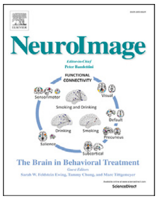
NeuroImage 179 (2018) 505–529

Contents lists available at ScienceDirect

NeuroImage

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



A generative model of whole-brain effective connectivity

Stefan Frässle [a,*,1], Ekaterina I. Lomakina [a,b,1], Lars Kasper [a,c], Zina M. Manjaly [a,d], Alex Leff [e,f], Klaas P. Pruessmann [c], Joachim M. Buhmann [b], Klaas E. Stephan [a,e]

[a] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich & ETH Zurich, 8032 Zurich, Switzerland
[b] Department of Computer Science, ETH Zurich, 8032 Zurich, Switzerland
[c] Institute for Biomedical Engineering, ETH Zurich & University of Zurich, 8092 Zurich, Switzerland
[d] Dept. of Neurology, Schulthess Clinic, 8008 Zurich, Switzerland
[e] Wellcome Trust Centre for Neuroimaging, University College London, London WC1N 3BG, United Kingdom
[f] Institute of Cognitive Neuroscience, University College London, London WC1N 3AZ, United Kingdom

ARTICLE INFO

Keywords:
Bayesian regression
Dynamic causal modeling
Generative model
Effective connectivity
Sparsity
Connectomics

ABSTRACT

The development of whole-brain models that can infer effective (directed) connection strengths from fMRI data represents a central challenge for computational neuroimaging. A recently introduced generative model of fMRI data, regression dynamic causal modeling (rDCM), moves towards this goal as it scales gracefully to very large networks. However, large-scale networks with thousands of connections are difficult to interpret; additionally, one typically lacks information (data points per free parameter) for precise estimation of all model parameters.

This paper introduces sparsity constraints to the variational Bayesian framework of rDCM as a solution to these problems in the domain of task-based fMRI. This *sparse rDCM* approach enables highly efficient effective connectivity analyses in whole-brain networks and does not require *a priori* assumptions about the network's connectivity structure but prunes fully (all-to-all) connected networks as part of model inversion. Following the derivation of the variational Bayesian update equations for sparse rDCM, we use both simulated and empirical data to assess the face validity of the model. In particular, we show that it is feasible to infer effective connection strengths from fMRI data using a network with more than 100 regions and 10,000 connections. This demonstrates the feasibility of whole-brain inference on effective connectivity from fMRI data – in single subjects and with a run-time below 1 min when using parallelized code. We anticipate that sparse rDCM may find useful application in connectomics and clinical neuromodeling – for example, for phenotyping individual patients in terms of whole-brain network structure.

Introduction

The human brain comprises multiple levels of organization, with cognitive functions arising from the interplay of functional specialization and integration (Sporns, 2013; Tononi et al., 1994). With the advent of non-invasive neuroimaging techniques, such as functional magnetic resonance imaging (fMRI), researchers have begun to systematically study these fundamental properties of human brain organization. While early neuroimaging studies focused on localizing cognitive processes to specific brain regions, contemporary studies are commonly concerned with functional integration of these regions; this requires analyzing brain connectivity (Smith, 2012). In addition to analyses of structural (anatomical) connections, assessments of functional connectivity (statistical dependencies between network nodes) play a major role. Particularly, functional connectivity has been used frequently for studying the functional organization of large (whole-brain) networks both in tasks and in the "resting state" (i.e., unconstrained cognition in the absence of external perturbations). Various methods have been proposed (for a comprehensive review, see Karahanoglu and Van De Ville, 2017), ranging from conventional correlation or coherence analyses which assume stationarity (Fox et al., 2005) to sliding-window correlation analyses that can capture dynamic fluctuations in functional connectivity (Chang and Glover, 2010). Recently, more sophisticated methods have been introduced to characterize brain organization such as (sparsely

[*] Corresponding author. University of Zurich & ETH Zurich, Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, Wilfriedstrasse 6, 8032 Zurich, Switzerland.
E-mail address: stefanf@biomed.ee.ethz.ch (S. Frässle).
[1] Contributed equally to this work.

https://doi.org/10.1016/j.neuroimage.2018.05.058
Received 1 December 2017; Received in revised form 16 May 2018; Accepted 24 May 2018
Available online 25 May 2018
1053-8119/© 2018 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

coupled) Hidden Markov models (HMM; Bolton et al., 2018; Karahanoglu and Van De Ville, 2015; Vidaurre et al., 2017), and approaches from statistical mechanics that rely on entropy maximization (Ashourvan et al., 2017). Other functional connectivity methods have focused on sparsity (Bielczyk et al., 2018; Eavani et al., 2015; Ryali et al., 2012), including generative models that can exploit anatomical information (Hinne et al., 2014). While functional connectivity affords valuable insights into the dynamics of brain networks both in health and disease (Buckner et al., 2013; Bullmore and Sporns, 2009; Fornito et al., 2015), it provides undirected measures of coupling (Friston, 2011), without an explicit model of the system of interest (Stephan, 2004). By contrast, effective connectivity refers to directed interactions among neuronal populations and rests on a model describing both dynamics within the (neuronal) system and how activity of local network nodes is transformed into observations (Friston, 2011; Valdes-Sosa et al., 2011).

Estimates of effective connectivity typically derive from a generative model that provides a forward mapping from hidden (latent) neuronal circuit dynamics to observable brain signals (Friston et al., 2013). While different models of effective connectivity have been proposed (for a comprehensive summary, see Valdes-Sosa et al., 2011), to our knowledge, none has so far enabled estimates of connection-specific strengths for networks derived from typical whole-brain parcellation schemes with more than 100 nodes (Glasser et al., 2016; Tzourio-Mazoyer et al., 2002). For instance, biophysical network models (BNMs) combine mean-field models of local neuronal dynamics with anatomical data on long-range connections, capturing many structural and physiological details of whole-brain networks (Deco et al., 2013a; Jirsa et al., 2016). However, the complexity of these models has made parameter estimation computationally extremely challenging. Consequently, applications have typically focused on simulations under fixed parameters (Deco et al., 2013a; Honey et al., 2007) or on simplified models that allow for the estimation of a global scaling parameter (Deco et al., 2013b, 2014a, 2014b). Only recently has a novel variant of BNMs been proposed in which brain dynamics during the "resting state" are described by an Ornstein-Uhlenbeck process and which provided maximum likelihood estimates of directed connection strengths in networks comprising up to 68 regions (Gilson et al., 2016, 2017; Rolls et al., 2018; Senden et al., 2018). Additionally, other variants of connectivity analyses have been proposed that might potentially enable the investigation of larger networks (Ambrogioni et al., 2017; Prando et al., 2017).

In contrast, dynamic causal modeling (DCM) uses a Bayesian framework to compute the posterior distribution over effective connectivity parameters. DCM was originally devised for fMRI (Friston et al., 2003) and later extended to other modalities like M/EEG (David et al., 2006). Critically, in order to render model inversion computationally feasible, a central limitation of DCM (for review, see Daunizeau et al., 2011) is that models are restricted to relatively small networks, on the order of 10 regions. Consequently, in its classical formulation, DCM cannot accommodate whole-brain networks. Recently, cross-spectral DCM (Friston et al., 2014a) has been combined with an approach to constrain the effective number of parameters (Seghier and Friston, 2013) to invert networks comprising 36 brain regions (Razi et al., 2017).

Regression DCM (rDCM) has recently been introduced as a novel variant of DCM for fMRI that has promising potential for the application to large (whole-brain) neural networks (Frässle et al., 2017). In brief, rDCM converts the numerically costly problem of estimating coupling parameters in differential equations (of a linear DCM in the time domain) into an efficiently solvable Bayesian linear regression in the frequency domain. Under some generic assumptions, analytic variational Bayesian update equations can be derived for the model parameters and hyperparameters (i.e., noise precision), enabling computationally highly efficient model inversion. The current implementation of rDCM is designed to work with experimentally controlled perturbations (the driving inputs in a linear DCM) and hence task fMRI (although it can, in principle, also be applied to "resting state" data, see Discussion). rDCM scales gracefully (polynomially, as opposed to exponentially) with the number of nodes (with the inversion of a covariance matrix being the computationally most expensive operation), and previous simulations demonstrated that it can be applied to networks that are (at least) one order of magnitude larger than the ones afforded in classical DCM implementations (Frässle et al., 2017). In particular, we showed that rDCM can accurately estimate effective connectivity in a network comprising 66 brain regions and 300 free (neuronal) connectivity parameters, based on a realistic human structural connectome (Hagmann et al., 2008).

However, for such large-scale networks, the number of model parameters quickly outweighs the number of acquired data points, a statistical problem known as the "large-*p*-small-*n*" scenario (Buhlmann and van de Geer, 2011). In this situation, precise inference of parameters may no longer be feasible (a common rule of thumb suggests that ten data points per free parameter are required; Penny and Roberts, 1999). Furthermore, even when enough data points are available, large-scale effective connectivity patterns are difficult to interpret due to the vast amount of parameters. A principled solution to both problems rests on regularized or sparse regression methods that introduce some form of penalty on the number of parameters. Established sparse regression methods in the field of statistics include LASSO (Tibshirani, 1996), elastic net regularization (Zou and Hastie, 2005), or Spike-and-Slab priors for Bayesian linear regression (Hernandez-Lobato et al., 2013). All these methods implicitly assume that only few predictors convey useful information for explaining the measured data, whereas all others are negligible and should be excluded from the model – thus, effectively reducing the dimensionality of the problem. For models of functional integration, this assumption seems tenable, given that functional characteristics (e.g., small-world) and economics (metabolic costs of maintaining long-range myelinated axons) imply a non-trivial degree of sparsity in whole-brain networks (Bullmore and Sporns, 2009; Kötter and Stephan, 2003; Sporns et al., 2000; Sporns and Zwi, 2004).

In this paper, we augment the original rDCM framework with sparsity constraints to enable automatic "pruning" of fully connected networks to the most essential connections. In brief, this *sparse rDCM* rests on introducing binary indicator variables that are embedded into the likelihood function of the generative model and mediate feature selection as part of model inversion. This yields the posterior density over model parameters under a parsimonious representation of the whole-brain graph.

In what follows, we first introduce the theoretical foundation of sparse rDCM and outline how sparsity constraints can be embedded into the original framework. We then derive variational Bayesian update equations for the neuronal connectivity, noise precision, and binary indicator variables. We subsequently use simulations to demonstrate the face validity of sparse rDCM for effective connectivity analyses in a large network, comprising up to 66 brain regions and 300 (known) parameters. For this, we evaluate the accuracy of the framework to recover the data-generating network architecture (i.e., the "true" sparse effective connectivity pattern). For these simulations, we used network architectures of varying complexity and biological realisms, starting with a relatively simple model (*grid-like* DCM) which allows for easy visualization of the performance of sparse rDCM. We then turned to models (*small-world* DCMs) based on the S50 network structure tested in Smith et al. (2011) that, as highlighted by those authors, captures the small-world characteristics of the brain. Finally, we tested a biologically more plausible network architecture (*connectome-based* DCM) derived from a whole-brain atlas and the structural human connectome provided by the diffusion-weighted imaging work by Hagmann et al. (2008), which has been frequently used in previous studies on large-scale models of effective connectivity (e.g., Deco et al., 2014a, 2014b, 2013b; Gilson et al., 2017; Honey et al., 2009; Ponce-Alvarez et al., 2015a, 2015b). We then proceed to empirical fMRI datasets in order to demonstrate the practical utility of sparse rDCM. First, we focus on small networks (of typical size for conventional DCM analyses) and established fMRI datasets (Büchel and Friston, 1997; Schofield et al., 2012), and compare sparse rDCM solutions to those obtained by classical DCM. These datasets were chosen because they (i) have been extensively studied with other methods of connectivity (“attention-to-motion” dataset), or (ii) are associated with a known ground truth in terms of the group membership of subjects, i.e., stroke patients and healthy controls (“aphasia” dataset). Finally, we provide a proof-of-principle that sparse rDCM enables whole-brain inference of effective connectivity from fMRI data, using a single-subject dataset from a hand movement paradigm. The choice of this dataset was motivated by the simple and robust nature of the task, and the extensive knowledge available on the cerebral network supporting visually paced hand movements (e.g., Ledberg et al., 2007; Rizzolatti and Luppino, 2001). Here, we used a whole-brain parcellation scheme with more than 100 regions, resulting in more than 10,000 connections that span the entire cortex and cerebellum.

Methods and materials

Dynamic causal modeling

Dynamic causal modeling (DCM; Friston et al., 2003) is a generative modeling framework for inferring hidden (latent) neuronal states from measured neuroimaging data. For fMRI, DCM describes the dynamics in neuronal activity in brain regions as a function of the effective (i.e., directed) connectivity among neuronal populations:

$$\frac{dx}{dt} = \left(A + \sum_{j=1}^{m} u_j B^{(j)}\right)x + Cu \tag{1}$$

where A denotes network connectivity in the absence of experimental manipulations (endogenous connectivity), B represents perturbations of the endogenous connectivity (modulatory influences) by experimental manipulations u_j (e.g., sensory stimulation, task demands), and C describes how these experimental manipulations directly influence neuronal activity (driving inputs). This neuronal model is coupled to a hemodynamic forward model that maps neuronal dynamics to observed BOLD signal time series (Buxton et al., 1998; Friston et al., 2000; Stephan et al., 2007). Inversion of this generative model rests on a variational Bayesian approach under the Laplace approximation (VBL), i.e., posterior densities over model parameters and hyperparameters are assumed to have a Gaussian form (Bishop, 2006; Friston et al., 2007). Comprehensive reviews of DCM for fMRI can be found elsewhere (Daunizeau et al., 2011; Friston et al., 2013).

While DCM has proven useful for studying the functional integration in small networks (in the order of 10 regions), the model does not scale to large (whole-brain) networks for several reasons: First, evaluating the likelihood function of DCMs requires one to integrate the differential equations in both the neuronal (Eq. (1)) and the hemodynamic model. This integration is computationally costly, especially for long time series. Second, since classical DCM treats the data as a region × time-series vector, the error covariance matrix can become prohibitively large with increasing numbers of regions.

Regression DCM

Regression DCM (rDCM) was recently introduced as a novel variant of DCM for fMRI to address this bottleneck (Frässle et al., 2017). The approach rests on several modifications of the original DCM framework. In brief, these include (i) translating state and observation equations from time to frequency domain, (ii) linearizing the hemodynamic forward model, (iii) a mean field assumption across regions (i.e., independence of connectivity parameters targeting different regions), and (iv) specifying conjugate priors to enable analytic update equations. We only briefly outline these steps here; for details, see Frässle et al. (2017).

The neuronal state equation of rDCM corresponds to a linear DCM:

$$\frac{dx}{dt} = Ax + Cu \tag{2}$$

Using the differential property of the Fourier transform and a fixed hemodynamic convolution kernel h with additive Gaussian noise as forward model, one obtains an algebraic equation in the frequency domain:

$$\left(e^{2\pi i\frac{m}{N}} - 1\right)\frac{\widehat{y}}{T} = A\widehat{y} + C\widehat{h}\widehat{u} + \upsilon \tag{3}$$

Here, y denotes the measured BOLD signal, N is the number of data points, T is the time interval between subsequent points, m is a vector of frequency indices, υ is the observation noise, and the hat symbol $\widehat{()}$ represents the discrete Fourier transform (DFT). In Eq. (3), the observation noise υ has a complicated form that motivates a mean field approximation across regions; this assumes that the approximate posterior factorizes into sets of connections entering different nodes. Under conjugate priors, namely a Gaussian prior on the neuronal connectivity parameters and a Gamma prior on the noise precision, rDCM can then be formalized as a Bayesian linear regression model:

$$\begin{aligned}
p(Y|\theta,\tau,X) &= \prod_{r=1}^{R}\mathcal{N}\left(Y_r; X\theta_r, \tau_r^{-1} I_{N\times N}\right)\\
Y_r &= \left(e^{2\pi i\frac{m}{N}} - 1\right)\frac{\widehat{y}_r}{T}\\
X &= \left[\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_R, \widehat{h}\widehat{u}_1, \widehat{h}\widehat{u}_2, \ldots, \widehat{h}\widehat{u}_K\right]\\
\theta_r &= \left[a_{r,1}, a_{r,2}, \ldots, a_{r,R}, c_{r,1}, c_{r,2}, \ldots, c_{r,K}\right]
\end{aligned} \tag{4}$$

Here, Y_r is the signal in region r that is explained as a linear mixture of afferent connections from other regions and direct (driving) inputs of unspecified origin, y_r is the measured regional BOLD signal, X is the design matrix (i.e., a set of regressors or explanatory inputs), u_k is the k^{th} experimental input, θ_r represents the parameter vector comprising all connections and inputs targeting region r, τ_r denotes the noise precision parameter for region r, and $I_{N\times N}$ is the identity matrix. Note that the model in Eq. (4) factorizes across regions due to the mean field approximation mentioned above. This implies that variational (Bayesian) update equations for the sufficient statistics of the posterior density can be derived for each region separately. The final iterative update scheme for rDCM then takes the form:

$$\begin{aligned}
\Sigma_{\theta|y} &= \left(\frac{\alpha_{\tau|y}}{\beta_{\tau|y}}X^T X + \Sigma_0^{-1}\right)^{-1}\\
\mu_{\theta|y} &= \Sigma_{\theta|y}\left(\frac{\alpha_{\tau|y}}{\beta_{\tau|y}}X^T Y + \Sigma_0^{-1}\mu_0\right)\\
\alpha_{\tau|y} &= \alpha_0 + \frac{N}{2}\\
\beta_{\tau|y} &= \beta_0 + \frac{1}{2}\left(Y - X\mu_{\theta|y}\right)^T\left(Y - X\mu_{\theta|y}\right) + \frac{1}{2}\mathrm{tr}\left(X^T X\Sigma_{\theta|y}\right)
\end{aligned} \tag{5}$$

Where tr represents the trace of a matrix. In Eq. (5), $\mu_{\theta|y}$ and $\Sigma_{\theta|y}$ denote the posterior mean and covariance of the Gaussian distribution over (neuronal) connectivity parameters, respectively. Similarly, $\alpha_{\tau|y}$ and $\beta_{\tau|y}$ represent the posterior shape and rate parameters of the Gamma distribution on noise precision, respectively. The update equations in Eq. (5) are mutually dependent on each other; hence, the optimal posterior distributions must be obtained by iterating over these update equations until convergence.

Furthermore, one can derive an expression for the negative (variational) free energy (Friston et al., 2007), which represents a lower-bound approximation to the log model evidence (i.e., the probability of the data given the model). The negative free energy provides a trade-off between the accuracy and complexity of a model, and serves as a principled metric for testing competing hypotheses (models) of the network architecture by means of Bayesian model comparison (Penny, 2012; Stephan et al., 2009a). A comprehensive derivation of the rDCM framework can be found elsewhere (Frässle et al., 2017; Lomakina, 2016).

In summary, rDCM can be seen as transforming a linear DCM in the time domain into a Bayesian linear regression model in the frequency domain. The advantage of this reformulation is that the likelihood function takes an algebraic form and no longer requires time-consuming integration. Furthermore, a consequence of the mean field approximation is that model inversion can take place one region at a time. Together with the analytic variational Bayesian update equations, this yields a dramatic increase in computational efficiency compared to classical DCM implementations. As the compute time scales polynomially with the number of regions and the dimensionality of the error covariance matrix for each region is fixed (independent of the number of regions), rDCM is capable, in principle, of analyzing very large networks.

Sparse regression DCM

In this section, we augment the standard rDCM approach described above with sparsity constraints. This *sparse rDCM* rests on introducing an additional set of binary random variables ζ that encode network structure. This effectively renders model inversion equivalent to pruning a full (all-to-all connected) network to a degree of optimal sparsity (that minimizes variational free energy, see below).

In sparse rDCM, for each region r, we define a diagonal matrix Z_r of these binary indicator variables as follows

$$Z_r^{(i,j)} = \begin{cases} \zeta_i, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Where $\zeta_i \in \{0, 1\}$. While there are several ways how these region-wise indicator matrices could be embedded into the probabilistic model of rDCM (for a comprehensive treatment of different possibilities, see Lomakina, 2016), a natural way is to cast Z_r as a feature selector in the likelihood function. Hence, extending the likelihood of rDCM specified in Eq. (4), the (Bayesian) sparse linear regression takes the form:

$$\begin{aligned} p(Y|\theta,\tau,X) &= \prod_{r=1}^{R} \mathcal{N}\left(Y_r; XZ_r\theta_r, \tau_r^{-1} I_{N\times N}\right) \\ Y_r &= (e^{2\pi i\frac{m}{N}} - 1)\frac{\widehat{y}_r}{T} \\ X &= \left[\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_R, \widehat{h}\widehat{u}_1, \widehat{h}\widehat{u}_2, \ldots, \widehat{h}\widehat{u}_K\right] \\ \theta_r &= \left[a_{r,1}, a_{r,2}, \ldots, a_{r,R}, c_{r,1}, c_{r,2}, \ldots, c_{r,K}\right] \\ Z_r &= \text{diag}\left(\left[\xi_{r,1}, \xi_{r,2}, \ldots, \xi_{r,R}, \xi_{r,R+1}, \ldots, \xi_{r,R+K}\right]\right) \end{aligned} \tag{7}$$

where each binary indicator variable ζ_i relates to a specific connectivity parameter i in a fully connected linear model and equals 1 if the connection is present – and thus contributes to explaining the measured BOLD data – and 0 if the connection is not involved in generating the observed data.

Generative model of sparse regression DCM

In order to turn the sparse linear regression model in Eq. (7) into a full generative model, we specify prior distributions over parameters, hyperparameters and binary indicator variables. In line with Frässle et al. (2017), we used a Gaussian prior on the neuronal connectivity parameters θ and a (conjugate) Gamma prior on the precision of the measurement noise τ. Furthermore, we assumed a Bernoulli prior on the binary indicator variables ζ_i. The prior densities then take the following form:

$$\begin{aligned} p(\theta) = \mathcal{N}(\theta; \mu_0, \Sigma_0) &= \frac{1}{\sqrt{(2\pi)^D |\Sigma_0|}} \exp\left(-\frac{1}{2}(\theta - \mu_0)^T \Sigma_0^{-1} (\theta - \mu_0)\right) \\ p(\tau) = \text{Gamma}(\tau; \alpha_0, \beta_0) &= \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \tau_i^{\alpha_0 - 1} \exp(-\beta_0 \tau) \\ p(\zeta_i) = \text{Bern}\left(\zeta_i; p_0^i\right) &= \left(p_0^i\right)^{\zeta_i} \left(1 - p_0^i\right)^{1-\zeta_i} \end{aligned} \tag{8}$$

Where D is the number of afferent connections and direct (driving) inputs entering the region, μ_0 and Σ_0 are the mean and covariance of the Gaussian prior on (neuronal) connectivity parameters, α_0 and β_0 are the shape and rate parameters of the Gamma prior on noise precision, Γ is the Gamma function (Bishop, 2006), and p_0^i is the parameter of the Bernoulli prior of the binary indicator variable for connectivity parameter *i*.

Intuitively, the p_0^i encode the *a priori* knowledge or belief about the connectivity pattern's degree of sparseness. In this initial work, we assume that the *a priori* probability of a connection being present is identical for all connections (but see the Discussion for potential ways of informing individual prior probabilities), with the exception of the inhibitory self-connections (i.e., diagonal entries of the A matrix), whose existence was enforced by setting $p_0^i = 1$. In simulations shown below, we systematically varied p_0^i to evaluate the impact of this choice on the accuracy of sparse rDCM. We propose a principled approach to optimizing p_0^i for empirical datasets (where the "true" degree of network sparsity is unknown) that performs model inversion for different p_0^i values and selects the one that yields the highest negative free energy.

In this paper, the prior densities over connectivity parameters and noise precision were defined as in Frässle et al. (2017). Specifically, we used the standard neuronal priors from DCM10 as implemented in the Statistical Parametric Mapping software package SPM8 (version R4290; www.fil.ion.ucl.ac.uk/spm). Additionally, we used $\alpha_0 = 2$ and $\beta_0 = 1$ as the shape and rate parameter of the Gamma distribution on noise precision, respectively.

Since sparse rDCM rests on the same generative model as the original rDCM framework, the model also factorizes across regions. Hence, the posterior distribution over model parameters for a single region r and for the entire model, respectively, takes the following form:

$$\begin{aligned} p(\theta_r, \tau_r, \zeta_r | Y_r, X) &\propto p(Y_r | X, \theta_r, \tau_r, Z_r) p(\theta_r) p(\tau_r) \prod_{i=1}^{D} p\left(\zeta_{r,i}\right) \\ p(\theta, \tau, \zeta | X, Y) &\propto \prod_{r=1}^{R} p(Y_r | X, \theta_r, \tau_r, Z_r) \prod_{r=1}^{R} \left(p(\theta_r) p(\tau_r) \prod_{i=1}^{D} p\left(\zeta_{r,i}\right)\right) \end{aligned} \tag{9}$$

Since there is no closed-form analytical solution for the posterior distribution, Eq. (9) cannot be solved exactly and one needs to resort to approximate inference. Here, we derive a variational Bayesian scheme for model inversion, which yields estimates of three quantities simultaneously: (i) an approximation to the posterior density over neuronal connectivity and noise precision parameters, (ii) a lower-bound approximation to the log model evidence, and (iii) a posterior belief about the sparsity of the network.

Variational Bayes for sparse regression DCM

This section presents the variational Bayesian update equations for the sufficient statistics of the variational (approximate) posterior distribution $q(\theta, \tau, \zeta | X, Y)$. To facilitate derivation of update equations for finding the optimal q, we assume a mean field approximation to the variational density. As already expressed in Eq. (4), this mean field approximation assumes independence across connection sets targeting different regions; additionally, it assumes mutual independence of parameters, hyperparameters and binary indicator variables. The approximate posterior for a single region r then takes the form:

$$q(\theta_r, \tau_r, \zeta_r | Y_r, X) \approx q(\theta_r | Y_r, X) q(\tau_r | Y_r, X) \prod_{i=1}^{D} q\left(\zeta_{r,i} | Y_r, X\right) \tag{10}$$

Under this factorization, one can derive the variational update equations by making use of the fundamental principle from variational calculus (Bishop, 2006):

$$\ln q(\vartheta_i) = \langle \ln p(\vartheta, y) \rangle_{q(\vartheta_{\backslash i})} \tag{11}$$

where ϑ denotes all model parameters. Eq. (11) states that the logarithm of the approximate marginal posterior over a particular set of model parameters *i* is given by the expected energy of the system (i.e., log joint) under the variational distributions over all other sets of parameters.

In the following, we present the variational update equations for the different parameter classes in sparse rDCM (neuronal connectivity, noise precision, and binary indicator variables). A detailed derivation of these equations is provided in the Appendix (A.1-A.7). Importantly, under the mean field approximation of sparse rDCM, optimization can be performed for each region independently. Hence, we restrict the update equations to a single region (and drop the subscript *r* to keep the notation uncluttered).

Update equation of θ

$$\begin{aligned}\ln q^{*}(\theta|Y,X) &= \langle \ln p(\theta,\tau,\zeta,Y|X)\rangle_{q(\tau,\zeta)} \\ &= -\frac{1}{2}\theta^{T}\left(\frac{\alpha_{\tau|y}}{\beta_{\tau|y}}\left(P_{\zeta|y}X^{T}XP_{\zeta|y}+(X^{T}X)\circ\left(P_{\zeta|y}-P_{\zeta|y}^{2}\right)\right)+\Sigma_{0}^{-1}\right)\theta \\ &\quad +\theta^{T}\left(\frac{\alpha_{\tau|y}}{\beta_{\tau|y}}P_{\zeta|y}X^{T}Y+\Sigma_{0}^{-1}\mu_{0}\right)+c\end{aligned} \tag{12}$$

Here, all terms independent of θ were absorbed into the constant term *c*. Additionally, we have made use of $\langle\tau\rangle_{q(\tau)}=\frac{\alpha_{\tau|y}}{\beta_{\tau|y}}$ and $\langle Z\rangle_{q(\zeta)}=P_{\zeta|y}$, with $\langle\cdot\rangle_{q}$ denoting the expected value with respect to the variational density $q(\tau)$ and $q(\zeta)$, respectively. A detailed derivation is given in Eq. (A.1) in the Appendix. Comparing Eq. (12) to the logarithm of the multivariate normal distribution, one can derive the update equations for the sufficient statistics of the variational density over (neuronal) connectivity parameters $q(\theta|X,Y)$. The respective expressions for the posterior mean and covariance are summarized in the final iterative update scheme in Eq. (15).

Update equation of τ

$$\begin{aligned}&\ln q^{*}(\tau|Y,X)=\langle\ln p(\theta,\tau,\zeta,Y|X)\rangle_{q(\theta,\ \zeta)} \\ &=-\frac{\tau}{2}\left(\left(Y-XP_{\zeta|y}\mu_{\theta|y}\right)^{T}\left(Y-XP_{\zeta|y}\mu_{\theta|y}\right)+\mathrm{tr}\left(P_{\zeta|y}X^{T}XP_{\zeta|y}\Sigma_{\theta|y}\right)\right) \\ &-\frac{\tau}{2}\left(\mu_{\theta|y}^{T}\left((X^{T}X)\circ\left(P_{\zeta|y}-P_{\zeta|y}^{2}\right)\right)\mu_{\theta|y}+\mathrm{tr}\left(\left((X^{T}X)\circ\left(P_{\zeta|y}-P_{\zeta|y}^{2}\right)\right)\Sigma_{\theta|y}\right)\right) \\ &+\frac{N}{2}\ln\tau+(\alpha_{0}-1)\ln\tau-\beta_{0}\tau+c\end{aligned} \tag{13}$$

Here, $\circ$ denotes the element-wise product of two matrices. All terms independent of τ were absorbed into the constant term *c*. Additionally, we made use of $\langle\theta\rangle_{q(\theta)}=\mu_{\theta|y}$, with $\langle\cdot\rangle_{q(\theta)}$ denoting the expected value with respect to the variational density $q(\theta)$. A detailed derivation is given in Eq. (A.2) in the Appendix. Comparing Eq. (13) to the logarithm of the Gamma distribution, one can derive update equations for the sufficient statistics of the variational density over noise precision $q(\tau|X,Y)$. Again, the respective expressions for the posterior shape and rate parameters are summarized in the final iterative scheme in Eq. (15).

Update equation of ζ_i

$$\begin{aligned}\ln q^{*}(\zeta_{i}|Y,X) &= \langle\ln p(\theta,\tau,\zeta,Y|X)\rangle_{q(\theta,\tau,\ \zeta_{\backslash i})} \\ &= \zeta_{i}\left(\frac{\alpha_{\tau|y}}{\beta_{\tau|y}}\mu_{\theta|y}\upsilon_{i}-\frac{\alpha_{\tau|y}}{2\beta_{\tau|y}}\left(\left(\mu_{\theta|y}^{i}\right)^{2}W_{ii}+2\mu_{\theta|y}^{i}\sum_{j\neq i}p_{\zeta|y}^{j}\mu_{\theta|y}^{j}W_{ij}\right)\right) \\ &\quad -\zeta_{i}\left(\frac{\alpha_{\tau|y}}{2\beta_{\tau|y}}W_{ii}\Sigma_{\theta|y}^{ii}+\frac{\alpha_{\tau|y}}{\beta_{\tau|y}}\sum_{j\neq i}p_{\zeta|y}^{j}W_{ij}\Sigma_{\theta|y}^{ij}-\ln\left(\frac{p_{0}^{i}}{1-p_{0}^{i}}\right)\right)+c\end{aligned} \tag{14}$$

Here, all terms independent of ζ_i were absorbed into the constant term *c*. Note that we made use of the fact that $\langle\zeta_j\rangle_{q(\ \zeta_{\backslash i})}$ is a constant with respect to ζ_i for all terms $j\neq i$. Additionally, we have set $\upsilon=X^{T}Y$ and $W=X^{T}X$ to ease readability. A detailed derivation is given in Eqs. (A.3)-(A.4) in the Appendix. Comparing Eq. (14) to the logarithm of the Bernoulli distribution, one can derive update equations for the sufficient statistic of the approximate posterior density over indicator variables $q(\zeta_i|X,Y)$.

Final iterative scheme:

$$\Sigma_{\theta|y}=\left(\frac{\alpha_{\tau|y}}{\beta_{\tau|y}}\left(P_{\zeta|y}X^{T}XP_{\zeta|y}+(X^{T}X)\circ\left(P_{\zeta|y}-P_{\zeta|y}^{2}\right)\right)+\Sigma_{0}^{-1}\right)^{-1}$$

$$\mu_{\theta|y}=\Sigma_{\theta|y}\left(\frac{\alpha_{\tau|y}}{\beta_{\tau|y}}P_{\zeta|y}X^{T}Y+\Sigma_{0}^{-1}\mu_{0}\right)$$

$$\alpha_{\tau|y}=\alpha_{0}+\frac{N}{2}$$

$$\begin{aligned}\beta_{\tau|y}&=\beta_{0}+\frac{1}{2}\left(\left(Y-XP_{\zeta|y}\mu_{\theta|y}\right)^{T}\left(Y-XP_{\zeta|y}\mu_{\theta|y}\right)+\mathrm{tr}\left(P_{\zeta|y}X^{T}XP_{\zeta|y}\Sigma_{\theta|y}\right)\right) \\ &+\frac{1}{2}\left(\mu_{\theta|y}^{T}\left((X^{T}X)\circ\left(P_{\zeta|y}-P_{\zeta|y}^{2}\right)\right)\mu_{\theta|y}+\mathrm{tr}\left(\left((X^{T}X)\circ\left(P_{\zeta|y}-P_{\zeta|y}^{2}\right)\right)\Sigma_{\theta|y}\right)\right)\end{aligned}$$

$$p_{\zeta|y}^{i}=\frac{1}{1+\exp(-g_{i})}$$

$$\begin{aligned}g^{i}&=\frac{\alpha_{\tau|y}}{\beta_{\tau|y}}\mu_{\theta|y}^{i}\upsilon_{i}-\frac{\alpha_{\tau|y}}{2\beta_{\tau|y}}\left(\left(\mu_{\theta|y}^{i}\right)^{2}W_{ii}+2\mu_{\theta|y}^{i}\sum_{j\neq i}p_{\zeta|y}^{j}\mu_{\theta|y}^{j}W_{ij}\right) \\ &-\frac{\alpha_{\tau|y}}{2\beta_{\tau|y}}\left(W_{ii}\Sigma_{\theta|y}^{ii}+2\sum_{j\neq i}p_{\zeta|y}^{j}W_{ij}\Sigma_{\theta|y}^{ij}\right)+\ln\left(\frac{p_{0}^{i}}{1-p_{0}^{i}}\right)\end{aligned} \tag{15}$$

From Eq. (15), we see that the update equations for the sufficient statistics of $q(\theta|X,Y)$, $q(\tau|X,Y)$ and $q(\zeta_i|X,Y)$ are mutually dependent on each other; hence, the optimal posterior distribution must be obtained by iterating over these update equations until convergence. In the current version of sparse rDCM, the iterative scheme proceeds until the change in τ falls below a critical threshold (i.e., 10^{-10}) or the number of iterations exceeds a pre-specified upper limit (i.e., 500 iterations).

Negative free energy for sparse regression DCM

Having derived the variational Bayesian update equations for the posterior densities, we now provide an expression for the negative free energy. The negative free energy can be cast as the sum of the expected energy of the system (log joint) under the variational density *q* and the entropy of *q* (Bishop, 2006; Friston et al., 2007):

$$\begin{aligned}F &= \langle\ln p(\theta,\tau,\zeta,Y|X)\rangle_{q(\theta,\tau,\zeta)}-\langle\ln q(\theta,\tau,\zeta|Y,X)\rangle_{q(\theta,\tau,\zeta)} \\ &= \langle\ln p(Y|\theta,\tau,\zeta,X)\rangle_{q(\theta,\tau,\zeta)}+\langle\ln p(\theta)\rangle_{q(\theta)}-\langle\ln q(\theta)\rangle_{q(\theta)}+\langle\ln p(\tau)\rangle_{q(\tau)} \\ &\quad -\langle\ln q(\tau)\rangle_{q(\tau)}+\sum_{i=1}^{D}\left(\langle\ln p(\zeta_{i})\rangle_{q(\zeta_{i})}-\langle\ln q(\zeta_{i})\rangle_{q(\zeta_{i})}\right)\end{aligned} \tag{16}$$

In the following, we present the expressions for the individual components of the negative free energy for a single region (see Lomakina, 2016):

Expectation of the likelihood

$$\begin{aligned}&\langle \ln p(\theta,\tau,\zeta,Y|X)\rangle_{q(\theta,\tau,\zeta)} = \left\langle \ln\mathscr{N}\left(Y;XZ\theta,\tau^{-1}I_{N\times N}\right)\right\rangle_{q(\theta,\tau,\zeta)}\\ &= -\frac{N}{2}\ln 2\,\pi + \frac{N}{2}\left(\Psi\left(\alpha_{\tau|y}\right) - \ln\beta_{\tau|y}\right)\\ &-\frac{\alpha_{\tau|y}}{2\beta_{\tau|y}}\left(\left(Y - XP_{\zeta|y}\mu_{\theta|y}\right)^T\left(Y - XP_{\zeta|y}\mu_{\theta|y}\right) - \mathrm{tr}\left(P_{\zeta|y}WP_{\zeta|y}\Sigma_{\theta|y}\right)\right)\\ &-\frac{\alpha_{\tau|y}}{2\beta_{\tau|y}}\left(\mu_{\theta|y}^T W\circ\left(P_{\zeta|y} - \left(P_{\zeta|y}\right)^2\right)\mu_{\theta|y} - \mathrm{tr}\left(W\circ\left(P_{\zeta|y} - \left(P_{\zeta|y}\right)^2\right)\Sigma_{\theta|y}\right)\right)\end{aligned} \tag{17}$$

Here, Ψ is the digamma function and all other variables are defined as above.

Expectation of the prior on θ:

$$\begin{aligned}&\langle \ln p(\theta)\rangle_{q(\theta)} = \langle \ln\mathscr{N}(\theta;\mu_0,\Sigma_0)\rangle_{q(\theta)}\\ &= -\frac{D}{2}\ln 2\,\pi - \frac{1}{2}\ln|\Sigma_0| - \frac{1}{2}\left(\mu_{\theta|y} - \mu_0\right)^T\Sigma_0^{-1}\left(\mu_{\theta|y} - \mu_0\right) - \frac{1}{2}\mathrm{tr}\left(\Sigma_0^{-1}\Sigma_{\theta|y}\right)\end{aligned} \tag{18}$$

Expectation of the prior on τ:

$$\begin{aligned}\langle \ln q(\tau)\rangle_{q(\tau)} &= \langle \mathrm{Gamma}(\tau;\alpha_0,\beta_0)\rangle_{q(\tau)}\\ &= \alpha_0\ln\beta_0 - \ln\Gamma(\alpha_0) + (\alpha_0 - 1)\left(\Psi\left(\alpha_{\tau|y}\right) - \ln\beta_{\tau|y}\right) - \beta_0\frac{\alpha_{\tau|y}}{\beta_{\tau|y}}\end{aligned} \tag{19}$$

Expectation of the prior on ζ_i:

$$\begin{aligned}\langle \ln p(\zeta_i)\rangle_{q(\zeta_i)} &= \left\langle \ln\mathrm{Bern}\left(\zeta_i;p_0^i\right)\right\rangle_{q(\zeta_i)}\\ &= \ln\left(1 - p_0^i\right) + p_{\zeta|y}^i\ln\frac{p_0^i}{1 - p_0^i}\end{aligned} \tag{20}$$

Entropy of θ:

$$\begin{aligned}-\langle \ln q(\theta)\rangle_{q(\theta)} &= -\left\langle \mathscr{N}\left(\theta;\mu_{\theta|y},\Sigma_{\theta|y}\right)\right\rangle_{q(\theta)}\\ &= \frac{D}{2}(1 + \ln 2\,\pi) + \frac{1}{2}\ln\left|\Sigma_{\theta|y}\right|\end{aligned} \tag{21}$$

Entropy of τ:

$$\begin{aligned}-\langle \ln q(\tau)\rangle_{q(\tau)} &= -\left\langle \mathrm{Gamma}\left(\tau;\alpha_{\tau|y},\beta_{\tau|y}\right)\right\rangle_{q(\tau)}\\ &= \alpha_{\tau|y} - \ln\beta_{\tau|y} + \ln\Gamma\left(\alpha_{\tau|y}\right) - \left(\alpha_{\tau|y} - 1\right)\Psi\left(\alpha_{\tau|y}\right)\end{aligned} \tag{22}$$

Entropy of ζ_i:

$$\begin{aligned}-\langle \ln q(\zeta_i)\rangle_{q(\zeta_i)} &= -\left\langle \ln\mathrm{Bern}\left(\zeta_i;p_{\zeta|y}^i\right)\right\rangle_{q(\zeta_i)}\\ &= -p_{\zeta|y}^i\ln p_{\zeta|y}^i - \left(1 - p_{\zeta|y}^i\right)\ln\left(1 - p_{\zeta|y}^i\right)\end{aligned} \tag{23}$$

A detailed derivation of the above equations is presented in Eqs. (A.5)-(A.11) in the Appendix. The negative free energy for each individual region is obtained by summing Eqs. 17–23. Summing over all regions of the model then yields the negative free energy for the complete model.

Classifying connections after model inversion

In contrast to classical DCM, our approach not only provides the posterior density over connectivity and hemodynamic parameters but also a posterior belief (a Bernoulli distribution parameterized by $p_{\zeta|y}^i$) about whether a connection i exists. For the interpretation of sparse rDCM results, this raises the question when to classify a connection as present or absent, given $p_{\zeta|y}^i$. Here, we base this decision on the posterior odds ratio

$$\frac{p_{\zeta=1|y}^i}{p_{\zeta=0|y}^i} = \frac{p_{\zeta=1|y}^i}{1 - p_{\zeta=1|y}^i} \tag{24}$$

and take a ratio larger than 10 (corresponding to >90.9% posterior probability of presence) as decisive evidence for a connection being present. Conversely, if the ratio is below 0.1 (corresponding to >90.9% posterior probability of absence), we consider this as decisive evidence in favor of pruning connection *i* from the model. For posterior odds ratios that fall in between these boundaries, decisions are less clear. In this "grey area", classifying a connection as present or absent essentially boils down to whether one wishes to maximize sensitivity or specificity. This resembles the situation in classical (frequentist) hypothesis testing where the significance threshold dictates sensitivity and specificity of a null hypothesis test.

For any given connection, the posterior odds ratio not only depends on the data, but also on the choice of the Bernoulli prior $p(\zeta_i)$; the latter is parameterized by p_0^i, the prior probability of connection *i* being present. Below, we report simulation results that demonstrate how the choice of p_0^i and the SNR of the data impact on sensitivity and specificity of our method. In order to choose p_0^i in a principled manner and ensure an adequate level of sparsity in rDCM, we use a standard procedure for choosing hyperparameters in a Bayesian setting. This is known in the statistics literature as maximum likelihood II estimation or empirical Bayes (Berger, 1985; Gelman et al., 2004; Murphy, 2012). This procedure rests on maximizing the log marginal likelihood (log model evidence) that is obtained by integrating over the model parameters (Bishop, 2006). In other words, the optimal value of the sparsity hyperparameter p_0^i maximizes the log model evidence (which, in this work, is approximated by the negative free energy).

Furthermore, to illustrate the impact of the above (uniform) decision rule on the sensitivity and specificity of our model, we compared both options – that is, interpreting all connections with "grey area" values for $p_{\zeta|y}^i$ as either present or absent, respectively.

Computational complexity

Inspection of the variational Bayesian update equations of sparse rDCM (Eq. (15)) reveals that the computationally most expensive operation is the matrix inversion for computing the posterior covariance matrix $\Sigma_{\theta|y}$. Efficient algorithms, such as the Coppersmith-Winograd algorithm (Coppersmith and Winograd, 1990) or the Optimized CW-like algorithm (Davie and Stothers, 2013) allow matrix inversion with complexity $O(n^{2.4})$, where n is the number of columns/rows of a square matrix. For sparse rDCM, $n = R + K$ with R representing the number of regions and K the number of driving inputs. We can assume $n \approx R$ since the number of driving inputs is typically much smaller than the number of regions. The update equations in Eq. (15) apply for a single region and thus have to be repeated for all regions to obtain the posterior densities for the complete model. Consequently, the complexity of sparse rDCM for the inversion of a fully (all-to-all) connected model is approximately $O(n^{3.4})$. It is worth pointing out that the optimal posterior distribution for each region is obtained by iterating over the variational Bayesian update equations in Eq. (15) until convergence. However, since our implementation of sparse rDCM sets a fixed upper limit on the number of iterations, this enters only as a constant factor into the complexity analysis and therefore does not show up in the O-notation.

Based on this complexity analysis, one can provide a rough estimate of the run-time of sparse rDCM for a model with a far higher number of regions than examined in this paper. For example, assuming that model inversion runs in parallel on 16 cores (as in our current implementation), sparse rDCM would take approximately 36 h to infer the effective connectivity in a network comprising 1000 brain regions. This compares favorably to other large-scale models, such as the 36-region spectral DCM by Razi et al. (2017) for which the authors reported 24 h of runtime on a high-performance cluster.

Synthetic data

To demonstrate the face validity of sparse rDCM, we performed systematic simulation studies to assess how recovery of known network architecture in large (whole-brain) networks depends on the signal-to-noise ratio (SNR) of synthetic fMRI data and the choice of prior probability p_0^i (the parameter of the Bernoulli prior on binary indicator variables). For comparison and completeness, we also provide simulations for small neural networks (of the typical size for conventional DCM analyses) in the Supplementary Material.

Here, we constructed four different linear DCMs. First, a *grid-like* DCM comprising 66 brain regions was constructed, where each network node was connected to its direct neighbors – thus, four connections entered any given brain region (Fig. 1A). While the systematic structure of the grid-like DCM makes it easy to visualize the scalability of sparse rDCM to large networks, the model lacks typical network characteristics of the human brain, for instance, with regard to small-world architecture, node degree, path length, centrality of nodes, or modularity (Bullmore and Sporns, 2009). Two additional *small-world* DCMs, comprising 50 brain regions, were therefore created based on the simulation study by Smith et al. (2011). Specifically, we used their S50 network which consists of 10 local sub-networks (each comprising 5 nodes connected in a ring – although not with cyclic directionality of influences) that were connected via one long-range connection to model the small-world architecture of the human brain (Fig. 1B). Notably, the S50 model assumes only unidirectional connections among network nodes and thus neglects reciprocal interactions which are known to play an important role for functional integration (e.g., forward and backward connections in cortical hierarchies; Felleman and Van Essen, 1991; Zeki and Shipp, 1988). We therefore created an additional variant of the S50 model that replaced all unidirectional with reciprocal connections (Supplementary Figure S1). Finally, to further increase the biological realism of the synthetic DCMs, a

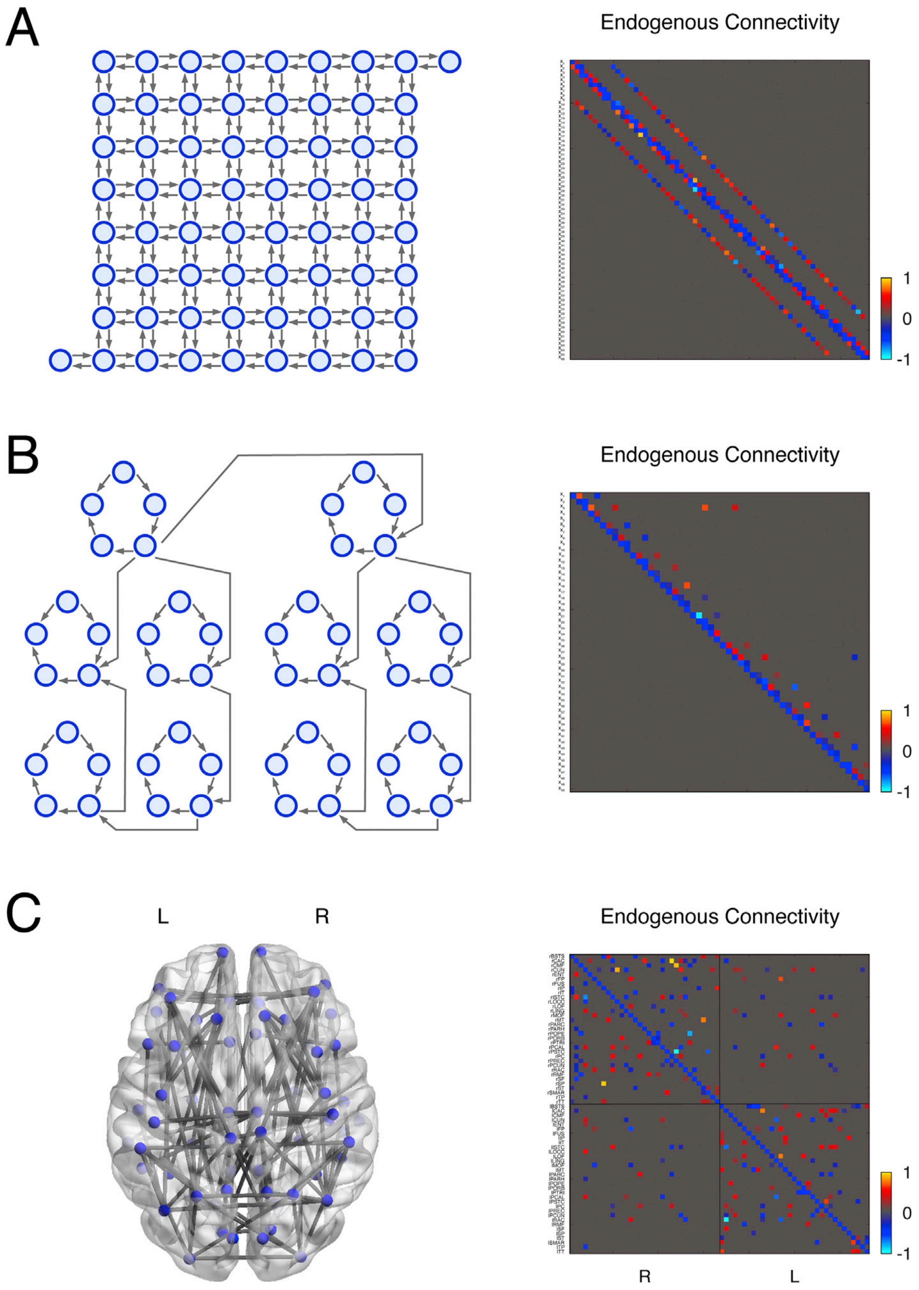


Fig. 1. Connectivity architecture of the large-scale networks used for simulations. (A) Grid-like DCM, comprising 66 network nodes, where each node is connected only to its direct neighbors, yielding merely four connections entering any given region (*left*). An actual instance of the effective connectivity structure, generated by sampling connection strengths from the prior density on the A matrix parameters (*right*). (B) The respective illustrations for the original small-world DCM, comprising 50 network nodes, which was initially introduced as the S50 model in Smith et al. (2011). This model consists of 10 local sub-networks (each comprising 5 nodes connected in a ring via unidirectional links – although not with cyclic directionality of influences) that were connected via one long-range connection to mimic the small-world architecture of the human brain. (C) The respective illustrations for the connectome-based DCM, comprising 66 network nodes, whose structure was based on a real human structural connectome given by Hagmann et al. (2008). Specifically, connections were restricted to the most pronounced structural links by only selecting those connections for which an average inter-regional fiber density larger than 0.06 has been reported. The brain network was visualized using the BrainNet Viewer (Xia et al., 2013), which is freely available for download (http://www.nitrc.org/projects/bnv/).

66-regions DCM was constructed based on a whole-brain atlas and the human connectome provided by the diffusion-weighted imaging work by Hagmann et al. (2008). For the endogenous connectivity structure (A matrix) of this *connectome-based* DCM, all connections from the matrix of average inter-regional fiber densities with a weight larger than 0.06 were included (Fig. 1C). This threshold was chosen to ensure that the system remained stable (i.e., satisfied the Lyapunov stability criterion) under random sampling from the prior density.

For all models, block input regressors were then used as driving inputs to half of the brain regions to ensure that the effect of experimental manipulations was pronounced in all network nodes. In total, this resulted in 327 (grid-like DCM), 136 and 197 (original and reciprocal small-world DCM, respectively), and 345 (connectome-based DCM) non-zero model parameters (A- and C-parameters).

For all models, we then systematically evaluated the accuracy of sparse rDCM to recover the true network architecture as a function of the SNR of fMRI data and the prior probability p_0^i. We simulated synthetic BOLD signal time series under various settings of the SNR (1, 3, 5, 10, and 100) and a fixed TR = 0.5s. For each of the two models and each SNR setting, we generated 20 different sets of BOLD signal time series (synthetic "subjects") by sampling the "true" (data-generating) parameter values from the prior distribution over connections (A matrix) and driving input (C matrix) parameters. Synthetic DCMs were then evaluated as follows: sparse rDCM initially assumed a fully connected network – that is, all brain regions were linked by reciprocal connections. This yielded a total of 4356 and 2500 free connectivity parameters (including those of self-connections) for the 66-region and 50-region networks, respectively, to be estimated. Model inversion was performed under various settings of the prior probability p_0^i (0.05–0.95, with step size of 0.05) for all connections, except for the inhibitory self-connections; the existence of the latter was enforced by setting $p_0^i = 1$. Furthermore, to focus our examination of model inversion on the connectivity, driving inputs were fixed to the true target regions by setting p_0^i to 1 for the true driving input parameters, and to 0 for all other entries of the C matrix. In other words, the pattern of driving inputs was assumed to be known *a priori*, whereas connections had to be inferred from the data by automatically pruning the fully connected A matrix.

Empirical data: small networks

In a next step, we applied sparse rDCM to empirical fMRI data. First, we restricted our analyses to small networks to evaluate the utility of the approach for models that are of typical size for conventional DCM analyses. We used two previously published fMRI datasets: First, the "attention-to-motion" dataset, which has been employed to introduce various methodological developments, including structural equation models (SEM; Büchel and Friston, 1997; Penny et al., 2004b) and several variants of DCM for fMRI (Friston et al., 2003, 2014a; Li et al., 2011; Marreiros et al., 2008; Penny et al., 2004b; Stephan et al., 2008). Second, a dataset of stroke patients with aphasia, which has been used for model-based classification by generative embedding (Brodersen et al., 2011). In what follows, we briefly summarize the most relevant information for both datasets – details can be found elsewhere (attention-to-motion: Büchel and Friston, 1997; aphasia: Brodersen et al., 2011; Schofield et al., 2012).

Attention to motion

The "attention-to-motion" fMRI data are from a single subject in a visual attention study. The experiment included four conditions: (i) fixation only, (ii) presentation of stationary dots (*static*) (iii) passive observation of dots moving radially (i.e., away from the center), at a fixed speed (*passive*), and (iv) attention to non-existent changes in the speed of the radially moving dots (*attention*). The order of the conditions alternated between fixation and visual stimulation (i.e., static, passive, or attention). In all conditions, the subject had to fixate the center of the screen and no overt responses were required.

A total of 360 functional images were acquired on a 2-T MR scanner (Siemens Magnetom VISION) using a T_2*-weighted gradient echo-planar-imaging (EPI) sequence (TR = 3220 ms, TE = 40 ms, 32 axial slices, voxel size $3 \times 3 \times 3$ mm^3). fMRI data were analyzed using a first-level General Linear Model (GLM; Friston et al., 1995) with the following three regressors: (i) "photic" (static + passive + attention), (ii) "motion" (passive + attention), and (iii) "attention". Consistent with previous analyses of the same dataset, three regions of interest (ROIs) were defined, representing primary visual cortex (V1), motion-sensitive area V5, and attention-sensitive superior parietal cortex (SPC). From these ROIs, BOLD signal time series were extracted as the principal eigenvariate, which then entered effective connectivity analyses.

For the sparse rDCM analyses, we assumed a fully connected network – that is, all three regions were connected reciprocally. Driving inputs were specified as follows: (i) photic input elicited activity in V1, (ii) motion input drove activity in V5, and (iii) attention input targeted SPC. Model inversion using sparse rDCM then pruned the fully connected network to a sparser representation.

Aphasia

Twenty-six right-handed subjects with normal hearing abilities and no history of neurological disease (12 female, mean age: 54.1 years, age range: 26–72 years), and eleven patients with moderate aphasia due to left-hemisphere stroke (1 female, mean age: 66.1 years, age range: 45–90 years) participated in an auditory fMRI paradigm. The patients' aphasia profile was based on the Comprehensive Aphasia Test (CAT; Swinburn et al., 2004). All subjects listened to auditory stimuli consisting of word pairs either presented in normal (forward) or time-reversed sequence. While time-reversed stimuli contained the same (low-level) characteristics as normal speech stimuli (e.g., speaker identity, spectral complexity), they were incomprehensible and served as control condition. To ensure engagement throughout the experiment, subjects were assigned an incidental task, asking them to report the gender of the speaker for each stimulus via button press.

For each subject, a total of 488 functional images were acquired on a 1.5-T MR scanner (Siemens Sonata) using a T_2*-weighted EPI sequence (TR = 3150 ms, TE = 50 ms, 35 axial slices, voxel size $3 \times 3 \times 2$ mm^3, inter-slice gap 1 mm). For each subject, BOLD activations were then analyzed by means of a first-level GLM (Friston et al., 1995) with the following regressors: (i) all auditory events (i.e., normal and time-reversed speech), and (ii) intelligibility (i.e., normal vs. time-reversed speech) as a parametric modulation. From the "all auditory" contrast based on the first regressor, six regions of interest (ROIs), representing key components in the auditory hierarchy, were defined – namely, bilateral medial geniculate body (MGB), Heschl's gyrus (HG), and planum temporale (PT). Hence, the DCMs concerned processing of acoustic stimuli *per se*, not differentiating between normal and time-reversed speech. Importantly, lesions in the aphasic patients were located outside the neural network underlying speech processing and thus did not affect the regions included in the DCM analysis (Brodersen et al., 2011; Schofield et al., 2012). From the individual ROIs, time series were extracted as the principal eigenvariate.

For the sparse rDCM analyses, we assumed that all six regions of interest (i.e., MGB, HG, and PT, each in both hemispheres) were fully connected via reciprocal connections and that the driving input (auditory stimulation) elicited activity in all six regions. Starting from these full A and C matrices, sparse rDCM was used to prune network structure. The ensuing posterior means from each subject were used to create a generative score space for a discriminative classifier. Within this generative score space, a linear kernel, representing the inner product $k(x_i,x_j) = \langle x_i, x_j \rangle$, was used to compare two instances (subjects). A support vector machine (SVM) was then used for classification (LIBLINEAR; Fan et al., 2008). In order to pinpoint which DCM parameters enabled discrimination between healthy controls and aphasic patients, we used an l_1-regularizer for the SVM. This entails sparse solutions by using only a minimal subset of data features for classification and thus fosters a straightforward analysis of the most discriminative connectivity and driving input parameters. Classification performance was evaluated by means of a leave-one-out cross-validation procedure similar to the one described in Brodersen et al. (2011).

Empirical data: whole-brain DCM

In a final analysis, we road-tested the utility of sparse rDCM for inferring the sparse effective connectivity pattern in a realistic whole-brain network based on empirical data. For this, we used a single-subject dataset from an fMRI study with a simple task that activates a well-known network, thus allowing us to assess the plausibility of the sparse whole-brain connectivity pattern provided by our method.

The fMRI data were acquired using a block design, asking subjects to perform visually synchronized whole-hand fist closings with either their left or right hand. At the beginning of each block, an arrow informed subjects which hand to use in the upcoming block. The arrow then started to blink at a rate of 1.25 Hz for 16 s, dictating the rhythm of subjects' hand movements (i.e., 20 fist closings per block). Subsequent blocks were interleaved with a resting period of the same length where subjects did not perform hand movements and kept visual fixation. The experiment consisted of two separate sessions and each session comprised only hand movements of the same condition (i.e., left or right). This renders the present dataset a particularly suitable candidate for testing the current implementation of sparse rDCM since no modulatory influences are necessary. In this proof of concept, we analyzed data from the left-hand movement session of a single representative healthy subject (a comprehensive analysis of the whole-brain effective connectivity for the entire group dataset will be presented in future work, Frässle et al., *in preparation*).

A total of 230 functional images were acquired on a 7-T MR scanner (Philips Achieva) using a T_2*-weighted EPI sequence ($TR = 2000$ ms, $TE = 25$ ms, 36 axial slices, voxel size $1.77 \times 1.77 \times 3\ mm^3$). fMRI data were analyzed using a first-level GLM (Friston et al., 1995) with a regressor encoding left-hand fist closing movements. We used the Automated Anatomical Labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002) as a whole-brain parcellation scheme to define anatomical regions for subsequent effective connectivity analyses. This resulted in 104 regions from which BOLD signal time series were extracted as the principal eigenvariate (after removing signal mean and correcting for head movements), which then entered sparse rDCM analyses.

We assumed a fully connected network, with all 104 regions coupled to each other via reciprocal connections. Additionally, the driving input (representing visually cued left-hand fist closing movements) was allowed to elicit activity in all regions. In other words, we assumed full A and C matrices, yielding more than 10,000 free parameters to be estimated. Starting from this fully connected network, sparse rDCM was then used to automatically prune both connections and driving inputs, leading to a sparse whole-brain effective connectivity pattern during left-hand movements.

Results

Synthetic data

First, we tested how accurately sparse rDCM could recover a known network architecture for large (whole-brain) models under various settings of SNR and the prior probability p_0^i, the parameter of the Bernoulli prior on binary indicator variables (for an analysis of the face validity for small networks, see Supplementary Material). We evaluated the sensitivity and specificity of identifying the known network architecture (i.e., the connections that were present in the data-generating model). As explained in the Methods, the decision of whether a given connection existed was based on the posterior odds ratio, using thresholds >10 and < 0.1 for declaring a connection as present or absent, respectively. For all other connections, with posterior odds ratios in a “grey zone” between 10 and 0.1, we applied a uniform decision rule, either interpreting these connections as present (which increases sensitivity) or as absent (which increases specificity).

Grid-like DCM

We constructed a *grid-like* network where each network node was coupled to its direct neighbors (Fig. 1A). This resulted in a regular connectivity pattern with four afferent connections for any given brain region. We deliberately chose such a rather simple network architecture as a starting point in order to allow for easy visualization of the performance of sparse rDCM. It is worth reiterating that sparse rDCM analyses initially assume that all regions are reciprocally connected – hence, this particular model with its 66 areas contained 4356 free connectivity parameters that were to be estimated.

Supplementary Figure S2 shows sensitivity and specificity of sparse rDCM for the grid-like DCM as a function of SNR and p_0^i. Panels A and B differ in the decision rule, i.e., connections with posterior odds ratios in a “grey zone” between 10 and 0.1 (see Methods) were either interpreted as present (Supplementary Figure S2A and Supplementary Table S1) or absent (Supplementary Figure S2B and Supplementary Table S2). In both panels, the expected dependency of sensitivity and specificity on SNR can be seen clearly when p_0^i is close to the true degree of sparseness of the grid-like DCM (approx. 0.07). For this value of p_0^i the decision rule did not matter much: For very high SNR ($SNR = 100$), sparse rDCM showed very high sensitivity (0.91 ± 0.03; Supplementary Figure S2A-B, top) and near-perfect specificity (0.97 ± 0.02; Supplementary Figure S2A-B, bottom). For more realistic data quality ($SNR = 3$), the sensitivity was reduced (0.58 ± 0.04; Supplementary Figure S2A-B, top), while the specificity remained close to perfect (0.99 ± 0.003; Supplementary Figure S2A-B, bottom). For challenging scenarios with low signal-to-noise ($SNR = 1$), sparse rDCM frequently failed to detect the true connectivity parameters and provided an overly conservative explanation of the fMRI data by pruning most connections from the model and accounting for task-related variability in BOLD signals by adjusting the weights of the driving inputs (for which the prior probability p_0^i was set to 1).

As p_0^i was increased to intermediate values, the decision rule for “grey zone” connections exerted a more pronounced effect on the performance of sparse rDCM. This is because, in this intermediate regime, individual connections tended to have posterior probabilities $p_{\zeta|y}^i$ close to 0.5. Hence, depending on whether these connections were classified as present or absent, sparse rDCM yielded either dense (sensitive; Supplementary Figure S2A) or sparse (specific; Supplementary Figure S2B) solutions, respectively.

When p_0^i moved close to 1, the known effective connectivity pattern was identified with perfect sensitivity while specificity dropped to 0. In other words, in this case model inversion trivially resulted in a full graph where all connections were inferred to be present (regardless of SNR).

The simulations above illustrate that the performance of sparse rDCM for the grid-like DCM depends on how well the assumed sparsity (i.e., the hyperparameter p_0^i) matches the actual sparsity of the network. As highlighted above (see Methods), a principled way of selecting the optimal p_0^i is to choose the hyperparameter such that the log model evidence is maximized (maximum likelihood II or empirical Bayes; Berger, 1985; Gelman et al., 2004; Murphy, 2012). Here, we approximated the log evidence by the negative free energy (F) and compared its values obtained under different p_0^i settings by evaluating the sum of log evidences across the "synthetic" subjects (i.e., a fixed-effects analysis; Stephan et al., 2009a). Across all SNR levels, BMS assigned the highest posterior model probability to p_0^i values that were close to the true sparseness of the network (Supplementary Figure S2C). Notably, it is also these p_0^i values that were optimal with regard to the trade-off between sensitivity and specificity for recovering the data-generating effective connectivity pattern. This suggests that in a first step of empirical sparse rDCM analyses, p_0^i can be determined by inverting the model under different possible p_0^i values (using a suitably defined grid in the range 0..1) and performing BMS.

Small-world DCMs

In subsequent simulations, we created two additional whole-brain models (*small-world* DCMs) that derived from the most complex (S50) model from Smith et al. (2011), which captures the small-world architecture of the human brain. First, we used the original S50 model that assumes unidirectional connections among network nodes (Fig. 1B). Second, we replaced all unidirectional influences with bidirectional connections to account for the typically reciprocal nature of functional integration in the human brain (Supplementary Figure S1). The degree of sparseness of the original and reciprocal small-world DCM were approx. 0.05 and 0.07, respectively.

In brief, across all settings, the results for the original small-world DCM (Fig. 2 and Supplementary Tables S3-S4) and the reciprocal small-world DCM (Fig. 3 and Supplementary Tables S5-S6) were similar to (and overall better than) the grid-like DCM. Figs. 2 and 3 again show sensitivity and specificity as a function of SNR and p_0^i, with panels A and B differing in the decision rule (see Methods). As for the grid-like DCM, we observed the expected dependence on SNR and p_0^i. When p_0^i was close to the true degree of sparseness, the decision rule did not matter. For this setting, sparse rDCM showed high sensitivity (0.88 ± 0.04; Fig. 2A–B, top) and near-perfect specificity (0.99 ± 0.01; Fig. 2A–B, bottom) for the original small-world DCM for the case of realistic data quality ($\text{SNR} = 3$). For the reciprocal small-world DCM, sensitivity was slightly decreased but remained reasonably high (0.73 ± 0.07; Fig. 3A–B, top), and specificity was still close to perfect (0.98 ± 0.01; Fig. 3A–B, bottom).

As for the grid-like DCM, we tested whether the negative free energy would recover the known p_0^i of the small-world DCMs. Fixed-effects BMS selected – with posterior model probabilities equal to 1 – p_0^i values that were close to the true degree of sparseness of the effective connectivity pattern of the original (Fig. 2C) and reciprocal small-world DCM (Fig. 3C). Again, it was also these p_0^i values that were optimal with regard to the trade-off between sensitivity and specificity of sparse rDCM.

Finally, at the request of one reviewer, we compared the performance of sparse rDCM for the original small-world S50 network from Smith et al. (2011) to three other methods that infer directed interactions in large-scale brain networks from fMRI data: Multivariate Granger causality (MVGC; Goebel et al., 2003; Roebroeck et al., 2005; Seth, 2010), (ii) Fast Adjacency Skewness (FASK; Sanchez-Romero et al., 2018), and (iii) Fast Greedy Equivalence Search (FGES; Chickering, 2003; Ramsey et al., 2017; Ramsey et al., 2010). In brief, for most settings, sparse rDCM showed comparable or better sensitivity than these approaches. One exception was the low SNR case ($\text{SNR} = 1$), where MVGC showed significantly higher sensitivity. For higher SNR values ($\text{SNR} \geq 3$), sparse rDCM outperformed all other methods in terms of sensitivity, when it was allowed to estimate driving inputs (Supplementary Figure S11A, top); when it was not allowed to account for driving inputs, sparse rDCM performed equivalently to MVGC and superior to FGES and FASK (Supplementary Figure S11B, top). Specificity was relatively similar across methods, with a small advantage for sparse rDCM at low SNR (Supplementary Figure S11A-B, bottom). For a detailed description of how data were simulated and analyzed, please see the Supplementary Material.

Connectome-based DCM

The grid-like DCM lacks biological realism in that it does not capture typical network characteristics of the human brain (Bullmore and Sporns, 2009). The S50 variants represent an improvement as they reflect small-world topology; yet, they do not fully capture the anisotropic and irregular structure of the brain's connectivity. In a next simulation, we therefore created a whole-brain model using the structural connectome based on the diffusion-weighted imaging work by Hagmann et al. (2008) (Fig. 1C). This *connectome-based* DCM captures more closely the functional organization of the human brain since it derives from a biologically plausible whole-brain structural network. The connectome-based DCM had a near-identical degree of sparsity (approx. 0.07) as the grid-like DCM.

In brief, across all settings, the results for the connectome-based DCM (Fig. 4 and Supplementary Tables S7-S8) were again similar to the grid-like and small-world DCMs. Most notably, as before, when p_0^i was close to the true degree of sparseness, the decision rule did not matter, sensitivity (0.46 ± 0.04; Fig. 4A–B, top) was considerably lower than specificity (0.99 ± 0.003; Fig. 4A–B, bottom) for the case of realistic data quality ($\text{SNR} = 3$), and the expected dependence on SNR was clearly visible.

As for the other DCMs, we tested whether the negative free energy could recover the known p_0^i of the connectome-based DCM. Again, fixed-effects BMS selected p_0^i values that were close to the true degree of sparseness of the effective connectivity pattern with posterior model probabilities equal to 1 (Fig. 4C) and resulted in an optimal trade-off between sensitivity and specificity of sparse rDCM.

Estimation of connectivity and driving input architecture

In the above simulations, driving inputs were fixed to the true target regions by setting the sparsity hyperparameter p_0^i to 1 for the true driving input parameters and to 0 for all other entries of the C matrix, in order to focus our examination of model inversion on the network connectivity (A matrix). In a final simulation, we tested the accuracy of sparse rDCM for a scenario where the driving inputs were no longer fixed, but where the p_0^i of the driving input parameters was also varied from 0.05 to 0.95 (with step size of 0.05). In other words, this simulation assessed whether connections *and* driving inputs can be inferred simultaneously from the data by automatically pruning the full A and C matrices. For this, we re-used the synthetic data from the original small-world DCM, which corresponds to the S50 model in Smith et al. (2011) and captures small-world architecture as seen in the human brain (Fig. 1B).

In brief, across all settings, the driving input architecture for the original small-world DCM could be reliably recovered, without notably compromising the accuracy of identifying the connectivity structure (Supplementary Figure S3 and Supplementary Tables S9-S11), as compared to the results obtained under fixed driving inputs (cf. Fig. 2). Supplementary Figure S3 again shows sensitivity and specificity as a function of SNR and p_0^i. Note that only the results for the "sensitive" decision rule (see Methods) are shown here, because the decision rule again did not matter when p_0^i was close to the true degree of sparseness of the original small-world DCM.

For the p_0^i that was optimal in terms of the highest negative free energy, sparse rDCM showed reasonably high sensitivity for identifying the connectivity architecture (0.69 ± 0.05; Supplementary Figure S3A, top) and high sensitivity for the driving inputs (0.90 ± 0.07; Supplementary Figure S3B, top) for the case of realistic data quality ($\text{SNR} = 3$). Specificity remained close to perfect for identifying both the connectivity (0.99 ± 0.01; Supplementary Figure S3A, bottom) and the driving inputs (0.99 ± 0.01; Supplementary Figure S3B, bottom).

Overall, these simulation results for the grid-like, small-world and connectome-based DCMs suggest that sparse rDCM is a suitable, albeit conservative, tool for inferring sparse whole-brain effective connectivity patterns from fMRI data. Under an appropriately chosen p_0^i (as can be

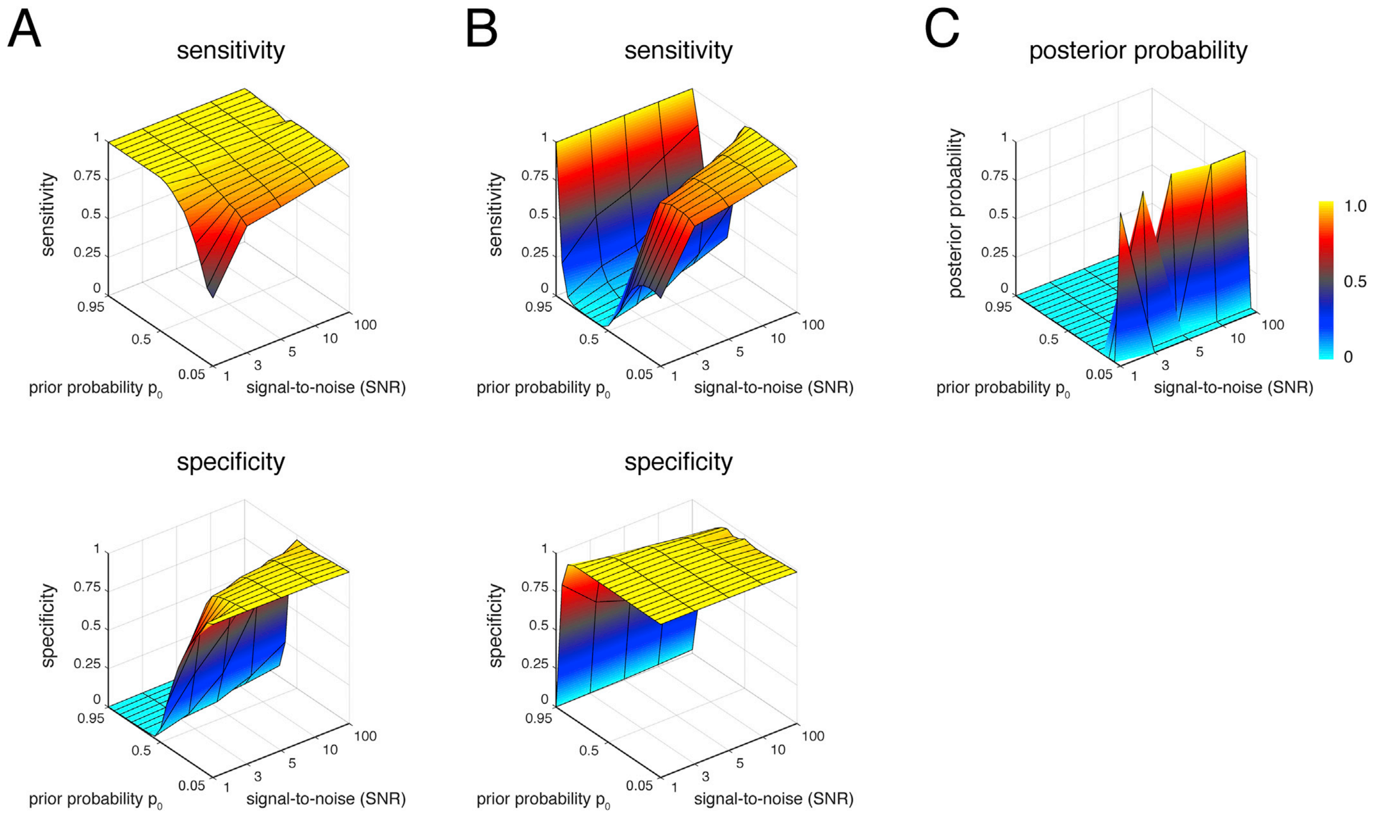


Fig. 2. Model architecture recovery of sparse rDCM in terms of the sensitivity and specificity for the original small-world DCM. (A) Sensitivity (*top*) and specificity (*bottom*) of identifying the true (data-generating) connections of the original small-world DCM, classifying a connection as present when its posterior odds ratio fell within the "grey zone" between 10 and 0.1 (see Methods). Sensitivity and specificity are shown along the z-axis as surface plots for various combinations of the signal-to-noise ratio (SNR) of the simulated fMRI data and the parameter p_0^i of the Bernoulli prior on binary indicator variables. The various SNR settings (1, 3, 5, 10, and 100) are shown along the x-axis of each subplot. The different p_0^i settings (0.05–0.95, with step size of 0.05) are shown along the y-axis of each subplot. The same plot is shown in (B) when classifying a connection as absent when its posterior odds ratio fell within the "grey zone" between 10 and 0.1 (see Methods). (C) Fixed effects Bayesian model selection results for the various SNR settings based on the negative free energy. The posterior model probabilities for all combinations of SNR and p_0^i settings are shown. See Fig. 1B for a visualization of the network structure of the original small-world DCM.

achieved by model selection), sensitivity of sparse rDCM was only moderate for realistic SNRs, but yielded close to perfect specificity, with hardly any false positives occurring. This suggests that not all effective connections may be identified by sparse rDCM; those connections that are detected, however, likely represent real effects. Our simulations further highlight the importance of SNR-boosting measures, such as optimized experimental designs, careful correction for physiological noise and head movements, specific scanner hardware (e.g., high-field MRI or magnetic field sensing; Bollmann et al., 2017) and/or optimized acquisition sequences (e.g., matched-filter acquisition; Kasper et al., 2014).

Computational burden

To illustrate the computational efficiency of sparse rDCM, we computed the run-times for the grid-like, small-world and connectome-based DCMs. We evaluated the run-times for all different settings of SNR, under a fixed TR of 0.5s. Notably, the reported run-times are only meant to provide a rough indication and depend on the computer hardware and software settings used. Here, each model was inverted on a single processor core (without parallelization) on the Euler cluster at ETH Zurich (for details, see https://scicomp.ethz.ch/wiki/Euler). Model inversion under sparse rDCM was highly efficient, taking between 2 and 13 min for the 50-regions small-world DCMs and between 4 and 18 min for the 66-regions grid-like and connectome-based DCMs (across the different SNR settings).

Empirical data: small networks

"Attention to motion" dataset

In a next step, we applied sparse rDCM to the "attention-to-motion" dataset (Büchel and Friston, 1997). We chose to use this dataset because it has been extensively studied with other methods of connectivity, including structural equation modeling (SEM; Büchel and Friston, 1997; Penny et al., 2004b), autoregressive models (Harrison et al., 2003), and different variants of DCM for fMRI (Friston et al., 2003, 2014a; Li et al., 2011; Marreiros et al., 2008; Penny et al., 2004b; Stephan et al., 2008). Here, sparse rDCM was used to invert a fully connected model (Fig. 5A). Reciprocal connections between V1 and SPC were pruned from the model, whereas all other connectivity and driving input parameter estimates had non-negligible values (Fig. 5B).

To test whether the automatic removal of V1-SPC connections was plausible, we tested whether the sparser model provided a more convincing explanation of the fMRI data than the full model. We inverted both the full and the sparser model (without V1-SPC connections) using the default VBL implementation in DCM10 (as implemented in SPM8, version R4290) and used the negative free energy for BMS (Penny et al.,

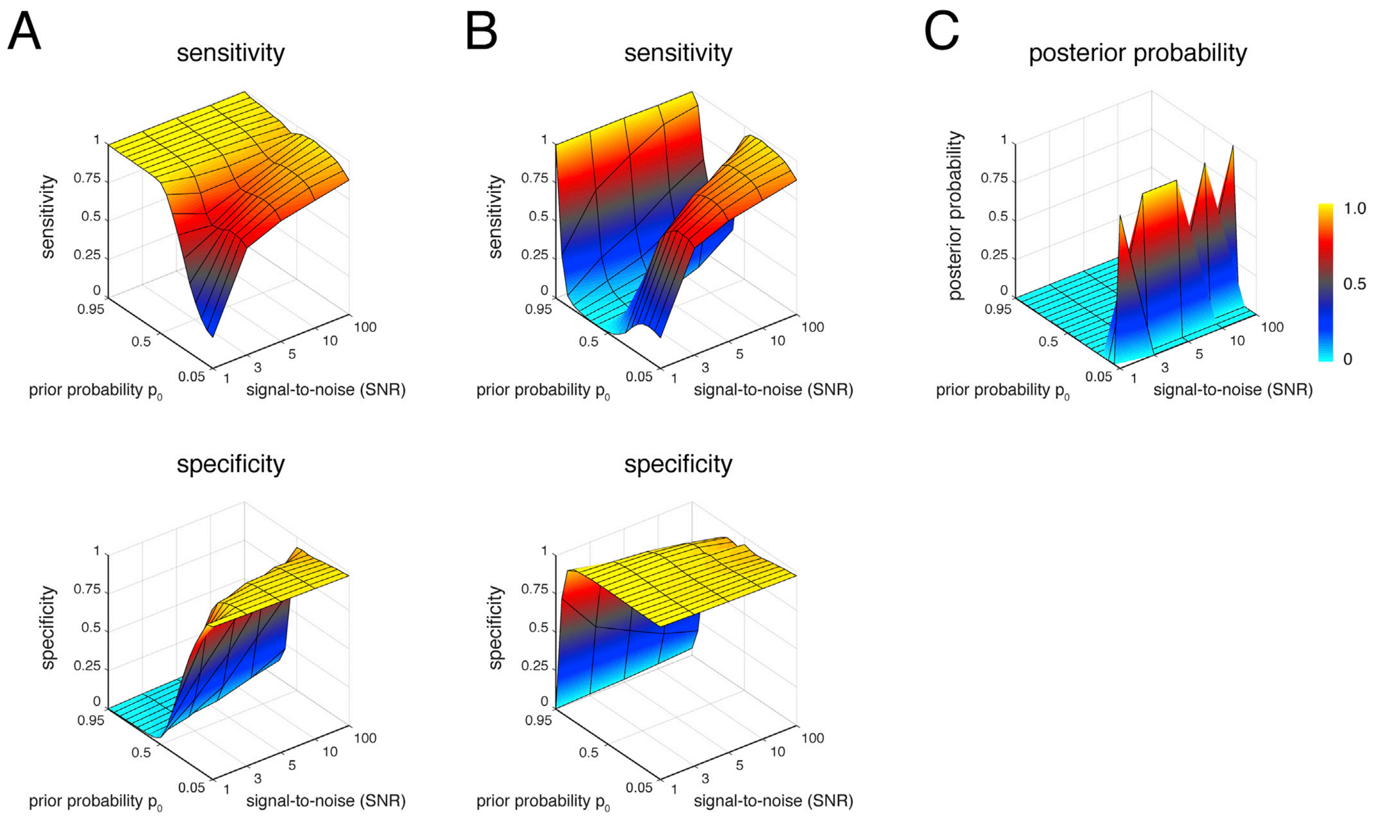


Fig. 3. Model architecture recovery of sparse rDCM in terms of the sensitivity and specificity for the reciprocal small-world DCM. (A) Sensitivity (*top*) and specificity (*bottom*) of identifying the true (data-generating) connections of the reciprocal small-world DCM, classifying a connection as present when its posterior odds ratio fell within the "grey zone" between 10 and 0.1 (see Methods). Sensitivity and specificity are shown along the z-axis as surface plots for various combinations of the signal-to-noise ratio (SNR) of the simulated fMRI data and the parameter p_0^i of the Bernoulli prior on binary indicator variables. The various SNR settings (1, 3, 5, 10, and 100) are shown along the x-axis of each subplot. The different p_0^i settings (0.05–0.95, with step size of 0.05) are shown along the y-axis of each subplot. The same plot is shown in (B) when classifying a connection as absent when its posterior odds ratio fell within the "grey zone" between 10 and 0.1 (see Methods). (C) Fixed effects Bayesian model selection results for the various SNR settings based on the negative free energy. The posterior model probabilities for all combinations of SNR and p_0^i settings are shown. See Supplementary Figure S1 for a visualization of the network structure of the reciprocal small-world DCM.

2004a; Stephan et al., 2009a). We found the sparse model to be decisively superior, with a posterior model probability of 0.96. This indicates that the reciprocal connections between V1 and SPC were correctly removed by sparse rDCM.

For the sparse (winning) model we compared the posterior parameter estimates from VBL (Fig. 5B, right) and sparse rDCM (Fig. 5B, left). The two inversion schemes yielded qualitatively identical effective connectivity patterns (with regard to their excitatory or inhibitory nature) but differed quantitatively. Specifically, parameter estimates deviated less strongly from their prior mean (zero) for sparse rDCM as compared to VBL. Similarly, sparse rDCM yielded slightly less accurate fits (Fig. 5C) as indicated by the coefficient of determination R^2 between predicted and measured BOLD signal (sparse rDCM: 0.65, VBL: 0.70). These quantitative differences are to be expected given the differences between the underlying generative models of the two frameworks (see Discussion).

Aphasia dataset

Furthermore, we applied sparse rDCM to a dataset from a passive auditory listening task (Schofield et al., 2012), which had previously been used to introduce generative embedding to neuroimaging (Brodersen et al., 2011). We chose this dataset because it is associated with a known ground truth in terms of group membership (i.e., stroke patients with aphasia and healthy controls) which we can challenge our method to detect. This represents an important evaluation of predictive validity, particularly because we can test whether connectivity estimates by our method perform better than conventional functional connectivity measures. Here, we adopted a similar generative embedding approach as in Brodersen et al. (2011) to evaluate the practical utility of sparse rDCM, testing whether the posterior parameter estimates could differentiate healthy controls from patients with moderate aphasia. To this end, sparse rDCM was used to invert the fully connected model (Fig. 6A) for each subject separately, resulting in sparse effective connectivity architectures and (approximate) posterior densities over model parameters. The individual MAP estimates then entered an l_1-regularized linear SVM, for which classification performance was assessed by means of leave-one-out cross-validation to obtain the posterior distribution of the balanced accuracy (Brodersen et al., 2010). Sparse rDCM achieved almost perfect classification performance with a balanced accuracy of 95% ($p < 0.001$), assigning 36 out of the 37 subjects to the correct disease/health state. This is only marginally below the balanced accuracy obtained using classical DCM (98%) as reported in Brodersen et al. (2011).[2] Furthermore, we compared the predictive accuracy of sparse rDCM with a more conventional classification approach operating on measures of functional connectivity (i.e., Pearson correlation coefficient). Functional connectivity measures also achieved a reasonable classification performance (balanced accuracy: 78%, $p < 0.001$). This classification result, however,

[2] The balanced accuracies for sparse rDCM and classical DCM are not identical because sparse rDCM misclassified a patient as a healthy control, whereas in Brodersen et al. (2010) a healthy control was misclassified as a patient.

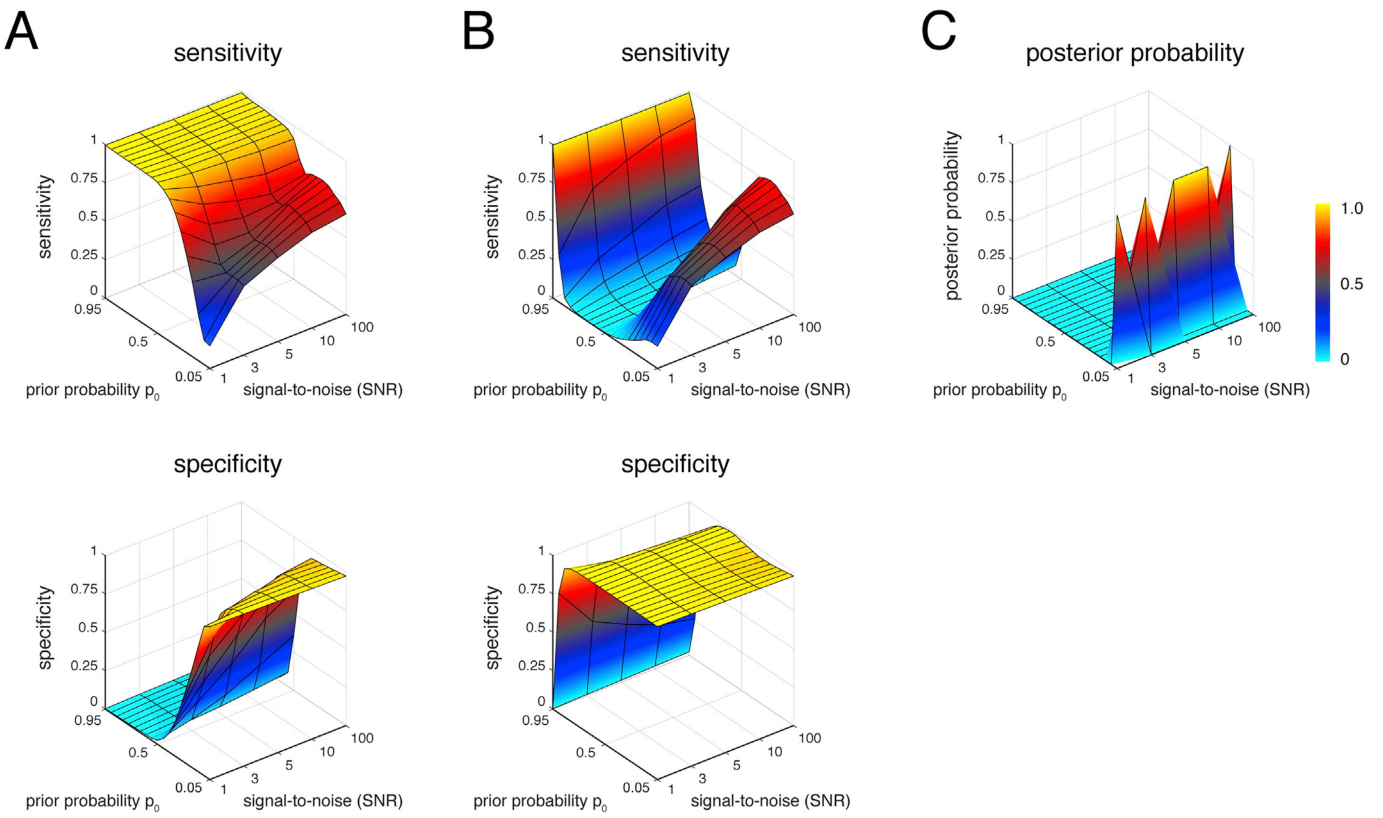


Fig. 4. Model architecture recovery of sparse rDCM in terms of the sensitivity and specificity for the connectome-based DCM. (A) Sensitivity (*top*) and specificity (*bottom*) of identifying the true (data-generating) connections of the connectome-based DCM, classifying a connection as present when its posterior odds ratio fell within the "grey zone" between 10 and 0.1 (see Methods). Sensitivity and specificity are shown along the z-axis as surface plots for various combinations of the signal-to-noise ratio (SNR) of the simulated fMRI data and the parameter p_0^i of the Bernoulli prior on binary indicator variables. The various SNR settings (1, 3, 5, 10, and 100) are shown along the x-axis of each subplot. The different p_0^i settings (0.05–0.95, with step size of 0.05) are shown along the y-axis of each subplot. The same plot is shown in (B) when classifying a connection as absent when its posterior odds ratio fell within the "grey zone" between 10 and 0.1 (see Methods). (C) Fixed effects Bayesian model selection results for the various SNR settings based on the negative free energy. The posterior model probabilities for all combinations of SNR and p_0^i settings are shown. See Fig. 1C for a visualization of the network structure of the connectome-based DCM.

was significantly worse than the 95% balanced accuracy achieved by sparse rDCM (paired-sample Wald test, $p = 0.008$).

In order to identify the connection and driving input parameters that jointly discriminated healthy controls and aphasic patients, we counted in how many cross-validation folds each feature was selected. We found that sparse sets of 5–10 (out of 42) model parameters consistently served as support vectors. The parameters that acted as support vectors in at least one cross-validation fold are displayed in Fig. 6B, showing some overlap with the results reported by Brodersen et al. (2011), where primarily connections mediating information transfer from the right to the left hemisphere were identified as discriminative features. When restricting our analysis to those parameters that were selected in more than 95% of the cross-validation folds we found only five model parameters. These comprised the auditory driving inputs to left and right PT, the auditory driving input to right HG, the inhibitory self-connection of right HG, as well as the inter-hemispheric endogenous connection from right MBG to left MGB.

Whole-brain analyses: hand movement fMRI data

Sparse effective connectivity

In a final step, we applied sparse rDCM to an fMRI dataset from a hand movement paradigm acquired at high magnetic field strength (7 T). Using data from a single subject performing visually paced (fist closing) movements with the left hand, we inferred a sparse representation of the whole-brain effective connectivity pattern. We chose to use this particular dataset for two reasons: (i) the data derive from an extremely simple task and the cerebral network supporting visually paced hand movements is well known (e.g., Ledberg et al., 2007; Rizzolatti and Luppino, 2001; Witt et al., 2008), and (ii) high SNR afforded by 7 T should facilitate network inference in this initial proof-of-concept analysis.

As expected, visually synchronized left-hand movements activated a widespread cortical network comprising the primary motor area (M1), premotor cortex (PMC), supplementary motor area (SMA), and the cerebellum (Fig. 7A). For each of 104 brain regions spanning the entire cortex and cerebellum in the AAL atlas (Tzourio-Mazoyer et al., 2002), we extracted the principal eigenvariate of BOLD signal time series. Sparse rDCM was used to prune the fully connected whole-brain model, which contained over 10,000 free connectivity parameters.

Model inversion resulted in a sparse graph with less than 10% of the possible connections and inputs. That is, only 940 non-negligible connections and driving inputs remained, whereas all other parameters were pruned from the model. Importantly, the inferred connectivity (Fig. 7B, left) and driving input patterns (Fig. 7B, right) were biologically plausible: during left-hand movements, we found strong excitatory driving inputs to left and right precentral cortex, bilateral SMA, and left cerebellum (Fig. 7B, right), all representing key components of the motor network mediating hand movements (Witt et al., 2008). Additionally,

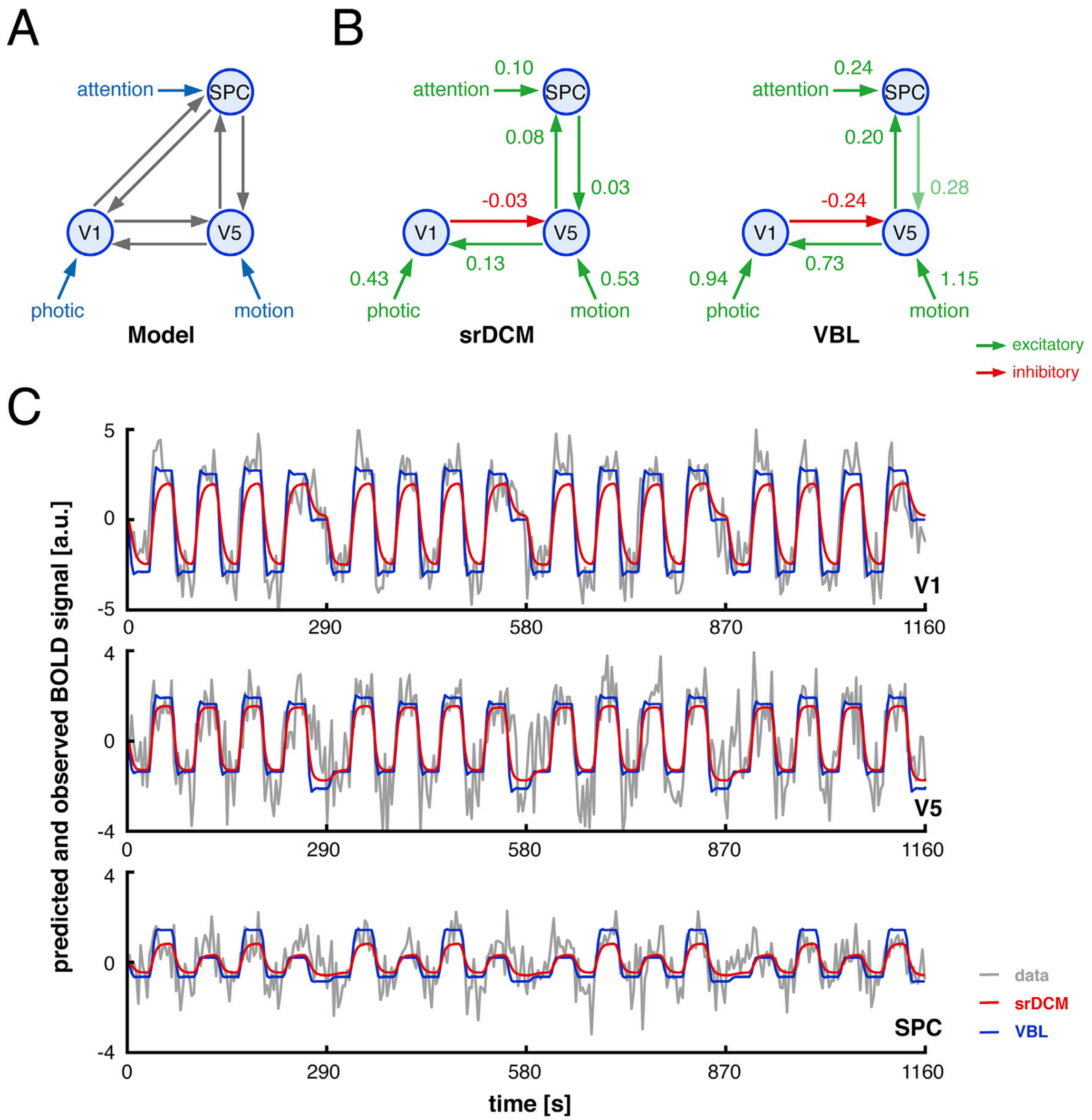


Fig. 5. Effective connectivity in a visual-attention network as assessed with sparse rDCM for the attention-to-motion fMRI dataset. (A) Effective connectivity structure serving as the starting point of sparse rDCM analyses. The model comprised primary visual areas V1 and V5, and superior parietal cortex (SPC). A full endogenous connectivity matrix (A matrix) was assumed initially. (B) Parameter estimates for the endogenous connectivity and driving inputs as estimated using sparse rDCM (*left*). Furthermore, parameter estimates are shown for VBL when using the sparse model structure suggested by sparse rDCM (*right*). Results are qualitatively consistent across the two methods. The coupling strength of each connection is displayed in [Hz]. Connections with a large effect size (i.e., posterior probability > 0.95 that parameter estimates were different from zero) are shown in full color; connections with moderate effect sizes (i.e., posterior probability < 0.95) are shown in faded colors. (C) Measured (grey) and predicted BOLD signal time series for sparse rDCM (red) and VBL (blue) in V1 (*top*), V5 (*middle*), and SPC (*bottom*).

driving inputs activated the cuneus and other occipital regions, consistent with the visual pacing of fist closings. Furthermore, excitatory driving inputs targeted regions in the postcentral gyrus and parietal cortex, regions activated by the somatosensory and proprioceptive aspects of the task and essential for visuomotor integration (Andersen, 1997; Culham and Kanwisher, 2001; Grefkes et al., 2004). Finally, we found driving inputs to regions in the frontal lobe of the right hemisphere (middle frontal gyrus and pars opercularis of the inferior frontal gyrus) that potentially engaged in top-down control and executive functioning, also representing key ingredients of visuomotor abilities (Fuster, 2003; Ledberg et al., 2007).

With regard to the endogenous connectivity (Fig. 7B, left), sparse rDCM revealed a prominent cluster of excitatory connections among bilateral pre- and postcentral gyrus and parietal regions. This was expected, given that these represent key cortical regions for the motor and somatosensory aspects of the task. We further observed strong excitatory influences from the left and right SMA to regions in the pre- and postcentral gyrus, consistent with the prominent role of SMA in initiating hand movements (e.g., Grefkes et al., 2008). Additionally, the ipsilateral (left) cerebellum exhibited pronounced excitatory connections with the contralateral (right) precentral gyrus, whereas the influence from ipsilateral cerebellum onto the ipsilateral precentral gyrus was inhibitory. Consistent with the visual pacing of hand movements, sparse rDCM revealed excitatory connections from the cuneus and other occipital regions to the motor areas mentioned above, indicating that visual information was sent via forward connections to regions involved in motor processes. Finally, frontal regions

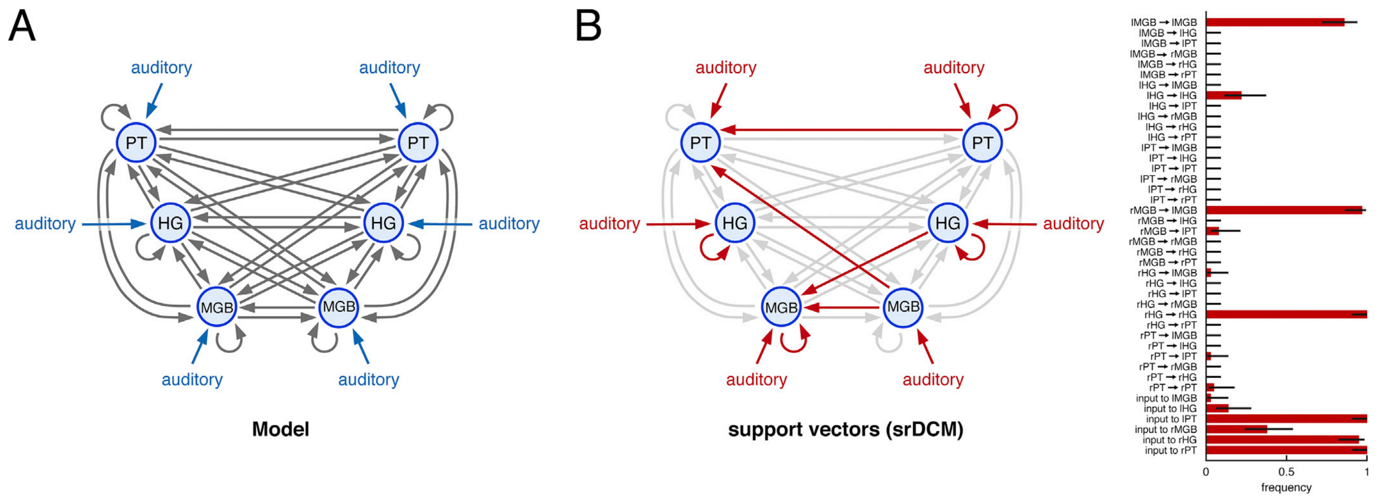


Fig. 6. Generative embedding to differentiate healthy subjects from moderately aphasic patients using sparse rDCM for the aphasia fMRI dataset (Schofield et al., 2012). (A) Effective connectivity structure serving as the starting point for the sparse rDCM analyses. The model comprised medial geniculate body (MGB), Heschl's gyrus (HG), and planum temporale (PT), each in both hemispheres. A full endogenous connectivity matrix (A matrix) was assumed initially. Additionally, the driving input representing auditory stimulation was assumed to elicit activity in all six regions. (B) All connectivity and driving input parameters that were discriminative between healthy controls and aphasic patients in the sense that they served as support vectors in at least one cross-validation fold (*left*). Frequency of how often a particular parameter was selected during cross validation to discriminate between the two groups.

exerted excitatory influences onto the motor network, possibly indicating top-down control required to adhere to task instructions (e.g., motor performance in synchrony with the visual cue).

An alternative graphical representation of the connectivity pattern is given by the connectogram, a circular representation of interdependencies among brain regions used frequently for visualization of connectomes (Irimia et al., 2012). This illustrates nicely the sparsity of the effective connectivity pattern detected by sparse rDCM (Fig. 7C). Regions in the frontal, parietal, and occipital lobe as well as the cerebellum exhibited abundant connections consistent with visually synchronized left-hand movements. By contrast, regions in the cingulate cortex, temporal lobe, and basal ganglia were substantially less involved. Furthermore, brain regions in the right hemisphere showed a higher node degree (i.e., total number of incoming and outgoing connections) than in the left hemisphere, consistent with the established contralateral hemispheric lateralization of unimanual hand movements (Roland and Zilles, 1996).

Finally, it can also be informative to inspect connectivity patterns when projected onto the whole-brain volume (Fig. 7D). Again, it is apparent that brain regions in the right hemisphere were more strongly connected than left-hemispheric regions. The edges of the connectivity graph in this representation are directed and signed, where excitatory and inhibitory influences are colored in green and red, respectively. We did not visualize connection strengths to keep the graphical representation uncluttered. Having said this, it is worth reiterating that the edges obtained from sparse rDCM are not only directed and signed, but also weighted (see Fig. 7B, left) – thus providing a full characterization of interactions within the whole-brain network.

Comparison with functional connectivity

We again compared the effective connectivity pattern obtained using sparse rDCM with a conventional measure of functional connectivity. For this, we computed the Pearson correlation coefficients among the BOLD signal time series of the 104 brain regions defined by the AAL atlas. In order to match the sparsity of the functional connectivity matrix with that of effective connectivity inferred by sparse rDCM, we thresholded the functional connectivity matrix such that only the 9.3% strongest connections (in absolute terms) were kept. This threshold represented precisely the degree of sparseness of the effective connectivity pattern obtained with sparse rDCM. Similar to sparse rDCM, thresholded functional connectivity revealed connections among motor regions (precentral, SMA, cerebellum), visual regions (cuneus, occipital), regions associated with the somatosensory and proprioceptive aspects of the task (postcentral, parietal), and frontal regions engaging in top-down control.

To quantify the similarity between functional and effective connectivity profiles, we binarized the two matrices and computed the association between these two binary matrices in terms of the simple matching coefficient (SMC; Dunn and Everitt, 1982). The SMC measures the proportion of pairs where the values of both matrices agree and ranges from 0 to 1; here, its value was 0.89. To test whether this was significantly different from chance, we generated a null distribution by computing the SMC for 100,000 randomly sampled functional connectivity matrices (Pearson correlation coefficients were sampled from a uniform distribution between −1 and 1), thresholded under the same criterion. Under this null distribution, the above value was highly unlikely ($p < 0.001$), suggesting that functional and effective connectivity profiles were more similar than what one would expect by chance. In summary, in this analysis, functional connectivity analyses and sparse rDCM yielded similar and biologically plausible connectivity profiles.

Computational burden

Concerning the computational efficiency of our method, running model inversion on a single processor core (without parallelization) on the Euler cluster at ETH Zurich, sparse rDCM took 624 s to infer the whole-brain effective connectivity pattern in this dataset for a single p_0^i value. This should only be treated as a rough indication, as run-times will depend on the specific hardware used. Notably, when parallelizing the code (by exploiting the fact that the VB equations apply to each region separately) and using 16 processor cores on the Euler cluster of ETH Zurich, this run-time could be reduced to 55 s.

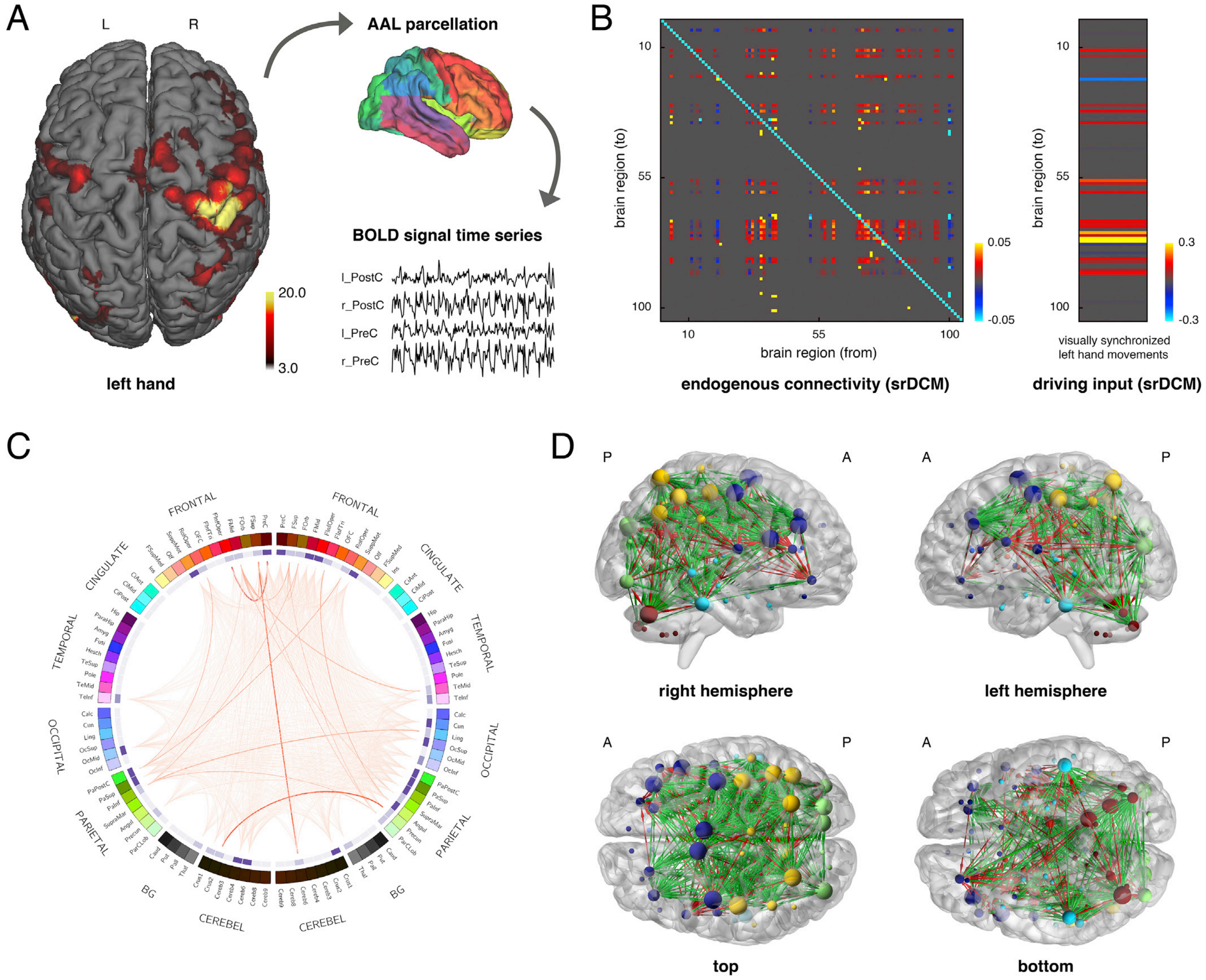


Fig. 7. Whole-brain sparse effective connectivity pattern underlying left-hand movements as assessed with sparse rDCM for an empirical fMRI dataset. (A) BOLD activation pattern of a single representative healthy subject showing regions that were activated during visually synchronized whole-hand fist closing movements with the left hand. Results are thresholded at $p < 0.05$ (FWE peak-level corrected). For the given BOLD activation pattern, the Automated Anatomical Labeling atlas (AAL; Tzourio-Mazoyer et al., 2002) was used as a whole-brain parcellation scheme. Region-wise BOLD signal time series were extracted as principal eigenvariates and entered effective connectivity analyses using sparse rDCM. (B) Posterior parameter estimates for connections (*left*) and driving inputs (*right*). (C) Estimated connectivity matrix graphically rendered as a connectogram. The labels on the outermost ring show the anatomical lobe for each of the nodes: frontal, cingulate, temporal, occipital, parietal, basal ganglia, and cerebellum. For each brain region defined by the AAL atlas, an abbreviation and color is defined. Inside the parcellation ring, the concentric circle represents the node degree (i.e., total number of incoming and outgoing connections) for each brain region. Finally, in the inner part of the connectogram, non-negligible connection strengths are displayed as edges, with the (absolute) connection strength being represented by the opacity of the line. The connectogram was created using Circos (Krzywinski et al., 2009), publicly available for download (http://www.circos.ca/software/). (D) Estimated connectivity matrix projected into the whole-brain volume. The size of each node represents the node degree for each brain region, whereas the node color indicates the lobe it belongs to: frontal (dark blue), cingulate (light blue), temporal (cyan), occipital (green), parietal (yellow), basal ganglia (orange), and cerebellum (red). Note that edges in this graphical representation are directed and differentiate between inhibitory (red) and excitatory (green) connections. Information on coupling strength of connections is not accounted for in order to keep the representation simple. The brain network was visualized using the BrainNet Viewer (Xia et al., 2013), also freely available for download (http://www.nitrc.org/projects/bnv/). L = left hemisphere; R = right hemisphere; A = anterior; P = posterior.

Discussion

In this paper, we introduced a novel extension to the rDCM framework which enables automatic pruning of fully (all-to-all) connected whole-brain graphs. This pruning rests on binary indicator variables that are embedded into the likelihood function and act as a feature selector. Using simulations, we first demonstrated the face validity of sparse rDCM for large (whole-brain) networks comprising up to 66 brain regions and 300 neuronal connectivity parameters. We then demonstrated the practical utility of sparse rDCM using empirical fMRI datasets. In particular, we demonstrated for the first time that sparse effective connectivity patterns can be inferred, with connection-specific estimates, in a whole-brain model with more than 100 regions and 10,000 connections – and within minutes of compute time on standard hardware.

Our simulations indicated that, as expected, the accuracy of sparse rDCM was dependent on the SNR of the data and the parameter p_0^i of the Bernoulli prior on binary indicator variables. For p_0^i settings close to the true degree of sparseness of the effective connectivity pattern, sparse rDCM identified the data-generating network architecture reasonably well. More precisely, while the sensitivity of sparse rDCM was only low to moderate in the case of an SNR of 3, the specificity of the approach was close to perfect with hardly any false positives occurring (regardless of SNR). For challenging noise settings (SNR = 1), sparse rDCM frequently failed to identify existing connections. Hence, for scenarios where the fMRI data is subject to inherently low SNR (e.g., subcortical regions), the current implementation of sparse rDCM likely shows poor sensitivity. This is not too surprising, considering that the initial version of sparse rDCM reported in this paper is based on the original rDCM implementation, which itself suffers from these limitations (Frässle et al., 2017). Furthermore, enforcing sparsity has an intrinsic tendency to favoring specificity at the expense of sensitivity. Overall, our simulation results suggest a tendency of our method towards conservativeness: while not all effective connections may be identified due to low sensitivity, the high specificity implies that detected connections likely represent real effects.

Our simulations also illustrate that the negative free energy (as a bound approximation to the log evidence) can recover the known p_0^i of the large-scale DCMs. Specifically, p_0^i values selected according to the largest negative free energy were identical or close to the true degree of sparseness of the effective connectivity pattern and resulted in an optimal trade-off between sensitivity and specificity of sparse rDCM. For applications of sparse rDCM to empirical data in practice, this suggests a simple procedure for determining an optimal p_0^i: inverting the model under different possible p_0^i values (using a suitably defined grid in the range 0..1) and selecting the value with the highest negative free energy. In a group setting, this can be done using random or fixed effects BMS procedures (see Stephan et al., 2009a).

We anticipate that advances in scanner hardware and sequences, which enable higher signal-to-noise ratios, will help boost the sensitivity of sparse rDCM. For example, high magnetic field strengths (Duyn, 2012; Redpath, 1998) boost SNR levels considerably. Recently developed magnetic field sensing techniques can further improve SNR by accounting for confounds due to magnetic field fluctuations (Barmet et al., 2008; Bollmann et al., 2017). Additionally, fast image acquisition has previously been identified as a key factor for improving the accuracy of rDCM (Frässle et al., 2017). Hence, recent methodological developments such as ultra-fast imaging (Stirnberg et al., 2017) and multiband EPI sequences (Moeller et al., 2010; Xu et al., 2013) are likely to become important as they make the acquisitions of whole-brain fMRI data at sub-second TRs feasible. Finally, improving the generative model of sparse rDCM will also help to enhance the sensitivity of the approach and thus constitutes a key target of forthcoming work (see below for potential future extensions of sparse rDCM).

Even at its current stage, however, our approach has practical utility, as demonstrated by analyses of three empirical fMRI datasets. First, we analyzed two small networks that had been the subject of previous studies using conventional DCM (Büchel and Friston, 1997; Schofield et al., 2012). Sparse rDCM yielded plausible connectivity estimates for the "attention-to-motion" dataset that were qualitatively similar to the results obtained using VBL in classical DCM implementations. Note that an exact match between sparse rDCM and VBL is not to be expected because of the differences in the generative models: (i) the introduction of binary indicator variables to enable automatic pruning of fully connected graphs, (ii) the use of a fixed HRF instead of the nonlinear hemodynamic model, (iii) the mean field approximation between parameters targeting different regions, and (iv) the use of a Gamma prior on noise precision instead of the log-normal prior.

For the aphasia dataset, sparse rDCM enabled discrimination of healthy controls and aphasic patients with high accuracy (95%), comparable to the classification results reported previously for classical DCM (Brodersen et al., 2011). It is worth highlighting that sparse rDCM achieved this performance without any prior assumptions on the connectivity structure of the network. That is, sparse rDCM started from the fully connected A and C matrices, which were then automatically pruned during model inversion, resulting in subject-specific sparse connectivity "fingerprints" that differentiated the two groups. This is remarkable considering that the predictive accuracies reported in Brodersen et al. (2011) rapidly declined when deliberately modifying the connectivity structure of the DCM used.

In a final step, we applied sparse rDCM to empirical data from a simple hand movement paradigm acquired at high magnetic field strength (7 T) and evaluated its utility for inferring whole-brain effective connectivity from fMRI data. Our analyses yielded plausible connections and driving input patterns that fit the known cerebral network underlying visually triggered motor actions (Ledberg et al., 2007).

Our model is not the only approach towards inferring effective connectivity in large-scale network models (for reviews, see Deco and Kringelbach, 2014; Stephan et al., 2015). One of its shortcomings is that it is not specifically designed to deal with random fluctuations in BOLD signals in non-task paradigms (although it can be applied to "resting state" data, see below), something that two recent developments in particular are designed to do (Gilson et al., 2017; Razi et al., 2017). Gilson et al. (2017) recently introduced a large-scale network model in which the dynamics followed an Ornstein-Uhlenbeck process and which allows for computing directed connection strengths. This model further differs from ours in that it does not include a forward model (from neuronal states to fMRI data) but directly operates on BOLD signals; additionally, it is not a generative (Bayesian) model but employs maximum likelihood estimation. This model was applied to fMRI data (using the same 66-area Hagmann parcellation as in our simulation study), but its compute time was not reported. In a second recent study, Razi et al. (2017) used cross-spectral DCM (Friston et al., 2014a) to invert networks consisting of 36 brain regions based on resting-state fMRI data. The approach exploits the principal components of the functional connectivity matrix to constrain the prior covariance matrix of the DCM, essentially reducing the effective number of free parameters by replacing the number of nodes with a (lower) number of modes (Seghier and Friston, 2013). While spectral DCM was computationally more efficient than stochastic DCM (Daunizeau et al., 2009), model inversion of the 36-region network was still computationally very demanding, with run-times between 21 and 42 h for a single model (i.e., 20 min per iteration, 64–128 iterations until convergence). It remains to be tested whether DCMs covering the entire brain remain computationally feasible for spectral DCM and yield reliable parameter estimates.

Our method compares favorably in terms of computational efficiency. Estimating the effective connectivity of a whole-brain network with nearly three times as many nodes and 10 times more connections, the run-time of sparse rDCM was roughly 10 min for a single p_0^i value without parallelization and just 1 min when running model inversion on 16 processor cores in parallel. It is worth highlighting in this context that the current implementation of sparse rDCM is not optimized for speed. Further acceleration is straightforward by, for instance, using faster programming languages (e.g., C/C++ instead of MATLAB).

However, as already pointed out in Frässle et al. (2017), the current implementation of rDCM – and thus also sparse rDCM – only represents a starting point and is still subject to major limitations, which will be addressed in forthcoming developments: First, we will replace the fixed hemodynamic response function with a more flexible hemodynamic model, using, for example, a linearized version of the hemodynamic model in DCM (Friston et al., 2000; Stephan et al., 2007). Second, extending the linear neuronal model to incorporate modulatory influences would represent an important methodological advance, enabling more sophisticated analyses related to task-induced changes in effective connectivity. Third, in its current form, rDCM can be applied to the "resting state" (i.e., unconstrained cognition in the absence of external perturbations) by "switching off" driving inputs (i.e., setting p_0^i to zero). This is possible because the measured data (in the Fourier domain) feature as predictors in the Bayesian linear regression model that constitutes the likelihood function in Eq. (4). However, the model does not explicitly account for endogenous fluctuations in neuronal activity; for example, it has no concept of stochastic "innovations" or similar ways how noise can drive activity intrinsically. We anticipate that including a mechanism to account for endogenous fluctuations will constitute an important future step to further increase the explanatory power of rDCM for "resting state" fMRI data.

An additional improvement of the generative model of sparse rDCM becomes apparent when closely inspecting the results from our simulations for the grid-like, small-world and connectome-based DCMs. Specifically, the overall sensitivity of sparse rDCM was slightly diminished for the connectome-based model as compared to the other models. This is likely due to the fact that for the connectome-based DCM, the optimal p_0^i is essentially different for each node as brain regions differ in the number of afferent connections: for example, hubs are more densely connected while other regions exhibit sparser connectivity profiles (Bullmore and Sporns, 2009). However, the current implementation of sparse rDCM assumed identical p_0^i for all connections in the model, which is likely to result in sub-optimal estimates for realistic networks. In future implementations of sparse rDCM, we will therefore explore the utility of specifying p_0^i for each region – or even individual connections – independently. For example, this could be achieved by informing p_0^i of each connection by subject-specific anatomical connectivity measures, as derived from diffusion-weighted imaging data (cf. anatomically informed priors in DCM; Stephan et al., 2009b). Alternatively, measures of the functional connectivity between two nodes (e.g., correlation, spectral coherence) could be used to inform the prior probabilities p_0^i of each connection – giving rise to an approach not unrelated to the one proposed by Seghier and Friston (2013). Taking advantage of multimodal neuroimaging data in this manner might represent a pragmatic way to increase the currently low sensitivity of sparse rDCM. In forthcoming work, we will therefore systematically explore the utility of anatomical and functional connectivity matrices for informing the prior probabilities of the Bernoulli prior on binary indicator variables.

The simulations and empirical analyses presented in this paper represent a first step to assess the validity and practical utility of sparse rDCM. In the simulations (Figs. 2–4, Supplementary Figs. S2, S6-S9), we varied three parameters of key importance for the model's performance: signal-to-noise ratio, sparsity assumptions (Bernoulli prior), and the values of data-generating model parameters. We also examined the model's performance separately for regions with different levels of indegree (Supplementary Figures S4-S5). By contrast, we did not vary TR (as this was assessed in previous work; Frässle et al., 2017). Clearly, more validation work can still be done. In future work, we would like to assess construct validity in more detail, comparing the performance of sparse rDCM to other emerging approaches for inferring whole-brain effective connectivity. In addition, we hope to conduct further tests of the predictive validity of sparse rDCM, extending the challenge of predicting independent variables (e.g., diagnostic status as shown in this paper) from brain connectivity to larger pharmacological and patient datasets.

Notably, embedding sparsity constraints into rDCM requires that the sparsity assumptions of the model match the actual sparsity of the network. In this paper, we use the log model evidence (as approximated by the negative free energy) as a principled criterion for selecting an optimal hyperparameter p_0^i that determines the sparsity of the network. When moving to more fine-grained parcellation schemes than the one utilized in the present study, new challenges may arise, for instance, with regard to spurious short-distance connections that originate from the inherent spatial smoothness of the BOLD signal (Power et al., 2011). Finally, it remains to be tested in future work whether sparsity constraints are equally appropriate for the "resting state" as compared to task-based paradigms.

In summary, the findings presented in this study indicate promising potential of sparse rDCM for estimating effective connectivity patterns in large (whole-brain) networks by automatically pruning fully connected graphs to the most essential connections. We conclude by highlighting two potential future applications of our approach. First, sparse rDCM may enable the application of graph-theoretical approaches (Bullmore and Sporns, 2009; Rubinov and Sporns, 2010) to whole-brain effective connectivity patterns. To date, the application of graph-theoretical measures to human brain data has typically been restricted to the analysis of undirected and unweighted graphs (Bullmore and Sporns, 2009). This is because both structural connectivity (diffusion-weighted imaging) and functional connectivity methods do not allow one to obtain directed estimates. Application of graph theory to whole-brain effective connectivity profiles may provide a more accurate characterization of principles of functional organization of the human brain, for example, by accommodating the functional asymmetries between forward and backward connections in cortical hierarchies (Felleman and Van Essen, 1991; Zeki and Shipp, 1988).

Finally, estimates of whole-brain effective connectivity may also advance our understanding of the pathophysiology of brain disorders – and thus make important contributions to the emerging fields of Computational Psychiatry, Computational Neurology and Computational Psychosomatics (Deco and Kringelbach, 2014; Friston et al., 2014b; Huys et al., 2016; Maia and Frank, 2011; Montague et al., 2012; Petzschner et al., 2017; Stephan and Mathys, 2014; Stephan et al., 2015). Sparse rDCM may become particularly useful in the context of disorders for which global dysconnectivity has been suggested, such as schizophrenia (Anticevic et al., 2015; Bullmore et al., 1997; Friston et al., 2016; Friston and Frith, 1995; Stephan et al., 2006). In these situations, the computational framework introduced in this paper may deliver global "fingerprints" of aberrant functional integration, bringing computational phenotyping of whole-brain effective connectivity patterns in individual patients within reach (Stephan et al., 2015).

Clearly, as discussed above, sparse rDCM is in its infancy and still subject to major limitations that have to be addressed in future developments. Similarly, the utility of sparse rDCM as a clinically relevant computational assay remains to be tested using pharmacological and patient datasets. We hope to follow these lines of research in forthcoming work.

Software note

A MATLAB implementation of the sparse regression dynamic causal modeling (sparse rDCM) approach introduced in the present paper will be made available as open source code in a future release of the **T**ranslational **A**lgorithms for **P**sychiatry-**A**dvancing **S**cience (TAPAS) Toolbox (www.translationalneuromodeling.org/software).

Acknowledgements

This work was supported by the ETH Zurich Postdoctoral Fellowship Program and the Marie Curie Actions for People COFUND Program (SF), as well as by the René and Susanne Braginsky Foundation (KES), the University of Zurich, and the German Research Foundation (Transregional Collaborative Research Center 134; KES).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.neuroimage.2018.05.058.

APPENDIX

A.1: Derivation of model parameters

In the following, we outline the derivation of the variational Bayesian update equations for the different parameter classes in sparse rDCM (neuronal connectivity, noise precision, and binary indicator variables). Importantly, under the mean field approximation of sparse rDCM, optimization can be performed for each region independently. Hence, we restrict the ensuing derivation of update equations to a single region.

Update equation of θ:

$$\begin{aligned}
\ln q^*(\theta|Y,X) &= \langle \ln p(\theta,\tau,\zeta,Y|X)\rangle_{q(\tau,\,\zeta)} \\
&= \left\langle \ln\mathcal{N}\left(Y;XZ\theta,\tau^{-1}I_{N\times N}\right) + \ln\mathcal{N}(\theta;\mu_0,\Sigma_0)\right\rangle_{q(\tau,\,\zeta)} + c \\
&= \left\langle -\frac{1}{2}\theta^T(\tau Z^TX^TXZ)\theta + \theta^T\tau Z^TX^TY\right\rangle_{q(\tau,\,\zeta)} + \left\langle -\frac{1}{2}\theta^T\Sigma_0^{-1}\theta + \theta^T\Sigma_0^{-1}\mu_0\right\rangle_{q(\tau,\,\zeta)} + c \\
&= -\frac{\alpha_{\tau|y}}{2\beta_{\tau|y}}\theta^T\left\langle ZX^TXZ\right\rangle_{q(\zeta)}\theta + \frac{\alpha_{\tau|y}}{\beta_{\tau|y}}\theta^T\langle Z\rangle_{q(\zeta)}X^TY - \frac{1}{2}\theta^T\Sigma_0^{-1}\theta + \theta^T\Sigma_0^{-1}\mu_0 + c \\
&= -\frac{1}{2}\theta^T\left(\frac{\alpha_{\tau|y}}{\beta_{\tau|y}}\left\langle ZX^TXZ\right\rangle_{q(\zeta)} + \Sigma_0^{-1}\right)\theta + \theta^T\left(\frac{\alpha_{\tau|y}}{\beta_{\tau|y}}\langle Z\rangle_{q(\zeta)}X^TY + \Sigma_0^{-1}\mu_0\right) + c \\
&= -\frac{1}{2}\theta^T\left(\frac{\alpha_{\tau|y}}{\beta_{\tau|y}}\left(P_{\zeta|y}X^TXP_{\zeta|y} + (X^TX)\circ\left(P_{\zeta|y} - P_{\zeta|y}^2\right)\right) + \Sigma_0^{-1}\right)\theta \\
&\quad +\theta^T\left(\frac{\alpha_{\tau|y}}{\beta_{\tau|y}}P_{\zeta|y}X^TY + \Sigma_0^{-1}\mu_0\right) + c
\end{aligned} \tag{A.1}$$

Here, all terms independent of θ were absorbed into the constant term c. Additionally, we have made use of $\langle\tau\rangle_{q(\tau)} = \frac{\alpha_{\tau|y}}{\beta_{\tau|y}}$ and $\langle Z\rangle_{q(\zeta)} = P_{\zeta|y}$, with $\langle\cdot\rangle_q$ denoting the expected value with respect to the variational density $q(\tau)$ and $q(\zeta)$, respectively. Furthermore, we utilized the expression for $\langle Z^TX^TXZ\rangle_{q(\zeta)}$ derived in appendix A.3.

Update equation of τ:

$$\begin{aligned}
\ln q^*(\tau|Y,X) &= \langle \ln p(\theta,\tau,\zeta,Y|X)\rangle_{q(\theta,\,\zeta)} \\
&= \left\langle \ln\mathcal{N}\left(Y;XZ\theta,\tau^{-1}I_{N\times N}\right) + \ln\mathrm{Gamma}(\tau;\alpha_0,\beta_0)\right\rangle_{q(\theta,\,\zeta)} + c \\
&= \frac{N}{2}\ln\tau - \frac{\tau}{2}\left\langle (Y-XZ\theta)^T(Y-XZ\theta)\right\rangle_{q(\theta,\,\zeta)} + (\alpha_0-1)\ln\tau - \beta_0\tau + c \\
&= \frac{N}{2}\ln\tau - \frac{\tau}{2}\left(Y^TY - 2\left\langle\theta^TZ^TX^TY\right\rangle_{q(\theta,\,\zeta)} + \left\langle\theta^TZ^TX^TXZ\theta\right\rangle_{q(\theta,\,\zeta)}\right) + (\alpha_0-1)\ln\tau \\
&\quad -\beta_0\tau + c \\
&= \frac{N}{2}\ln\tau - \frac{\tau}{2}\left(Y^TY - 2\mu_{\theta|y}^TP_{\zeta|y}^TX^TY + \left\langle\theta^TZ^TX^TXZ\theta\right\rangle_{q(\theta,\,\zeta)}\right) + (\alpha_0-1)\ln\tau - \beta_0\tau + c \\
&= -\frac{\tau}{2}\left(Y^TY - 2\mu_{\theta|y}^TP_{\zeta|y}^TX^TY + \left\langle\theta^T\left(P_{\zeta|y}X^TXP_{\zeta|y} + (X^TX)\circ\left(P_{\zeta|y}-P_{\zeta|y}^2\right)\right)\theta\right\rangle_{q(\theta)}\right) \\
&\quad +\frac{N}{2}\ln\tau + (\alpha_0-1)\ln\tau - \beta_0\tau + c \\
&= -\frac{\tau}{2}\left(\left(Y - XP_{\zeta|y}\mu_{\theta|y}\right)^T\left(Y - XP_{\zeta|y}\mu_{\theta|y}\right) + \mathrm{tr}\left(P_{\zeta|y}X^TXP_{\zeta|y}\Sigma_{\theta|y}\right)\right) \\
&\quad -\frac{\tau}{2}\left(\mu_{\theta|y}^T\left((X^TX)\circ\left(P_{\zeta|y}-P_{\zeta|y}^2\right)\right)\mu_{\theta|y} + \mathrm{tr}\left(\left((X^TX)\circ\left(P_{\zeta|y}-P_{\zeta|y}^2\right)\right)\Sigma_{\theta|y}\right)\right) \\
&\quad +\frac{N}{2}\ln\tau + (\alpha_0-1)\ln\tau - \beta_0\tau + c
\end{aligned} \tag{A.2}$$

Here, $\circ$ denotes the element-wise product of two matrices. All terms independent of τ were absorbed into the constant term c. Additionally, we made use of $\langle\theta\rangle_{q(\theta)} = \mu_{\theta|y}$, with $\langle\cdot\rangle_{q(\theta)}$ denoting the expected value with respect to the variational density $q(\theta)$. Furthermore, we utilized the expression for $\langle\theta^T Z^T X^T XZ\theta\rangle_{q(\theta,\ \zeta)}$ derived in appendix A.4.

Update equation of ζ_i:

$$\begin{aligned}
\ln q^*(\zeta_i|Y,X) &= \langle \ln p(\theta,\tau,\zeta,Y|X)\rangle_{q(\theta,\tau,\ \zeta_{\setminus i})} \\
&= \left\langle \ln\mathcal{N}\left(Y;XZ\theta,\tau^{-1}I_{N\times N}\right) + \sum_{j=1}^{D}\ln\mathrm{Bern}\left(\zeta_j;p_0^j\right)\right\rangle_{q(\theta,\tau,\ \zeta_{\setminus i})} + c \\
&= \left\langle -\frac{\tau}{2}(Y-XZ\theta)^T(Y-XZ\theta) + \sum_{j=1}^{D}\zeta_j\ln p_0^j + \left(1-\zeta_j\right)\ln\left(1-p_0^j\right)\right\rangle_{q(\theta,\tau,\ \zeta_{\setminus i})} + c \\
&= -\frac{\alpha_{\tau|y}}{2\beta_{\tau|y}}\left\langle(Y-XZ\theta)^T(Y-XZ\theta)\right\rangle_{q(\theta,\ \zeta_{\setminus i})} + \zeta_i\ln\left(\frac{p_0^i}{1-p_0^i}\right) + c \\
&= -\frac{\alpha_{\tau|y}}{2\beta_{\tau|y}}\left\langle\left(Y-XZ\mu_{\theta|y}\right)^T\left(Y-XZ\mu_{\theta|y}\right) + \mathrm{tr}\left(ZX^TXZ\Sigma_{\theta|y}\right)\right\rangle_{q(\ \zeta_{\setminus i})} \\
&\quad +\zeta_i\ln\left(\frac{p_0^i}{1-p_0^i}\right) + c \\
&= -\frac{\alpha_{\tau|y}}{2\beta_{\tau|y}}\left(Y^TY - 2\mu_{\theta|y}\langle Z\rangle_{q(\ \zeta_{\setminus i})}X^TY + \mu_{\theta|y}\left\langle ZX^TXZ\right\rangle_{q(\ \zeta_{\setminus i})}\mu_{\theta|y}\right) \\
&\quad -\frac{\alpha_{\tau|y}}{2\beta_{\tau|y}}\left\langle\mathrm{tr}\left(ZX^TXZ\Sigma_{\theta|y}\right)\right\rangle_{q(\ \zeta_{\setminus i})} + \zeta_i\ln\left(\frac{p_0^i}{1-p_0^i}\right) + c
\end{aligned} \tag{A.3}$$

Here, all terms independent of ζ_i were absorbed into the constant term c. Note that we made use of the fact that $\langle\zeta_j\rangle_{q(\ \zeta_{\setminus i})}$ is a constant with respect to ζ_i for all terms $j \neq i$. The expression in Eq. (A.3) can be further simplified by making use of the results in the appendices A.5-A.7. The final expression for the approximate posterior over binary indicator variables then takes the form:

$$\begin{aligned}
\ln q^*(\zeta_i|Y,X) &= \zeta_i\left(\frac{\alpha_{\tau|y}}{\beta_{\tau|y}}\mu_{\theta|y}\upsilon_i - \frac{\alpha_{\tau|y}}{2\beta_{\tau|y}}\left(\left(\mu_{\theta|y}^i\right)^2 W_{ii} + 2\mu_{\theta|y}^i\sum_{j\neq i}p_{\zeta|y}^j\mu_{\theta|y}^j W_{ij}\right)\right) \\
&\quad -\zeta_i\left(\frac{\alpha_{\tau|y}}{2\beta_{\tau|y}}W_{ii}\Sigma_{\theta|y}^{ii} + \frac{\alpha_{\tau|y}}{\beta_{\tau|y}}\sum_{j\neq i}p_{\zeta|y}^j W_{ij}\Sigma_{\theta|y}^{ij} - \ln\left(\frac{p_0^i}{1-p_0^i}\right)\right) + c
\end{aligned} \tag{A.4}$$

where we have set $\upsilon = X^TY$ and $W = X^TX$.

A.2: Derivation of negative free energy

Having derived the variational Bayesian update equations for the posterior densities, we now derive the expressions for the individual components of the negative free energy for a single region.

Expectation of the likelihood:

$$\begin{aligned}
\langle\ln p(\theta,\tau,\zeta,Y|X)\rangle_{q(\theta,\tau,\zeta)} &= \left\langle\ln\mathcal{N}\left(Y;XZ\theta,\tau^{-1}I_{N\times N}\right)\right\rangle_{q(\theta,\tau,\zeta)} \\
&= \left\langle -\frac{N}{2}\ln 2\,\pi - \frac{1}{2}\ln\left|\tau^{-1}I_{N\times N}\right| - \frac{\tau}{2}(Y-XZ\theta)^T(Y-XZ\theta)\right\rangle_{q(\theta,\tau,\zeta)} \\
&= -\frac{N}{2}\ln 2\,\pi + \frac{N}{2}\ln\langle\tau\rangle_{q(\tau)} - \frac{\langle\tau\rangle_{q(\tau)}}{2}\left\langle(Y-XZ\theta)^T(Y-XZ\theta)\right\rangle_{q(\theta,\zeta)} \\
&= -\frac{N}{2}\ln 2\,\pi + \frac{N}{2}\left(\Psi\left(\alpha_{\tau|y}\right) - \ln\beta_{\tau|y}\right) \\
&\quad -\frac{\alpha_{\tau|y}}{2\beta_{\tau|y}}\left(\left(Y-XP_{\zeta|y}\mu_{\theta|y}\right)^T\left(Y-XP_{\zeta|y}\mu_{\theta|y}\right) - \mathrm{tr}\left(P_{\zeta|y}WP_{\zeta|y}\Sigma_{\theta|y}\right)\right) \\
&\quad -\frac{\alpha_{\tau|y}}{2\beta_{\tau|y}}\left(\mu_{\theta|y}^T W\circ\left(P_{\zeta|y}-\left(P_{\zeta|y}\right)^2\right)\mu_{\theta|y} - \mathrm{tr}\left(W\circ\left(P_{\zeta|y}-\left(P_{\zeta|y}\right)^2\right)\Sigma_{\theta|y}\right)\right)
\end{aligned} \tag{A.5}$$

Expectation of the prior on θ:

$$\begin{aligned}
\langle \ln p(\theta)\rangle_{q(\theta)} &= \langle \ln \mathcal{N}(\theta;\mu_0,\Sigma_0)\rangle_{q(\theta)} \\
&= -\frac{D}{2}\ln 2\,\pi - \frac{1}{2}\ln|\Sigma_0| - \frac{1}{2}\left\langle(\theta-\mu_0)^T\Sigma_0^{-1}(\theta-\mu_0)\right\rangle_{q(\theta)} \\
&= -\frac{D}{2}\ln 2\,\pi - \frac{1}{2}\ln|\Sigma_0| - \frac{1}{2}\left(\mu_{\theta|y}-\mu_0\right)^T\Sigma_0^{-1}\left(\mu_{\theta|y}-\mu_0\right) - \frac{1}{2}\mathrm{tr}\left(\Sigma_0^{-1}\Sigma_{\theta|y}\right)
\end{aligned} \tag{A.6}$$

Expectation of the prior on τ:

$$\begin{aligned}
\langle \ln q(\tau)\rangle_{q(\tau)} &= \langle \mathrm{Gamma}(\tau;\alpha_0,\beta_0)\rangle_{q(\tau)} \\
&= \alpha_0 \ln\beta_0 - \ln\Gamma(\alpha_0) + \langle(\alpha_0-1)\ln\tau - \beta_0\tau\rangle_{q(\tau)} \\
&= \alpha_0 \ln\beta_0 - \ln\Gamma(\alpha_0) + (\alpha_0-1)\langle\ln\tau\rangle_{q(\tau)} - \beta_0\frac{\alpha_{\tau|y}}{\beta_{\tau|y}} \\
&= \alpha_0 \ln\beta_0 - \ln\Gamma(\alpha_0) + (\alpha_0-1)\left(\Psi\left(\alpha_{\tau|y}\right) - \ln\beta_{\tau|y}\right) - \beta_0\frac{\alpha_{\tau|y}}{\beta_{\tau|y}}
\end{aligned} \tag{A.7}$$

Expectation of the prior on ζ_i:

$$\begin{aligned}
\langle \ln p(\zeta_i)\rangle_{q(\zeta_i)} &= \left\langle \ln\mathrm{Bern}\left(\zeta_i;p_0^i\right)\right\rangle_{q(\zeta_i)} \\
&= \left\langle \zeta_i \ln p_0^i + (1-\zeta_i)\ln\left(1-p_0^i\right)\right\rangle_{q(\zeta_i)} \\
&= p_{\zeta|y}^i \ln p_0^i + \left(1-p_{\zeta|y}^i\right)\ln\left(1-p_0^i\right) \\
&= \ln\left(1-p_0^i\right) + p_{\zeta|y}^i \ln\frac{p_0^i}{1-p_0^i}
\end{aligned} \tag{A.8}$$

Entropy of θ:

$$\begin{aligned}
-\langle \ln q(\theta)\rangle_{q(\theta)} &= -\left\langle \mathcal{N}\left(\theta;\mu_{\theta|y},\Sigma_{\theta|y}\right)\right\rangle_{q(\theta)} \\
&= \frac{D}{2}\ln 2\,\pi + \frac{1}{2}\ln\left|\Sigma_{\theta|y}\right| + \frac{1}{2}\left\langle\left(\theta-\mu_{\theta|y}\right)^T\Sigma_{\theta|y}^{-1}\left(\theta-\mu_{\theta|y}\right)\right\rangle_{q(\theta)} \\
&= \frac{D}{2}\ln 2\,\pi + \frac{1}{2}\ln\left|\Sigma_{\theta|y}\right| + \frac{1}{2}\left(\mu_{\theta|y}-\mu_{\theta|y}\right)^T\Sigma_{\theta|y}^{-1}\left(\mu_{\theta|y}-\mu_{\theta|y}\right) + \frac{1}{2}\mathrm{tr}\left(\Sigma_{\theta|y}^{-1}\Sigma_{\theta|y}\right) \\
&= \frac{D}{2}(1+\ln 2\,\pi) + \frac{1}{2}\ln\left|\Sigma_{\theta|y}\right|
\end{aligned} \tag{A.9}$$

Entropy of τ:

$$\begin{aligned}
-\langle \ln q(\tau)\rangle_{q(\tau)} &= -\left\langle \mathrm{Gamma}\left(\tau;\alpha_{\tau|y},\beta_{\tau|y}\right)\right\rangle_{q(\tau)} \\
&= -\alpha_{\tau|y}\ln\beta_{\tau|y} + \ln\Gamma\left(\alpha_{\tau|y}\right) - \left\langle\left(\alpha_{\tau|y}-1\right)\ln\tau - \beta_{\tau|y}\tau\right\rangle_{q(\tau)} \\
&= -\alpha_{\tau|y}\ln\beta_{\tau|y} + \ln\Gamma\left(\alpha_{\tau|y}\right) - \left(\alpha_{\tau|y}-1\right)\langle\ln\tau\rangle_{q(\tau)} - \beta_{\tau|y}\frac{\alpha_{\tau|y}}{\beta_{\tau|y}} \\
&= \alpha_{\tau|y} - \ln\beta_{\tau|y} + \ln\Gamma\left(\alpha_{\tau|y}\right) - \left(\alpha_{\tau|y}-1\right)\Psi\left(\alpha_{\tau|y}\right)
\end{aligned} \tag{A.10}$$

Entropy of ζ_i:

$$\begin{aligned}
-\langle \ln q(\zeta_i)\rangle_{q(\zeta_i)} &= -\left\langle \ln\mathrm{Bern}\left(\zeta_i;p_{\zeta|y}^i\right)\right\rangle_{q(\zeta_i)} \\
&= -\left\langle \zeta_i\ln p_{\zeta|y}^i + (1-\zeta_i)\ln\left(1-p_{\zeta|y}^i\right)\right\rangle_{q(\zeta_i)} \\
&= -p_{\zeta|y}^i\ln p_{\zeta|y}^i - \left(1-p_{\zeta|y}^i\right)\ln\left(1-p_{\zeta|y}^i\right)
\end{aligned} \tag{A.11}$$

A.3 Expression for $\langle Z^TX^TXZ\rangle_{q(\zeta)}$:

First, we define two matrices

$$\begin{aligned}
&G = \left\langle Z^TX^TXZ\right\rangle_{q(\zeta)} = \left\langle ZX^TXZ\right\rangle_{q(\zeta)} \\
&H = P_{\zeta|y}^TX^TXP_{\zeta|y} = P_{\zeta|y}X^TXP_{\zeta|y}
\end{aligned} \tag{A.12}$$

where we made used of $P_{\zeta|y} = \langle Z\rangle_{q(\zeta)}$. Note that the right sides of Eq. (A.12) follow directly from the fact that the diagonal matrix Z is symmetric and thus $Z^T = Z$. The matrix G is the term we like to compute and we aim to do this by expressing it by making using of matrix H. For this, we examine the individual elements of each matrix

$$\begin{aligned} G_{ij} &= \left\langle \zeta_i \zeta_j \sum_{k=1}^{N} x_{ik} x_{kj} \right\rangle_{q(\zeta)} = \left\langle \zeta_i \zeta_j \right\rangle_{q(\zeta)} \sum_{k=1}^{N} x_{ik} x_{kj} \\ H_{ij} &= p^i_{\zeta|y} p^j_{\zeta|y} \sum_{k=1}^{N} x_{ik} x_{kj} \end{aligned} \tag{A.13}$$

where $p^i_{\zeta|y}$ is the posterior probability of the Bernoulli distribution over connection *i*.

$$\left\langle \zeta_i \zeta_j \right\rangle_{q(\zeta)} = \begin{cases} \langle \zeta_i \rangle_{q(\zeta_i)} \langle \zeta_j \rangle_{q(\zeta_j)} = p^i_{\zeta|y} p^j_{\zeta|y} & i \neq j \\ \langle \zeta_i^2 \rangle_{q(\zeta_i)} = \langle \zeta_i \rangle_{q(\zeta_i)} = p^i_{\zeta|y} & i = j \end{cases} \tag{A.14}$$

which follows from the fact that for any binary variable ξ, we have $\xi^2 = \xi$. Hence, we see that off-diagonal element of G and H are equivalent, whereas the terms on the diagonal differ. This leads to the following expression for the elements of G:

$$G_{ij} = \begin{cases} p^i_{\zeta|y} p^j_{\zeta|y} \sum_{k=1}^{N} x_{ik} x_{kj} = H_{ij} & i \neq j \\ p^i_{\zeta|y} \sum_{k=1}^{N} x_{ik} x_{ki} = H_{ii} + \sum_{k=1}^{N} x_{ik} x_{kj} \left(p^i_{\zeta|y} - \left(p^i_{\zeta|y} \right)^2 \right) & i = j \end{cases} \tag{A.15}$$

From Eq. (A.15) we find that the entire matrix G is given by:

$$\left\langle Z X^T X Z \right\rangle_{q(\zeta)} = P_{\zeta|y} X^T X P_{\zeta|y} + (X^T X) \circ \left(P_{\zeta|y} - P^2_{\zeta|y} \right) \tag{A.16}$$

where $\circ$ denotes the element-wise product of two matrices.

A.4 Expression for $\langle \theta^T Z^T X^T X Z \theta \rangle_{q(\theta,\, \zeta)}$:

Next, we derive an expression for the term $\langle \theta^T Z^T X^T X Z \theta \rangle_{q(\theta,\, \zeta)}$. For this, we first compute the expectation with respect to the approximate distribution over θ. Due to the quadratic from, this yields

$$\left\langle \theta^T Z^T X^T X Z \theta \right\rangle_{q(\theta,\, \zeta)} = \left\langle \mu^T_{\theta|y} Z^T X^T X Z \mu_{\theta|y} + \mathrm{tr}\left(Z^T X^T X Z \Sigma_{\theta|y} \right) \right\rangle_{q(\zeta)} \tag{A.17}$$

We can then compute the expectation with respect to $q(\zeta)$ and find:

$$\begin{aligned} \left\langle \theta^T Z^T X^T X Z \theta \right\rangle_{q(\theta,\, \zeta)} &= \mu^T_{\theta|y} \left\langle Z^T X^T X Z \right\rangle_{q(\zeta)} \mu_{\theta|y} + \left\langle \mathrm{tr}\left(Z^T X^T X Z \Sigma_{\theta|y} \right) \right\rangle_{q(\zeta)} \\ &= \mu^T_{\theta|y} \left\langle Z^T X^T X Z \right\rangle_{q(\zeta)} \mu_{\theta|y} + \mathrm{tr}\left(\left\langle Z^T X^T X Z \right\rangle_{q(\zeta)} \Sigma_{\theta|y} \right) \\ &= \mu^T_{\theta|y} \left(P_{\zeta|y} X^T X P_{\zeta|y} + (X^T X) \circ \left(P_{\zeta|y} - P^2_{\zeta|y} \right) \right) \mu_{\theta|y} \\ &\quad + \mathrm{tr}\left(\left(P_{\zeta|y} X^T X P_{\zeta|y} + (X^T X) \circ \left(P_{\zeta|y} - P^2_{\zeta|y} \right) \right) \Sigma_{\theta|y} \right) \\ &= \mu^T_{\theta|y} P_{\zeta|y} X^T X P_{\zeta|y} \mu_{\theta|y} + \mathrm{tr}\left(P_{\zeta|y} X^T X P_{\zeta|y} \Sigma_{\theta|y} \right) \\ &\quad + \mu^T_{\theta|y} (X^T X) \circ \left(P_{\zeta|y} - P^2_{\zeta|y} \right) \mu_{\theta|y} + \mathrm{tr}\left(\left((X^T X) \circ \left(P_{\zeta|y} - P^2_{\zeta|y} \right) \right) \Sigma_{\theta|y} \right) \end{aligned} \tag{A.18}$$

where we made us of Eq. (A.16) and the fact that the trace is a linear operator and hence, the order of taking the expectation and the trace can be exchanged.

A.5 Expression for $\mu_{\theta|y} \langle Z \rangle_{q(\zeta_{\backslash i})} X^T Y$:

One can show that:

$$\begin{aligned} \mu_{\theta|y} \langle Z \rangle_{q(\zeta_{\backslash i})} X^T Y &= \mu_{\theta|y} \langle Z \rangle_{q(\zeta_{\backslash i})} \upsilon \\ &= \sum_{j=1}^{D} \mu^j_{\theta|y} \left\langle \zeta_j \right\rangle_{q(\zeta_{\backslash i})} \upsilon_j \\ &= \mu^i_{\theta|y} \zeta_i \upsilon_i + \sum_{j \neq i} \mu^j_{\theta|y} p^j_{\zeta|y} \upsilon_j \\ &= \mu^i_{\theta|y} \zeta_i \upsilon_i + c \end{aligned} \tag{A.19}$$

where $\upsilon = X^T Y$.

A.6 Expression for $\mu_{\theta|y}^T \langle ZX^T XZ \rangle_{q(\zeta_{\backslash i})} \mu_{\theta|y}$:

We re-write all matrices and inner products as sums:

$$\mu_{\theta|y}^T \left\langle ZX^T XZ \right\rangle_{q(\zeta_{\backslash i})} \mu_{\theta|y} = \mu_{\theta|y}^T \langle ZWZ \rangle_{q(\zeta_{\backslash i})} \mu_{\theta|y} = \sum_{j=1}^{D} \sum_{k=1}^{D} \mu_{\theta|y}^j \mu_{\theta|y}^k \left\langle \zeta_j \zeta_k \right\rangle_{q(\zeta_{\backslash i})} W_{jk} \tag{A.20}$$

where $W = X^T X$. Next, we can inspect Eq. (A.20) for all possible combinations of j and k. This yields the following expressions as a function of ζ_i:

$$\mu_{\theta|y}^j \mu_{\theta|y}^k \left\langle \zeta_j \zeta_k \right\rangle_{q(\zeta_{\backslash i})} W_{jk} = \begin{cases} \left(\mu_{\theta|y}^i\right)^2 \zeta_i W_{ii} & i = j = k \\ \mu_{\theta|y}^i \mu_{\theta|y}^k \zeta_i p_{\zeta|y}^k W_{ik} & i = j \neq k \\ \mu_{\theta|y}^i \mu_{\theta|y}^j \zeta_i p_{\zeta|y}^j W_{ij} & i = k \neq j \\ \text{constant with respect to } \zeta_i & i \neq j, i \neq k \end{cases} \tag{A.21}$$

Inserting Eq. (A.21) into Eq. (A.20), we get the final expression

$$\mu_{\theta|y}^T \left\langle ZX^T XZ \right\rangle_{q(\zeta_{\backslash i})} \mu_{\theta|y} = \zeta_i \left(\left(\mu_{\theta|y}^i\right)^2 W_{ii} + 2\mu_{\theta|y}^i \sum_{j \neq i} \mu_{\theta|y}^j p_{\zeta|y}^j W_{ij} \right) + c \tag{A.22}$$

where all terms independent of ζ_i were absorbed into the constant term c.

A.7 Expression for $\langle tr(ZX^T XZ\Sigma_{\theta|y}) \rangle_{q(\zeta_{\backslash i})}$:

First, we can write

$$\left\langle \mathrm{tr}\left(ZX^T XZ\Sigma_{\theta|y}\right) \right\rangle_{q(\zeta_{\backslash i})} = \mathrm{tr}\left(\left\langle ZX^T XZ\Sigma_{\theta|y} \right\rangle_{q(\zeta_{\backslash i})} \right) = \mathrm{tr}\left(\left\langle ZX^T XZ \right\rangle_{q(\zeta_{\backslash i})} \Sigma_{\theta|y} \right) \tag{A.23}$$

which follows directly from the fact that the trace is a linear operator. We can now use the result from Eq. (A.22) and find the following expression:

$$\left\langle \mathrm{tr}\left(ZX^T XZ\Sigma_{\theta|y}\right) \right\rangle_{q(\zeta_{\backslash i})} = \zeta_i \left(W_{ii} \Sigma_{\theta|y}^{ii} + 2 \sum_{j \neq i} p_{\zeta|y}^j W_{ij} \Sigma_{\theta|y}^{ij} \right) + c \tag{A.24}$$

where all terms independent of ζ_i were absorbed into the constant term c.

References

Ambrogioni, L., Hinne, M., van Gerven, M., Maris, E., 2017. GP CaKe: Effective Brain Connectivity with Causal Kernels arXiv:1705.05603.

Andersen, R.A., 1997. Multimodal integration for the representation of space in the posterior parietal cortex. Philos. Trans. R. Soc. Lond. B Biol. Sci. 352, 1421–1428.

Anticevic, A., Hu, X., Xiao, Y., Hu, J., Li, F., Bi, F., Cole, M.W., Savic, A., Yang, G.J., Repovs, G., Murray, J.D., Wang, X.J., Huang, X., Lui, S., Krystal, J.H., Gong, Q., 2015. Early-course unmedicated schizophrenia patients exhibit elevated prefrontal connectivity associated with longitudinal change. J. Neurosci. 35, 267–286.

Ashourvan, A., Gu, S., Mattar, M.G., Vettel, J.M., Bassett, D.S., 2017. The energy landscape underpinning module dynamics in the human brain connectome. Neuroimage 157, 364–380.

Barmet, C., De Zanche, N., Pruessmann, K.P., 2008. Spatiotemporal magnetic field monitoring for MR. Magn. Reson. Med. 60, 187–197.

Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis. Springer.

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.M., 2006. Pattern Recognition and Machine Learning, vol. 12. Springer, New York., p. 105, 13, 47.

Bollmann, S., Kasper, L., Vannesjo, S.J., Diaconescu, A.O., Dietrich, B.E., Gross, S., Stephan, K.E., Pruessmann, K.P., 2017. Analysis and correction of field fluctuations in fMRI data using field monitoring. Neuroimage 154, 92–105.

Bolton, T.A.W., Tarun, A., Sterpenich, V., Schwartz, S., Van De Ville, D., 2018. Interactions between large-scale functional brain networks are captured by sparse coupled HMMs. IEEE Trans. Med. Imag. 37, 230–240.

Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution. In: Proceedings of the 20th International Conference on Pattern Recognition IEEE Computer Society, pp. 3121–3124.

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 e1002079.

Büchel, C., Friston, K.J., 1997. Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modelling and fMRI. Cerebr. Cortex 7, 768–778.

Buckner, R.L., Krienen, F.M., Yeo, B.T., 2013. Opportunities and limitations of intrinsic functional connectivity MRI. Nat. Neurosci. 16, 832–837.

Buhlmann, P., van de Geer, S., 2011. Statistics for high-dimensional data: methods, theory and applications. Statistics for high-dimensional data: methods. Theory and Applications, 1-+.

Bullmore, E., Sporns, O., 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. Nat. Rev. Neurosci. 10, 186–198.

Bullmore, E.T., Frangou, S., Murray, R.M., 1997. The dysplastic net hypothesis: an integration of developmental and dysconnectivity theories of schizophrenia. Schizophr. Res. 28, 143–156.

Buxton, R., Wong, E., Frank, L., 1998. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. Magn. Reson. Med. 39, 855–864.

Chang, C., Glover, G.H., 2010. Time-frequency dynamics of resting-state brain connectivity measured with fMRI. Neuroimage 50, 81–98.

Chickering, D., 2003. Optimal structure identification with greedy search. J. Mach. Learn. Res. 3, 507–554.

Coppersmith, D., Winograd, S., 1990. Matrix multiplication via arithmetic progressions. J. Symbolic Comput. 9, 251–280.

Culham, J.C., Kanwisher, N.G., 2001. Neuroimaging of cognitive functions in human parietal cortex. Curr. Opin. Neurobiol. 11, 157–163.

Daunizeau, J., David, O., Stephan, K., 2011. Dynamic causal modelling: a critical review of the biophysical and statistical foundations. Neuroimage 58, 312–322.

Daunizeau, J., Friston, K., Kiebel, S., 2009. Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models. Phys. Nonlinear Phenom. 238, 2089–2118.

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.

Davie, A.M., Stothers, A.J., 2013. Improved bound for complexity of matrix multiplication. Proceedings of the Royal Society of Edinburgh Section a-Mathematics 143, 351–369.

Deco, G., Jirsa, V.K., McIntosh, A.R., 2013a. Resting brains never rest: computational insights into potential cognitive architectures. Trends Neurosci. 36, 268–274.

Deco, G., Kringelbach, M.L., 2014. Great expectations: using whole-brain computational connectomics for understanding neuropsychiatric disorders. Neuron 84, 892–905.

Deco, G., McIntosh, A.R., Shen, K., Hutchison, R.M., Menon, R.S., Everling, S., Hagmann, P., Jirsa, V.K., 2014a. Identification of optimal structural connectivity using functional connectivity and neural modeling. J. Neurosci. 34, 7910–7916.

Deco, G., Ponce-Alvarez, A., Hagmann, P., Romani, G.L., Mantini, D., Corbetta, M., 2014b. How local excitation-inhibition ratio impacts the whole brain dynamics. J. Neurosci. 34, 7886–7898.

Deco, G., Ponce-Alvarez, A., Mantini, D., Romani, G.L., Hagmann, P., Corbetta, M., 2013b. Resting-state functional connectivity emerges from structurally and dynamically shaped slow linear fluctuations. J. Neurosci. 33, 11239–11252.

Dunn, G., Everitt, B., 1982. An Introduction to Mathematical Taxonomy. Cambridge University Press.

Duyn, J.H., 2012. The future of ultra-high field MRI and fMRI for study of the human brain. Neuroimage 62, 1241–1248.

Eavani, H., Satterthwaite, T.D., Filipovych, R., Gur, R.E., Gur, R.C., Davatzikos, C., 2015. Identifying Sparse Connectivity Patterns in the brain using resting-state fMRI. Neuroimage 105, 286–299.

Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J., 2008. LIBLINEAR: a library for large linear classification. J. Mach. Learn. Res. 9, 1871–1874.

Felleman, D.J., Van Essen, D.C., 1991. Distributed hierarchical processing in the primate cerebral cortex. Cerebr. Cortex 1, 1–47.

Fornito, A., Zalesky, A., Breakspear, M., 2015. The connectomics of brain disorders. Nat. Rev. Neurosci. 16, 159–172.

Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., Raichle, M.E., 2005. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proc. Natl. Acad. Sci. U.S.A. 102, 9673–9678.

Frässle, S., Lomakina, E.I., Razi, A., Friston, K.J., Buhmann, J.M., Stephan, K.E., 2017. Regression DCM for fMRI. Neuroimage 155, 406–421.

Friston, K., Brown, H.R., Siemerkus, J., Stephan, K.E., 2016. The dysconnection hypothesis (2016). Schizophr. Res. 176, 83–94.

Friston, K., Harrison, L., Penny, W., 2003. Dynamic causal modelling. Neuroimage 19, 1273–1302.

Friston, K., Holmes, A., Poline, J., Grasby, P., Williams, S., Frackowiak, R., Turner, R., 1995. Analysis of fMRI time-series revisited. Neuroimage 2, 45–53.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W., 2007. Variational free energy and the Laplace approximation. Neuroimage 34, 220–234.

Friston, K., Moran, R., Seth, A., 2013. Analysing connectivity with Granger causality and dynamic causal modelling. Curr. Opin. Neurobiol. 23, 172–178.

Friston, K.J., 2011. Functional and effective connectivity: a review. Brain Connect. 1, 13–36.

Friston, K.J., Frith, C.D., 1995. Schizophrenia: a disconnection syndrome? Clin. Neurosci. 3, 89–97.

Friston, K.J., Kahan, J., Biswal, B., Razi, A., 2014a. A DCM for resting state fMRI. Neuroimage 94, 396–407.

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.

Friston, K.J., Stephan, K.E., Montague, R., Dolan, R.J., 2014b. Computational psychiatry: the brain as a phantastic organ. Lancet Psychiatry 1, 148–158.

Fuster, J.M., 2003. Cortex and Mind: Unifying Cognition. Oxford University Press, Oxford.

Gelman, A., Charlin, J.B., Stern, H.S., Rubin, D.B., 2004. Bayesian Data Analysis. Chapman and Hall.

Gilson, M., Deco, G., Friston, K.J., Hagmann, P., Mantini, D., Betti, V., Romani, G.L., Corbetta, M., 2017. Effective connectivity inferred from fMRI transition dynamics during movie viewing points to a balanced reconfiguration of cortical interactions. Neuroimage.

Gilson, M., Moreno-Bote, R., Ponce-Alvarez, A., Ritter, P., Deco, G., 2016. Estimation of directed effective connectivity from fMRI functional connectivity hints at asymmetries of cortical connectome. PLoS Comput. Biol. 12 e1004762.

Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., Smith, S.M., Van Essen, D.C., 2016. A multi-modal parcellation of human cerebral cortex. Nature 536, 171–178.

Goebel, R., Roebroeck, A., Kim, D.S., Formisano, E., 2003. Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping. Magn. Reson. Imaging 21, 1251–1261.

Grefkes, C., Eickhoff, S., Nowak, D., Dafotakis, M., Fink, G., 2008. Dynamic intra- and interhemispheric interactions during unilateral and bilateral hand movements assessed with fMRI and DCM. Neuroimage 41, 1382–1394.

Grefkes, C., Ritzl, A., Zilles, K., Fink, G.R., 2004. Human medial intraparietal cortex subserves visuomotor coordinate transformation. Neuroimage 23, 1494–1506.

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C.J., Wedeen, V.J., Sporns, O., 2008. Mapping the structural core of human cerebral cortex. PLoS Biol. 6, e159.

Harrison, L., Penny, W.D., Friston, K., 2003. Multivariate autoregressive modeling of fMRI time series. Neuroimage 19, 1477–1491.

Hernandez-Lobato, D., Hernandez-Lobato, J., Dupont, P., 2013. Generalized spike-and-slab priors for bayesian group feature selection using expectation propagation. J. Mach. Learn. Res. 14, 1891–1945.

Hinne, M., Ambrogioni, L., Janssen, R.J., Heskes, T., van Gerven, M.A., 2014. Structurally-informed Bayesian functional connectivity analysis. Neuroimage 86, 294–305.

Honey, C.J., Kötter, R., Breakspear, M., Sporns, O., 2007. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. Proc. Natl. Acad. Sci. U. S. A. 104, 10240–10245.

Honey, C.J., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J.P., Meuli, R., Hagmann, P., 2009. Predicting human resting-state functional connectivity from structural connectivity. Proc. Natl. Acad. Sci. U. S. A. 106, 2035–2040.

Huys, Q.J., Maia, T.V., Frank, M.J., 2016. Computational psychiatry as a bridge from neuroscience to clinical applications. Nat. Neurosci. 19, 404–413.

Irimia, A., Chambers, M.C., Torgerson, C.M., Van Horn, J.D., 2012. Circular representation of human cortical networks for subject and population-level connectomic visualization. Neuroimage 60, 1340–1351.

Jirsa, V.K., Proix, T., Perdikis, D., Woodman, M.M., Wang, H., Gonzalez-Martinez, J., Bernard, C., Bénar, C., Guye, M., Chauvel, P., Bartolomei, F., 2016. The Virtual Epileptic Patient: individualized whole-brain models of epilepsy spread. Neuroimage.

Karahanoglu, F.I., Van De Ville, D., 2015. Transient brain activity disentangles fMRI resting-state dynamics in terms of spatially and temporally overlapping networks. Nat. Commun. 6, 7751.

Karahanoglu, F.I., Van De Ville, D., 2017. Dynamics of large-scale fMRI networks: deconstruct brain activity to build better models of brain function. Current Opinion in Biomedical Engineering 3, 28–36.

Kasper, L., Haeberlin, M., Dietrich, B.E., Gross, S., Barmet, C., Wilm, B.J., Vannesjo, S.J., Brunner, D.O., Ruff, C.C., Stephan, K.E., Pruessmann, K.P., 2014. Matched-filter acquisition for BOLD fMRI. Neuroimage 100, 145–160.

Kötter, R., Stephan, K., 2003. Network participation indices: characterizing componet roles for information processing in neural networks. Neural Network. 16, 1261–1275.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A., 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19, 1639–1645.

Ledberg, A., Bressler, S.L., Ding, M., Coppola, R., Nakamura, R., 2007. Large-scale visuomotor integration in the cerebral cortex. Cerebr. Cortex 17, 44–62.

Li, B., Daunizeau, J., Stephan, K.E., Penny, W., Hu, D., Friston, K., 2011. Generalised filtering and stochastic DCM for fMRI. Neuroimage 58, 442–457.

Lomakina, E.I., 2016. Machine Learning in Neuroimaging: Methodological Investigations and Applications to FMRI. PhD thesis. ETH Zurich. https://doi.org/10.3929/ethz-a-010639985.

Maia, T., Frank, M., 2011. From reinforcement learning models to psychiatric and neurological disorders. Nat. Neurosci. 14, 154–162.

Marreiros, A., Kiebel, S., Friston, K., 2008. Dynamic causal modelling for fMRI: a two-state model. Neuroimage 39, 269–278.

Moeller, S., Yacoub, E., Olman, C.A., Auerbach, E., Strupp, J., Harel, N., Uğurbil, K., 2010. Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. Magn. Reson. Med. 63, 1144–1153.

Montague, P., Dolan, R., Friston, K., Dayan, P., 2012. Computational psychiatry. Trends Cognit. Sci. 16, 72–80.

Murphy, K.P., 2012. Machine learning: a probabilistic perspective. Mach. Learn.: A Probabilistic Perspective, 1–1067.

Penny, W., 2012. Comparing dynamic causal models using AIC, BIC and free energy. Neuroimage 59, 319–330.

Penny, W., Roberts, S., 1999. Bayesian neural networks for classification: how useful is the evidence framework? Neural Network. 12, 877–892.

Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004a. Comparing dynamic causal models. Neuroimage 22, 1157–1172.

Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004b. Modelling functional integration: a comparison of structural equation and dynamic causal models. Neuroimage 23 (Suppl. 1), S264–S274.

Petzschner, F.H., Weber, L.A.E., Gard, T., Stephan, K.E., 2017. Computational Psychosomatics and computational psychiatry: toward a joint framework for differential diagnosis. Biol. Psychiatr.

Ponce-Alvarez, A., Deco, G., Hagmann, P., Romani, G.L., Mantini, D., Corbetta, M., 2015a. Resting-state temporal synchronization networks emerge from connectivity topology and heterogeneity. PLoS Comput. Biol. 11 e1004100.

Ponce-Alvarez, A., He, B.J., Hagmann, P., Deco, G., 2015b. Task-driven activity reduces the cortical activity space of the brain: experiment and whole-brain modeling. PLoS Comput. Biol. 11 e1004445.

Power, J.D., Cohen, A.L., Nelson, S.M., Wig, G.S., Barnes, K.A., Church, J.A., Vogel, A.C., Laumann, T.O., Miezin, F.M., Schlaggar, B.L., Petersen, S.E., 2011. Functional network organization of the human brain. Neuron 72, 665–678.

Prando, G., Zorzi, M., Bertoldo, A., Chiuso, A., 2017. Estimating Effective Connectivity in Linear Brain Network Models arXiv:1703.10363.

Ramsey, J., Glymour, M., Sanchez-Romero, R., Glymour, C., 2017. A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. Int J Data Sci Anal 3, 121–129.

Ramsey, J.D., Hanson, S.J., Hanson, C., Halchenko, Y.O., Poldrack, R.A., Glymour, C., 2010. Six problems for causal inference from fMRI. Neuroimage 49, 1545–1558.

Razi, A., Seghier, M.L., Zhou, Y., McColgan, P., Zeidman, P., Park, H.-J., Sporns, O., Rees, G., Friston, K.J., 2017. Large-scale DCMs for resting state fMRI. Network Neuroscience.

Redpath, T.W., 1998. Signal-to-noise ratio in MRI. Br. J. Radiol. 71, 704–707.

Rizzolatti, G., Luppino, G., 2001. The cortical motor system. Neuron 31, 889–901.

Roebroeck, A., Formisano, E., Goebel, R., 2005. Mapping directed influence over the brain using Granger causality and fMRI. Neuroimage 25, 230–242.

Roland, P.E., Zilles, K., 1996. Functions and structures of the motor cortices in humans. Curr. Opin. Neurobiol. 6, 773–781.

Rolls, E.T., Cheng, W., Gilson, M., Qiu, J., Hu, Z., Ruan, H., Li, Y., Huang, C.C., Yang, A.C., Tsai, S.J., Zhang, X., Zhuang, K., Lin, C.P., Deco, G., Xie, P., Feng, J., 2018. Effective connectivity in depression. Biol Psychiatry Cogn Neurosci Neuroimaging 3, 187–197.

Rubinov, M., Sporns, O., 2010. Complex network measures of brain connectivity: uses and interpretations. Neuroimage 52, 1059–1069.

Ryali, S., Chen, T., Supekar, K., Menon, V., 2012. Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty. Neuroimage 59, 3852–3861.

Sanchez-Romero, R., Ramsey, J.D., Zhang, K., Glymour, M.R.K., Huang, B., Glymour, C., 2018. Causal discovery of feedback networks with functional magnetic resonance imaging. https://doi.org/10.1101/245936.

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.

Seghier, M.L., Friston, K.J., 2013. Network discovery with large DCMs. Neuroimage 68, 181–191.

Senden, M., Reuter, N., van den Heuvel, M.P., Goebel, R., Deco, G., Gilson, M., 2018. Task-related effective connectivity reveals that the cortical rich club gates cortex-wide communication. Hum. Brain Mapp. 39, 1246–1262.

Seth, A.K., 2010. A MATLAB toolbox for Granger causal connectivity analysis. J. Neurosci. Meth. 186, 262–273.

Smith, S., 2012. The future of FMRI connectivity. Neuroimage 62, 1257–1266.

Smith, S.M., Miller, K.L., Salimi-Khorshidi, G., Webster, M., Beckmann, C.F., Nichols, T.E., Ramsey, J.D., Woolrich, M.W., 2011. Network modelling methods for FMRI. Neuroimage 54, 875–891.

Sporns, O., 2013. Network attributes for segregation and integration in the human brain. Curr. Opin. Neurobiol. 23, 162–171.

Sporns, O., Tononi, G., Edelman, G.M., 2000. Theoretical neuroanatomy: relating anatomical and functional connectivity in graphs and cortical connection matrices. Cerebr. Cortex 10, 127–141.

Sporns, O., Zwi, J.D., 2004. The small world of the cerebral cortex. Neuroinformatics 2, 145–162.

Stephan, K., Baldeweg, T., Friston, K., 2006. Synaptic plasticity and dysconnection in schizophrenia. Biol. Psychiatr. 59, 929–939.

Stephan, K., Kasper, L., Harrison, L., Daunizeau, J., den Ouden, H., Breakspear, M., Friston, K., 2008. Nonlinear dynamic causal models for fMRI. Neuroimage 42, 649–662.

Stephan, K., Mathys, C., 2014. Computational approaches to psychiatry. Curr. Opin. Neurobiol. 25, 85–92.

Stephan, K., Penny, W., Daunizeau, J., Moran, R., Friston, K., 2009a. Bayesian model selection for group studies. Neuroimage 46, 1004–1017.

Stephan, K.E., 2004. On the role of general system theory for functional neuroimaging. J. Anat. 205, 443–470.

Stephan, K.E., Iglesias, S., Heinzle, J., Diaconescu, A.O., 2015. Translational perspectives for computational neuroimaging. Neuron 87, 716–732.

Stephan, K.E., Tittgemeyer, M., Knösche, T.R., Moran, R.J., Friston, K.J., 2009b. Tractography-based priors for dynamic causal models. Neuroimage 47, 1628–1638.

Stephan, K.E., Weiskopf, N., Drysdale, P.M., Robinson, P.A., Friston, K.J., 2007. Comparing hemodynamic models with DCM. Neuroimage 38, 387–401.

Stirnberg, R., Huijbers, W., Brenner, D., Poser, B.A., Breteler, M., Stocker, T., 2017. Rapid whole-brain resting-state fMRI at 3 T: Efficiency-optimized three-dimensional EPI versus repetition time-matched simultaneous-multi-slice EPI. Neuroimage 163, 81–92.

Swinburn, K., Porter, G., Howard, D., 2004. Comprehensive Aphasia Test. Psychology Press, New York.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society Series B-Methodological 58, 267–288.

Tononi, G., Sporns, O., Edelman, G.M., 1994. A measure for brain complexity: relating functional segregation and integration in the nervous system. Proc. Natl. Acad. Sci. U. S. A. 91, 5033–5037.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15, 273–289.

Valdes-Sosa, P.A., Roebroeck, A., Daunizeau, J., Friston, K., 2011. Effective connectivity: influence, causality and biophysical modeling. Neuroimage 58, 339–361.

Vidaurre, D., Abeysuriya, R., Becker, R., Quinn, A.J., Alfaro-Almagro, F., Smith, S.M., Woolrich, M.W., 2017. Discovering dynamic brain networks from big data in rest and task. Neuroimage.

Witt, S.T., Laird, A.R., Meyerand, M.E., 2008. Functional neuroimaging correlates of finger-tapping task variations: an ALE meta-analysis. Neuroimage 42, 343–356.

Xia, M., Wang, J., He, Y., 2013. BrainNet Viewer: a network visualization tool for human brain connectomics. PLoS One 8 e68910.

Xu, J., Moeller, S., Auerbach, E.J., Strupp, J., Smith, S.M., Feinberg, D.A., Yacoub, E., Uğurbil, K., 2013. Evaluation of slice accelerations using multiband echo planar imaging at 3 T. Neuroimage 83, 991–1001.

Zeki, S., Shipp, S., 1988. The functional logic of cortical connections. Nature 335, 311–317.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. B Stat. Meth. 67, 301–320.