ESTIMATING FUNCTIONAL CONNECTIVITY IN FUNCTIONAL MAGNETIC RESONANCE IMAGING

TUE HERLAU ETC. ETC.

Abstract. etc. etc.

1. Comments

The overall message is

- All being equal, MCMC sampling is preferable over VB
- A simpler model which behaves as well as a complex model is preferable; we offer a simpler model with on par or better performance to DCM
- We make a few specific recommendations over a basic AR(1) model which appear to be important
- Reliably testing methods for functional connectivity is difficult; principled way is multi-subject analysis and comparison.
- (if works) estimate of functional connectivity for HCP/whatever.

Informal statement of questions likely to be raised by reviewers (main goal of introduction+theory section is to preempt these in a less direct way)

- DCM based on general stochastic differential equation+1st order Taylor expansion. This is more general than what you are doing and makes model parameters in DCM more "physical".

  > *Basic structure of DCM and MDS is mathematically equivalent except minor differences in noise structure. DCM may benefit from modelling hemodynamics better, however benefits of this must be demonstrated experimentally before they can be claimed*

- A benefit of DCM is it allows estimation of marginal likelihood and thereby comparison of models/selection between "hypothesis". This is not true for your model

  > *It is important to understand that selecting between models using the marginal likelihood is a procedure/practice applied to the DCM and not a part of the DCM model. The procedure of using marginal likelihood to select between models is simply an application of Bayes theorem and an assumption we can rule out a great number of models. If this information is available MDS-MCMC too can take advantage of this information, albeit slightly differently, but with the advantage we avoid the VB approximation of the marginal likelihood which introduce a source of uncertainty in model selection procedure that is not quantified. We show this can lead to the wrong model being selected in an example*

- Didn't Friston argue an AR(1) model is not the way to go?

Friston point out AR(1) models is a larger class of models than the Ornstein-Uhlenberg process which is the basis of the DCM. This is not an argument for or against using either

- Didn't Friston argue sampling is prohibitedly expensive?

  Sampling IS more expensive, however computers have also grown stronger, and sampling has the potential of overcoming fundamental limitations of VB (such as local minima). If these advantages can be cached in in practice must be tested by showing the sampler converges.
- Your method is stochastic because of sampling and must run an infinite amount of time. That makes the results unscientific and method too cumbersome to be of use

  Sampling-based methods are used in many contexts for scientific data analysis. VB arguably have worse issues than sampling since VB is not guaranteed to converge. The "nice" behaviour of VB in e.g. the DCM is a result of a deterministic (and somewhat subjective) initialization being used.
- Method fundamentally incompairable to DCM: DCM selects between hypothesis, your method fit parameters. This tells us nothing

  DCM select between models based on an estimate of the marginal likelihood without a way of checking how good that estimate is. This produces a simpler answer, but the answer is confounded by factors which do not have to do with data or the model; our method estimates a probability density of connectivity matrices which, at least in principle, can be used to quantify normal/abnormal connectivity. The situation should be compared to logistic regression and determination of which factors are important for the problem: Standard methods estimate a distribution of the weight of each factor and the determination if a factor should be included in the model relies on whether that distribution significanly overlap with 0 or not. DCM-type selection on factors based on marginal likelihood estimates could be used (with the benefit the estimate of the marginal likelihood is better for logistic regression than VB!), however this is not common practice.

2. Introduction

A prominent goal of neuroscience is understanding how interaction between brain structures supports cognitive tasks. It is widely believed such an understanding must go beyond localization regions whose activity correlate with task-related stimuli but should consider functional interaction of spatially segregated areas and their dependence on context such as task or cognetive state.

According to this view, the brains function should be understood as the flow of information between brain regions and the functional role of any brain component is defined by it's *connections* to other brain regions McIntosh [2000]. It is worth distinguishing between thee different types of connnectivity in this context. (i) structual connectivity as defined by axonal connectivity in brain tissue (ii) functional connectivity as defined by a statistical relationship between the time-series of two areas without making assumptions of influence (iii) effective connectivity defined as influence one region exert over another.

The first goal is accomplished using invasive techniques such as dissection or specialized computational methods such as diffisuion tensor imaging. The second and third goal may appear superficially similar but comes with important differences. Prominent examples for functional connectivity involves simple linear correlation or more theoreticallly motivated measures such as Granger causal analysisGranger [1969] which quantify functional connectivity between two time series as how well knowing one time series aids our ability to predict the future of the other Goebel et al. [2003] or the related information-theoretic quantity of transfer entropy Schreiber [2000].

Models for effectivity connectivity explicitly constructs a biophysically motivated generative model of the observed fMRI data. Deriving from how the model is formulated, it contains one or more matrices encoding the "functional interaction" between different brain regions and, when the model is trained on fMRI data, can then be used to quantify the functional dependency. The most prominent example of models in this class is Dynamic Causal Modeling (DCM) Friston et al. [2003, 2008a, 2010] and the related Multivariate Dynamical Systems (MDS) model Ryali et al. [2011] which is very similar to our proposed procedure.

The main advantage of models for effective connectivity such as the DCM over models of functional connectivity such as GCA is that the estimated parameters will, assuming the model is biophysically accurate, be expected to grossly reflect our belief of biophysically relevant quantities such as synaptic efficacies given the available data. This difference can be compared to describing the (observed) orbit of a planet using the statistics of various quantities (orbit time, maximum velocity, etc.) and a model based on Newtons laws. The advantage of the later model is not as much in describing the phenomena, but that it allows us to derive distributions of physically relevant quantities such as the gravitational constant or the energy which are presumably of greater scientific interest.

What allows this interpretation is two things. Firstly, that the model accurately describes both the dynamics (i.e. functional dependences) of the system and statistical properties of noise sources and acquisition biases, secondly, that inference in the model is handled using Bayes theorem. Both of these assumptions can be called into question for the DCM and MDS model. Regarding the first item there are obvious limitations in both the assumption of how the neural signal evolves, the hemodynamic model as well as various acquisition/processing biases and regarding the second, neither DCM or MDS relies on exact Bayesian inference but rather approximate Variational Bayesian inference. These concerns are well known and will apply to nearly any real-world Bayesian modeling situation in which VB is used.

As theoreticians, we are interested in making better models. From a Bayesian perspective it is evidence how we should proceed: By making our models better conform to the phenomena we model, which amounts to making the model more elaborate. Examples of this for the DCM includes ... . Our work differs in two important respects. Firstly, our focus is to avoid the use of an approximate inference scheme, but replace this with asymptotically exact Markov-chain Monte-Carlo inference. This will allow us to know what a given model *actually* says about the data rather than the VB approximation. This may seem like an obtuse point, however we will illustrate this approximation have qualitatively important consequences in terms of what structures the method prefers. Secondly, rather than seeking a complicated model, our proposed method can be seen as a simplified variant of the DCM with a few tweeks. This simplification is required to make MCMC inference feasible, however understanding how well simpler models can solve a task before considering more elaborate procedures (which contain more degrees of freedom) is crucial to justify additional complexity. Any such discussion hinge on a fair evaluation of the model. We propose a data-driven validation procedure based on 3 multi-subject datasets which is free of subjective interpretations or assumptions about a known ground-truth. Surprisingly, we show our simpler method performs on par or better than the DCM and MDS, which calls the justification of the additional complexity of these models into question.

NOTES:. Only having an approximately correct model is not unique to fMRI data analysis but applies to *any* real-world modeling situation and the remedy is to express due concern and proceed. However for a purely unsupervised task such as recovery of functional interaction in fMRI data

despite the limitations of the model for evolution of the neural signal and the hemodynamic model,

it is much easier to interpret causality results from DCM in the standard framework of neuronal networks, despite obvious current limitations on the neuronal and hemodynamic models. This is because DCM connectivity parameters grossly resemble synaptic efficacies of conductance-based neural models. In contrast, GCA of fMRI time series is not sensitive to lack of biophysical knowledge, as DCM can be, but is then entirely dependent on the acquisition modality and associated biases. (...) Only biophysical modelling, such as the one proposed in DCM or other generative frameworks, that tries to correct for experimental biases will ensure to stick to the core of biological processes that are the true events of interest.

The main difference between these models

Despite many successfull applications of these models, there remains open questions about DCM-type models Lohmann et al. [2012].

This paper offers two contributions. Firstly, we will revisit some of the concerns about DCM, and propose a variant of the MDS model which has as it's main advantage that it use Markov-Chain Monte-Carlo (MCMC) sampling for inference as well as a few other minor modifications. The proposed model is simpler than DCM, yet offers similar power in terms of resolving underlying structure. The second contribution is the evaluation and comparison of our proposed model (MDS-MCMC) against MDS and DCM, where we highlight limitations of existing validation procedures.

Granger [1969], Goebel et al. [2003], Schreiber [2000]

These tools can be divided into tow major groups. The first group are the model-free methods. These methods attempt to quantify functional dependency by considering the signal in few, isolated regions. Prominent examples are

3. Assumptions and limitations of causal modeling

Treating the DCM, MDS and our proposed model in a unified fashion is difficult because the underlying mathematical framework is different (continuous-time stochastic processes vs. a simple autoregressive process) and the special-purpose Variational Bayes framework used by the DCM. We will therefore initially focus on a simplified variant of the DCM and MDS where non-important differences are not considered. We will initially focus on the DCM and then explain how it relates to the MDS and our proposed method.

Consider the following setup. After processing, we have available M time-series, $y_m(t)$, such that $y_m(t)$ is the (observed) BOLD activation of region m at time t. The signal y is generated from an (unobserved) underlying neural signal $s_m(t)$ and it is assumed these evolves according to the stochastic differential equations:

$$\dot{s}(t) = f(s(t)) + dW, \quad y(t) = h(z(t)), \; \dot{z}(t) = g(z_m(t), s_m(t)) \tag{3.1}$$

In the DCM, it is assumed f is linear and thus we are left with the *Ornstein-Uhlenbeck* process defined by the Langevin equation

{{dcmL}}

$$\frac{d}{dt}s(t) = \mathcal{C}s(t) + \epsilon(t), \quad \epsilon(t) \sim \mathrm{WN}(0, \Sigma_0) \tag{3.2}$$

The matrix $\mathcal{C}$ is what we are interested in and commonly has the interpretation that $\mathcal{C}_{ij}$ is the strength with which region j influence region i. In a practical situation we observe the signal $y(t)$ at discrete time points $0, \Delta$, 2Δ and so on. It is reasonable to consider how a realizations of the Ornstein-Uhlenbeck process behaves at these time points and the answer turns out to be simple. If we define $s_t = s(\Delta t)$ then this sample forms an AR(1) process

{{dcmAR}}

$$s_{t+1} = C s_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \Sigma) \tag{3.3}$$

Where importantly $C = e^{\Delta C_0}$ and for completeness $\Sigma = \int_0^\Delta e^{C_0 t'} \Sigma_0 e^{C_0^T t'} dt'$. The MDS Ryali et al. [2011] takes the AR(1) process eq. (3.3) as a starting point and provides an estimate of C.

This relationship is naturally well-known in the DCM litterature Friston [2011], however there seems little principled reason to prefer one formulation over another. One can argue that eq. (3.2) is more principled, stemming from a Tayler expansion of a general stochastic differential equation as the signal and this makes $\mathcal{C}$ more accurately reflect "influence" between the signals. However the true interpretation of $\mathcal{C}$ or C stem from what they actually do, in eq. (3.2)and eq. (3.3), and both apparently allows us to identify an element from C or $\mathcal{C}$ as the influence of j over i.

A point brought up by Friston [2011] is that while every OU process corresponds to an AR(1), the converse is not true (consider for instance the AR(1) process $s_{t-1} = -s_t$). This is true, however it is essentially an argument about the prior distribution of C and, since the prior of C or $\mathcal{C}$ is likely to be chosen based on analytical convenience rather than real knowledge, it is far from clear which if any practical ramifications this has. A person could just as well argue we should prefer the more general AR(1) model class. A final argument is that a discretization of the OU process to an AR(1) process must also discretize the observational differential equation g and something may be lost when this is carried out. This is true, however the observational differntial equation has a fairly simple behavior and is itself an approximation of what the true hemodynamic is which is both subject and region specific. Simply put, it is not known what is lost by discretizing the hemodynamic function if anything.

The bottom line is that while the continious process may appear superior mathematically, it represents a smaller model class and we do not know any a-priori arguments why it should be considered superior to the simpler AR(1) and rely on more familiar estimation approaches.

{sec:modselect}

3.1. **Model selection and the DCM.** An important aspect of the litterature on DCM and studies applying DCM is its use for selecting a "preferred" model within a set of specified candidate models; in the context of our discussion, a model is a sparsity pattern for $\mathcal{C}$ (i.e. which elements are different than 0) which represents one plausible idea of how the considered regions interact. This use of DCM will not play a role in our evaluation of DCM and it may therefore appear surprising we discuss it, however since the use of DCM to select between different models has become an intrinsic part of what DCM is believed to *be* as well as the source of the most fierce controversy Lohmann et al. [2012], Friston et al. [2013] we will here elaborate on this point and why we feel justified in avoiding it.

To be more specific, when DCM is applied for model selection the user propose a set of models $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_K$ representing different (this set can range into the hundreds) and DCM is used to select between them. Suppose $\mathbf{y}$ denote all observed data $\{y_{mt}\}$, $\mathbf{x}$ all (latent) parameters such as s, non-zero elements of C, etc. The marginal likelihood (how plausible the observed fMRI data is on a given model) is then defined as:

$$p(\mathbf{y}|\mathcal{M}_k) = \int p(\mathbf{y}|\mathbf{x}, \mathcal{M}_k) p(\mathbf{x}|\mathcal{M}_k) d\mathbf{x}. \tag{3.4}$$

Bayes theorem then gives

$$p(\mathcal{M}_k|\mathbf{y}) = \frac{p(\mathbf{y}|\mathcal{M}_k)p(\mathcal{M}_k)}{\sum_{k'=1}^{K} p(\mathbf{y}|\mathcal{M}_{k'})p(\mathcal{M}_{k'})} \tag{3.5}$$

Assuming all models are equally plausible a-priori this suggests the model with the highest marginal likelihood is preferable.

In a review of a previously published DCM study, Lohmann et al. [2012] brings up the point that the set of all possible model is combinatoriallarge and therefore unfeasible to explore and furthermore, that when additional models are introduced to the set of models originally tested DCM will prefer a biologically implausible model. We cannot summarize the response here fully, however Friston et al. [2013] stress this critique is a misunderstanding of how the DCM is supposed to be applied and the way models are not grouped may be misleading.

To meaningfully discuss this subject it is important to stress what the DCM is and is not. The formula eq. (3.5) is general and can be applied just as well to e.g. a logistic regression model or anything else. When this is not done more often it is because the estimators of the marginal likelihood is a difficult problem and selecting between models is therefore usually done by assuming "the model" (in this case the sparsity pattern) is another element of the variable space $\mathbf{x}$ and learn it explicitly during inference; slap-and-spike priors are examples of models which take this approach. The use of eq. (3.5) therefore do not represent anything specific to DCM but rather two additional assumptions: Firstly, a belief the marginal likelihood can be estimated with sufficient reliability and secondly, that we *know* a-priori that some models can be *ruled out*, specifically those with a sparsity pattern which do not match any of the models. In other words, the second assumption is nothing more than saying we have a particular prior for $\mathcal{C}$ and the goodness of this assumption is subject to the usual consideration for any prior in a Bayesian model.

Assuming we *can* estimate the marginal likelihood accurately, this allows us to state exactly when this procedure should and can be applied: If we *know* some models can be ruled out, then ruling them out is correct, otherwise ruling them out is not correct. This is *purely* a statement about our prior belief and therefore, whether the model selection procedure eq. (3.5) should be applied, can *only* be answered by consulting what we know of a given problem rather than any general consideration.

Keeping this in mind we can better make sense of what to make of the result that an "unphysical" model is preferable to a physical model according to eq. (3.5) as Lohmann et al. [2012] found. Assume for a second our notion of "physical" model is true (i.e. we do know what parts of the brain should physically interact with others for a given problem), it can mean one of two things: Either that the model is bad, such that even if given an abundance of data it will converge to the wrong result, or that the model is good but the amount of data is so small we only obtain reliable results if we use all our prior information. This may be worrying but not wrong. The situation can be compared to a image recognition system to determine which animal is in an image where the system can determine if an image of a cat contain a cat (and not a dog) if given only these two options, but if allowed to select between more animals it believes the picture contains an elephant. This does not mean the system cannot give the right answer, assuming we limit ourselves to cats and dogs, but it certainly do not instill much confidence the underlying model is sound.

To summarize, when using eq. (3.5) we are simply making the additional assumptions we can estimate the marginal likelihood fairly robustly and that we have additional information available that allows us to rule out all models except those matching $\mathcal{M}_1, \dots, \mathcal{M}_K$. It is therefore fair to divorce our evaluation of DCM from the model-selection procedure as this is simply equivalent to assuming we have *less* information available than what might ideally be the case. This will likely make the inference problem harder (which can be compensated for by consider simpler problems or use more data), however it does not correspond to any changes to the underlying model. We will therefore compare models according to how well they estimate C without assuming a known sparsity pattern.

3.2. **Limitations.** Both DCM and MDS use variational Bayes to infer the parameters of the model. Using our notation from above, suppose $p(\mathbf{y}, \mathbf{x})$ represent our model. VB assumes we have access to a (simple) density over $\mathbf{x}$ $q_\varphi(\mathbf{x})$ where φ is a set of parameters; the typical choice is a multivariate normal distribution. Variational Bayes then make use of the relationship:

$$(3.6) \quad \{\{\texttt{eq:KL2}\}\} \log p(\mathbf{y}) = \mathcal{L}(\varphi) + D_{\mathrm{KL}}(q_\varphi | p), \quad \mathcal{L}(\varphi) = -\int q_\varphi(\mathbf{x}) \log \frac{q_\varphi(\mathbf{x})}{p(\mathbf{y}, \mathbf{x})} d\mathbf{x}$$

Since the Kullback-Leibner divergence is always positive the value of φ which maximize $\mathcal{L}$ will also minimize the Kullback-Leibner divergence.

The advantage of VB is that it transforms the inference problem into an optimization problem which, provided the integral in (3.6) is tractable, can be solved using standard methods. Furthermore $\mathcal{L}(\varphi)$ provides a lower bound on the true (but in nearly all instances, uncomputable marginal likelihood) this estimate can in turn be used for model selection. This treatment of VB is deliberately vague and leave out important details relating to how the optimization problem over φ is to be carried out or how VB should be applied to a continuous process, however

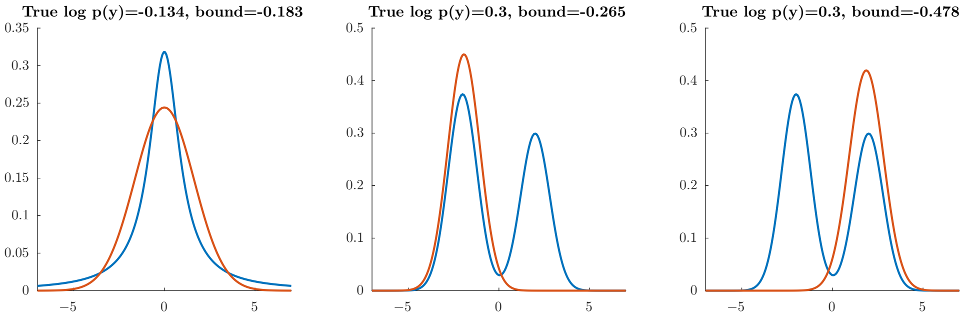


Figure 1. See `Latex/matlab/VBfigs.m` for code {fig:VB1}

it allows us to appreciate the following two issues which persist in more elaborate applications such as DCM

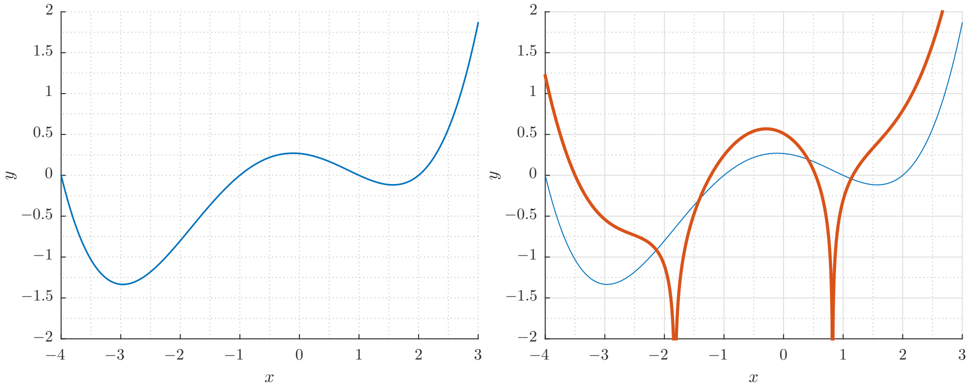
Firstly, being an approximate inference method, it will inherently give an approximate answer, i.e. a posterior distribution over the parameters of interest which does not coincide with the true distribution. This is illustrated in the right-hand pane of fig. 1 where a Gaussian variational distribution q_φ (red line) is fitted to the true posterior $p(\mathbf{y}, \mathbf{x})$ (blue line) using eq. (3.6). As a rule of thumb, VB underestimate the variability in the posterior providing overconfident answers as evident in the figure, however especially if the posterior is multi-modal the VB approximation can be arbitrarily bad.

Secondly, as VB is formulated as an optimization method, the result will depend on the specifics of the VB objective function, the initialization of the optimization method and which optimization method is used. This is illustrated in fig. 1 (middle and right) where the VB approximation was initialized differently and gave divergent results. The result produced by VB is therefore not simply a function of the data and model but also depends on how the method is initialized.

3.3. **VB and DCM.** Recall from the discussion in section 3.1, DCM use the estimate of the marginal likelihood to select between different models according to eq. (3.5). The problem is the quality of this estimate depend on the specifics of p and q_φ. In fig. 1 this is illustrated by computing the true marginal likelihoods numerically according to which model $\mathcal{M}_2$ is preferred over $\mathcal{M}_1$. However the VB estimates of the marginal likelihoods prefer $\mathcal{M}_1$ over $\mathcal{M}_2$ and also depend on the initialization. As the marginal likelihood is intractable for a realistic problem it is very difficult to say how far or how close the estimate is to the true marginal likelihood and this use of the bound therefore provides a source of uncertainty in the model selection procedure which seem difficult to address. The one thing which can perhaps be said with some confidence is that VB is plausibly biased towards less complex models relative to an exact application of Bayes theorem. To see this, assume q_φ and p both factorize fully. In this case:

$$p(\mathbf{y}) = \mathcal{L}(\varphi) + \sum_{i=1}^{K} D_{\mathrm{KL}}(q_\varphi(x_i)|p(x_i)) \approx \mathcal{L}(\varphi) + K D_0 \tag{3.7}$$

where in the last line we have assumed each KL divergence is equal to D_0 for simplicity. Thus we see that the larger K is, the further from $p(\mathbf{y})$ the lower bound $\mathcal{L}(\varphi)$ can be expected to be, creating a bias towards simpler models.



{dcm:GF1B} FIGURE 2. See DCM_deconvolve/GF_plots.m for code

Finally, this discussion have focused on generic aspects of VB which are certainly relevant to DCM or MDS, however neither actually use the "simple" VB objective eq. (3.6). MDS applies a combination of VB updates for some variables and maximization of others which has the important property the marginal likelihood is not estimated and the posterior will not agree with the "true" VB posterior. DCM too applies a unique VB optimization scheme. The software package spm provides two methods for estimating a DCM, DEM Friston et al. [2008b] and generalized filtering (GF) Friston et al. [2010] which is the default. These methods are much more sophisticated than the technique used by MDS and we cannot fully describe them here, however, for instance Generalized Filtering applies a Laplace Approximation to the likelihood to obtain the underlying objective function to be minimized (see Eq. (2.6) of Friston et al. [2010]).

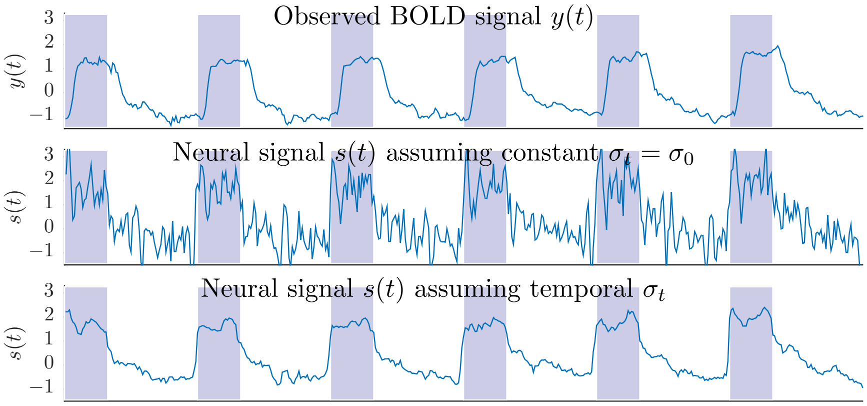
$$\mathcal{F}(u) = \tilde{\mathcal{L}}(u) + \frac{1}{2}\log|\tilde{\mathcal{L}}_{uu}| - \frac{n}{2}\log(2\pi)$$

Where $\tilde{\mathcal{L}}_{uu}$ is the second derivative after the vector parameter u of the Gibbs free energy $\tilde{\mathcal{L}}$.

An issue is the logarithmic term which, for non-trivial function, can become zero leading to the log-term diverging. Since we attempt to minimize the free energy estimate $\mathcal{F}(u)$ these points are energetically favorable. The (generalized filtering) objective is shown in fig. 2 (right)

As indicated, the objective function contains singularities corresponding to the inflection points of the likelihood function. The point is not that the the estimation method typically used in DCM is flawed, but that owning to the specifics of GF it is not simply an application of VB but rather VB with additional, specific assumptions.

To summarize, a standard application of VB to estimate a Bayesian model is not uncontroversial but have a series of known issues (dependency on initialization and the inherent approximate nature of VB). These issues are compounded by the use of specific VB estimation frameworks for both the DCM and MDS; it is not known how well the DCM estimate the marginal likelihood but this estimate is, by the very nature of VB, an approximation of the true marginal likelihood which may easily be confounded by e.g. multimodaility of the posterior which (we we will see



{rat_stimconcept}

FIGURE 3. See `tests/tests01B.m` for code

in a moment) is a very real property of both the MDS and DCM model. Similar comments applies to the MDS with the caveat it does not estimate the marginal likelihood at all and so allows no way of selecting between results obtained by different initializations or different models.

4. Proposed method

The model we will consider can be described as MDS with a few simple extensions and Markov-Chain Monte-Carlo sampling. The starting point is eq. (3.3) where we include external stimulis u_{mt}. We use $I + \Delta C_t$ for the interaction matrix to make it correspond more closely to C_0 (to see why, note $C = e^{\Delta C_0}$ and Taylor expand)

$$\mathbf{s}_{t+1} = (\mathbf{I} + \Delta\mathbf{C}_t)\mathbf{s}_t + \mathbf{U}u_{mt} + \mathrm{diag}\begin{bmatrix}\sigma_{1t} & \sigma_{2t} & \cdots & \sigma_{Mt}\end{bmatrix}\varepsilon_{mt} \tag{4.1}$$

$$y_{kt} = \sum_{l=1}^{L} \phi_{lt} y_{k,t-l+1} + \sigma_m \varepsilon_t + \mu_k^{(y)} \tag{4.2}$$

The convolution with the vector Φ_k allows a region-specific Hemodynamic response. A difference between the proposed model and the MDS is that the noise variance is σ_{mt} is not stationary. This is a simple change but appears to be important. The convolution of the hemodynamic kernel ϕ_k acts as a low-pass filter dramatically smoothening the bold signal $\mathbf{y}$. All being equal, we should thus expect $\mathbf{s}$ to have much more rapid fluctuations than $\mathbf{y}$. Since $\mathbf{C}$ is learned from $\mathbf{s}$, these rapid fluctuations will tend to make this problem more difficult. An example of this effect is shown in fig. 3 where the top pane reflects a real-world BOLD signal $y(t)$ and the middle figure the recovered neural signal $s(t)$ assuming constant $\sigma_t = \sigma_0$ and the lower pane assuming σ_t depend on time. We observed large variance in $s(t)$ was associated with poor behavior of the overall model and selecting priors for σ_t which favored lower values did not have an appreciable effect. Simply but, sometimes the neural signal *must* change appreciably due to driving events such as stimuli onset.

In the MDS model, ϕ_k is obtained as a linear combination of 3 pre-selected basis-responses $\phi = \phi_1 b_1 + \phi_2 b_2 + \phi_3 b_3$. We tried a similar approach, however the learned linear combinations would often not correspond to realistic kernels and was associated with poor behavior of the overall model. We believe this is less as an issue for MDS because the VB optimization will converge to local minima before **b** is changed appreciably from their initial value. Naturally, not being able to model a region-specific hemodynamic response should put or model at a disadvantage, and we will return to this question in section 5

4.1. **Stability and restrictions on** C**.** We implement restrictions on $\mathbf{C}$ for two reasons. Firstly, as discussed in section 3.1, it is widely believed by practitioners of DCM information about sparsity pattern of $\mathbf{C}$ is available and our model should be able to make us of this information.

Secondly, not all $\mathbf{C}$ matrices correspond to convergent processes and being able to rule out non-physical processes will greatly benefit numerical stability of the method. To illustrate this point we will briefly review some basic properties of AR(1) processes. Consider the simple autogressive form of the MDS model $s_{t+1} = (\mathbf{I} + \Delta\mathbf{C})s_t + \varepsilon_t$. The expected value of $\mathbf{s}_t$ can be considered as the superposition of K 1-d AR(1) processes s_t^k with correlated noise. Letting $\lambda_1, \dots, \lambda_M$ be the eigenvalues of $\mathbf{I} + \Delta\mathbf{C}$ the expectation of each process evolves as:

$$\langle s_{t+1}^k \rangle = e^{-\frac{t}{\tau_k}} e^{\arg \lambda_k it} \langle s_t^k \rangle \tag{4.3}$$

$$\text{with } \textit{time scale} \qquad \tau_k = \Delta \frac{-1}{\log |\lambda_k|} \tag{4.4}$$

$$\text{and } \textit{period} \qquad T_k = \Delta \frac{2\pi}{|\Im(\log \lambda_k)|} \tag{4.5}$$

The Δ terms arise because the time scale is (naturally) measured in time steps and should be converted back into units of actual walltime Δ.

For the process to be stable it must have finite time scale. This requires $0 < |\tau_k| < \infty$; assume for a moment C is diagonal in which case $\tau_k = 1 + \Delta C_{kk}$ and therefore $C_{kk} < 0$. Any realistic $\mathbf{C}$ must satisfy these requirements on the eigenvalues, however specifying a prior for the eigenvalues of $\mathbf{C}$ directly would be quite cumbersome and dramatically complicated the inference procedure. As a somewhat crude alternative we assume $\mathbf{C}$ is diagonal and specify a maximal time scale $\bar{\tau}$ such that for all k: $\tau_k \leq \bar{\tau}$ implying

$$\bar{\tau} \geq \Delta \frac{-1}{\log |\lambda_k|} \Rightarrow e^{-\frac{\Delta}{\bar{\tau}}} \geq |\lambda_k| = 1 + \Delta C_{kk} \Rightarrow C_{kk} \leq \frac{1}{\Delta}\left(e^{\frac{-\Delta}{\bar{\tau}}} - 1\right). \tag{4.6}$$

$$C_{kk} \leq \frac{1}{\Delta}\left(e^{\frac{-\Delta}{\bar{\tau}}} - 1\right). \tag{4.7}$$

In our experiments, we adopted a maximum time scale of $\bar{\tau} = 10$s. To handle the constraint of a known sparsity pattern we assumed it was possible to constrain certain values of $\mathbf{C}$ to be exactly zero corresponding to them being blocked out. The advantage with these two constraints is that they correspond to an affine restriction on C and allowed an efficient implementation without significant overhead (see section 8)

4.2. **Inference.** Suppose we are given an unnormalized density $\pi(\mathbf{x})$ and wish to obtain samples $\mathbf{x}_1, \mathbf{x}_2, \dots$ from the normalized density $p(\mathbf{x}) = \frac{1}{Z}\pi(\mathbf{x})$, $Z = \int \pi(\mathbf{x}')dx'$. This problem is in general very difficult to solve when the normalization constant Z is intractable which is generally the case in Baysian machine learning. Markov-Chain Monte-Carlo sampling solves this problem by relaxing the restriction the samples $\mathbf{x}_1, \mathbf{x}_2, \dots$ have to be i.i.d. from p but allows them to be correlated, but in such a way they reliably allows us to compute expectations wrt. p i.e.

$$\sum_{t=1}^{N} f(\mathbf{x}_i) \to \mathbb{E}_p[f] = \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x} \quad \text{for } N \to \infty. \tag{4.8}$$

This is accomplished by iteratively generating $\mathbf{x}_{t+1}$ from $\mathbf{x}_t$ by the following two-step procedure. First we generate a candidate sample from a distribution of our choice $\mathbf{x}' \sim q(\cdot|\mathbf{x}_t)$ and then we set $\mathbf{x}_{t+1} = \mathbf{x}'$ with probability

$$\min\left\{1, \frac{\pi(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')}{\pi(\mathbf{x})q(\mathbf{x}'|\mathbf{x})}\right\} \tag{4.9}$$

And otherwise setting $\mathbf{x}_{t+1} = \mathbf{x}_t$. This procedure is guaranteed to converge to the true distribution p in the sense of eq. (4.8) under very modest conditions however the convergence can be slow. A number of important extensions exist such as selecting between different transitions kernels q affecting a subset of the variables $\mathbf{x}$.

A particular simple choice for q is to sample a variable from it's marginal distribution assuming this marginal distribution has a simple form. Note for the MDS-MCMC model $s(t)$ and C (ignoring the normalization contraint for a moment) are marginally normally distributed. This allows us to sample the entire path $s(t)$ from the appropriate normal distribution and sampling C is also easily done using the method of ... . The other quantities of interest are sampled using random-walk MCMC sampling. (+ more description and cleanup; move to appendix)

{sec:simulations}

5. Simulations

Evaluating unsupervised learning methods is inherently difficult because ground-truth is often either unavailable or to some extend subjective.

6. Artificial dataset

We generated artificial datasets by first setting the "stimulated" region $s_1(t)$ to a fixed box-response function, then applied forward simulation according to the update equation

$$s(t+1) = s(t) + \Delta_t C s(t)$$

Then we added noise to the signal

$$s(t) \leftarrow s(t) + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma I),$$

then applied a bold convolution to obtain $\hat{y}(t)$ and finally added more noise to obtain the observed signal

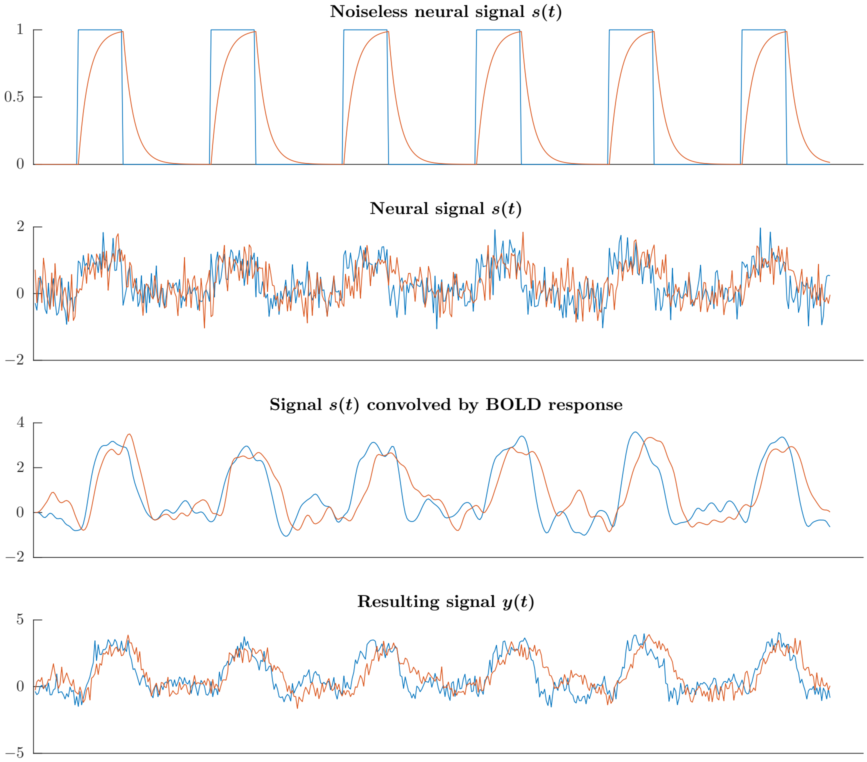
$$y(t) = \hat{y}(t) + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma I)$$

An example of this procedure where $\Delta_t = \sigma = \frac{3}{4}$ and

$$C = \frac{1}{5}\left(\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} - I\right)$$

is shown in fig. 4. The three other examples used can be seen in fig. 5

6.1. **fMRI data.** Master results



{DCMadata1A}

FIGURE 4. See `MDSadata/MDS_adata_basicplots.m` for code

TABLE 1. Datasets, see `latex/matlab/tbl_datasets.m` for code

{tbl:datasets}

Name	Timepoints	Nr.ROIs	Subjects	TR	Regions
HCP	405	5	122	0.72	IAI, rAI, rMFG, IFEF, rFEF
Yale	295	3	59	2	rACC, rAI, rDLPFC, rIFG, rPPC
StopGo	756	3	14	0.49	lM1, lPut, lSMA, lVis, rVis
Opto-rats	480	3	3	0.75	M1, Thalamus, Insula
Oddball	200	5	15	2	rAI, rdACC, rVLPFC, rPPC, rDLPFC

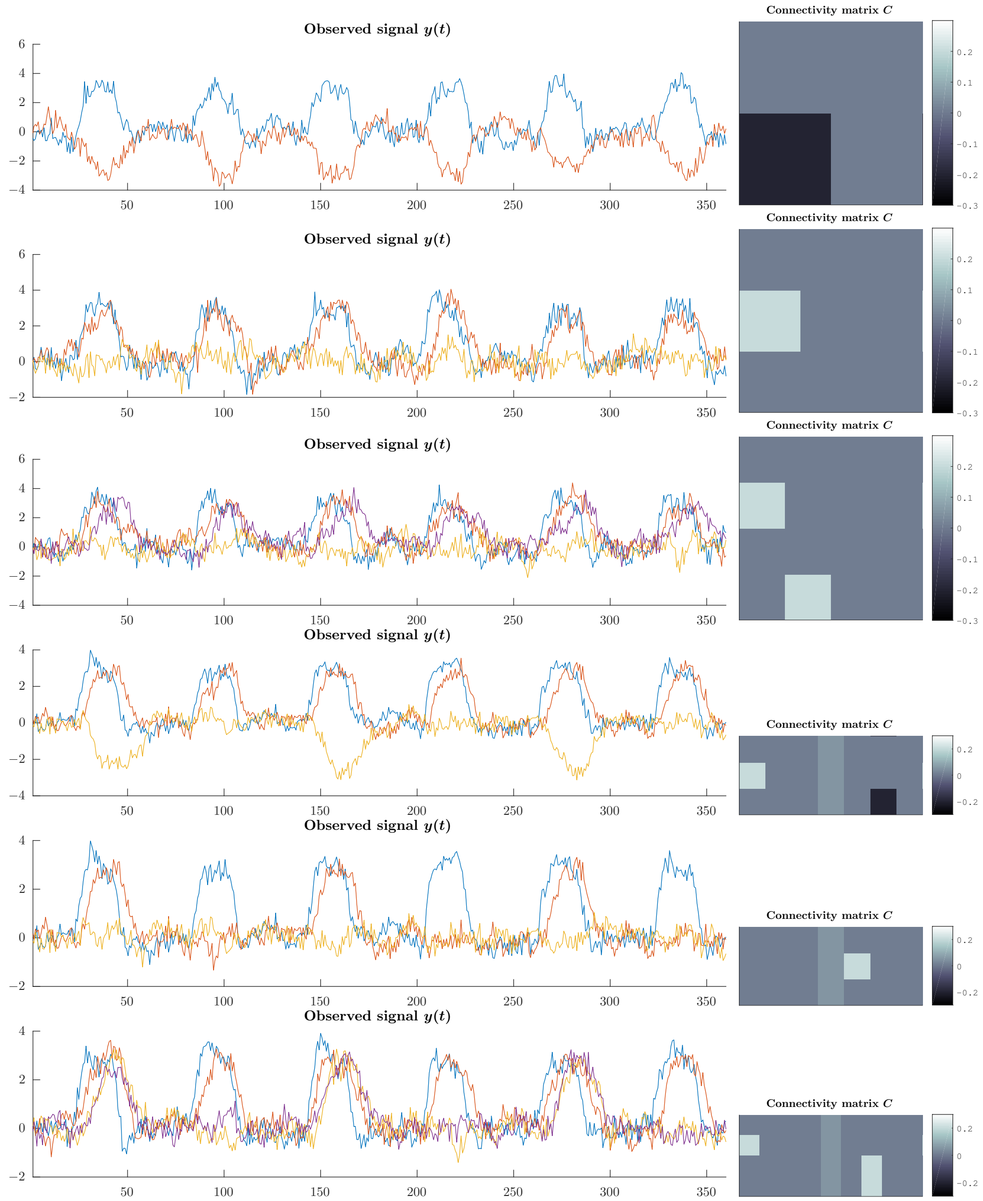
7. Opto-Rat data

We analyzed the Rat-dataset in the canonical settting (i.e. unknown stimulus patterns) using both each of the three rats individually and all jointly. The "true" interaction is region 1 to region 2 and we compared this stimulus pattern to the ones found using AUC score. Results are shown in fig. 7 with the first three rows being the three rats independently analysed and the last row when the three rats are jointly analyzed.

Concept figure of rats fig. 8

7.1. **Stopgo.**

7.2. **Oddball.**



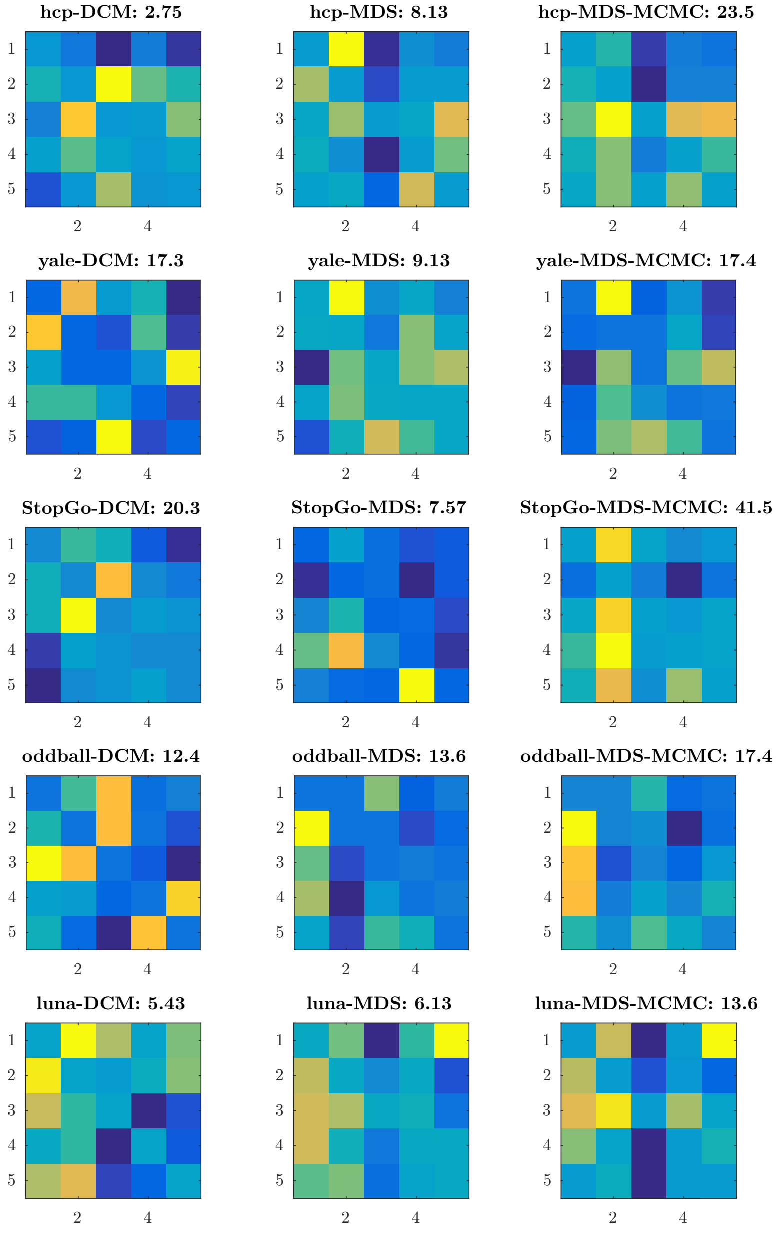
{DCMadata1B}

FIGURE 5. See `MDSadata/MDS_adata_basicplots.m` for code

7.3. **HCP.** Results on HCP data can be seen in fig. 11

7.4. **Yale.**

7.5. **Evaluation.** (Describe how consistency is evaluated here.) See table 2 for results.



hcp-DCM: 2.75

hcp-MDS: 8.13

hcp-MDS-MCMC: 23.5

yale-DCM: 17.3

yale-MDS: 9.13

yale-MDS-MCMC: 17.4

StopGo-DCM: 20.3

StopGo-MDS: 7.57

StopGo-MDS-MCMC: 41.5

oddball-DCM: 12.4

oddball-MDS: 13.6

oddball-MDS-MCMC: 17.4

luna-DCM: 5.43

luna-MDS: 6.13

luna-MDS-MCMC: 13.6

{allmatrices} FIGURE 6. See `plottools/Kendalmatrix2.m` for code

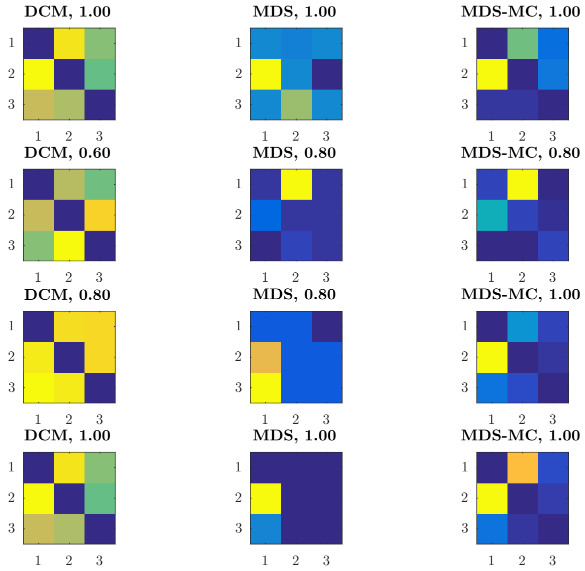
{sec:supplementaryA}

8. SUPLEMENTARY A

PLACEDHOLDER (implementation details.)

8.1. **Details on pars.** As a prior for the elements of C we selected a centered Normal-Gamma distribution where $C_{ij} \sim \mathcal{N}(0, \sigma^2_{C,ij})$ and for each precision parameter:

$$\sigma^2_C \sim \text{Gamma}(\alpha_C, \beta_C)$$



{rat_fig1} FIGURE 7. See `rats/rats_DCM_MDS_grid_plot.m` for code

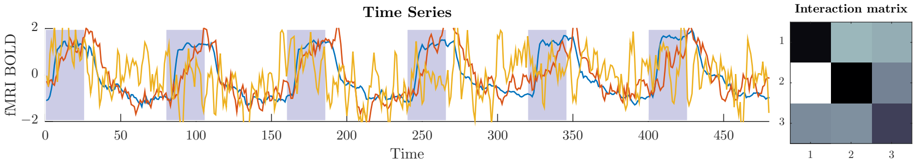


FIGURE 8. See `rats/rats_gridA_plot_single.m` for code {rat_concept}

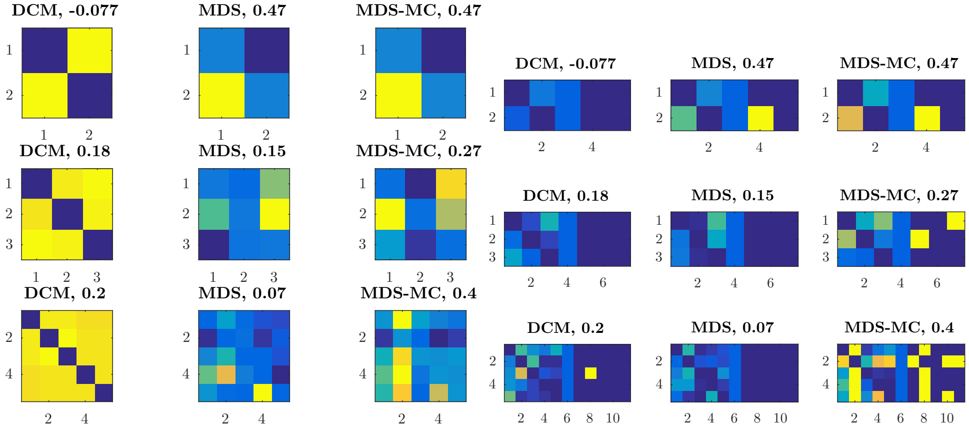


FIGURE 9. See `StanfordStopGo/stopgo_DCM_MDS_gridB_plot.m` for code {stopgoA}

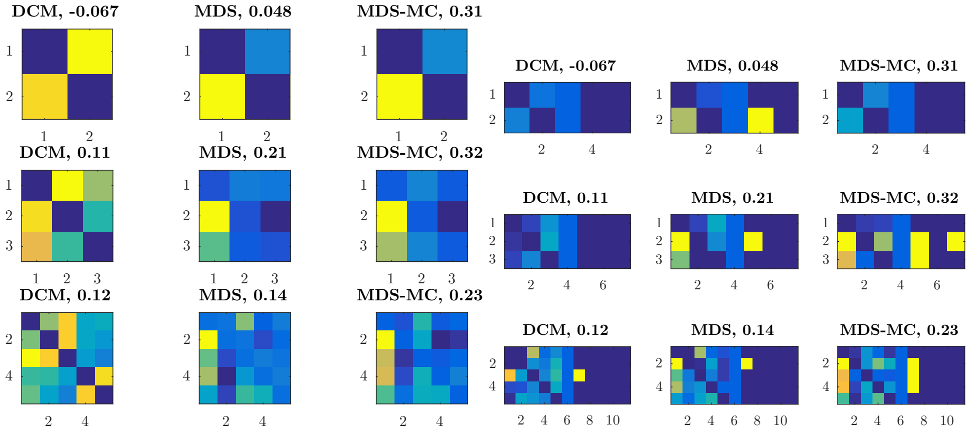


FIGURE 10. See `oddballl/oddball_DCM_MDS_gridB_plot.m` for code {oddballA}

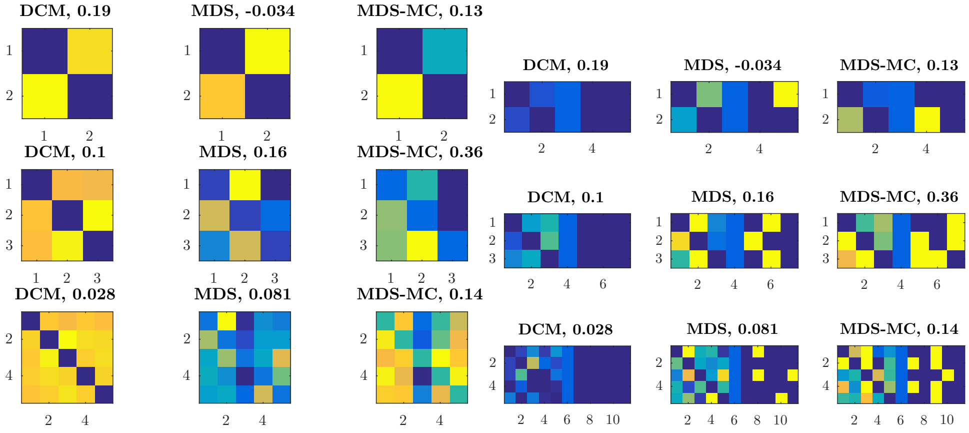
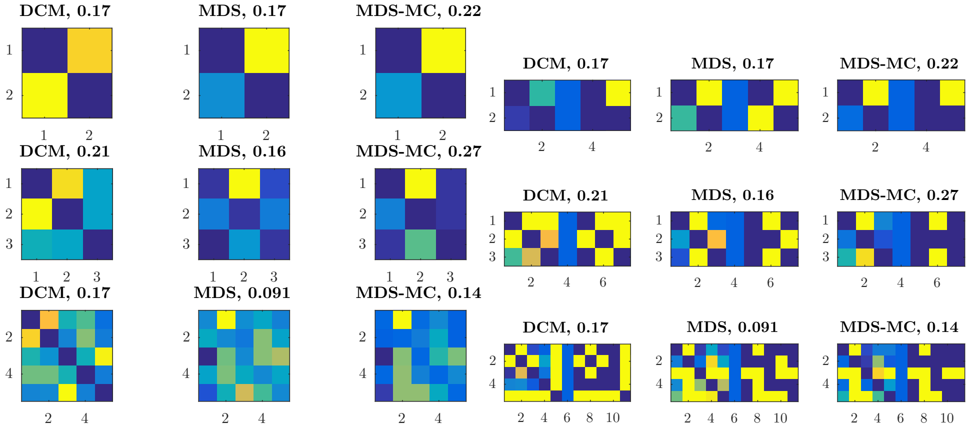


FIGURE 11. See `hcp/hcp_DCM_MDS_gridB_plot.m` for code {hcpA}



{yaleA} FIGURE 12. See `Yale/yale_DCM_MDS_gridB_plot.m` for code

TABLE 2. Consistency results, see `plottools/KendallMatrix.m` for details.

{tbl:consistency}

Dataset	Maximum (debug)			i, j fixed			Bootstrap		
	DCM	MDS	MDS-MC	DCM	MDS	MDS-MC	DCM	MDS	MDS-MC
oddball ($M = 3$)	0.11	0.21	**0.32**	0.11	0.21	**0.32**	0.11	0.21	**0.22**
oddball ($M = 5$)	0.12	0.14	**0.23**	0.12	0.14	**0.23**	0.12	0.14	**0.23**
Yale ($M = 3$)	0.21	0.16	**0.27**	0.21	0.16	**0.27**	0.21	0.16	**0.26**
Yale ($M = 5$)	**0.17**	0.091	0.13	**0.17**	0.091	0.11	**0.17**	0.091	0.092
stopgo ($M = 3$)	0.18	0.15	**0.27**	0.18	0.15	**0.27**	0.18	0.15	**0.26**
stopgo ($M = 5$)	0.2	0.07	**0.4**	0.2	0.07	**0.38**	0.2	0.07	**0.38**
HCP ($M = 3$)	0.1	0.16	**0.36**	0.1	0.16	**0.36**	0.1	0.16	**0.24**
HCP ($M = 5$)	0.028	0.081	**0.14**	0.028	0.081	**0.13**	0.028	0.081	**0.11**

TABLE 3. Consistency results for $J = 2$, see `plottools/KendallMatrix.m` for details.

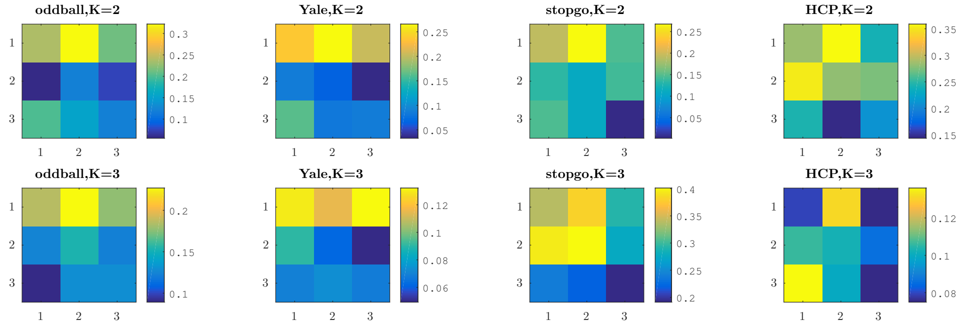
{tbl:consistencyJ2}

Dataset	$C_{J=1}$			$C_{J=2}$		
	DCM	MDS	MDS-MCMC	DCM	MDS	MDS-MCMC
HCP	0.01	0.09	**0.2**	0.02	-0.01	**0.11**

TABLE 4. Consistency results for $J = 1$, see `plottools/KendallMatrix2.m` for details. These results are with sampling of the hyper-parameters but with the old (false square) parameterization

{tbl:consistencyJ1}

Dataset	Maximum			Fixed i, j		
	DCM	MDS	MDS-MCMC	DCM	MDS	MDS-MCMC
hcp	0.03	0.08	**0.24**	0.03	0.08	**0.21**
yale	0.17	0.09	**0.17**	**0.17**	0.09	0.15
StopGo	0.2	0.08	**0.42**	0.2	0.08	**0.4**
oddball	0.12	0.14	**0.17**	0.12	0.14	**0.16**
luna	0.05	0.06	**0.14**	0.05	0.06	**0.13**



{kendall_all}

FIGURE 13. See `plottools/KendallMatrix.m` for code

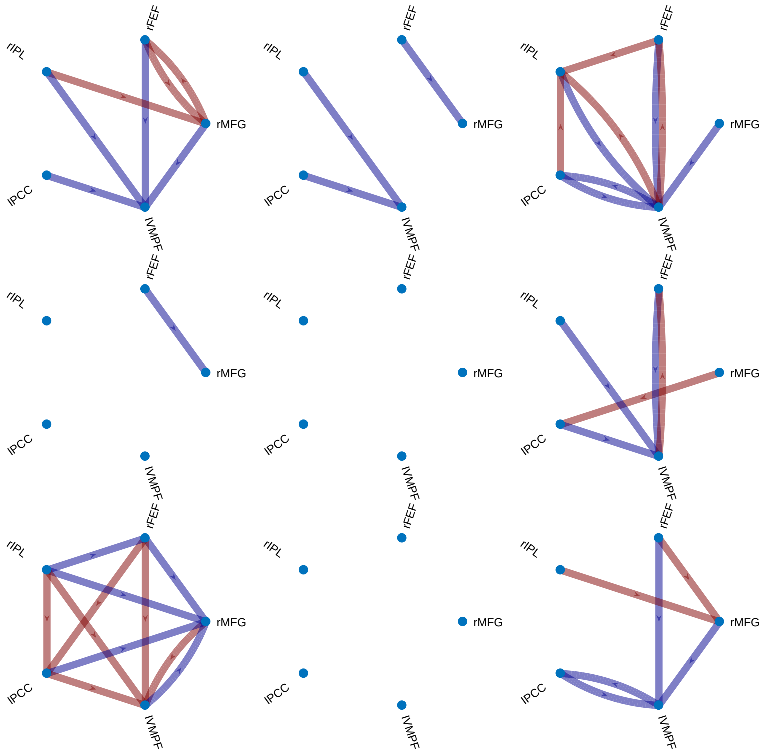


FIGURE 14. See `hpc/hpc_GPMDS_gridJ3_plot_graphs.m` for code. Horizontally: The three methods considered (DCM, MDS, MDS-MCMC; vertically the three different graphs with the first row corresponding to the "base interactions" and the second and third row 0back and 1back respectivelly (remember to confirm experiment labels with Weidong). Comments by Weidong on figure: This kind of figure is as a rule difficult to interpret. MDS-MCMC seems to group nodes involved in DMN and task-related networks by edges. (top/bottom nodes). Good there are more crosstalk in bottom row. Figure consistent with idea MDS-MCMC does something useful, but this is about as much as can possibly be said. {hpc_graphs1}

with $\alpha_C = 1$. To bound eigenvalues of C we prefer σ_C^2 to be low. In our case $\sigma_C \approx \frac{1}{50}$ seems reasonable. A reasonable way to encode this kind of prior is a gamma distribution with mean $m = \frac{1}{50}$ and standard deviation some fraction ρ of m. We write this as:

$$\mathrm{Gamma}_\rho(m) = \mathrm{Gamma}(\alpha = \rho^{-2}, \beta = \rho^{-2} m^{-1})$$

and typically select $\rho = \frac{1}{4}$. For β_C we assume $\beta_C \sim \mathrm{Gamma}_{\frac{1}{4}}(50^2)$.

Similarly, w_{kt} is also Normal-Gamma distributed as $w_{kt} \sim \mathcal{N}(0, \sigma_{w,kt})$ and

$$\sigma_w^2 \sim \mathrm{Gamma}(\alpha_w, \beta_w), \quad \alpha_w = 1 \tag{8.1}$$

Table 5. Significant between-session agreement for blocks of 5 subjects. See `WDNG/w_kendallmatrix.m` for code

{w_kendall_J3}

Dataset	$C_{J=1}$			$C_{J=2}$			$C_{J=3}$		
	ses$_1$	ses$_2$	ses$_{12}$	ses$_1$	ses$_2$	ses$_{12}$	ses$_1$	ses$_2$	ses$_{12}$
yale, DCM	0.21	0.17	0.2	0.05	0.04	0.07	0.01	0.04	0.05
yale, MDS	0.3	0.37	0.33	0.05	0.0	0.04	-0.01	-0.01	-0.0
yale, MDSMC	**0.37**	**0.59**	**0.48**	0.04	**0.14**	0.09	0.05	**0.14**	**0.12**
sstxui, DCM	0.21	0.05	0.2	0.12	-0.04	0.08	0.22	0.1	0.11
sstxui, MDS	0.25	**0.43**	0.26	0.12	0.0	0.03	0.1	0.11	-0.05
sstxui, MDSMC	0.41	0.26	0.36	0.02	-0.01	0.05	0.03	-0.06	0.03
flanker, DCM	0.03	-0.06	0.1	0.04	0.04	0.12	0.07	-0.0	-0.02
flanker, MDS	-0.03	-0.05	0.07	0.09	-0.04	0.05	-0.02	0.06	-0.02
flanker, MDSMC	**0.42**	**0.28**	**0.39**	0.21	**0.26**	**0.24**	**0.28**	0.09	**0.18**

Table 6. p-values for (significant) positive correlation at $\alpha = 0.05$ on (net) causal outflow between the two scan sessions. Computed on blocks of 5 subjects. See `WDNG/w_J3_plot_graphs.m` for code

{w_corr_J3}

Dataset	$C_{J=1}$	$C_{J=2}$	$C_{J=3}$	$C_{J=3} - C_{J=2}$
sstxui DCM	**0.001**	0.807	0.825	0.559
sstxui MDS	**0.0**	0.947	0.433	0.501
sstxui MDSMC	**0.0**	0.114	**0.007**	**0.036**
flanker DCM	**0.002**	**0.0**	0.205	0.262
flanker MDS	**0.025**	0.069	0.982	0.356
flanker MDSMC	**0.0**	**0.0**	**0.0**	0.379
yale DCM	**0.0**	**0.002**	**0.001**	0.336
yale MDS	**0.0**	0.236	0.402	0.225
yale MDSMC	**0.0**	**0.0**	**0.001**	**0.022**

The scale on β_w control the variance of the autoregressive model. Recall for the autoregressive process: $X_{t+1} = \phi X_t + w$ the variance is:

$$\mathrm{var}[X_t] = \frac{\sigma_w^2}{1-\phi^2} = \frac{\sigma_w^2}{1-(1+\Delta C)^2}$$

This implies:

$$\mathrm{std}[X_t] \leq \frac{1}{\beta_w} \frac{1}{\sqrt{1-(1+\Delta C_{\mathrm{cut}})^2}} \tag{8.2}$$

Where C_{cut} is the miminal cutoff. Tuning β_w is thus equivalent (in this simplified setting) to selecting a preferred variance of the process. Thus, we finally have

$$w_t \sim \mathcal{N}(0, \sigma_t^2) \tag{8.3}$$

$$\sigma^2 \sim \mathrm{Gamma}(\alpha_w, \beta_w) \tag{8.4}$$

$$\beta_w \sim \mathrm{Gamma}_\rho(m_w) \tag{8.5}$$

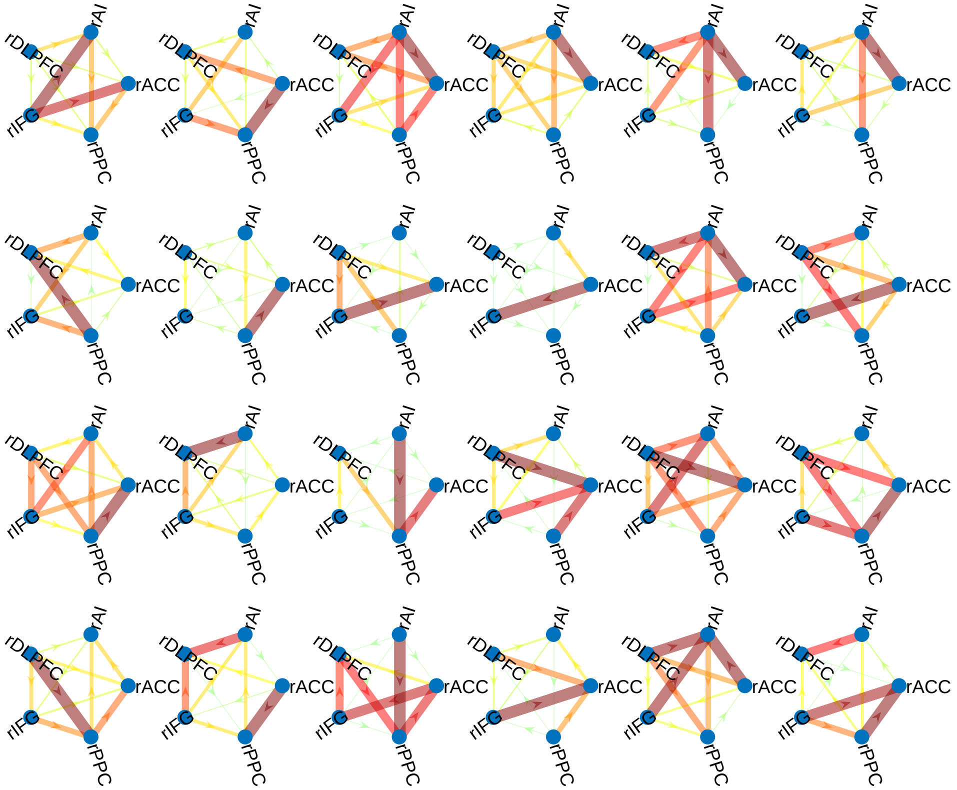


FIGURE 15. SSTXui: See `WDNG/w_J3_graphs.m` for code. Horizontally: 3 methods each having two columns corresponding to the two sessions. Vertically C_1, C_2, C_3 and $C_2 - C_3$. Graph show net causal outflow

{w_graphs1}

Let us therefore assume that $\text{std}[X_t] = B_0$ is some fixed, pre-specified constant and let $K_C = 1 - (1 + \Delta_t C_{\text{cut}})^2$. Then we get:

$$B_0^2 = \frac{\sigma_w^2}{K_C} = \frac{\alpha_w}{\beta_w K_C} \tag{8.6}$$

$$\text{or} \quad \beta_w = \frac{\alpha_w}{B_0^2 K_C} \approx m_w \tag{8.7}$$

We select distribution of β_C to reflect this by setting β_C Priors for w: Normal-gamma,

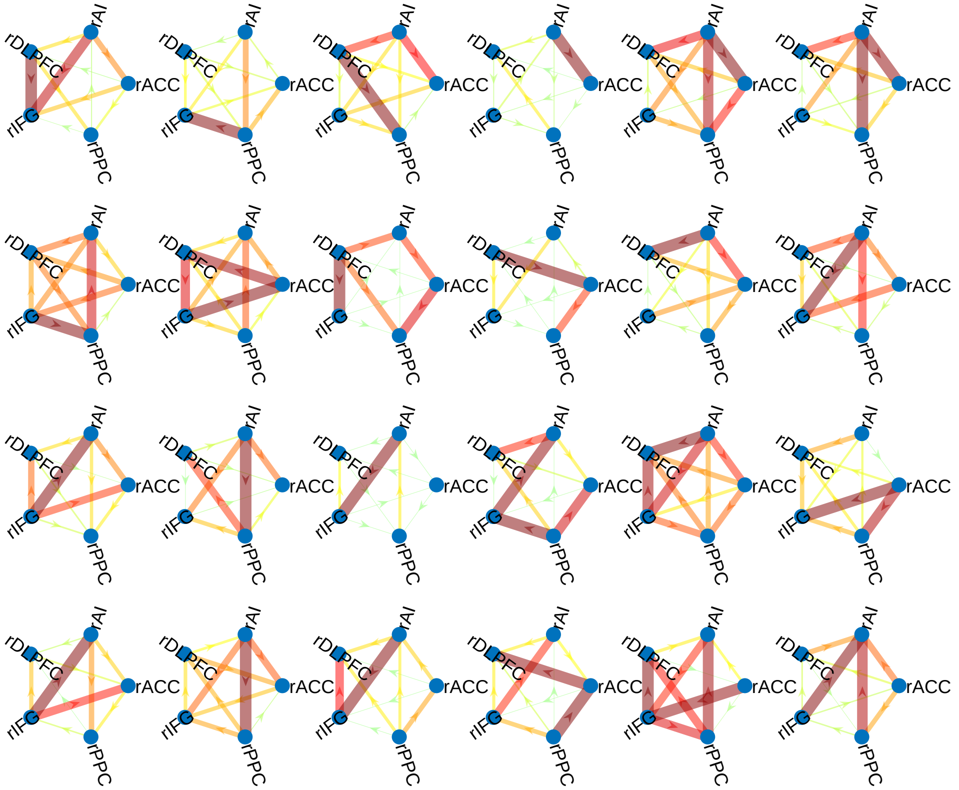
$$\sigma^2 \sim \text{Gamma}(\alpha, \beta)$$

$\alpha = 1, \beta = 5.6$ (ca) Mean for later process is:

$$\sigma_w = \sqrt{\frac{\alpha_w}{\beta_w}}$$

For the autoregressive process: $X_{t+1} = \phi X_t + \varepsilon$ the variance is:

$$\text{var}[X_t] = \frac{\sigma^2}{1 - \phi^2} = \frac{\sigma^2}{1 - (1 + \Delta C)^2}$$



{w_graphs1} FIGURE 16. Flanker: See `WDNG/w_J3_graphs.m` for code. Horizontally: 3 methods each having two columns corresponding to the two sessions. Vertically C_1, C_2, C_3 and $C_2 - C_3$. Graph show net causal outflow

9. SUPPLEMENTARY 2, LINEARIZED MODEL

Basic model defined as:

$$s_{t+1} = Cs_t + \varepsilon_t$$

Inference can be done according to Fasen [2013] as

$$sN = CsM + \varepsilon$$

This provides

$$\hat{C} = (sNM^Ts^T)(sMM^Ts^T)^{-1}$$

If we write this as

$$\hat{C} = (sNz^T)(zz^T)^{-1}$$

Taking the derivatives we obtain:

$$\begin{aligned}\frac{\partial \hat{C}}{\partial s_{it}} &= \frac{\partial (sNz^T)}{\partial s_{it}}(zz^T)^{-1} - (sNz^T)(zz^T)^{-1}\frac{\partial (zz^T)}{\partial s_{it}}(zz^T)^{-1} \\ &= \left[\frac{\partial s}{\partial s_{it}}Nz^T + sN\left(\frac{\partial z}{\partial s_{it}}\right)^T - (sNz^T)(zz^T)^{-1}\left[\frac{\partial z}{\partial s_{it}}z^T + z\left(\frac{\partial z}{\partial s_{it}}\right)^T\right]\right](zz^T)^{-1}\end{aligned}$$

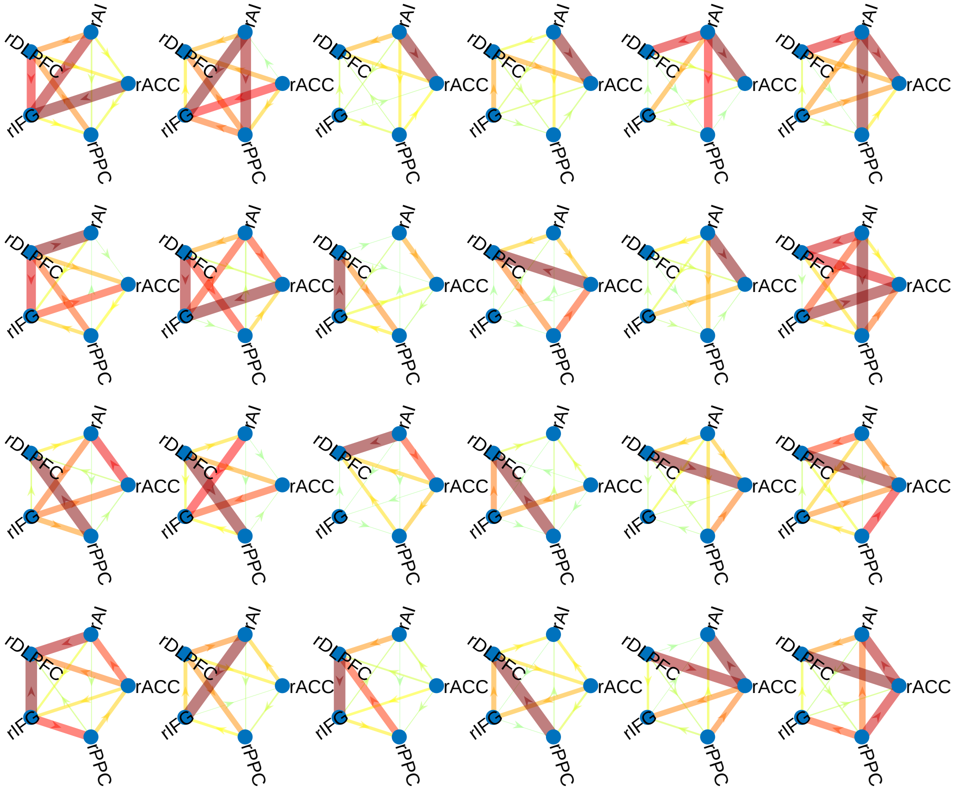


FIGURE 17. Yale: See `WDNG/w_J3_graphs.m` for code. Horizontally: 3 methods each having two columns corresponding to the two sessions. Vertically C_1, C_2, C_3 and $C_2 - C_3$. Graph show net causal outflow

{w_graphs1}

This is used for taking the derivative of the overall error: $E = g(C)$:

$$\frac{\partial E}{\partial s_{it}} = \frac{\partial g(C)}{\partial s_{it}} = \mathrm{Tr}\left[\frac{\partial C}{\partial s_{it}}\left(\frac{\partial g(C)}{\partial C}\right)^T\right]$$

Assuming C is time-dependent, $C_t = \sum_k C_k v_{kt}$. Then this can be written as:

$$sN = \sum_k C_k s V_k M = \begin{bmatrix} C_1 & \cdots & C_J \end{bmatrix} \sum_{j=1}^{J} e_j \otimes s V_j M \tag{9.1}$$

$$= C \sum_{i=1}^{J} (I_{J\times J} \otimes s)(e_j \otimes V_j M) \tag{9.2}$$

$$= C(I \otimes s)Q = Cz, \quad z = (I \otimes s)Q \tag{9.3}$$

This gives

$$\frac{\partial z}{\partial s_{it}} = (I \otimes I_{it})Q$$

Let us assume we have multiple subjects. Writing

$$q_u = Cz_u, \quad q_u = s_u N. \tag{9.4}$$

The cost function is then:

$$E = \frac{\lambda_C}{2} \operatorname{Tr}\left[C^T C\right] + \sum_{u=1}^{S} \left(\frac{\lambda_y}{2} \left\| y_u - s_u \Phi \right\|^2 + \frac{\lambda_q}{2} \left\| q_u - C z_u \right\|^2 + \operatorname{Tr}\left[s_u \Sigma^{-1} s_u^T \right] \right) \tag{9.5}$$

Maximizing wrt. C we obtain the maximum value of C as:

$$\begin{aligned} \hat{C} &= \left(\sum_u q_u z_u^T \right) \left(\lambda_C I + \sum_u z_u z_u^T \right)^{-1} \\ &= \left(\sum_u q_u z_u^T \right) Z^{-1} \end{aligned}$$

Which needs to be plugged back into E before derivatives are taken. Then, turning to C, we have:

$$\operatorname{Tr}\left[A \frac{\partial \hat{C}}{\partial s_{it}} B \right] = ((N z^T \bar{B})(A))^T + \left(\sum_{j=1}^{J} \left(V_j M \left[z_u^T \bar{B}(sNz^T Z^{-1})) + (sN + (sNz^T)Z^{-1}z)^T (\bar{B})^T \right] (e_j \otimes I) \right) \right)^T \tag{9.6}$$

Then we need some trace relations. If we have:

$$\begin{aligned} \operatorname{Tr}\left[A ds B\right] &= \operatorname{Tr}\left[ds B A\right] \\ &= (BA)^T \\ \operatorname{Tr}\left[A \frac{\partial z}{\partial s_{it}} B \right] &= \sum_j \sum_i \operatorname{Tr}\left[(e_j \otimes ds V_j M)(e_i^T \otimes (BA(e_i \otimes I))) \right] \\ &= \sum_j \operatorname{Tr}\left[ds V_j M B A (e_j \otimes I)) \right] \\ &= \sum_j \left[V_j M B A (e_j \otimes I)) \right]^T \\ \operatorname{Tr}\left[A \left(\frac{\partial z}{\partial s_{it}} \right)^T B \right] &= \operatorname{Tr}\left[B^T \frac{\partial z}{\partial s_{it}} A^T \right] \end{aligned}$$

References

Anthony Randal McIntosh. Towards a network theory of cognition. *Neural Networks*, 13(8):861–870, 2000.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

Rainer Goebel, Alard Roebroeck, Dae-Shik Kim, and Elia Formisano. Investigating directed cortical interactions in time-resolved fmri data using vector autoregressive modeling and granger causality mapping. *Magnetic resonance imaging*, 21 (10):1251–1261, 2003.

Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2): 461, 2000.

Karl J Friston, Lee Harrison, and Will Penny. Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302, 2003.

Karl J. Friston, Nelson J. Trujillo-Barreto, and Jean Daunizeau. DEM: A variational treatment of dynamic systems. *NeuroImage*, 41(3):849–885, 2008a. URL `http://dblp.uni-trier.de/db/journals/neuroimage/neuroimage41.html#FristonTD08`.

Karl Friston, Klaas Stephan, Baojuan Li, and Jean Daunizeau. Generalised filtering. *Mathematical Problems in Engineering*, 2010, 2010.

Srikanth Ryali, Kaustubh Supekar, Tianwen Chen, and Vinod Menon. Multivariate dynamical systems models for estimating causal interactions in fmri. *Neuroimage*, 54(2):807–823, 2011.

Gabriele Lohmann, Kerstin Erfurth, Karsten Müller, and Robert Turner. Critical comments on dynamic causal modelling. *Neuroimage*, 59(3):2322–2329, 2012.

Karl Friston. Dynamic causal modeling and granger causality comments on: The identification of interacting networks in the brain using fmri: Model selection, causality and deconvolution. *Neuroimage*, 58(2):303–305, 2011.

Karl Friston, Jean Daunizeau, and Klaas Enno Stephan. Model selection and gobbledygook: Response to lohmann et al. *Neuroimage*, 75:275–278, 2013.

Karl J Friston, N Trujillo-Barreto, and Jean Daunizeau. Dem: a variational treatment of dynamic systems. *Neuroimage*, 41(3):849–885, 2008b.

Vicky Fasen. Statistical estimation of multivariate ornstein–uhlenbeck processes and applications to co-integration. *Journal of Econometrics*, 172(2):325–337, 2013.