumor patients. Neuroimage 103, 462–475.
Lartillot, N., Philippe, H., 2006. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55, 195–207.
Lindquist, M.A., Krishnan, A., López-Solà, M., Jepma, M., Woo, C.-W., Koban, L., Roy, M., Atlas, L.Y., Schmidt, L., Chang, L.J., Reynolds Losin, E.A., Eisenbarth, H., Ashar, Y.K., Delk, E., Wager, T.D., 2017. Group-regularized individual prediction: theory and application to pain. Neuroimage 145, 274–287.
Lomakina, E.I., Paliwal, S., Diaconescu, A.O., Brodersen, K.H., Aponte, E.A., Buhmann, J.M., Stephan, K.E., 2015. Inversion of hierarchical Bayesian models using Gaussian processes. Neuroimage 118, 133–145.
MacKay, D.J.C., 2004. In: Information Theory, Inference, and Learning Algorithms, Repr. With Corr. Ed. Univ. Press, Cambridge.
Mandke, K., Meier, J., Brookes, M.J., O'Dea, R.D., Van Mieghem, P., Stam, C.J., Hillebrand, A., Tewarie, P., 2018. Comparing multilayer brain networks between groups: introducing graph metrics and recommendations. Neuroimage 166, 371–384.
Manning, C.D., Raghavan, P., Schütze, H., 2009. Introduction to Information Retrieval, Repr. Ed. Cambridge University Press, Cambridge.
Mathys, C.D., Lomakina, E.I., Daunizeau, J., Iglesias, S., Brodersen, K.H., Friston, K.J., Stephan, K.E., 2014. Uncertainty in perception and the hierarchical Gaussian filter. Frontiers in human. Neuroscience 8.
Paquet, U., 2008. Bayesian Inference for Latent Variable Models. University of Cambridge, Computer Laboratory.
Penny, W.D., Friston, K.J., Ashburner, J.T., Kiebel, S.J., Nichols, T.E. (Eds.), 2007. Statistical Parametric Mapping: the Analysis of Functional Brain Images. Academic Press, London, UK.
Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004. Comparing dynamic causal models. Neuroimage 22, 1157–1172.
Piray, P., den Ouden, H.E.M., van der Schaaf, M.E., Toni, I., Cools, R., 2017. Dopaminergic modulation of the functional Ventrodorsal architecture of the human striatum. Cerebr. Cortex 27, 485–495.

Rae, C.L., Hughes, L.E., Anderson, M.C., Rowe, J.B., 2015. The prefrontal cortex achieves inhibitory control by facilitating subcortical motor pathway connectivity. J. Neurosci. 35, 786–794.
Raftery, A., Newton, M., Satagopan, J., Krivitsky, P., 2007. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In: Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M. (Eds.), Bayesian Statistics 8. Oxford University Press, Oxford, pp. 1–45.
Raman, S., Deserno, L., Schlagenhauf, F., Stephan, K.E., 2016. A hierarchical model for integrating unsupervised generative embedding and empirical Bayes. J. Neurosci. Meth. 269, 6–20.
Richardson, S., Green, P.J., 1997. On bayesian analysis of mixtures with an unknown number of components (with discussion). J. Roy. Stat. Soc. B 59, 731–792.
Richiardi, J., Eryilmaz, H., Schwartz, S., Vuilleumier, P., Van De Ville, D., 2011. Decoding brain states from fMRI connectivity graphs. Neuroimage 56, 616–626.
Sanyal, N., Ferreira, M.A.R., 2012. Bayesian hierarchical multi-subject multiscale analysis of functional MRI data. Neuroimage 63, 1519–1531.
Schofield, T.M., Penny, W.D., Stephan, K.E., Crinion, J.T., Thompson, A.J., Price, C.J., Leff, A.P., 2012. Changes in auditory feedback connections determine the severity of speech processing deficits after stroke. J. Neurosci. 32, 4260–4270.
Stephan, Klaas E., Iglesias, S., Heinzle, J., Diaconescu, Andreea O., 2015. Translational perspectives for computational neuroimaging. Neuron 87, 716–732.
Stephan, K.E., Kasper, L., Harrison, L.M., Daunizeau, J., den Ouden, H.E.M., Breakspear, M., Friston, K.J., 2008. Nonlinear dynamic causal models for fMRI. Neuroimage 42, 649–662.
Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J., 2009. Bayesian model selection for group studies. Neuroimage 46, 1004–1017.
Stephan, K.E., Schlagenhauf, F., Huys, Q.J.M., Raman, S., Aponte, E.A., Brodersen, K.H., Rigoux, L., Moran, R.J., Daunizeau, J., Dolan, R.J., Friston, K.J., Heinz, A., 2017. Computational neuroimaging strategies for single patient predictions. NeuroImage 145 (Part B), 180–199.
Stephan, K.E., Weiskopf, N., Drysdale, P.M., Robinson, P.A., Friston, K.J., 2007. Comparing hemodynamic models with DCM. Neuroimage 38, 387–401.
van Leeuwen, T.M., den Ouden, H.E.M., Hagoort, P., 2011. Effective connectivity determines the nature of subjective experience in grapheme-color synesthesia. J. Neurosci. 31, 9879–9884.
Vidaurre, D., Smith, S.M., Woolrich, M.W., 2017. Brain network dynamics are hierarchically organized in time. Proc. Natl. Acad. Sci. Unit. States Am. 114, 12827–12832.
Welvaert, M., Rosseel, Y., 2013. On the definition of signal-to-noise ratio and contrast-to-noise ratio for fMRI data. PLoS One 8 e77089.
Wiecki, T.V., Poland, J., Frank, M.J., 2015. Model-based cognitive neuroscience approaches to computational psychiatry. Clini. Psychol. Sci. 3, 378–399.